Vision-Language Models are Fragile Multilingual Associators

Anonymous ACL submission

Abstract

Vision-language models must associate visual entities with textual attributes. Whether these associations or concept bindings remain stable when the language of the input changes is unexplored. We introduce M^2BIND, a benchmark varying the language of the context and query across multiple languages. We evaluate binding both extrinsically through task performance metrics and intrinsically through causal interventions. We find that binding is not language-invariant: cross-family and cross-script settings trigger significant binding collapse, with the model's internal binding computation shifting to later layers and losing causal strength. Closely related languages preserve associations comparatively better. In a broader sense, our findings indicate how VLMs deployed globally in multilingual settings cannot be assumed to maintain the same association quality observed in monolingual evaluation.

1 Introduction

A reader who speaks **French** and **English** does not pause to translate when informed about something "yellow" and asked about something "jaune". Humans fuse what we see with what we read into a single concept, and we do so without effort no matter which of the world's languages the words arrive in (Fig. 1) (Bonner and Epstein, 2021; Xie et al., 2017). These multilingual associations come innately with our understanding of languages and how they describe (here, in the visual sense) the world around us.

Vision-language models (VLMs) (OpenAI et al., 2024; Team et al., 2025) now operate in exactly these settings, answering questions about images for users spread across the languages and scripts of the map in Fig. 1. The languages depicted (top) are chosen to span multiple scripts and language families across several continents, so our findings speak to globally diverse deployment rather than

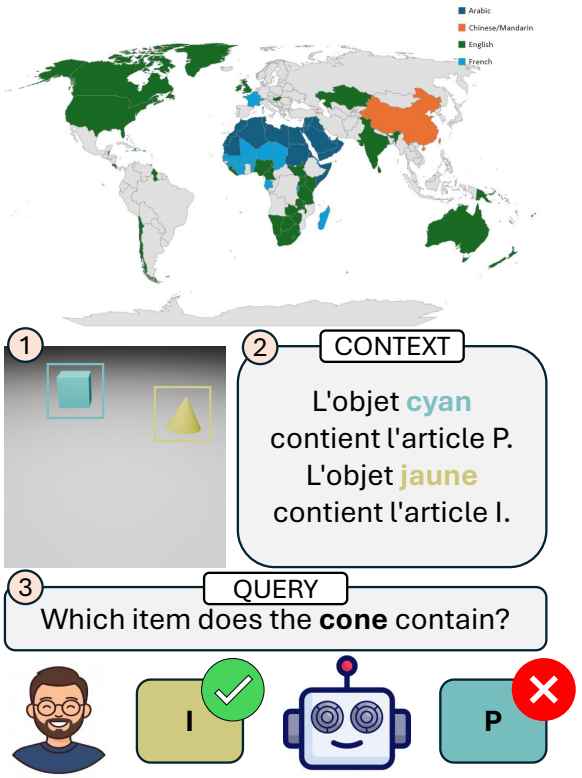

Figure 1: **Concept binding across languages. (Top)** The world map situates the languages in our study across continents, families, and scripts. **(Bottom)** A **French** speaker who also reads **English** effortlessly maps jaune onto the same concept as yellow, and so links the 1 image, the 2 **French** context, and the 3 **English** query to answer **cone** → I (✓). VLMs frequently fail this association (✗). We test whether such bindings survive when the context and query languages differ.

to any single region. If the concept associations a VLM forms internally are tethered to the surface form of **English**, then every non-**English** user inherits a silent degradation that monolingual evaluation never exposes. Such associations are governed by *binding*, which in its simplest causal form asks : `given X:Y and Y:Z, can the model infer X:Z?` (see Eq. (1)). Jiahai Feng and Jacob Steinhardt (2024) showed that language models at-

tach latent Binding IDs to entities, representing this structure in their activations without it ever appearing in the prompt; Saravanan et al. (2025) extended the result to VLMs, where image tokens bind to their textual references. We subject this in-context binding to a novel test, where we establish the binding in one language and retrieve it in another. We organize our study around three questions:

RQ1: Do visual-textual associations survive across languages?
RQ2: What inside the VLM drives their degradation?
RQ3: Does linguistic closeness aid binding transfer?

Contributions. We (i) introduce M^{2}BIND, a task and benchmark that varies the context and query language independently over eight languages along two axes of linguistic distance; (ii) evaluate binding both extrinsically and intrinsically; (iii) trace the behaviour to linguistic distance, model tokenizer and decoder computations. To our knowledge, this is the first work of its kind.

Sec. 2 defines our task, data, and metrics; Sec. 3 discusses results and Sec. 4 concludes the paper.

2 Methodology: M^{2}BIND

Task. We introduce the task of **Multilingual Multimodal BINDing**: for a visual scene and a textual context that assigns attributes to objects in language ℓ_{ctx}, can a VLM retrieve the correct association when queried in a different language ℓ_q?

