

MultiInpaintCOCO: Fine-Grained Visual Understanding in Multiple Languages

Andrei-Alexandru Manea¹ and Jindřich Libovický¹ and Philipp J. Rösch²

¹Faculty of Mathematics and Physics, Charles University, Czech Republic

²Bundeswehr University Munich, Germany

{manea, libovicky}@ufal.mff.cuni.cz philipp.roesch@unibw.de

Abstract

Previous research shows that vision-language (VL) models struggle to understand fine-grained image details and tend to focus on the overall layout. Moreover, any weakness in English tends to be further amplified in the multilingual setting. To study that, we extend InpaintCOCO, an English dataset focused on fine-grained visual details, to 13 typologically diverse languages across five language families, creating **MultiInpaintCOCO** – a new multilingual benchmark of 953 minimal-pair examples. We further test several state-of-the-art models in multilingual multimodal retrieval, considering both light VL CLIP-like encoders and large language models. Our results show that performance degrades substantially for lower-resource languages, with the strongest predictor being the data resources available for the languages rather than typological similarity to English. In cross-lingual retrieval, most models exhibit a systematic bias toward English: when given a correct caption in a non-English language and an incorrect one in English, models preferentially select the English caption. This bias varies considerably across architectures. Among the tested models, Gemma 3 shows the strongest English preference while Qwen 3 is the least biased, despite both reaching similar monolingual accuracy across languages. The code is available at <https://github.com/ufal/multi-inpaint-coco>.

1 Introduction

Vision-language (VL) models take both images and text as input. Depending on the task, two main types of models are currently used. *Dual encoders*, such as CLIP (Radford et al., 2021) and SigLIP 2 (Tschannen et al., 2025), encode images and texts independently into a shared embedding space; they are the standard choice for retrieval and similarity tasks. *Decoder-only models*, such as Gemma 3 (Team et al., 2025) and Qwen 2.5-VL (Bai et al.,

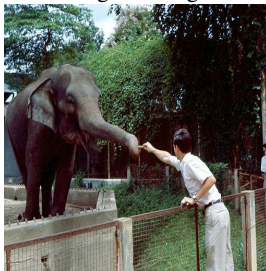
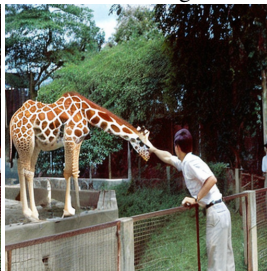
	Original image	Edited image
		
en	A man is feeding an elephant over a fence	A man is feeding a giraffe over a fence
cs	Muž krmí slona přes plot	Muž krmí žirafu přes plot
jp	男性が柵越しに象に餌をあげている	男性が柵越しにキリンに餌をあげている
az	Bir kişi hasarın üstündən bir fili yemləyir	Bir kişi hasarın üstündə zürafəni qidalandırır

Figure 1: Example from the MultiInpaintCOCO dataset. The left image and English caption are from the COCO dataset, the right equivalent part is from InpaintCOCO.

2025b), are decoder-based and generate text conditioned on both visual and textual input. Both model types are pre-trained on large image-text corpora using objectives (contrastive or generative) that do not require distinguishing between images or captions that differ only in fine-grained details.

Several studies have shown that VL models fail to capture fine-grained visual information and instead rely on coarse features, such as global scene layout or the most salient objects (Fu et al., 2025). To probe this limitation precisely, Rösch et al. (2024) proposes minimal pairs: each example consists of two images that differ by a single inpainted object and two captions that differ by exactly one keyword, so that the correct answer cannot be reached without identifying the specific visual detail. A similar strategy applied at the caption level underlies benchmarks such as SugarCrepe (Dumpala et al., 2024). The image-editing approach has a key advantage: it introduces genuine visual ambiguity, because nei-

ther the image nor the caption alone is sufficient to determine the correct match.

All of this work is English-only. It is well established that VL model performance drops substantially when moving from English to lower-resource languages (Liu et al., 2021; Bugliarello et al., 2022a), and that multilingual models are not equally calibrated across languages. Errors in less-resourced languages may therefore not only be more frequent but also qualitatively different. The combination of fine-grained visual understanding and multilingual evaluation is therefore almost entirely unexplored. To close this gap, we extend InpaintCOCO (Rösch et al., 2024) to 13 typologically diverse languages (Azerbaijani, Arabic, Czech, German, Greek, Hindi, Italian, Japanese, Romanian, Russian, Slovak, Ukrainian, Vietnamese), producing the MultiInpaintCOCO dataset. Captions were first translated automatically and then post-edited by native speakers.

We evaluate several state-of-the-art dual-encoder and generative decoder models on MultiInpaintCOCO and report two main findings. First, performance on lower-resource languages is substantially worse than on English and other high-resource languages, confirming that fine-grained retrieval is harder in the multilingual setting. Second, and more surprisingly, when a model is given one correct caption in a non-English language and one incorrect caption in English, most models systematically prefer the English caption. This bias appears in both dual-encoder and generative decoder-only models, with Azerbaijani the least favored language. The strength of the bias varies considerably across models: Gemma 3 exhibits the strongest preference for English, while Qwen 3 is the least biased. This is not just a reflection of lower overall accuracy in weaker languages. It reveals a systematic preference for English that persists even when the model is otherwise capable of making the correct match.

2 Dataset Description

MultiInpaintCOCO extends the English InpaintCOCO dataset (Rösch et al., 2024) to 13 more languages, producing a multilingual minimal-pair benchmark for fine-grained vision-language retrieval. The dataset covers 14 languages in total and spans five language families, with translations produced automatically and then post-edited by native speakers.

2.1 InpaintCOCO

InpaintCOCO (Rösch et al., 2024) is a minimal-pair dataset built on top of COCO (Lin et al., 2014). For each example, a keyword is replaced in the original caption, and the corresponding image region is edited using Stable Diffusion inpainting (Rombach et al., 2022), producing a pair of images that differ in a single localized detail and a pair of captions that reflect that difference. This construction ensures that a model can only retrieve the correct image-caption match by attending to fine-grained visual content rather than overall scene similarity.

Before translation, we manually reviewed all 1,260 examples in InpaintCOCO and removed or corrected low-quality instances. We discarded examples in which the inpainted image contained visible artifacts or the edit was too subtle to warrant a different caption, and we corrected captions that were inaccurate or inconsistent with the edited image. After this review, 953 examples remained.

2.2 Translation Process

For each language, captions were first translated automatically and then post-edited by a native speaker. Three translation systems were available to the annotators: Hunyuan (Zheng et al., 2025) (the WMT25 winner; Kocmi et al., 2025), Google Translate, and GPT-4. Annotators were free to choose whichever system they found most useful as a starting point and were instructed to correct any errors in the output. All annotators were native speakers of the target language with an excellent command of English. The translation into Russian was post-edited, with German as the source language. More details about the annotation process are in the Appendix A.

2.3 The MultiInpaintCOCO Dataset

The final dataset covers 14 languages across five language families: Indo-European, Japonic, Turkic, Austroasiatic, and Afro-Asiatic, with Indo-European being the most represented (10 languages). The full list of languages with their metadata is given in Table 7. The languages vary substantially in their resource levels: following the tiers of Joshi et al. (2020), the dataset includes four tier-5 languages (English, German, Japanese, Arabic), five tier-4 languages, four tier-3 languages, and one tier-1 language (Azerbaijani). Morphologically, the majority are fusional (10), with two agglutinative languages (Japanese, Azerbaijani) and two isolating languages (English, Vietnamese).

The disparity in the availability of pretraining data across our languages is stark. In FineWeb2 (Penedo et al., 2025), a 20 TB multilingual web corpus covering over 1,000 languages, Russian alone accounts for 6,400 GB, while German and Japanese each contribute around 1,650 GB. At the other extreme, Azerbaijani has only 29 GB of data, roughly 220 times less than Russian. Slovak (92 GB) and Hindi (130 GB) also fall well below the high-resource languages despite having tens of millions of native speakers. This orders-of-magnitude gap in raw text data directly motivates our expectation (and findings) of performance differences across languages in our experiments.

Table 9 places MultiInpaintCOCO in the context of existing multilingual vision-language benchmarks (Liu et al., 2021; Nooralahzadeh and Senrich, 2023; Bugliarello et al., 2022b; Romero et al., 2024; Pfeiffer et al., 2022; Schneider and Sitaram, 2024). Most existing benchmarks are either translation-based with a narrow language selection or culturally grounded but limited in scale. MultiInpaintCOCO occupies a distinct position: it is translation-based, enabling direct cross-lingual comparison, and the only benchmark that combines caption-level and image-level minimal pairs, allowing retrieval errors to be attributed to either the text or the visual representation. In terms of linguistic diversity, measured by the Grambank feature inclusion index (GBI; Ploeger et al., 2024), our dataset scores 79, which is higher than all other translation-based benchmarks in Table 9, such as xFlickrCo (Bugliarello et al., 2022b) (65) or XTD (Aggarwal and Kale, 2020) (71), and comparable to culturally grounded ones such as KnowRecall (Wang et al., 2025) (77.94) and CVQA (Romero et al., 2024) (88). Still, MultiInpaintCOCO’s 14 languages represent only 19% of the world’s population when considering languages spoken as first languages.

3 Multilingual Experiments

We evaluate both dual-encoder and decoder-only models on MultiInpaintCOCO across all 14 languages to assess how well current VL models handle fine-grained visual details in a multilingual setting. Beyond raw model performance, we measure consistency across languages and across models.

3.1 Experimental Setup

Task formulation. Each example in MultiInpaintCOCO consists of an original image–caption pair

and an edited image–caption pair. We construct two retrieval subtasks from this structure. In the *two-image* task (*2img*), the model is given one caption and two images and must select the image that matches the caption. In the *two-caption* task (*2txt*), the model is given one image and two captions and must select the caption that describes the image. In both cases, either the original or the edited item serves as the query, while the two remaining items of the other modality serve as the context.

Dual encoders. Dual encoders map images and texts independently into a shared embedding space. For each example, we embed both images and both captions and compute all pairwise cosine similarities. The predicted match is the image (or caption) that is most similar to the query. This evaluation protocol follows Rösch et al. (2024).

