Do Language Models Benefit from Seeing Text? Visual Representations for Low-Resource Languages

Fred Philippy
Interdisciplinary Centre for Security, Reliability and Trust (SnT), Luxembourg
University of Luxembourg, Luxembourg
fred.philippy@uni.lu

Relevant UniDive working groups: WG3

1 Introduction

Large language models have transformed natural language processing, yet their capabilities remain heavily skewed toward high-resource languages (Joshi et al., 2020; Blasi et al., 2022). A key factor behind this imbalance is the reliance on subword tokenization, which often fails for languages with limited training data or non-Latin writing systems (Rust et al., 2021; Petrov et al., 2023; Ahia et al., 2023). Even minor orthographic differences between closely related languages, such as diacritics, special characters, or slight spelling variations, can produce entirely disjoint token sequences. This fragmentation obscures surface-level lexical overlap that would otherwise serve as a valuable signal for cross-lingual transfer learning (Pires et al., 2019; Wu and Dredze, 2019; Patil et al., 2022). For example, the Luxembourgish sentence *Ech liese Bicher* and its German equivalent *Ich lese Bücher* (EN: *I read many books*) are visually near-identical, yet a standard tokenizer may split them into completely different subword sequences, preventing the model from recognizing their similarity.

At the same time, recent progress in multimodal learning has shown that neural systems can effectively integrate textual and visual information (Wang et al., 2024). This raises a natural question: can images of rendered text preserve orthographic structure, including character shapes and visual similarities across scripts, that tokenization discards?

2 Related Work

Prior work has explored visual text representations in various forms, from CNN-based glyph features for Chinese (Liu et al., 2017; Dai and Cai, 2017; Sun et al., 2021) to the pixel-based PIXEL model (Rust et al., 2023), as well as extensions to historical documents (Borenstein et al., 2023), autoregressive generation (Tai et al., 2024), multilingual translation (Salesky et al., 2023), and dialect transfer (Muñoz-Ortiz et al., 2025; Kesen et al., 2025). However, these approaches typically rely on specialized architectures pretrained from scratch. In contrast, this work investigates whether existing, off-the-shelf vision-language models can extract useful information from rendered text without additional training or model adaptation.

3 Approach

Rather than proposing a new architecture or performing any training, this work leverages existing models to test this hypothesis. Textual embeddings from four text embedding models[1] are combined with visual embeddings from three image embedding models[2]. For each sentence, both a text embedding and an image embedding (derived from a rendered image of the sentence) are computed and ℓ_2-normalized. The final similarity between two sentences is then computed as $s(x, y) = (1 - \alpha) \cdot \cos(\mathbf{t}_x, \mathbf{t}_y) + \alpha \cdot \cos(\mathbf{v}_x, \mathbf{v}_y)$, where $\mathbf{t}$ and $\mathbf{v}$ denote text and visual embeddings respectively, and $\alpha \in [0, 1]$ controls the weighting. The evaluation uses the FloresBitextMining task from MTEB (Enevoldsen et al., 2025), which spans 203 languages covering a broad range of scripts and typologically diverse language families. Performance is measured using Recall@1.

4 Key Findings

The results reveal that incorporating visual embeddings generally improves bitext mining performance at moderate α values (Figure 1).

When languages in a pair use different scripts (***cross-script***), those written in Cyrillic or CJK tend to gain the most from incorporating image embeddings. However, when both languages share the same script (***same-script***), Latin-script languages see the largest improvements from image embeddings, followed by Cyrillic languages, while CJK languages show almost no improvement.

[1] LaBSE (Feng et al., 2022); embeddinggemma-300m (Vera et al., 2025); jina-embeddings-v3 (Sturua et al., 2024); Qwen3-VL-Embedding-2B (Bai et al., 2025)

[2] Qwen3-VL-Embedding-2B (Bai et al., 2025); SigLIP 2 So400m (Tschannen et al., 2025); clip-ViT-B-32 (Radford et al., 2021)