Problem formulation. (Fig. 1) A scene $s = (v, \mathcal{O})$ consists of an image 1 v and a set of objects $\mathcal{O} = \{o_1, o_2\}$. Each object o_j is characterised by a shape $\sigma_j \in \Sigma$ and a colour $\gamma_j \in \Gamma$, with $\Sigma = \{\text{cone}, \text{cube}, \text{cylinder}, \text{sphere}\}$ and $\Gamma = \{\text{red}, \text{blue}, \text{green}, \text{yellow}, \text{cyan}, \text{purple}\}$. A textual context 2 c refers to each object by its colour and assigns it an item symbol $\iota_j \in \mathcal{I}$, thereby defining a colour-to-item binding $b : \gamma_j \mapsto \iota_j$. A query 3 q singles out a target object by its shape $\sigma_\star$ and asks for the item it contains. Producing the correct answer requires the model to compose two associations,

$$\sigma_\star \xrightarrow[v]{\text{ground}} \gamma_\star \xrightarrow[c]{\text{bind}} \iota_\star = b(\gamma_\star), \qquad (1)$$

first grounding the queried shape to a colour through the image, then resolving that colour to its bound item through the context (Eq. (1)). Crucially, associations must be carried by *latent binding variables* in the model's activations (Jiahai Feng and Jacob Steinhardt, 2024; Saravanan et al., 2025).

Linguistic conditioning. We render the context in a language ℓ_{ctx} and the query in a language ℓ_q while holding the image v fixed, so an instance is the triple $x = (v, c^{\ell_{\text{ctx}}}, q^{\ell_q})$. Because the image, the underlying scene, and the gold item $\iota_\star$ are identical across all conditions, any change in the model's response isolates the effect of *language* on binding rather than of perception or task difficulty.

Data and Setup. We build on the Shapes binding task of Saravanan et al. (2025), in which each instance pairs a Blender-rendered image of two 3D objects with a context that binds each object's colour to an item symbol and a query that targets one object by its shape. We extend the task by independently varying ℓ_{ctx} and ℓ_q over 8 languages selected along two axes of linguistic distance[1]: **(i) cross-family**: **English**, **French**, **Mandarin**, **Arabic**, spanning four families and three scripts (Asher and Moseley, 2018; Littell et al., 2017); and **(ii) within-family**, comprising a **Germanic** cluster (**English**, **Dutch**, **German**) and a **Romance** cluster (**French**, **Italian**, **Spanish**) that share the Latin script but differ in genealogical proximity (Campbell and Grondona, 2008; Gooskens et al., 2018)[2].

Model. We evaluate LLaVA-1.5-OV-7B (An et al., 2025) a SigLIP vision encoder, an MLP projector, and Qwen2 decoder following Saravanan et al. (2025) as a representative open-weight VLM. Given an instance x, the model returns a conditional distribution over candidate items at the answer position under greedy decoding. All reported values are averaged over 3 runs on a single NVIDIA A40 (48GB) GPU.

Sequence score. We compare the two candidate items $\mathcal{C} = \{\iota_\star, \iota_{\text{swap}}\}$, where $\iota_\star$ is bound to the queried object and ι_{swap} to the distractor. For instance x_i, we score a candidate I by the log-probability the model assigns to its token sequence $y_{1:m(I)}(I)$ at the answer position:

$$S_i(I) = \sum_{t=1}^{m(I)} \log p_\theta\big(y_t(I) \mid x_i,\, y_{<t}(I)\big). \qquad (2)$$

[1]We discuss linguistic choices in Sec. B.
[2]Combinations of these languages yield 16 arrangements for cross-family and 9 for within-family.

Accuracy and Factorization Margin. **Accuracy** is the fraction of instances for which $S_i(\iota_\star) > S_i(\iota_{\text{swap}})$ under the score of Eq. (2). Because accuracy saturates before binding fully degrades, we additionally report the **Factorization Margin** (FM), the mean score separation between the correct and swapped items (Jiahai Feng and Jacob Steinhardt, 2024):

$$\text{FM} = \frac{1}{N}\sum_{i=1}^{N}\big[S_i(\iota_{\star,i}) - S_i(\iota_{\text{swap},i})\big]. \quad (3)$$

A large positive FM indicates that the model cleanly separates the bound item from the distractor, whereas $\text{FM} \to 0$ signals a collapsed binding even when accuracy remains high.