Decoder-only models. Decoder-only models process inputs sequentially. We present the model with both options (*two-image* or *two-caption*) and a prompt asking which option is correct. The instruction prompt (see Appendix B) is written in English, regardless of the language of the captions. We are aware that this practice might amplify the English bias. To explore this, we also test with French (a language not in the current dataset), and the *strict accuracy* is generally lower (Table 23), but the differences between languages follow the same patterns. We therefore decide to stick with the English prompts. Because decoder-only models are known to be sensitive to input order (Wardle and Susnjak, 2024; Tian et al., 2025), we evaluate each example in both orders and require both to be correct, yielding eight comparison configurations per example in total (see Table 11). A detailed evaluation of the position bias of the models used in this paper is provided in Appendix C.

Model performance. We report *strict accuracy* as a primary metric. A dataset item is counted as correct only if the model answers correctly across all comparison configurations: four for order-invariant dual encoders and eight for decoder-only models.

Formally, each example consists of an original image–caption pair (I_o, T_o) and an edited pair (I_e, T_e) . We construct four binary correctness indicators: $c(I_o, \{T_o, T_e\})$ and $c(I_e, \{T_o, T_e\})$ for the *2txt* task, and $c(T_o, \{I_o, I_e\})$ and $c(T_e, \{I_o, I_e\})$ for the *2img* task, where $c(q, \{a, b\}) = 1$, if the model selects the correct item from the pair given the

query q . For the sake of notation simplicity, we consider that $c(q, \{a, b\})$ contains the results in both input orders for decoder-only models; therefore, $c(q, \{a, b\}) = c(q, (a, b)) \cdot c(q, (b, a))$.

The *2img* and *2txt* accuracies are the averages of the corresponding indicators over all examples. *Strict accuracy* requires all four to be correct:

$$c_{2img} = c(T_o, \{I_o, I_e\}) \cdot c(T_e, \{I_o, I_e\}) \quad (1)$$

$$c_{2txt} = c(I_o, \{T_o, T_e\}) \cdot c(I_e, \{T_o, T_e\}) \quad (2)$$

$$c_{strict} = c_{2img} \cdot c_{2txt} \quad (3)$$

and *strict accuracy* is the average of c_{strict} over all examples.

Factors influencing model performance. Because we do not have access to the training data or methodology of the models we study, we can only partially explain cross-lingual performance differences. To assess that, we compute the Spearman correlation between *strict accuracy* and a set of language-level variables listed in Table 2: resource tier (Joshi et al., 2020), number of native speakers according to Wikidata¹ (Vrandečić and Krötzsch, 2014), typological similarity to English and Chinese from the URIEL+ database (Khan et al., 2025), number of Wikipedia articles, and the amount of available text in the FineWeb2 corpus (Penedo et al., 2025). Further, we consider the number of tokens into which our dataset is tokenized per language (see Table 10 in the Appendix).

Language agreement. Similar accuracy scores across languages do not necessarily mean that models make the same errors. To measure how similar the error patterns are across languages, we compute Cohen’s κ , a standard metric for inter-annotator agreement, between the per-language model predictions. The agreement between two languages, l_1 and l_2 , inside the model m is represented by $C(m, l_1, l_2)$, and the average agreement across languages $A(l_1, l_2)$ is computed as follows:

$$C(m, l_1, l_2) = \kappa(\hat{y}_m(l_1), \hat{y}_m(l_2)), \quad (4)$$

$$A(l_1, l_2) = \frac{1}{|M|} \sum_m C(m, l_1, l_2) \quad (5)$$

where $\hat{y}_m(l)$ is the output of the model m , in language l . A value of 0 indicates that two languages make errors on completely different examples, while a value of 1 means their error patterns are identical.

¹<https://www.wikidata.org/wiki/Property:P1098>

Model agreement. We use the same method to compare error patterns across models. For two models m_1 and m_2 , in the language l , we compute the individual agreement $C(m_1, m_2, l)$ and the average agreement $A(m_1, m_2)$ as:

$$C(m_1, m_2, l) = \kappa(\hat{y}_{m_1}(l), \hat{y}_{m_2}(l)), \quad (6)$$

$$A(m_1, m_2) = \frac{1}{|L|} \sum_l C(m_1, m_2, l). \quad (7)$$

3.2 Tested Models

We evaluate two model families: decoder-only generative VL models and dual-encoder CLIP-like models. Tables 12 and 8 provide additional model details.

Decoder-only models. We evaluate five open-weight VL models with a parameter count ranging from 2B to 12B. Gemma 3 (Team et al., 2025) (12B parameters), Qwen VL 2.5 (Bai et al., 2025b) (7B), Qwen VL 3 (Bai et al., 2025a) (8B), Aya Vision (Dash et al., 2025) (8B), and Jina VL (Koukounas et al., 2026) (2B), the smallest model in our evaluation.

Dual encoder models. We evaluate seven CLIP-like models (Radford et al., 2021), all based on SigLIP (Zhai et al., 2023). This includes NLLB-SigLIP (Visheratin, 2023) in base and large sizes, MEXMA-SigLIP 2, and SigLIP 2 (Tschannen et al., 2025) from base to giant sizes. The vision encoders of SigLIP 2 are also used in Gemma and Qwen. The NLLB versions consist of the first SigLIP model, pre-trained across all languages in NLLB-200 (Team et al., 2022). MEXMA-SigLIP 2 combines the MEXMA text encoder (Janeiro et al., 2024) with the SigLIP visual backbone.

3.3 Results

Model performance. We report the *strict accuracy* in Table 1. Table 14 in the appendix contains the accuracy for the *2img* task, and Table 15 for *2txt*. On average, deciding between two images is slightly harder than deciding between two texts for all the models. This is expected, given that in the *2txt* task, the query image is fixed across languages, and errors are driven mainly by shared visual ambiguity. However, in the *2img* setup, the query is language-specific, so the per-language text-encoding quality introduces additional variation. The Jina model has the highest modality performance gap, while Qwen 3 has the lowest. Azerbaijani (tier-1) has

Model	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
Gemma 3 12B	.581	.425	.507	.581	.587	.569	.523	.580	.510	.570	.610	.472	.594	.599	.551
Qwen 2.5 VL 7B	.595	.335	.586	.690	.571	.714	.528	.664	.607	.571	.686	.544	.630	.621	.596
Qwen 3 VL 8B	.677	.504	.654	.700	.705	.723	.695	.705	.686	.678	.698	.623	.677	.685	.672
Aya Vision 8B	.244	.125	.296	.410	.407	.496	.301	.366	.335	.359	.379	.225	.346	.424	.337
Jina VL	.017	.002	.005	.034	.015	.019	.007	.021	.022	.012	.010	.007	.015	.033	.016
NLLB-SigLIP Base	.665	.338	.381	.488	.661	.678	.651	.478	.635	.450	.682	.405	.654	.516	.549
NLLB-SigLIP Large	.681	.443	.570	.599	.689	.699	.698	.643	.656	.608	.685	.573	.671	.551	.626
MEXMA-SigLIP2	.653	.657	.668	.687	.654	.678	.684	.677	.642	.677	.679	.671	.674	.654	.668
SigLIP2 Base	.630	.339	.645	.686	.587	.719	.501	.674	.450	.591	.687	.620	.623	.641	.599
SigLIP2 Large	.639	.397	.672	.709	.634	.722	.617	.696	.460	.639	.698	.643	.655	.668	.632
SigLIP2 SO400M	.633	.427	.664	.705	.659	.705	.611	.689	.511	.661	.698	.659	.660	.653	.638
SigLIP2 Giant	.627	.379	.674	.679	.641	.684	.573	.672	.490	.655	.677	.659	.650	.660	.623
Avg.	.554	.364	.527	.581	.567	.617	.532	.572	.500	.539	.599	.508	.571	.559	.542

Table 1: *Strict accuracy* for bidirectional retrieval, i.e. c_{strict} from Equation (3). The upper part of the tables contains LLMs, while the lower part shows the dual encoders. The breakdown into *2img* and *2txt* accuracy is reported in the Appendix in Tables 14 and 15.

the lowest performance, which is correlated with it being the least-resourced language we have. On the other hand, with dual encoders that rely on the SigLIP 2 text encoder, we observe a performance drop for Japanese, a high-resource language (tier-5). Moreover, for the same models, the gap between the *2img/2txt* accuracies and *strict accuracy* is unusually large for Hindi. Finally, MEXMA-SigLIP 2 performs best on Azerbaijani.

Statistical significance. To verify the robustness of our findings, we estimated 95% confidence intervals using bootstrap resampling with 10K samples (Table 22). With 953 examples per language, the intervals are roughly ± 3 percentage points, and the main effects discussed above are statistically significant: the drop for Azerbaijani (with MEXMA-SigLIP 2 significantly outperforming all other models there), the drop for Japanese for SigLIP 2, and the English lead for Aya Vision. The overlapping intervals appear mostly among higher-resource languages and among the strongest models, where performance is similar: e.g., for Qwen 3, all languages except Azerbaijani, Czech, and Slovak overlap with English, and for MEXMA-SigLIP 2, all 14 languages overlap.

Factors influencing model performance. Table 2 shows Spearman correlations between *strict accuracy* and a set of language-level variables. The strongest predictor across models is the amount of text in the FineWeb2 corpus ($\bar{\rho} = .56$), an open-source web crawl derived from CommonCrawl that reflects the amount of data realistically available for a given language on the web. The number of Wikipedia articles follows closely ($\bar{\rho} = .53$) and

can be understood as a proxy for the same underlying quantity. Even though the exact amount and composition of the training data are unknown to the public, these two proxy variables suggest that the volume of pretraining data might be the dominant factor in cross-lingual performance differences.

Typological similarity to English correlates moderately ($\bar{\rho} = .46$) and less strongly than data-size proxies, indicating that the volume of training data matters more than cross-lingual transfer from English. The notably low correlation between English similarity and performance for NLLB-SigLIP models (.18 for both Base and Large) is consistent with the finding in Section 4 that these models are less English-centric than the rest. For models with a stronger Chinese component (Qwen family), similarity to Chinese correlates slightly more than similarity to English; averaged over all models, however, the correlation with similarity to Chinese is only .34.

Languages that require fewer tokens tend to perform better, especially in the SigLIP 2 family and Jina VL, but not in the MEXMA and NLLB families. Among the two largest decoder-only models, this effect is relatively strong for Gemma 3 but much weaker for Qwen 3 VL. However, the average correlation of .30 shows that token count remains only a rough proxy. A notable counterexample, shown in Table 10, is that the Qwen tokenizers overtokenize Greek and Hindi by $2.8\times$ compared to English, yet these languages perform comparably to other languages. Meanwhile, Azerbaijani, despite having relatively long tokens, performs much worse.