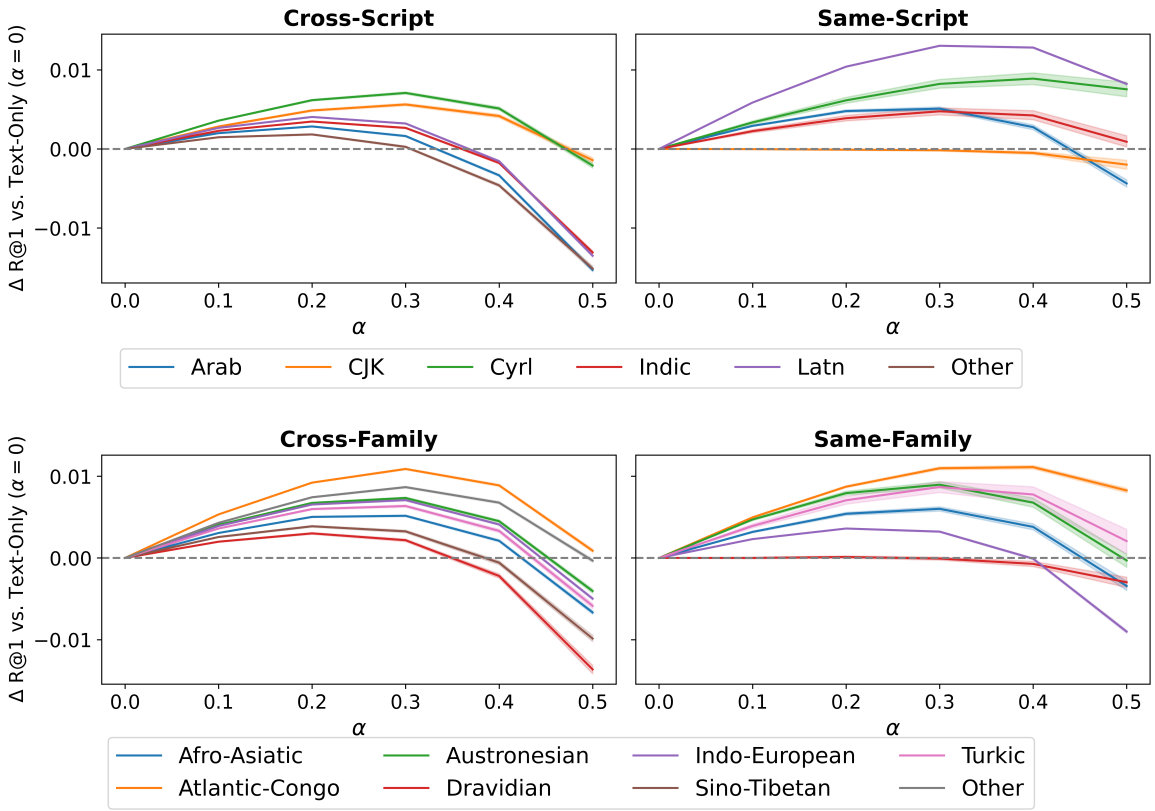

Figure 1: Average absolute change in bitext mining performance (R@1) relative to the text-only baseline (i.e., $\alpha = 0$), aggregated across models by major script families (top) and language families (bottom), as a function of the image embedding weight α. For clarity, we only report results for $\alpha \in [0, 0.5]$, as average performance consistently degrades beyond this range.

Regardless of whether the language pair belongs to the same or different families, Atlantic-Congo languages benefit the most from visual signals, while Dravidian languages benefit the least. Languages from smaller families grouped under "Other" tend to achieve some of the highest average gains in ***cross-family*** retrieval, as seen in pairs like Uralic with Atlantic-Congo, Nilotic with Indo-European, or Mande with Turkic. Additionally, Austronesian and Turkic languages gain slightly more from visual embeddings in ***same-family*** settings, whereas Indo-European languages improve more in cross-script settings.

Figure 2 illustrates how different text and image embedding combinations influence performance. While most combinations yield improvements, the magnitude varies: `Qwen2-VL` provides the largest gains as an image encoder but benefits the least when used for text embeddings. Despite these differences in scale, a high average Spearman correlation of 0.87 across all combinations confirms that the performance trends are robust and largely independent of the specific models chosen.

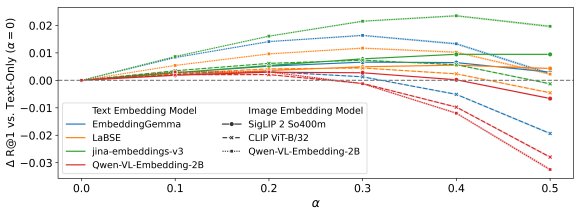

Figure 2: Average absolute change in bitext mining performance (R@1) relative to the text-only (i.e., $\alpha = 0$) baseline, aggregated across models by combinations of text and image embedding models, as a function of the image embedding weight α.