Causal localization. Accuracy and FM characterise binding only at the output. To locate *where* binding is consolidated inside the decoder, we apply an interchange intervention (Meng et al., 2022; Vig et al., 2020). Let $h_{k,j}$ denote the layer-k hidden state at the token position of object o_j. We re-run the forward pass with the two objects' states interchanged, $\text{do}(h_{k,1} \leftrightarrow h_{k,2})$, and define the **Intervention Causal Effect** (ICE) as the resulting drop in the correct-item score, averaged over instances:

$$\text{ICE}_k = \mathbb{E}_i\Big[S_i(\iota_\star) - S_i\big(\iota_\star \mid \text{do}(h_{k,1} \leftrightarrow h_{k,2})\big)\Big]. \quad (4)$$

A large positive ICE_k means that disrupting the layer-k representation of the two objects sharply lowers the score of the correct item, i.e. binding-relevant information is actively represented at layer k. This gives us a profile of where the associations are made.

3 Results and Discussions

(RQ1) Binding across language families. Binding strength is far from language-invariant: FM falls by 2.65, from 5.45 for monolingual **English** to 2.80 for the **Mandarin**–**Arabic** cross-script pair, less than half its best monolingual value. Tab. 1 reports accuracy and FM for all 16 context–query combinations over **English**, **French**, **Mandarin**, and **Arabic**. The monolingual diagonal ($\ell_{\text{ctx}}{=}\ell_{\text{q}}$) sets each language's upper bound: **English** is strongest at 5.45, while the two non-Latin scripts sit lowest even on their own diagonal (**Mandarin** 4.95, **Arabic** 4.55), so script already shapes binding before any language mixing. Cross-lingually,

$\text{L}_{\text{ctx}} \downarrow$ / $\text{L}_{\text{q}} \rightarrow$	En	Fr	Zh	Ar
En	0.995 / 5.45	0.990 / 5.20	0.955 / 3.85	0.940 / 3.40
Fr	0.990 / 5.10	0.995 / 5.25	0.945 / 3.60	0.930 / 3.25
Zh	0.965 / 4.10	0.955 / 3.85	0.985 / 4.95	0.900 / 2.80
Ar	0.950 / 3.75	0.940 / 3.50	0.905 / 2.85	0.975 / 4.55

Table 1: **Cross-family VLM associations.** Cross-lingual concept binding over **En**, **Fr**, **Zh**, and **Ar**; rows are context (ctx), columns are query (q). Each cell reports *Acc. / FM*; for both, higher (↑) is better. Worst cases are highlighted.

Lang	Mean tokens ↓	Trunc. (%) ↓	FM drop ↓
En	1450	0	0.00
Fr	1520	0	0.20
Zh	1720	5	1.70
Ar	1850	12	2.55

Table 2: **Impact of tokenization.** Values reported are for monolingual setting. We report mean tokens of the image+context+query, truncation (Trunc.) against the context limit, and the FM drop relative to **English**. For all metrics, lower (↓) is better.

FM tracks typological distance. The close **English**–**French** pair stays near its monolingual values, whereas mixing a Latin with a non-Latin script collapses FM to roughly 3.2–4.1, and the two non-Latin scripts together reach the 2.80 floor noted above with accuracy at its ∼0.90 floor. Degradation is also asymmetric: across the four Latin/non-Latin pairs, the non-Latin language on the query side is consistently harder than on the context side by about 0.25 FM (e.g. **English**→**Arabic** 3.40 vs. **Arabic**→**English** 3.75). This matches the Matrix Language Frame account of code-switching (Myers-Scotton, 2001), where ℓ_{ctx} builds the binding scaffold and ℓ_{q} must access it: access is more fragile than construction once the frames diverge. The trend is consistent with multilingual NLP (Lauscher et al., 2020; Pires et al., 2019).

(RQ2) Tokenization. Which component drives this fragility? Tokenization in particular plays a significant role since the same sentence can use a varying amount of tokens based on the language[3]. In the monolingual setting (Tab. 2), **Arabic** needs 27.6% more tokens than **English**, truncates 12% of prompts at the context limit, and loses 2.55 FM, with no cross-lingual transfer involved. Bind-

[3]This is referred to as the *fertility* of a tokenizer (Rust et al., 2021).

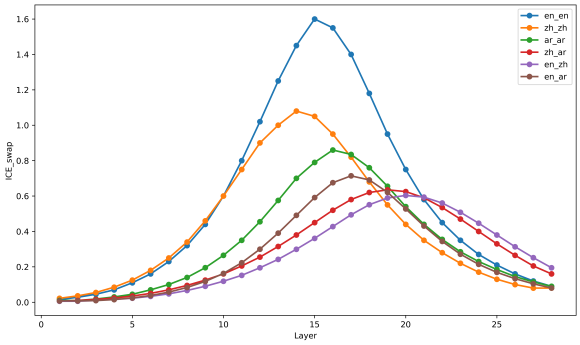

Figure 2: **Causal interventions for multilingual binding.** The y-axis is the change in log-probability (ICE) after intervention; the x-axis is decoder layer index. A higher peak implies stronger association strength at those layers. We show monolingual (**English**, **Mandarin**, **Arabic**) and cross-lingual combinations. Higher is better.