Finally, Joshi’s resource tiers are the weakest explanatory factor ($\bar{\rho} = .34$). This is likely for two

reasons: the tier classification partly reflects the availability of annotated linguistic resources (treebanks, lexicons, and so on), which are not directly relevant for vision-language pretraining; and the taxonomy is now six years old (Joshi et al., 2020) and might not reflect current data availability. The total number of native speakers is slightly more relevant ($\bar{\rho} = .37$).

Language agreement. The average agreement between languages is plotted in Table 3. The per-model breakdown is in the appendix (Table 19). Above the diagonal, the correlations within the *2img* task are considerably lower than those in the *2txt* task below. All the languages agree the least with Azerbaijani, making it an outlier language. All four SigLIP 2 models have a similar disregard for Japanese. On the other hand, the Qwen family of models shows the highest correlations, while Jina shows the lowest, which suggests a chaotic pattern in its output error. The individual model tables show no language outliers for Gemma 3, Qwen 3, and MEXMA-SigLIP 2. Jina pattern is totally random in the *2img* task, and Czech is an outlier in the NLLB-SigLIP base *2txt* setup.

Model agreement. The average agreement between models is plotted in Table 4. The per-language breakdown is in the appendix (Table 20 and Table 21). The average agreement is not as consistent as it is between languages. Gemma 3 and the Qwen family of models agree with each other. So do the SigLIP 2 models, but not with the NLLB-SigLIP. The highest correlation numbers are found in English, and the lowest in Azerbaijani. Moreover, in the Azerbaijani table, models do not agree with each other at all. In Japanese, the disagreement between SigLIP 2 models differs from that of other models.

4 Cross-lingual Experiments

Unlike the previous set of experiments, where we tested the models monolingually across all languages, here we test them in a more challenging *two-text* setup, in which each text is in a different language. This allows us to evaluate whether the models exhibit biases. We design a set of cross-lingual experiments to measure whether models are biased toward better-resourced languages.

4.1 Experimental Setup

In these experiments, the model is given a single image together with two captions in different lan-

guages (one correct and one incorrect) and must identify which caption matches the image. We use the same evaluation protocol for dual encoders and decoder-only models as in Section 3. In addition to accuracy (equivalent to the *2txt* task in Section 3), we report differences in per-language mistakes.

For two languages x and y , the mistake $M(x, y)$ measures how often the model makes an error by preferring y over x when x is correct. If the model makes mistakes randomly, $M(x, y) \approx M(y, x)$, and the difference $M_{\text{dif}}(x, y)$ is close to zero. A positive value for $M_{\text{dif}}(x, y)$ therefore indicates a systematic bias in favor of language x , while a negative value indicates a bias towards the language y . More concretely:

$$M(x, y) = \mathbb{0}_{c(I_o^x, \{T_o^x, T_e^y\})} + \mathbb{0}_{c(I_e^x, \{T_o^y, T_e^x\})} \quad (8)$$

$$M_{\text{dif}}(x, y) = M(y, x) - M(x, y) \quad (9)$$

where $\mathbb{0}$ is the error indicator function.

4.2 Results

We report the detailed *2txt* accuracy in this bilingual setup in Table 16 in the appendix. Table 5 presents the language-specific mistakes averaged across all models (above the diagonal; below are the negatives, making the matrix antisymmetric). The values in the table are $M_{\text{dif}}(x, y)$ divided by the total number of samples. For completeness, we also compute the same table with the French prompts in Table 24. We notice that the bias patterns are very similar to those in the English-prompt condition.

Overall trends. Averaged across models, there is a clear bias toward English: models systematically prefer an incorrect English caption over a correct one in another language. Azerbaijani is the most disfavored language overall, with Czech and Slovak also showing a mild negative bias.

Decoder-only models. The least biased model is Qwen 3 with close-to-zero language-specific mistakes in all languages, as displayed in Table 17. At the opposite extreme, Gemma 3 shows a stronger preference for English than other models, despite its competitive *strict accuracy*. Gemma disfavors languages using non-Latin scripts (Arabic, Japanese, Ukrainian, Greek, Hindi) as well as Azerbaijani. Aya and Qwen 2.5 show intermediate levels of English bias without strong script-related patterns. Jina VL exhibits an unusual pattern: in addition to preferring English, it strongly favors Hindi while systematically disfavoring Slavic languages, regardless of script.

Property	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	MEXMA-SigLIP2	NLLB-SigLIP Base	NLLB-SigLIP Large	SigLIP2 Base	SigLIP2 Large	SigLIP2 SO400M	SigLIP2 Giant	Average
FineWeb2 (GB)	.76	.58	.75	.45	.36	.21	.62	.39	.72	.68	.58	.60	.56
Wikipedia Articles	.73	.43	.70	.21	.29	.61	.47	.42	.66	.66	.56	.56	.53
Similarity to en	.60	.36	.42	.54	.25	.68	.18	.18	.49	.58	.62	.55	.46
Similarity to zh	.51	.46	.44	.45	.10	.53	.37	.40	.13	.19	.32	.20	.34
Number of tokens	.38	.27	.22	.14	.66	-.08	-.24	-.12	.64	.62	.53	.59	.30
Joshi's Tier	.52	.22	.50	-.04	.53	.08	.44	.33	.50	.42	.27	.35	.34
Speakers (M)	.70	.46	.61	-.01	.33	.30	.70	.63	.32	.22	.10	.12	.37

Table 2: Spearman correlation of the *strict accuracy* with the language properties.

	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
ar	—	.58	.73	.77	.76	.79	.75	.78	.72	.75	.80	.73	.78	.76
az	.64	—	.57	.58	.57	.58	.57	.59	.54	.58	.60	.59	.59	.56
cs	.76	.62	—	.76	.72	.76	.70	.76	.68	.74	.78	.79	.76	.72
de	.82	.64	.79	—	.77	.83	.75	.82	.73	.76	.83	.75	.80	.77
el	.82	.63	.76	.82	—	.78	.76	.77	.71	.74	.79	.71	.77	.76
en	.84	.64	.79	.87	.85	—	.76	.82	.74	.77	.84	.75	.80	.78
hi	.81	.64	.75	.80	.80	.83	—	.74	.72	.71	.76	.70	.75	.73
it	.83	.64	.78	.85	.82	.86	.80	—	.72	.77	.82	.75	.79	.77
ja	.78	.63	.71	.77	.77	.80	.78	.77	—	.68	.74	.67	.72	.71
ro	.80	.64	.77	.81	.80	.82	.78	.82	.74	—	.78	.74	.76	.73
ru	.84	.65	.80	.86	.84	.88	.82	.85	.80	.82	—	.77	.84	.78
sk	.78	.64	.81	.80	.78	.80	.76	.79	.73	.79	.81	—	.76	.71
uk	.83	.65	.79	.84	.83	.85	.81	.83	.78	.81	.88	.80	—	.76
vi	.82	.62	.75	.82	.81	.84	.79	.82	.76	.79	.83	.77	.82	—
Avg.	.80	.63	.75	.80	.78	.80	.76	.78	.72	.75	.79	.73	.76	.73

Table 3: Languages’ agreement as Cohen’s Kappa coefficient averaged over all models. Above the diagonal: *2img* task. Below the diagonal *2txt* task.

Dual encoder models. Dual encoders show less English bias overall (Table 18) but differ substantially from one another. The first point is likely because their text towers are trained with a symmetric cross-lingual contrastive objective, whereas decoder-only LMs are pretrained on English-dominant text with vision attached afterward, which plausibly induces a stronger English preference at generation time. NLLB-SigLIP Base is a strong outlier: it is biased against European languages, suggesting that its training on 200 languages with NLLB representations may have introduced an unusual skew. The Large variant is more balanced, though it retains a dispreference for Czech and Slovak. MEXMA-SigLIP 2 and the SigLIP 2 family (especially in the base and giant versions) show little bias toward English but exhibit a consistent tendency to penalize Hindi and Japanese, similar to the accuracy drop for Japanese

	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large	SigLIP2 SO400M	SigLIP2 Giant
Gemma 3 12B	—	.65	.70	.51	.09	.57	.61	.63	.58	.61	.62	.62
Qwen 2.5 VL 7B	.79	—	.76	.46	.09	.60	.65	.67	.60	.62	.64	.63
Qwen 3 VL 8B	.82	.83	—	.51	.08	.64	.69	.72	.64	.66	.68	.68
Aya Vision 8B	.71	.72	.74	—	.08	.45	.48	.49	.45	.47	.48	.48
Jina VL	.60	.67	.62	.56	—	.06	.06	.06	.06	.06	.06	.06
NLLB-SigLIP Base	.63	.65	.66	.60	.52	—	.70	.67	.55	.55	.57	.57
NLLB-SigLIP Large	.69	.71	.73	.65	.56	.71	—	.72	.59	.60	.62	.62
MEXMA-SigLIP2	.75	.76	.78	.68	.59	.68	.74	—	.60	.61	.63	.63
SigLIP2 Base	.62	.61	.64	.57	.49	.51	.57	.60	—	.74	.74	.74
SigLIP2 Large	.64	.63	.66	.58	.49	.53	.58	.61	.74	—	.78	.78
SigLIP2 SO400M	.63	.62	.65	.58	.48	.52	.57	.61	.73	.77	—	.80
SigLIP2 Giant	.61	.60	.63	.56	.47	.51	.56	.60	.70	.76	.78	—
Avg.	.68	.68	.69	.57	.36	.53	.57	.58	.57	.59	.60	.60

Table 4: Models’ agreement as Cohen’s Kappa coefficient averaged over all languages. Above the diagonal: *2img* task. Below the diagonal *2txt* task.

noted in Section 3.