5 Discussion and Implications

From a practical standpoint, the findings highlight a clear set of conditions in which visual input is most effective: low-resource languages that use Latin or Cyrillic scripts. This includes languages from families such as Atlantic-Congo, Austronesian, Mandé, and Uralic, where the Latin script is dominant. Languages using CJK writing systems can also see modest gains, but only when retrieving across different writing systems rather than within the CJK group itself. In contrast, lan-

guages written in abugida or abjad scripts show little benefit from visual representations when using off-the-shelf models.

A central design choice in this study was to keep the experimental setup intentionally simple. By using linear embedding fusion without extensive hyperparameter tuning, we ensure that observed improvements reflect the contribution of the visual signal rather than engineering optimizations. The reported gains should therefore be seen as a conservative lower bound, and more advanced multimodal fusion, script-aware rendering, or larger vision-language models would likely yield even greater benefits.

More broadly, the findings show that current vision-language models already possess enough visual capability to support low-resource languages without requiring specialized architectures or extra fine-tuning. This highlights the promise of multimodal approaches that rethink how textual information is encoded and processed, with the broader aim of making NLP systems more equitable across the world's linguistic diversity.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-vl technical report.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Nadav Borenstein, Phillip Rust, Desmond Elliott, and Isabelle Augenstein. 2023. PHD: Pixel-based language modeling of historical documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 87–107, Singapore. Association for Computational Linguistics.

Falcon Dai and Zheng Cai. 2017. Glyph-aware embedding of Chinese characters. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 64–69, Copenhagen, Denmark. Association for Computational Linguistics.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. Mmteb: Massive multilingual text embedding benchmark.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Ilker Kesen, Jonas F. Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. 2025. Multilingual pretraining for pixel language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29594–29611, Suzhou, China. Association for Computational Linguistics.

Frederick Liu, Han Lu, Chieh Lo, and Graham Neubig. 2017. Learning character-level compositionality with visual features. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2059–2068, Vancouver, Canada. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, Verena Blaschke, and Barbara Plank. 2025. Evaluating pixel language models on non-standardized languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6412–6419, Abu Dhabi, UAE. Association for Computational Linguistics.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36, pages 36963–36990. Curran Associates, Inc.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels.

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.

Elizabeth Salesky, Neha Verma, Philipp Koehn, and Matt Post. 2023. Multilingual pixel representations for translation and effective cross-lingual transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13845–13861, Singapore. Association for Computational Linguistics.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. ChineseBERT: Chinese pretraining enhanced by glyph and Pinyin information. 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 2065–2075, Online. Association for Computational Linguistics.

Yintao Tai, Xiyang Liao, Alessandro Suglia, and Antonio Vergari. 2024. PIXAR: Auto-regressive language modeling in pixel space. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14673–14695, Bangkok, Thailand. Association for Computational Linguistics.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, Weiyi Wang, Zhe Li, Gus Martins, Jinhyuk Lee, Mark Sherwood, Juyeong Ji, Renjie Wu, Jingxiao Zheng, Jyotinder Singh, Abheesht Sharma, Divyashree Sreepathihalli, Aashi Jain, Adham Elarabawy, AJ Co, Andreas Doumanoglou, Babak Samari, Ben Hora, Brian Potetz, Dahun Kim, Enrique Alfonseca, Fedor Moiseev, Feng Han, Frank Palma Gomez, Gustavo Hernández Ábrego, Hesen Zhang, Hui Hui, Jay Han, Karan Gill, Ke Chen, Koert Chen, Madhuri Shanbhogue, Michael Boratko, Paul Suganthan, Sai Meher Karthik Duddu, Sandeep Mariserla, Setareh Ariafar, Shanfeng Zhang, Shijie Zhang, Simon Baumgartner, Sonam Goenka, Steve Qiu, Tanmaya Dabral, Trevor Walker, Vikram Rao, Waleed Khawaja, Wenlei Zhou, Xiaoqi Ren, Ye Xia, Yichang Chen, Yi-Ting Chen, Zhe Dong, Zhongli Ding, Francesco Visin, Gaël Liu, Jiageng Zhang, Kathleen Kenealy, Michelle Casbon, Ravin Kumar, Thomas

Mesnard, Zach Gleicher, Cormac Brick, Olivier Lacombe, Adam Roberts, Qin Yin, Yunhsuan Sung, Raphael Hoffmann, Tris Warkentin, Armand Joulin, Tom Duerig, and Mojtaba Seyedhosseini. 2025. Embeddinggemma: Powerful and lightweight text representations.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024. A comprehensive review of multimodal large language models: Performance and challenges across different tasks.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.