$L_{ctx} \downarrow$ / $L_q \rightarrow$	English	Dutch	German
English	0.995 / 5.45	0.990 / 5.30	0.985 / 5.10
Dutch	0.990 / 5.25	0.990 / 5.20	0.985 / 5.05
German	0.985 / 5.00	0.985 / 5.00	0.985 / 5.05

(a)

$L_{ctx} \downarrow$ / $L_q \rightarrow$	French	Italian	Spanish
French	0.995 / 5.25	0.990 / 5.15	0.990 / 5.05
Italian	0.980 / 5.10	0.990 / 5.15	0.985 / 5.00
Spanish	0.985 / 5.00	0.985 / 5.00	0.990 / 5.10

(b)

Table 3: **Within-family binding for Germanic and Romance families.** Worst cases are highlighted.

ing thus weakens within a single language from token budget alone, implicating tokenizer unfairness (Petrov et al., 2023) and supporting Rust et al. (2021) on tokenizer quality and multilingual performance.

(RQ2) Binding in VLM. While Tab. 2 explains impact on binding before actual associations, we ask how the internal representations of the model actually perform multilingual association. Fig. 2 provides some insights. All curves are unimodal, depicting a localized region of layers where multilingual association is observed the strongest. For monolingual cases all curves peak around the middle layers ($\sim$15), indicating that the model has already committed to the correct association before the final generation layers. Cross-lingual instances undergo a shift towards the right, peaking around the late layers of the decoder ($\sim$20). This points to extra computation reconciling the context- and query-language representations. Interestingly, binding is reduced for cross-lingual cases as compared to monolingual curves. The combination of **Mandarin**-**Arabic**, which produced the weakest FM in Tab. 1, also produces the flattest and weakest curve in Fig. 2. This means that even after extra computations in the decoder, the binding is less causally separable , i.e, the swap intervention has less effect because the two items' representations are less distinct.

(RQ3) Binding within families. Instead of languages that do not seemingly overlap in mutual intelligibility, how do VLMs behave when the languages are closer to each other[4]? Tab. 3 shows concept binding for the two language families mentioned in Sec. 2. Overall, every within-family FM stays above 5, the worst drop being of 0.45 from **German** to other languages. In terms of accuracy, within-family excels with a lower bound around 0.98, as compared to a floor of 0.90 for cross-family interactions in Tab. 1. Tab. 3a shows results for the Germanic family. **English**-**Dutch** is akin to **English** in its monolingual setting, closer than **English**-**French**, despite both being non-English languages. Interestingly, **German**-**Dutch** pairs perform really well, generalizing beyond **English**. Tab. 3b shows results for Romance family, where **Italian** demonstrates to be better than **Spanish** as ctx or q. **Spanish**-**Italian** as a pair shows an FM of 5, showing associations work within the family, without the need of **French** as an anchor language, supporting Conneau et al. (2020).

4 Conclusion

We present M²BIND, a benchmark that varies the context and query language independently to test whether VLMs maintain entity-attribute bindings across languages. Binding dissociates with linguistic distance: distant languages and non-Latin scripts collapse both binding strength and its causal localization, so near-ceiling task accuracy is not a reliable predictor of cross-lingual binding fidelity. Language-invariant grounding, effortless for humans, remains unsolved for current VLMs.

[4]Consider a native **Dutch** speaker, they might have some understanding upon encountering a query in **German**.

5 Limitations

The Shapes task uses procedurally generated Blender images; we use this since our interest was to show a simple task where multilingual associations are not performed properly. Still, noticing an even further degradation of this concept binding for real images remains, beyond two objects or simple shapes. Further we show our work primarily on a LLaVA VLM, along with additional results on Qwen2.5 VL in Appendix Sec. C. A similar look into other commercial models could potentially illustrate our generalizations further.

References

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, and 4 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *Preprint*, arXiv:2509.23661.

Ronald E Asher and Christopher Moseley. 2018. *Atlas of the World's Languages*. Routledge.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Michael F Bonner and Russell A Epstein. 2021. Object representations in the human brain reflect the co-occurrence statistics of vision and language. *Nature communications*, 12(1):4081.

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

Lyle Campbell and Verónica Grondona. 2008. Ethnologue: Languages of the world. 15th edn. ed. by raymond g. gordonjr.. dallas: Sil international, 2005. pp. 1,272. isbn 155671159x. 80(*hb*). *Language*, 84(3):636–641.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. 2024. Representational analysis of binding in language models. *arXiv preprint arXiv:2409.05448*.