Examples of English preferences. We further analyze a subset of the mistakes: for every sample in the bilingual setup with English and any other target language, we count the number of times the model considers the wrong caption in English over the correct one in the target language. Figure 3 contains the sample with one of the highest English preferences, $ep_{536} = 161$, showing that the difference between *small* and *tiny* is not as clear in the target languages as it is in English. On the opposite pole, there are 16 samples for which the models display no such preference at all; one example is shown in Figure 4, which demonstrates that the models are capable of distinguishing between animals. A similar phenomenon can be observed in the headline picture (Figure 1), where $ep_{110} = 13$. Additionally,

Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.17	.04	.01	-.01	-.05	-.02	.00	.00	.00	-.02	.05	.01	-.01	.01
az	-.17	—	-.16	-.17	-.21	-.24	-.21	-.18	-.18	-.20	-.19	-.14	-.17	-.20	-.19
cs	-.04	.16	—	-.05	-.07	-.13	-.06	-.05	-.04	-.05	-.06	.03	-.04	-.07	-.04
de	-.01	.17	.05	—	-.02	-.09	-.02	.00	.00	.00	-.02	.07	.01	-.02	.01
el	.01	.21	.07	.02	—	-.06	-.01	.02	.03	.02	.00	.09	.03	.00	.03
en	.05	.24	.13	.09	.06	—	.05	.09	.08	.09	.05	.15	.08	.05	.09
hi	.02	.21	.06	.02	.01	-.05	—	.02	.02	.01	.00	.08	.03	-.01	.03
it	.00	.18	.05	.00	-.02	-.09	-.02	—	.01	.00	-.02	.07	.01	-.02	.01
ja	.00	.18	.04	.00	-.03	-.08	-.02	-.01	—	-.01	-.02	.06	.01	-.02	.01
ro	.00	.20	.05	.00	-.02	-.09	-.01	.00	.01	—	-.02	.08	.01	-.02	.01
ru	.02	.19	.06	.02	.00	-.05	.00	.02	.02	.02	—	.09	.03	.00	.03
sk	-.05	.14	-.03	-.07	-.09	-.15	-.08	-.07	-.06	-.08	-.09	—	-.06	-.09	-.06
uk	-.01	.17	.04	-.01	-.03	-.08	-.03	-.01	-.01	-.01	-.03	.06	—	-.03	.00
vi	.01	.20	.07	.02	.00	-.05	.01	.02	.02	.02	.00	.09	.03	—	.03

Table 5: Difference between language-specific mistakes, $M_{\text{dif}}(x, y)$ as defined in Equation (9) averaged over all models. Languages on the vertical axis correspond to x , while languages on the horizontal axis correspond to y .

for the image from the annotators’ instruction (Figure 2), $ep_{672} = 64$ shows that the models find it difficult to distinguish between color palettes.

Random pairs ablation. As a final ablation study, we aim to check the bilingual retrieval task with random pairs of captions. Table 25 and Table 26 display the accuracy and mistakes in this setup, showing that the task is almost trivial for our models, with high accuracies and few mistakes.

5 Related Work

Minimal pairs benchmarks. Minimal pairs, items differing in exactly one controlled dimension, are a standard diagnostic methodology in NLP. The most notable examples include BLiMP (Warstadt et al., 2020) and MultiBLiMP (Jumelet et al., 2025), which are sentence pairs isolating specific contrasts in syntax, morphology, and semantics. HANS (McCoy et al., 2019) probes whether natural language inference models exploit syntactic heuristics rather than genuine reasoning; WinoBias (Zhao et al., 2018) with Winograd-style coreference pairs measures gender bias in NLP systems.

The methodology has been adapted to vision-language models. Winoground (Thrush et al., 2022) tests whether models can match two images to two captions containing identical words in different orders. ARO (Yuksekgonul et al., 2023) scales this to 50,000 pairs, targeting relational and attributive understanding. SugarCrepe (Dumpala et al., 2024) addresses artifacts in earlier benchmarks by using an LLM to generate fluent hard-negative captions. All of these works focus exclusively on English. InpaintCOCO (Rösch et al., 2024) extends the paradigm by modifying both the caption and the

image via inpainting, introducing genuine visual ambiguity. We extend InpaintCOCO to 13 more languages.

Multilingual vision-language benchmarks. A survey by Manea and Libovický (2025) identified 23 benchmarks for multilingual vision-language tasks. We refer the reader to that section for a broader overview.

Most existing multilingual benchmarks are either translation-based (XVNLi, xGQA, xFlickrCo; Bugliarello et al. 2022b; Nooralahzadeh and Sennrich 2023) or with region-specific imagery (MARVL, Liu et al. 2021; CVQA, Romero et al. 2024). MultiInpaintCOCO is translation-based, enables direct cross-lingual comparison, and is the only benchmark that combines caption-level and image-level minimal pairs, allowing retrieval errors to be attributed to either text or vision understanding.

Language bias. Language bias in NLP refers to the tendency of multilingual models to favor high-resource languages over lower-resource ones. Early work showed that multilingual BERT encodes a structural bias toward English, with the grammatical properties of higher-resource languages bleeding into the representations of lower-resource ones (Papadimitriou et al., 2023). Nowadays, this bias is particularly pronounced in retrieval settings: in multilingual retrieval-augmented generation, models systematically prefer high-resource language sources, especially English, even when lower-resource language documents are more relevant (Wu et al., 2024; Ki et al., 2025).

In vision-language models, related phenomena have been observed, such as models defaulting to

English responses regardless of the query language when an image is present (Hinck et al., 2024), and degraded understanding of non-English text within images (Geigle et al., 2025).

However, to the best of our knowledge, none of the previous works conducted a direct study of language bias in a *multilingual vision-language retrieval* setup.

6 Conclusions

We introduced MultiInpaintCOCO, a multilingual minimal-pair benchmark for fine-grained vision-language retrieval, extending InpaintCOCO to 13 more typologically diverse languages via automatic translation followed by native-speaker post-editing. The benchmark combines caption-level and image-level minimal pairs, enabling retrieval errors to be attributed to either the textual or visual representation, a combination absent from prior multilingual VL benchmarks.

We evaluated five decoder-only and seven dual-encoder models and reported two main findings. First, performance drops substantially for lower-resource languages, with Azerbaijani—the least-resourced language in our study—consistently being the hardest across all models. The strongest predictor of this gap is the volume of pretraining data available for a language, as reflected by web corpus size, rather than typological proximity to English. Second, in cross-lingual retrieval, models exhibit a systematic preference for English captions over correct captions in other languages. This bias is not merely a consequence of lower monolingual accuracy; the tendency persists even in cases where a model would otherwise handle the non-English language well. The bias is strongest in Gemma 3 and weakest in Qwen 3, and it is present in both encoder and decoder architectures, though dual encoders tend to show less English preference overall.

These findings suggest that the English-centric nature of VL pretraining manifests in at least two separable ways: reduced representational capacity for lower-resource languages and an active preference for English during retrieval when English is one of the candidate options.

Limitations

MultiInpaintCOCO covers 14 typologically diverse languages across five language families. This remains a small fraction of the world’s approximately

7,000 languages, and our findings may not generalize beyond the languages included.

Automatic translation followed by native-speaker post-editing may introduce a bias toward grammatically correct but unnaturally formal phrasing, as post-editors tend to preserve the structure of machine-translated output rather than produce fully idiomatic text. Moreover, we did not evaluate the translation quality (e.g., for naturalness, consistency, or difficulty) of our annotated captions across languages. Therefore, we assumed that the low performance is the direct consequence of the models’ abilities.

Finally, the images in our dataset are drawn from COCO, which predominantly depicts scenes and objects from North America. This introduces a potential confound: strong performance on our benchmark does not necessarily imply that a model handles a given language well in its natural cultural context, since the visual content may be more familiar to models trained on Western-centric data, regardless of the language of the caption.

Acknowledgments

We thank our translators: Konstantinos Diamantopoulos, Turan Isgandarli, Nalin Kumar, Nam Luu, Yuka Miyake, Sabína Ságová, Tim Shchetilin, Ibrahim Qatato, and Gianluca Vico.

This research was supported by the Charles University project PRIMUS/23/SCI/023 and project CZ.02.01.01/00/23_020/0008518 of the Ministry of Education, Youth and Sports of the Czech Republic.

We used several AI assistants for coding (Microsoft Copilot) and paper editing (Grammarly and Writefull).

References

- Pranav Aggarwal and Ajinkya Kale. 2020. [Towards zero-shot cross-lingual image retrieval](#). *CoRR*, abs/2012.05107.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuezhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei

- Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. [Findings of the third shared task on multimodal machine translation](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.
- Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022a. [IGLUE: A benchmark for transfer learning across modalities, tasks, and languages](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR.
- Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022b. [IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages](#). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392, Baltimore, Maryland. PMLR.
- Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. [MaXM: Towards multilingual visual question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2667–2682, Singapore. Association for Computational Linguistics.
- Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, and 6 others. 2025. [Aya vision: Advancing the frontier of multilingual multimodality](#). *Preprint*, arXiv:2505.08751.
- Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos E. Milios, Sageev Oore, and Hassan Sajjad. 2024. [SUG-ARCREPE++ dataset: Vision-language model sensitivity to semantic and lexical alterations](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. 2025. [Hidden in plain sight: VLMs overlook their visual representations](#). *Preprint*, arXiv:2506.08008.
- Matthieu Futral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. [Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.
- 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 (Volume 1: Long Papers)*, pages 2831–2881, Vienna, Austria. Association for Computational Linguistics.
- Musashi Hinck, Carolin Holtermann, Matthew Lyle Olson, Florian Schneider, Sungduk Yu, Anahita Bhawandiwalla, Anne Lauscher, Shao-Yen Tseng, and Vasudev Lal. 2024. [Why do LLaVA vision-language models reply to images in English?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13402–13421, Miami, Florida, USA. Association for Computational Linguistics.
- João Maria Janeiro, Benjamin Piwowarski, Patrick Galinari, and Loïc Barrault. 2024. [Mexma: Token-level objectives improve sentence representations](#). *Preprint*, arXiv:2409.12737.
- 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.
- Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. [Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs](#). *CoRR*, abs/2504.02768.
- Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. [URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.
- Dayeon Ki, Marine Carpuat, Paul McNamee, Daniel Khashabi, Eugene Yang, Dawn J. Lawrie, and Kevin Duh. 2025. [Linguistic nepotism: Trading-off quality for language preference in multilingual RAG](#). *CoRR*, abs/2509.13930.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouгна, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25](#)

- general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Andreas Koukounas, Georgios Mastrapas, Florian Hönicke, Sedigheh Eslami, Guillaume Roncari, Scott Martens, and Han Xiao. 2026. *jina-vlm: Small multilingual vision language model*. Preprint, arXiv:2512.04032.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. *Microsoft COCO: common objects in context*. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.
- Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. *Visually grounded reasoning across languages and cultures*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Andrei-Alexandru Manea and Jindřich Libovický. 2025. *Multilingual vision-language models, A survey*. *CoRR*, abs/2509.22123.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. *Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Farhad Nooralahzadeh and Rico Sennrich. 2023. *Improving the Cross-lingual Generalisation in Visual Question Answering*. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023*, pages 13419–13427, Washington, DC, USA. AAAI Press.
- Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. *Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models*. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200, Dubrovnik, Croatia. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. *Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language*. In *Second Conference on Language Modeling*.
- 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, Dublin, Ireland. Association for Computational Linguistics.
- Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. *What is “typological diversity” in NLP?* In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700, Miami, Florida, USA. 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.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. *High-resolution image synthesis with latent diffusion models*. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesús-Germán Ortiz-Barajas, Emilio Villa-Cueva, Jinheon Baek, Soyeong Jeong, Injy Hamed, Zheng Xin Yong, Zheng Wei Lim, Paula Mónica Silva, Jocelyn Dunstan, Mélanie Joutiteau, David Le Meur, Joan Nwatu, Ganzorig Batnasan, and 57 others. 2024. *CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark*. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada*.
- Philipp J. Rösch, Norbert Oswald, Michaela Geierhos, and Jindřich Libovický. 2024. *Enhancing conceptual understanding in multimodal contrastive learning through hard negative samples*. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 102–115, Bangkok, Thailand. Association for Computational Linguistics.
- Florian Schneider and Sunayana Sitaram. 2024. *M5 – a diverse benchmark to assess the performance of large multimodal models across multilingual and multicultural vision-language tasks*. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4309–4345, Miami, Florida, USA. Association for Computational Linguistics.
- Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin