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. 2023. Discovering variable binding circuitry with desiderata. *arXiv preprint arXiv:2307.03637*.

Jiahai Feng and 1 others. 2025. Mixing mechanisms: How language models retrieve bound entities in-context. *arXiv preprint arXiv:2510.06182*.

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On drivers of multilingual ability of large vision-language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Aubrey L Gilbert, Terry Regier, Paul Kay, and Richard B Ivry. 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.

Charlotte Gooskens, Vincent J Van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte, and Stefanie Voigt. 2018. Mutual intelligibility between closely related languages in europe. *International Journal of Multilingualism*, 15(2):169–193.

Martin Haspelmath. 2009. The typological database of the world atlas of language structures. *The use of databases in cross-linguistic studies*, 41:283.

Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric attributes in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *International Conference on Learning Representations*, volume 2024, pages 10504–10526.

Jiahai Feng and Jacob Steinhardt. 2024. How do Language Models Bind Entities in Context? In *The Twelfth International Conference on Learning Representations*.

Paul Kay and Willett Kempton. 1984. What is the sapir-whorf hypothesis? *American anthropologist*, 86(1):65–79.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.

Andrei-Alexandru Manea and Jindřich Libovický. 2026. Multilingual vision-language models, a survey. *Preprint*, arXiv:2509.22123.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Carol Myers-Scotton. 2001. The matrix language frame model: Developments and responses. *Codeswitching worldwide II*, 2:23.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. In *Advances in Neural Information Processing Systems*.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Rebecca Posner. 1996. The romance languages.

Nikhil Prakash, Tamar Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, volume 2024, pages 8057–8082.

Terry Regier and Paul Kay. 2009. Language, thought, and color: Whorf was half right. *Trends in cognitive sciences*, 13(10):439–446.

Håkan Ringbom and Scott Jarvis. 2009. The importance of cross-linguistic similarity in foreign language learning. *The handbook of language teaching*, pages 106–118.

Don Ringe. 2017. *From Proto-Indo-European to Proto-Germanic*, volume 1. Oxford University Press.

Debi Roberson, Ian Davies, and Jules Davidoff. 2000. Color categories are not universal: replications and new evidence from a stone-age culture. *Journal of experimental psychology: General*, 129(3):369.

David Romero, Chenyang Lyu, Haryo A Wibowo, Teresa Lynn, Injy Hamed, Aditya N Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu L Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Advances in Neural Information Processing Systems*, 37:11479–11505.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Darshana Saravanan, Makarand Tapaswi, and Vineet Gandhi. 2025. Investigating Mechanisms for In-Context Vision Language Binding . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4852–4856, Los Alamitos, CA, USA. IEEE Computer Society.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Guillaume Thierry, Panos Athanasopoulos, Alison Wiggett, Benjamin Dering, and Jan-Rouke Kuipers. 2009. Unconscious effects of language-specific terminology on preattentive color perception. *Proceedings of the National Academy of Sciences*, 106(11):4567–4570.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Benjamin Lee Whorf. 2012. *Language, thought, and reality: Selected writings of Benjamin Lee Whorf*. MIT press.

Jonathan Winawer, Nathan Witthoft, Michael C Frank, Lisa Wu, Alex R Wade, and Lera Boroditsky. 2007. Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, 104(19):7780–7785.

Yuanjun Xie, Yuanyuan Xu, Chen Bian, and Min Li. 2017. Semantic congruent audiovisual integration during the encoding stage of working memory: an erp and sloreta study. *Scientific Reports*, 7(1):5112.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, JEAN NYANDWI, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. Pangea: A fully open multilingual multimodal llm for 39 languages. In *International Conference on Learning Representations*, volume 2025, pages 47758–47811.

A Related Works

Multilingual vision-language models. Vision-language models have advanced rapidly (OpenAI et al., 2024; Team et al., 2025), yet most are trained on predominantly English data and lose accuracy on non-English input (Geigle et al., 2025). This gap has driven a wave of multilingual and multicultural systems and evaluations. On the modelling side, efforts such as Centurio (Geigle et al., 2025) and Pangea (Yue et al., 2025) study how language coverage and data mixture shape cross-lingual ability while preserving English performance. On the evaluation side, cross-lingual visual question answering benchmarks such as xGQA (Pfeiffer et al., 2022) and culturally grounded suites such as CVQA (Romero et al., 2024) measure how well models transfer across languages and cultures. A recent survey catalogues 31 models and 21 benchmarks and identifies a persistent tension between language neutrality and faithful cross-lingual behaviour (Manea and Libovický, 2026). These efforts share a common lens, namely they measure multilingual ability through end-task accuracy. We instead probe an internal property, the stability of the entity–attribute binding itself, and show that aggregate accuracy can stay near ceiling while binding strength degrades, which makes downstream scores an incomplete diagnostic for multilingual deployment.

Language for cross-modal understanding. The question of whether the language one uses shapes how one perceives the world predates modern NLP, originating in the linguistic relativity tradition (Whorf, 2012; Kay and Kempton, 1984). A long line of psycholinguistic work has made this concrete in the visual domain, the same setting VLMs operate in. Speakers whose language lexicalises a colour distinction discriminate those colours faster, as shown for Russian light and dark blue (Winawer et al., 2007) and Greek blues (Thierry et al., 2009), with category effects that are lateralised to the language-dominant hemisphere (Gilbert et al., 2006) and that appear even pre-attentively(Roberson et al., 2000; Regier and Kay, 2009). Beyond colour, the language of thought has been argued to influence conceptions of time and other abstract domains (Boroditsky, 2001). Complementarily, neuroscientific evidence indicates that human object representations themselves reflect the co-occurrence statistics of vision and language (Bonner and Epstein, 2021) and that congruent visual and verbal signals are integrated during perception (Xie et al., 2017). Together these findings establish that, for humans, language and visual concepts are deeply entangled yet the underlying concept survives a change of language. Our work asks the machine analogue of this question, namely whether a VLM's internal visual-textual associations are likewise preserved when the language of the input changes.

In-context binding in language and vision-language models. Associating an attribute with the correct entity rather than a competing one is a prerequisite for in-context reasoning. Jiahai Feng and Jacob Steinhardt (2024) formalised this as the *Binding ID* mechanism, showing through causal interventions that language models tag co-referring entity and attribute tokens with a shared latent identifier occupying a low-rank subspace of the residual stream. Subsequent work has refined and extended this picture, localising an ordering component that determines binding behaviour (Dai et al., 2024), decomposing retrieval into distinct mechanisms the model mixes as contexts grow complex (Feng et al., 2025), recovering the underlying circuitry through learned component masks (Davies et al., 2023; Prakash et al., 2024), and relating it to the broader linearity of relation decoding and attribute representation in transformers (Hernandez et al., 2024; Heinzerling and Inui, 2024). Saravanan et al. (2025) extended the Binding ID analysis from text to vision, demonstrating on a synthetic Shapes task that VLMs assign a common identifier to an object's image tokens and its textual mentions. We adopt their task and representative open-weight model as our starting point, and use the same interchange-intervention methodology (Meng et al., 2022; Vig et al., 2020) to locate where binding is computed. Where all of this work establishes binding within a single (English) language, we vary the context and query language independently to test

whether the binding is language-invariant. Part of the fragility we uncover originates before the decoder, in tokenizer unfairness toward non-Latin and morphologically rich scripts (Petrov et al., 2023; Rust et al., 2021), and part inside it, which our layer-wise interventions separate.

B Linguistic Choices in Detail

B.1 Cross-family: Maximal Diversity

Our four cross-family languages were chosen to differ on every major typological axis. We characterize diversity using the World Atlas of Language Structures (WALS; Haspelmath 2009), a database of structural properties of languages compiled from descriptive grammars, covering features such as canonical word order, morphological type, writing system, and segmentation conventions. Fig. 3 shows how each cross-family language is classified along four WALS dimensions. Cells are colored by whether the value is shared with the majority of the set (blue) or divergent (orange); no two languages share the same profile across all four axes.

	En	Fr	Zh	Ar
Word Order	SVO	SVO	SVO	VSO / SVO
Morphology	Weak fusional	Fusional	Analytic	Root-and-pattern
Script	Latin	Latin	Logographic	Abjad (RTL)
Segmentation	Whitespace	Whitespace	Unsegmented	Whitespace

Figure 3: **Typological profile of the four cross-family languages along WALS dimensions.** Each cell shows the categorical classification; traits are either shared with the majority or divergent. No two languages match on all four axes.

English and **French** share SVO (subject-verb-object) order, Latin script, and whitespace segmentation, but differ in morphological type (English has largely lost its inflectional system; French retains richer agreement and gender). **Mandarin** shares SVO order with **English**/**French** but diverges on every other axis: it is analytic (no inflection), logographic (each character maps to a morpheme), and unsegmented (no whitespace between words). **Arabic** diverges most broadly: it permits VSO order, uses root-and-pattern morphology where three-consonant roots are modified by vocalic templates, writes in a right-to-left abjad, yet segments with whitespace like **English**/**French**. These four languages belong to four separate families: Germanic and Romance (both Indo-European), Sinitic (Sino-Tibetan), and Semitic (Afro-Asiatic). Typological distance vectors from `lang2vec` (Littell et al., 2017) confirm near-maximal dispersion in the joint syntactic–phonological–genetic feature space. A speaker of one cross-family language cannot partially comprehend another without formal training (Ringbom and Jarvis, 2009), unlike within-family pairs where shared vocabulary and structure enable partial comprehension (Gooskens et al., 2018).