- Mahmood, Hao Feng, Zhen Zhao, Yangfan He, Kuan Lu, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. 2025. [Mtvqa: Benchmarking multilingual text-centric visual question answering](#). *Preprint*, arXiv:2405.11985.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, and 1 others. 2025. [Gemma 3 Technical Report](#). *Preprint*, arXiv:2503.19786.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. [Crossmodal-3600: A massively multilingual multimodal evaluation dataset](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.
- Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. [Identifying and mitigating position bias of multi-image vision-language models](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 10599–10609. Computer Vision Foundation / IEEE.
- 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](#). *Preprint*, arXiv:2502.14786.
- Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadgign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2025. [All languages matter: Evaluating llms on culturally diverse 100 languages](#). *Preprint*, arXiv:2411.16508.
- Alexander Visheratin. 2023. [Nllb-clip – train performant multilingual image retrieval model on a budget](#). *Preprint*, arXiv:2309.01859.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wikidata: a free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Hao Wang, Pinzhi Huang, Jihan Yang, Saining Xie, and Daisuke Kawahara. 2025. [Traveling across languages: Benchmarking cross-lingual consistency in multimodal llms](#). *Preprint*, arXiv:2505.15075.
- Grant Wardle and Teo Susnjak. 2024. [Image first or text first? optimising the sequencing of modalities in large language model prompting and reasoning tasks](#). *Preprint*, arXiv:2410.03062.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mo-hananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Suhang Wu, Jialong Tang, Baosong Yang, Ante Wang, Kaidi Jia, Jiawei Yu, Junfeng Yao, and Jinsong Su. 2024. [Not all languages are equal: Insights into multilingual retrieval-augmented generation](#). *CoRR*, abs/2410.21970.
- Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2025. [Pangea: A fully open multilingual multimodal llm for 39 languages](#). *Preprint*, arXiv:2410.16153.
- Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. [When and why vision-language models behave like bags-of-words, and what to do about it?](#) In *International Conference on Learning Representations*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). *Preprint*, arXiv:2303.15343.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. [Gender bias in coreference resolution: Evaluation and debiasing methods](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.
- Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. [Shy-hunyuan-MT at WMT25 general machine translation shared task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 607–613, Suzhou, China. Association for Computational Linguistics.

	ar	az	cs	de	el	hi	it	ja	ro	ru	sk	uk	vi
MT	4	4	4	4	4	4	G	H	4	4	4	4	H
Hours	10	6	10	6	6	6	6	12	10	6	6	6	6

Table 6: Machine translation system used for post-editing (G: Google Translate, H: Hunyuan, 4: GPT-4) and self-reported time spent by post-editing.

A Annotation Process

The filtered dataset was translated into 13 additional languages. The filtering and corrections of the original dataset were carried out by researchers who are authors of this paper.

Translation into Czech, German, and Romanian was done by researchers. Translation into Russian was done by a language teacher of Russian. Translation into Arabic, Azerbaijani, Greek, Hindi, Italian, Japanese, Slovak, Ukrainian, and Vietnamese was done by university students. The students were recruited in classes. The students were paid an hourly rate that corresponds to 223% of the minimum wage and 125% of the median wage in the country where the translation was conducted.

We used Google Docs as the annotation interface (see Figure 2). The annotators were presented with the images, original English captions, and the machine translation and were instructed to edit the translation without changing the formatting. We asked the annotators to keep a separate list of issues that they encountered in the dataset. Self-reported postediting times (biased by telling annotators it should take 6 hours) and the preferred MT system are reported in Table 6.

B Hyperparameters

For the decoder-only models, we use the following instruction prompts:

English:

2img Based on two images, which follow the caption [C] ? Answer with 'left' or 'right' and nothing else.

2txt Based on the 2 captions: [NEWLINE] (left) [C₁] [NEWLINE] (right) [C₂] [NEWLINE] Which one follows the image? Answer with 'left' or 'right' and nothing else.

French:

2img À partir des deux images, qui suivent la description [C] ? Répondez par 'gauche' ou 'droite' et rien d'autre.



COCO en 672: a green orange and silver train and some cars

COCO trans 672: हरा, नारंगी और चांदी रंग की ट्रेन और कुछ कारें



Inpaint en 672: a blue train and some cars

Inpaint trans 672: नीली ट्रेन और कुछ कारें

Figure 2: A screenshot from the post-editing document for Hindi.

2txt À partir des deux descriptions suivantes: [NEWLINE] (gauche) [C₁] [NEWLINE] (droite) [C₂] [NEWLINE] Laquelle correspond à l'image? Répondez par 'gauche' ou 'droite', et rien d'autre.

In the *2txt* task, the newline was necessary for the captions written in a right-to-left system (e.g. Arabic). Otherwise, the keywords left and right were confusing to the LLMs, causing them to produce a random input.

For encoder models, we note one small limitation of the SigLIP 2 text encoder: the maximum input length is 64 tokens. This impacts a maximum of 4 input captions, the highest number being in Greek, with the tokenizer of the base version.

C Position Bias

Since the image order can introduce a small amount of noise in the results, we analyze how significant the noise is. We compute the number of mistakes for both query texts when the images are given in the direct order, M_{dir} , and in the swapped order, M_{swap} .

$$M_{\text{dir}} = \mathbb{O}_{c(T_o, (I_o, I_e))} + \mathbb{O}_{c(T_e, (I_o, I_e))} \quad (10)$$

$$M_{\text{swap}} = \mathbb{O}_{c(T_o, (I_e, I_o))} + \mathbb{O}_{c(T_e, (I_e, I_o))} \quad (11)$$

$$M_{\text{order}} = M_{\text{dir}} - M_{\text{swap}} \quad (12)$$

Language		Re- source tier	Native speak- ers (M)	Morphology	Family	FineWeb (GB)	Wiki articles (M)	WALS sim. w/	
								en	zh
English	en	5	380	Isolating	Indo-Europ.	2 · 10 ⁸	6.90	—	.451
Czech	cs	4	11	Fusional	Indo-Europ.	222	0.56	.436	.385
Romanian	ro	3	24	Fusional	Indo-Europ.	200	0.43	.537	.469
Vietnamese	vi	4	76	Isolating	Austroasiatic	343	1.31	.525	.423
German	de	5	77	Fusional	Indo-Europ.	1 655	2.98	.597	.450
Slovak	sk	3	6	Fusional	Indo-Europ.	92	0.26	.436	.385
Italian	it	4	65	Fusional	Indo-Europ.	794	1.86	.524	.421
Azerbaijani	az	1	24	Agglut.	Turkic	29	0.22	.451	.362
Japanese	ja	5	128	Agglut.	Japonic	1 650	1.41	.421	.459
Ukrainian	uk	3	27	Fusional	Indo-Europ.	274	1.33	.541	.465
Arabic	ar	5	335	Fusional	Afro-Asiatic	315	1.22	.432	.348
Greek	el	3	15	Fusional	Indo-Europ.	238	0.21	.500	.403
Hindi	hi	4	341	Fusional	Indo-Europ.	130	1.62	.513	.463
Russian	ru	4	154	Fusional	Indo-Europ.	6 400	1.99	.503	.469

Table 7: Language metadata for the 14 languages in this study. ISO codes follow ISO 639-1 and ISO 639-3. Resource tier by (Joshi et al., 2020) (0 = lowest, 5 = highest). FineWeb2 corpus size is given in gigabytes. Linguistic similarity (WALS. sim.) is $1 - d_{\text{featural}}$ computed via URIEL+ (Khan et al., 2025), where higher values indicate greater typological similarity; EN = vs. English, ZH = vs. Mandarin Chinese.

Model	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Notes
Gemma 3 12B	?	?	?	?	?	?	?	?	?	?	?	?	?	?	> 140 languages
Qwen 2.5 VL 7B	✓	?	?	✓	?	✓	?	✓	✓	?	✓	?	?	✓	10 languages (OCR data)
Qwen 3 VL 8B	✓	?	?	✓	?	✓	?	✓	✓	?	✓	?	?	?	119 languages and dialects
Aya Vision 8B	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	23 languages
Jina VL	✓	-	-	✓	-	✓	✓	✓	✓	-	✓	-	-	✓	17 languages
NLLB SigLIP Family	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	200 languages
MEXMA SigLIP2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	80 languages
SigLIP 2 Family	?	?	?	?	?	?	?	?	?	?	?	?	?	?	109 languages

Table 8: Overview of training data availability across the evaluated models for the 14 languages included in our dataset. Symbols indicate whether the training data is available (✓), the model was not trained on the language (-), or the exact data composition remains undisclosed (?).

where the $\mathbb{O}_{c(q,(a,b))}$ represents the number of times that $c(q, (a, b)) = 0$. We report, for each model, the difference M_{order} : a negative value indicates a bias towards the dataset order, while a positive value indicates a bias towards the swap.

We report the position bias in Table 13. The impact of the generative decoder models is visible but not highly significant. However, we can still see that the lowest values are for the Aya model, indicating a preference for the swapped order. On the opposite pole, the Jina model has the highest values, preferring the dataset image order. For the dual encoders, we do not plot the values because they are always 0. This is because the modality similarity is obtained from an outer product, except in the case when the similarity is equal.

D Additional Results

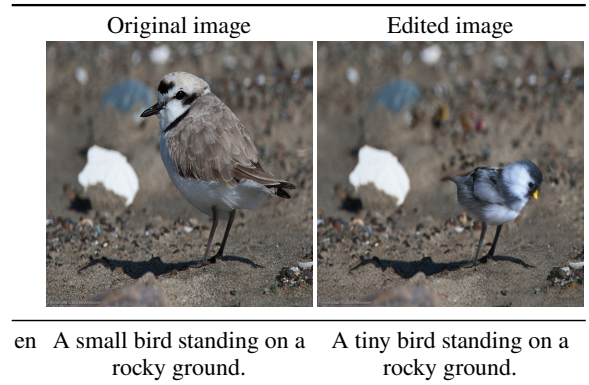


Figure 3: Sample from the MultiInpaintCOCO dataset, with a high number of English preference ($ep_{536} = 161$).

	Dataset	# lang.	GBI	# captions	# images	% population
MARVL	Liu et al. (2021)	5	56.41	5.7K	4.9K	13.79%
M5 (VGR,VLOD)	Schneider and Sitaram (2024)	12	74.87	1.4K	6.8K	16.79%
ALM-bench	Vayani et al. (2025)	99	93.84	22K	3K	59.92%
XM3600	Thapliyal et al. (2022)	36	91.79	261K	3.6K	49.18%
MaXM	Changpinyo et al. (2023)	7	64.61	2K	0.3K	21.45%
MTVQA	Tang et al. (2025)	10	77.94	28.6K	8.7K	16.39%
CVQA	Romero et al. (2024)	31	88.2	10.3K	5.2K	40.89%
XVNLI	Bugliarello et al. (2022b)	5	49.74	1.1k	357	16.89%
xGQA	Nooralahzadeh and Sennrich (2023)	8	63.58	2.4M	85K	21.76%
xFlickrCo	Bugliarello et al. (2022b)	8	65.12	2k	2k	27.54%
Multi30k	Barrault et al. (2018)	4	33.84	620K	31K	6.60%
XTD	Aggarwal and Kale (2020)	7	71.28	7K	1K	22.13%
CoMMuTE	Futeral et al. (2023)	6	66.15	1.9K	0.3K	18.61%
xMMMU	Yue et al. (2025)	7	69.23	2.6K	300	18.04%
KnowRecall	Wang et al. (2025)	15	77.94	45K	1.5K	35.26%
MultiInpaintCOCO		14	78.97	26K	1.9K	19.26%

Table 9: Dataset language coverage overview. The first half contains datasets with cultural knowledge, whereas the second ones are created by translating English samples.

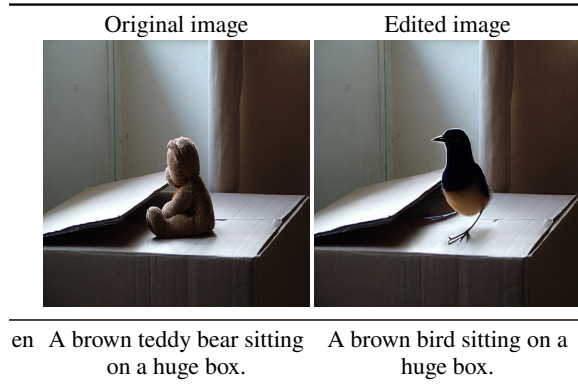


Figure 4: Sample from the MultiInpaintCOCO dataset, with a high number of English preference ($ep_{185} = 0$).

Model	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
Gemma 3 12B	16.6 4-56	19.5 5-78	17.2 4-63	15.5 5-51	25.9 8-74	12.3 4-40	16.3 4-37	16.1 5-50	13.5 3-46	19.2 5-63	16.6 6-50	17.9 4-69	19.4 5-65	18.3 6-60
Qwen 2.5 VL 7B	17.7 2-62	22.0 4-89	20.5 4-83	16.7 4-55	50.9 13-144	11.5 3-39	49.0 9-96	18.2 4-56	15.8 3-59	21.1 4-69	20.6 6-62	21.1 4-82	26.1 4-85	19.4 6-69
Qwen 3 VL 8B	17.7 2-62	22.0 4-89	20.5 4-83	16.7 4-55	50.9 13-144	11.5 3-39	49.0 9-96	18.2 4-56	15.8 3-59	21.1 4-69	20.6 6-62	21.1 4-82	26.1 4-85	19.4 6-69
Aya Vision	16.8 4-54	20.2 4-83	15.5 4-54	15.7 4-50	20.0 5-57	12.4 4-40	34.3 7-74	15.8 5-49	14.4 4-51	16.9 5-57	16.1 4-45	18.1 4-66	17.6 5-58	17.7 6-60
Jina VL	17.7 2-62	22.0 4-89	20.5 4-83	16.7 4-55	50.9 13-144	11.5 3-39	49.0 9-96	18.2 4-56	15.8 3-59	21.1 4-69	20.6 6-62	21.1 4-82	26.1 4-85	19.4 6-69
NLLB-SigLIP Base	17.9 5-59	17.1 5-63	17.7 5-63	18.6 6-58	22.9 8-66	14.8 5-46	18.8 4-40	17.9 6-55	15.3 5-48	19.5 6-67	19.1 7-55	17.6 5-65	19.8 5-63	19.7 7-62
NLLB-SigLIP Large	17.9 5-59	17.1 5-63	17.7 5-63	18.6 6-58	22.9 8-66	14.8 5-46	18.8 4-40	17.9 6-55	15.3 5-48	19.5 6-67	19.1 7-55	17.6 5-65	19.8 5-63	19.7 7-62
MEXMA-SigLIP2	16.5 5-52	16.1 5-60	16.5 5-59	17.1 6-53	20.9 7-62	15.0 5-47	19.2 4-40	17.2 5-53	16.8 6-53	18.4 5-62	17.4 6-53	16.6 5-62	18.2 5-61	19.3 7-62
SigLIP2 Base	16.9 4-56	19.5 5-80	16.8 5-62	14.5 4-48	26.4 8-76	12.3 4-40	22.7 6-48	15.1 5-46	12.6 3-43	18.8 5-61	15.7 5-50	17.4 5-67	19.2 5-65	17.9 6-62
SigLIP2 Large	16.9 4-56	19.5 5-80	16.8 5-62	14.5 4-48	26.4 8-76	12.3 4-40	22.7 6-48	15.1 5-46	12.6 3-43	18.8 5-61	15.7 5-50	17.4 5-67	19.2 5-65	17.9 6-62
SigLIP2 SO400M	16.9 4-56	19.5 5-80	16.8 5-62	14.5 4-48	26.4 8-76	12.3 4-40	22.7 6-48	15.1 5-46	12.6 3-43	18.8 5-61	15.7 5-50	17.4 5-67	19.2 5-65	17.9 6-62
SigLIP2 Giant	16.9 4-56	19.5 5-80	16.8 5-62	14.5 4-48	26.4 8-76	12.3 4-40	22.7 6-48	15.1 5-46	12.6 3-43	18.8 5-61	15.7 5-50	17.4 5-67	19.2 5-65	17.9 6-62

Table 10: Tokenizer statistics of the tested models for the MultiInpaintCOCO dataset. The table shows the average number of tokens per model-language pair and the range of text lengths.

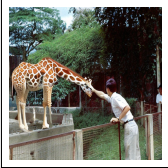
#	Query	Option (1)	Option (2)
1		A man is feeding an elephant over a fence	A man is feeding a giraffe over a fence
2		A man is feeding a giraffe over a fence	A man is feeding an elephant over a fence
3		A man is feeding an elephant over a fence	A man is feeding a giraffe over a fence
4		A man is feeding a giraffe over a fence	A man is feeding an elephant over a fence
5	A man is feeding an elephant over a fence	 (1) correct	
6	A man is feeding an elephant over a fence		 (2) correct
7	A man is feeding a giraffe over a fence		 (2) correct
8	A man is feeding a giraffe over a fence	 (1) correct	

Table 11: For decoder-only models, each example is evaluated in both input orders to account for order sensitivity, resulting in eight comparison configurations per example. Above – *2txt* configurations: the image is the query and the model selects the correct caption; correct answers are shown in **bold**. Below – *2img* configurations: the caption is the query and the model selects the correct image; the correct option is framed and labeled.

Model	Hugging Face Identifiers	Params (V + T + (X)= Total)	Img Size	Vocab Size
Gemma 3 12B	google/gemma-3-12b-it	417 M + 11.8 B = 12.2 B	896	262,144
Qwen 7B	Qwen/Qwen2.5-VL-7B-Instruct	677 M + 7.1 B = 7.8 B	3584	151,643
Qwen3 8B	Qwen/Qwen3-VL-8B-Instruct	576 M + 7.6 B = 8.1 B	3584	151,643
Aya Vision 8B	CohereLabs/aya-vision-8b	413 M + 8.0 B + 190.9 M = 8.6 B	364	255,000
Jina VL	jinaai/jina-vlm	449 M + 1.72 B + 311 M = 2.5 B	1176×910	151,643
NLLB-SigLIP Base	nllb-clip-base-siglip	93 M + 414 M = 508 M	384	256,204
NLLB-SigLIP Large	nllb-clip-large-siglip	428 M + 767 M = 1.2 B	384	256,204
MEXMA-SigLIP2	visheratin/mexma-siglip2	428 M + 573 M = 1B	512	250,002
SigLIP2 Base	google/siglip2-base-patch16-224	92 M + 282 M = 375 M	224	256,000
SigLIP2 Large	google/siglip2-large-patch16-256	316 M + 565 M = 881 M	256	256,000
SigLIP2 SO400M	google/siglip2-so400m-patch16-256	428 M + 708 M = 1.14 B	256	256,000
SigLIP2 Giant	google/siglip2-giant-opt-patch16-256	1.16 B + 708 M = 1.87 B	256	256,000

Table 12: Model overview with Hugging Face identifiers, parameter counts split into vision (V), text (T), and optional cross-modal (X) components, maximum image size, and tokenizer vocabulary size. Image sizes are square resolutions unless otherwise indicated.

Model	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
Gemma 3 12B	.01	.02	-.03	.00	-.02	-.06	-.02	.00	-.03	-.03	-.02	.01	-.03	-.02	-.01
Qwen 2.5 VL 7B	-.04	-.03	-.03	-.06	-.05	-.03	-.04	-.04	-.04	-.01	-.03	-.01	-.04	-.04	-.04
Qwen 3 VL 8B	-.03	-.05	-.02	-.04	-.03	-.06	-.03	-.05	-.03	-.03	-.05	-.03	-.02	-.03	-.04
Aya Vision 8B	-.07	-.04	-.06	-.06	-.07	-.07	-.10	-.10	-.07	-.06	-.06	-.06	-.07	-.09	-.07
Jina VL	.03	-.01	-.03	.05	.11	-.04	.04	.03	.01	.05	.03	.04	.05	.04	.03
Avg.	-.02	-.02	-.04	-.02	-.01	-.05	-.03	-.03	-.03	-.02	-.02	-.01	-.02	-.03	-.03

Table 13: Per-language position bias for each model, measured as M_{order} from Equation (12). Negative values indicate a preference for the dataset image order, whereas positive values indicate a preference for the swapped order.