B.2 Within-family: Controlled Similarity Gradient

To disentangle genealogical proximity from script difference, we add two clusters where all languages share the Latin script but vary in closeness to their anchor language.

Germanic (English, Dutch, German). All three descend from Proto-Germanic origins(Ringe, 2017). **Dutch** and **English** share the closer Ingvaeonic (North Sea Germanic) subgrouping, while **German** underwent the High German consonant shift that systematically altered its stop consonants (e.g., English *water* vs. German *Wasser*). **German** retains a four-case system, three grammatical genders, and verb-final order in subordinate clauses — all absent in modern **English**. **Dutch** occupies an intermediate position: it preserves two grammatical genders but has largely shed case marking and shares SVO order with **English**. Lexical similarity reflects this ordering: **English**–**Dutch** ∼75%, **Dutch**–**German** ∼75%, **English**–**German** ∼60% (Campbell and Grondona, 2008). Gooskens et al. (2018) confirm that **Dutch** speakers can partially comprehend written **English** without instruction, while **English**–**German** intelligibility is measurably lower.

Romance (French, Italian, Spanish). All three descend from Vulgar Latin (Posner, 1996). **Italian** is generally considered the most conservative major Romance language, retaining the greatest lexical and morphological continuity with the common ancestor. **French**–**Italian** lexical similarity (∼89%) is the highest pair in our study, while **French**–**Spanish** (∼75%) is lower due to divergences such as the Latin *f-* → **Spanish** *h-* shift and richer verb inflection in **Spanish** (Campbell and Grondona, 2008). Untrained readers can often extract the gist of a text in a related Romance language from shared Latinate vocabulary alone (Gooskens et al., 2018).

B.3 Lexical Similarity Structure

Fig. 4 shows the full lexical similarity matrix across all eight M^{2}BIND languages. The block-diagonal structure is immediately apparent: Germanic languages form one high-similarity cluster (60–75%), Romance languages form another (75–89%), and both **Mandarin** and **Arabic** show near-zero overlap ($\sim$1%) with every other language. The off-diagonal blocks between Germanic and Romance show modest similarities (20–27%), reflecting their shared but distant Indo-European ancestry. While **English** and **French** have low lexical similarity (27%), extensive Norman–French borrowing into English gives them a shared Latin-script vocabulary that exceeds what the raw percentage suggests, which may explain the relatively high FM (5.20) for this pair despite the cross-family classification.

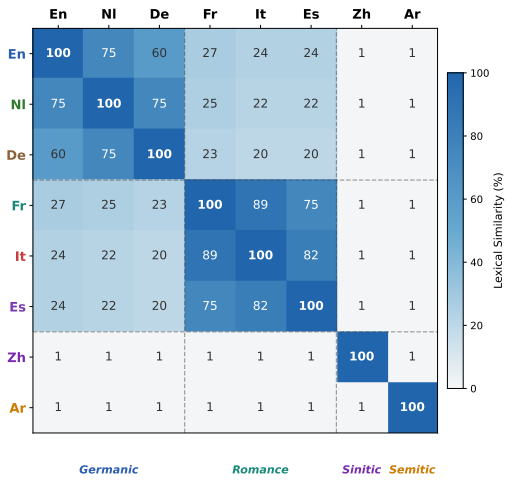

Figure 4: **Lexical similarity (%) across all eight languages in M^{2}BIND**. Dashed lines separate language families.

B.4 Connection to Multilingual Representations

Genealogically, six of our eight languages belong to Indo-European: **English**, **Dutch**, and **German** descend from Proto-Germanic (Ringe, 2017), while **French**, **Italian**, and **Spanish** descend from Vulgar Latin (Posner, 1996). **Mandarin** belongs to Sino-Tibetan and **Arabic** to Afro-Asiatic, placing them in entirely separate families with no shared ancestry. Work on multilingual transformers has shown that models trained on multilingual data develop internal representations where typologically related languages cluster together (Pires et al., 2019). Conneau et al. (2020) demonstrated that cross-lingual transfer improves with shared vocabulary and model scale, while Lauscher et al. (2020) showed that transfer quality degrades with increasing typological distance. These findings predict exactly the gradient visible in Fig. 4: if the VLM's embedding space reflects linguistic distance, within-family pairs should share sufficiently aligned representations to support binding transfer at the same decoder depth, while cross-family pairs should require additional late-layer reconciliation. Our ICE analysis (Fig. 2 in the main paper) confirms this at the mechanistic level.