Model	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
Gemma 3 12B	.639	.501	.562	.638	.652	.608	.578	.648	.571	.616	.663	.516	.650	.659	.607
Qwen 2.5 VL 7B	.651	.414	.641	.726	.614	.766	.615	.700	.643	.616	.724	.592	.672	.667	.646
Qwen 3 VL 8B	.745	.581	.739	.773	.771	.773	.769	.761	.747	.739	.770	.704	.761	.757	.742
Aya Vision 8B	.264	.163	.336	.459	.445	.548	.332	.412	.388	.387	.429	.268	.411	.470	.379
Jina VL	.023	.002	.008	.037	.018	.021	.016	.031	.026	.016	.016	.009	.017	.036	.020
NLLB-SigLIP Base	.768	.512	.548	.644	.754	.771	.759	.613	.726	.608	.775	.586	.748	.642	.675
NLLB-SigLIP Large	.766	.593	.701	.708	.774	.778	.771	.733	.740	.708	.770	.692	.767	.666	.726
MEXMA-SigLIP2	.759	.747	.756	.765	.749	.769	.757	.761	.742	.772	.763	.767	.771	.735	.758
SigLIP2 Base	.750	.495	.757	.814	.708	.817	.646	.794	.604	.700	.809	.745	.735	.757	.724
SigLIP2 Large	.765	.560	.783	.815	.744	.823	.709	.817	.619	.754	.818	.765	.775	.784	.752
SigLIP2 SO400M	.772	.600	.772	.832	.775	.827	.730	.827	.667	.785	.823	.793	.775	.784	.769
SigLIP2 Giant	.773	.566	.786	.825	.765	.839	.722	.815	.679	.774	.823	.776	.776	.801	.766
Avg.	.640	.478	.616	.670	.648	.695	.617	.659	.596	.623	.682	.601	.655	.646	.630

Table 14: Accuracy $2img$ for bidirectional retrieval, i.e. c_{2img} from Equation (1). The upper part of the tables contains LLMs, while the lower has the dual encoders.

Model	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
Gemma 3 12B	.737	.630	.705	.751	.733	.767	.724	.742	.711	.750	.760	.704	.742	.747	.729
Qwen 2.5 VL 7B	.719	.505	.720	.786	.741	.801	.677	.785	.772	.737	.788	.714	.765	.758	.733
Qwen 3 VL 8B	.791	.692	.761	.803	.787	.818	.789	.815	.793	.794	.802	.759	.780	.794	.784
Aya Vision 8B	.635	.355	.595	.648	.630	.688	.617	.660	.613	.644	.651	.508	.617	.646	.608
Jina VL	.445	.187	.417	.512	.448	.675	.418	.554	.447	.477	.485	.445	.449	.443	.457
NLLB-SigLIP Base	.756	.463	.488	.605	.763	.782	.750	.626	.742	.578	.775	.522	.759	.659	.662
NLLB-SigLIP Large	.765	.567	.672	.735	.779	.792	.784	.750	.759	.720	.779	.669	.764	.685	.730
MEXMA-SigLIP2	.764	.769	.793	.806	.774	.796	.797	.795	.763	.790	.803	.787	.784	.785	.786
SigLIP2 Base	.715	.489	.748	.762	.697	.793	.625	.753	.572	.674	.767	.726	.717	.715	.697
SigLIP2 Large	.737	.501	.748	.769	.729	.789	.722	.766	.610	.731	.772	.726	.749	.748	.721
SigLIP2 SO400M	.720	.541	.759	.766	.742	.772	.729	.749	.618	.743	.750	.737	.744	.733	.722
SigLIP2 Giant	.699	.482	.741	.732	.726	.735	.664	.732	.587	.730	.740	.738	.725	.730	.697
Avg.	.707	.515	.679	.723	.712	.767	.691	.727	.666	.697	.739	.669	.716	.704	.694

Table 15: Accuracy $2txt$ for bidirectional retrieval, i.e. c_{2txt} from Equation (2). The upper part of the tables contains LLMs, while the lower has the dual encoders.

	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.502	.622	.646	.665	.660	.621	.651	.600	.636	.683	.609	.660	.647	.657
az	.518	—	.508	.529	.520	.541	.490	.526	.475	.508	.545	.502	.531	.516	.551
cs	.601	.476	—	.626	.613	.630	.568	.621	.557	.614	.650	.652	.637	.602	.632
de	.644	.516	.645	—	.647	.669	.604	.655	.594	.647	.678	.630	.662	.632	.659
el	.663	.506	.628	.643	—	.680	.624	.654	.602	.645	.684	.617	.672	.653	.662
en	.661	.524	.648	.672	.690	—	.644	.669	.630	.670	.703	.631	.683	.672	.678
hi	.639	.489	.598	.616	.643	.652	—	.620	.591	.612	.657	.583	.639	.626	.640
it	.646	.506	.636	.654	.657	.670	.611	—	.599	.654	.672	.622	.660	.635	.659
ja	.625	.483	.600	.619	.632	.660	.611	.627	—	.612	.655	.582	.635	.620	.640
ro	.631	.487	.622	.640	.647	.662	.595	.653	.576	—	.668	.612	.652	.630	.648
ru	.675	.524	.661	.673	.680	.694	.640	.673	.625	.668	—	.644	.698	.665	.680
sk	.587	.464	.631	.612	.592	.612	.548	.612	.533	.594	.629	—	.626	.581	.616
uk	.655	.516	.639	.650	.669	.673	.621	.656	.606	.652	.695	.636	—	.645	.665
vi	.654	.507	.628	.636	.663	.680	.622	.648	.598	.645	.680	.612	.662	—	.660
Avg.	.657	.536	.647	.658	.665	.677	.628	.662	.613	.654	.686	.638	.673	.652	.646

Table 16: Accuracy of the bilingual retrieval, $2txt$ only.

Gemma 3 12B															
Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.14	-.12	-.15	.00	-.29	.04	-.11	-.03	-.18	-.10	-.10	-.04	-.11	-.08
az	-.14	—	-.33	-.37	-.22	-.47	-.16	-.32	-.22	-.36	-.28	-.29	-.20	-.35	-.28
cs	.12	.33	—	-.10	.04	-.29	.19	-.02	.12	-.05	-.02	.02	.06	-.07	.02
de	.15	.37	.10	—	.14	-.15	.25	.09	.19	.06	.04	.14	.12	.04	.12
el	.00	.22	-.04	-.14	—	-.26	.04	-.09	.01	-.12	-.09	-.02	-.01	-.11	-.05
en	.29	.47	.29	.15	.26	—	.37	.26	.35	.24	.18	.30	.24	.19	.28
hi	-.04	.16	-.19	-.25	-.04	-.37	—	-.22	-.11	-.29	-.18	-.16	-.11	-.21	-.15
it	.11	.32	.02	-.09	.09	-.26	.22	—	.11	-.02	-.03	.06	.06	-.03	.04
ja	.03	.22	-.12	-.19	-.01	-.35	.11	-.11	—	-.21	-.13	-.08	-.04	-.18	-.08
ro	.18	.36	.05	-.06	.12	-.24	.29	.02	.21	—	.02	.08	.11	.00	.09
ru	.10	.28	.02	-.04	.09	-.18	.18	.03	.13	-.02	—	.08	.07	-.01	.06
sk	.10	.29	-.02	-.14	.02	-.30	.16	-.06	.08	-.08	-.08	—	.01	-.11	-.01
uk	.04	.20	-.06	-.12	.01	-.24	.11	-.06	.04	-.11	-.07	-.01	—	-.10	-.03
vi	.11	.35	.07	-.04	.11	-.19	.21	.03	.18	.00	.01	.11	.10	—	.08

Qwen 3 VL 8B															
Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.08	-.01	-.05	-.02	-.09	-.01	-.04	-.08	-.02	-.01	.01	.02	-.08	-.02
az	-.08	—	-.15	-.14	-.12	-.17	-.13	-.14	-.18	-.11	-.08	-.12	-.05	-.15	-.12
cs	.01	.15	—	-.05	-.02	-.10	.00	.06	-.08	-.01	-.02	.04	.03	-.06	-.01
de	.05	.14	.05	—	.02	-.08	.03	.01	-.02	.04	.02	.07	.06	.02	.03
el	.02	.12	.02	-.02	—	-.05	.02	-.01	-.03	.01	.01	.04	.04	-.03	.01
en	.09	.17	.10	.08	.05	—	.06	.08	.05	.10	.07	.12	.09	.03	.08
hi	.01	.13	.00	-.03	-.02	-.06	—	-.04	.05	.01	-.01	.03	.04	.05	.00
it	.04	.14	.06	-.01	.01	-.08	.04	—	-.01	.04	.02	.08	.06	-.02	.03
ja	.08	.18	.08	.02	.03	-.05	.05	.01	—	.08	.05	.10	.09	.00	.06
ro	.02	.11	.01	-.04	-.01	-.10	-.01	-.04	-.08	—	.00	.04	.05	-.05	-.01
ru	.01	.08	.02	-.02	-.01	-.07	.01	-.02	-.05	.00	—	.03	.04	-.04	.00
sk	-.01	.12	-.04	-.07	-.04	-.12	-.03	-.08	-.10	-.04	-.03	—	.01	-.09	-.04
uk	-.02	.05	-.03	-.06	-.04	-.09	-.04	-.06	-.09	-.05	-.04	-.01	—	-.09	-.04
vi	.08	.15	.06	.02	.03	-.03	.05	.02	.00	.05	.04	.09	.09	—	.05