C Additional results on Qwen2.5

Our main experiments use LLaVA-1.5-OV-7B, whose decoder is itself a Qwen2 model. To test whether the binding behaviour we report depends on that particular stack, we repeat the cross-family protocol on Qwen2.5-VL-7B (Bai et al., 2025), a separately trained open-weight VLM with its own vision encoder and tokenizer. The images, scenes, gold items $\iota_\star$, and the 16 context–query conditions are identical to those in Sec. 3, and we reuse the same sequence score and Factorization Margin (FM).

Cross-family binding. Tab. 4 reports accuracy and FM for Qwen2.5-VL-7B. Two differences from Tab. 1 stand out. First, Qwen2.5-VL-7B is a stronger associator in absolute terms. Accuracy is at ceiling (1.000) across the **English** and **Mandarin** context rows and almost everywhere else, and FM exceeds the corresponding LLaVA cell in all but the two non-Latin monolingual conditions, reaching 6.67 for **English**→**Arabic**. Binding therefore does not collapse on this model in the way it does on LLaVA, and the headline accuracy gives little indication of any residual fragility. Second, the fragility that remains is visible only through the finer-grained signals. Accuracy dips below ceiling only in the **Arabic**-context row (0.950 in its three cross-lingual cells) and in **French**→**Mandarin** (0.975), which points to a non-Latin language, especially on the context side, as the setting that disturbs top-1 selection. In FM, the weakest binding is the monolingual **Mandarin** diagonal (3.71) together with the **Arabic**→**Mandarin** cross-script cell (3.76), so the two non-Latin scripts again mark the lower end of the binding-strength range even though their accuracy stays high.

Comparison with the main model. The agreement between the two models is qualitative rather

L_{ctx} ↓ / L_q →	En	Fr	Zh	Ar
En	1.000 / 5.85	1.000 / 6.15	1.000 / 4.98	1.000 / 6.67
Fr	1.000 / 5.88	1.000 / 6.00	0.975 / 4.60	1.000 / 6.15
Zh	1.000 / 5.31	1.000 / 5.80	1.000 / 3.71	1.000 / 4.85
Ar	0.950 / 4.79	0.950 / 4.52	0.950 / 3.76	0.975 / 4.43

Table 4: **Cross-family binding for Qwen2.5-VL-7B.** Rows are context (ctx), columns are query (q). Each cell reports *Acc. / FM*. For both, higher (↑) is better. Worst cases are highlighted.

	English	French	Mandarin	Arabic
{shape}	cone	cône	圆锥	مخروط
	cube	cube	立方体	مكعب
	cylinder	cylindre	圆柱	أسطوانة
	sphere	sphère	球体	كرة
{color}	red	rouge	红色	أحمر
	blue	bleu	蓝色	أزرق
	green	vert	绿色	أخضر
	yellow	jaune	黄色	أصفر
	cyan	cyan	青色	سماوي
	purple	violet	紫色	أرجواني

Figure 5: **Choices in template.** Colour and shape terms that fill the `{color}` and `{shape}` slots of the context and query, shown for the four cross-family languages (**English**, **French**, **Mandarin**, **Arabic**). The image and the gold item are held fixed across languages; only these surface forms change.

than numerical. Both place the floor of binding strength on the non-Latin scripts, **Mandarin** and **Arabic**, and both leave the **English**–**French** region untouched. The contrast is that LLaVA expresses this fragility as a large FM collapse and an accuracy drop toward 0.90, whereas Qwen2.5-VL-7B absorbs most of it into FM while holding accuracy near ceiling, with the largest accuracy cost appearing under **Arabic** context. This supports our central claim in two ways. The non-Latin scripts are the consistent locus of difficulty across independently trained VLMs, and accuracy alone is an unreliable indicator of binding fidelity, since on Qwen a reader of the accuracy column would conclude that multilingual binding is solved while FM shows the **Mandarin** and cross-script settings remain measurably weaker.

D Template choices in prompting

Each instance is a single prompt built from three parts, all derived from the same scene and illustrated earlier in Fig. 1. The **context** introduces every object by its colour and assigns it an item symbol, using one sentence per object. The **query** names a target object by its shape and asks which item it contains. The **answer prefix** restates the queried shape and ends just before the item, so that the model's next-token distribution over the candidate symbols is exactly the binding we score. The colour and shape words that fill these slots are listed for every cross-family language in Fig. 5.

For example, the scene in Fig. 1 pairs a cyan cube with item P and a yellow cone with item I, and queries the cone. In **English** this yields the context "The cyan object contains item P. The yellow object contains item I.", the query "Which item does the cone contain?", and the answer prefix "Answer: The cone contains item __", whose correct completion is I. The same scene rendered with ℓ_{ctx}=**French** and ℓ_q=**English** keeps the image and the gold answer fixed while swapping only the surface forms, for example the colour *cyan* becomes *cyan* and *yellow* becomes *jaune* in the context, isolating the effect of language on binding.