Jina VL															
Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.24	.05	.00	.03	-.06	-.15	-.05	-.03	-.01	.13	.02	.19	.01	.03
az	-.24	—	-.23	-.21	-.28	-.28	-.44	-.30	-.22	-.29	-.15	-.25	-.15	-.21	-.25
cs	-.05	.23	—	-.06	-.09	-.22	-.26	-.11	-.10	-.06	-.01	-.01	.02	-.08	-.06
de	.00	.21	.06	—	.01	-.16	-.19	-.08	-.08	.00	.14	.03	.16	-.09	.00
el	-.03	.28	.09	-.01	—	-.34	-.28	-.05	-.07	-.02	.04	.08	.08	-.07	-.02
en	.06	.28	.22	.16	.34	—	.12	.10	.00	.12	.17	.21	.24	.03	.16
hi	.15	.44	.26	.19	.28	.12	—	.16	.13	.15	.30	.23	.36	.04	.20
it	.05	.30	.11	.08	.05	-.10	-.16	—	-.02	.02	.18	.08	.21	.01	.06
ja	.03	.22	.10	.08	.07	.00	-.13	.02	—	.06	.14	.09	.18	.09	.07
ro	.01	.29	.06	.00	.02	-.12	-.15	-.02	-.06	—	.13	.05	.15	-.04	.02
ru	-.13	.15	.01	-.14	-.04	-.17	-.30	-.18	-.14	-.13	—	-.03	.03	-.11	-.09
sk	-.02	.25	.01	-.03	-.08	-.21	-.23	-.08	-.09	-.05	.03	—	.07	-.06	-.04
uk	-.19	.15	-.02	-.16	-.08	-.24	-.36	-.21	-.18	-.15	-.03	-.07	—	-.18	-.13
vi	-.01	.21	.08	.09	.07	-.03	-.04	-.01	-.09	.04	.11	.06	.18	—	.05

Qwen 2.5 VL 7B															
Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.14	-.07	-.12	-.06	-.18	.07	-.13	-.06	-.10	-.08	-.04	-.03	-.09	-.06
az	-.14	—	-.26	-.28	-.25	-.30	-.22	-.32	-.27	-.26	-.21	-.26	-.17	-.23	-.24
cs	.07	.26	—	-.08	-.05	-.18	.06	-.11	-.03	-.03	-.05	.05	.04	-.02	-.01
de	.12	.28	.08	—	.05	-.08	.14	.00	.05	.08	.04	.12	.12	.05	.08
el	.06	.25	.05	-.05	—	-.12	.05	-.03	.02	.00	-.02	.09	.04	.00	.03
en	.18	.30	.18	.08	.12	—	.17	.08	.12	.15	.08	.19	.14	.12	.15
hi	-.07	.22	.06	.14	-.05	-.17	—	-.16	-.10	-.09	-.11	-.01	-.04	.10	-.07
it	.13	.32	.11	.00	.03	-.08	.16	—	.08	.08	.04	.14	.12	.05	.09
ja	.06	.27	.03	-.05	-.02	-.12	.10	-.08	—	-.03	-.03	.06	.04	.00	.02
ro	.10	.26	.03	-.08	.00	-.15	.09	-.08	.03	—	-.02	.08	.07	-.02	.02
ru	.08	.21	.05	-.04	.02	-.08	.11	-.04	.03	.02	—	.09	.07	.01	.04
sk	.04	.26	.05	-.12	-.09	-.19	.01	-.14	-.06	-.08	-.09	—	.02	-.08	-.05
uk	.03	.17	-.04	-.12	-.04	-.14	.04	-.12	-.04	-.07	-.07	.02	—	-.08	-.04
vi	.09	.23	.02	-.05	.00	-.12	.10	-.05	.00	.02	-.01	.08	.08	—	.03

Aya Vision 8B															
Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.31	.11	.19	-.09	.21	-.22	.08	-.13	.04	-.03	.19	.03	-.08	.04
az	-.31	—	-.27	-.11	-.43	-.27	-.55	-.22	-.45	-.34	-.34	-.15	-.34	-.43	-.32
cs	-.11	.27	—	.06	-.19	-.10	-.29	-.01	.20	-.08	-.07	.12	-.07	.19	-.07
de	-.19	.11	.06	—	-.22	-.22	-.36	-.07	-.26	-.13	-.14	.02	-.14	-.21	-.14
el	.09	.43	.19	.22	—	.19	-.09	.14	-.01	.08	.07	.31	.10	.01	.13
en	-.21	.27	.10	.22	-.19	—	-.40	.07	-.06	.02	-.05	.15	-.01	.12	-.02
hi	.22	.55	.29	.36	.09	.40	—	.26	.08	.20	.19	.39	.19	.09	.26
it	-.08	.22	.01	.07	-.14	-.07	-.26	—	-.16	-.03	-.04	.12	-.03	.12	-.04
ja	.13	.45	.20	.26	.01	.06	-.08	.16	—	.11	.13	.31	.15	.03	.15
ro	-.04	.34	.08	.13	-.08	-.02	-.20	.03	-.11	—	.00	.21	.00	-.09	.02
ru	.03	.34	.07	.14	-.07	.05	-.19	.04	-.13	.00	—	.20	.01	-.10	.03
sk	-.19	.15	-.12	-.02	-.31	-.15	-.39	-.12	-.31	-.21	-.20	—	-.19	-.30	-.18
uk	.03	.34	.07	.14	-.10	.01	-.19	.03	.15	.00	-.01	.19	—	-.11	.02
vi	.08	.43	.19	.21	-.01	.12	-.09	.12	-.03	.09	.10	.30	.11	—	.13

Table 17: The difference between the language-specific mistakes, for decoder-only model.

Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.43	.40	.23	-.01	-.03	-.01	.29	.01	.18	-.01	.46	.00	.09	.16
az	-.43	—	-.06	-.21	-.47	-.49	-.45	-.19	-.46	-.33	-.46	.01	-.44	-.41	-.34
cs	-.40	.06	—	-.17	-.41	-.43	-.42	-.13	-.40	-.27	-.42	.09	-.41	-.39	-.28
de	-.23	.21	.17	—	-.26	-.29	-.25	.05	-.26	-.10	-.26	.24	-.24	.20	-.11
el	.01	.47	.41	.26	—	-.02	.01	.31	.04	.21	.01	.47	.03	.11	.18
en	.03	.49	.43	.29	.02	—	.02	.33	.05	.24	.02	.48	.06	.12	.20
hi	.01	.45	.42	.25	-.01	-.02	—	.31	.02	.20	.01	.47	.01	.09	.17
it	-.29	.19	.13	-.05	-.31	-.33	-.31	—	-.29	-.16	-.32	.20	-.30	-.26	-.16
ja	-.01	.46	.40	.26	-.04	-.05	-.02	.29	—	.19	-.01	.47	.00	.05	.15
ro	-.18	.33	.27	.10	-.21	-.24	-.20	.16	-.19	—	-.20	.35	-.18	-.14	-.02
ru	.01	.46	.42	.26	-.01	-.02	-.01	.32	.01	.20	—	.47	.02	.08	.17
sk	-.46	-.01	-.09	-.24	-.47	-.48	-.47	-.20	-.47	-.35	-.47	—	-.45	-.43	-.35
uk	.00	.44	.41	.24	-.03	-.06	-.01	.30	.00	.18	-.02	.45	—	.06	.15
vi	-.09	.41	.39	.20	-.11	-.12	-.09	.26	-.05	.14	-.08	.43	-.06	—	-.09

Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.04	.00	-.01	.01	-.04	.00	-.03	.08	.01	.00	-.02	.00	.05	.01
az	-.04	—	-.04	-.05	-.04	-.08	-.02	.05	.02	-.04	-.03	-.04	-.04	.00	-.03
cs	.00	.04	—	-.01	.01	-.04	.02	-.02	.07	.00	.00	.00	.00	.03	.01
de	.01	.05	.01	—	.02	-.02	.02	.00	.07	.02	.02	.01	.02	.05	.02
el	-.01	.04	-.01	-.02	—	-.05	-.01	-.02	.07	-.01	.00	-.01	.00	.03	.00
en	.04	.08	.04	.02	.05	—	.05	.02	.10	.04	.05	.04	.04	.07	.05
hi	.00	.02	-.02	-.02	.01	-.05	—	-.02	.07	.00	.00	-.01	.00	.03	.00
it	.03	.05	.02	.00	.02	-.02	.02	—	.09	.02	.02	.02	.02	.06	.03
ja	-.08	-.02	-.07	-.07	-.07	-.10	-.07	-.09	—	-.07	-.06	-.08	-.07	.05	-.07
ro	-.01	.04	.00	-.02	.01	-.04	.00	-.02	.07	—	.00	.01	.00	.04	.01
ru	.00	.03	.00	.02	.00	-.05	.00	-.02	.06	.00	—	-.01	.00	.03	.00
sk	.02	.04	.00	-.01	.01	-.04	.01	-.02	.08	.01	.01	—	.00	.04	.01
uk	.00	.04	.00	-.02	.00	-.04	.00	-.02	.07	.00	.00	.00	—	.04	.01
vi	-.05	.00	-.03	-.05	-.03	-.07	-.03	-.06	.05	-.04	-.03	-.04	-.04	—	-.03

Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.12	.00	-.02	.00	-.03	.01	-.02	.06	.00	-.02	.01	.00	-.01	.01
az	-.12	—	-.13	-.14	-.12	-.15	-.14	-.14	-.06	-.12	-.14	-.11	-.13	-.13	-.12
cs	.00	.13	—	-.01	.01	-.02	.02	-.01	.07	.01	-.01	.01	.00	.00	.01
de	.02	.14	.01	—	.02	-.01	.03	.00	.08	.02	.00	.02	.01	.01	.03
el	.00	.12	.01	-.02	—	-.03	.01	-.02	.06	.00	-.02	.00	-.01	.01	.00
en	.03	.15	.02	.01	.03	—	.04	.01	.09	.03	.01	.03	.02	.02	.04
hi	-.01	.11	-.02	-.03	-.01	-.04	—	-.03	.05	-.01	-.03	-.01	-.02	.02	.00
it	.02	.14	.01	.00	.02	-.01	.03	—	.08	.02	.00	.02	.01	.01	.03
ja	-.06	.06	.07	-.08	-.06	-.09	-.05	-.08	—	-.06	-.08	-.06	-.07	-.07	-.06
ro	.00	.12	-.01	-.02	.00	-.03	.01	-.02	.06	—	-.02	.00	-.01	.01	.01
ru	.02	.14	.01	.00	.02	-.01	.03	.00	.08	.02	—	.03	.01	.01	.03
sk	-.01	.11	-.01	-.02	.00	-.03	.01	-.02	.06	.00	-.03	—	-.01	.01	.00
uk	.00	.13	.00	-.01	.01	-.02	.02	-.01	.07	.01	-.01	.01	—	.00	.01
vi	.01	.13	.00	-.01	.01	-.02	.02	-.01	.07	.01	-.01	.01	.00	—	.01

Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.11	-.02	-.02	-.01	-.02	.02	-.02	.06	-.02	-.02	-.02	-.01	-.01	.00
az	-.11	—	-.13	-.13	-.12	-.13	-.10	-.13	-.06	-.13	-.13	-.13	-.13	-.13	-.12
cs	.02	.13	—	.00	.01	.00	.04	.00	.08	.00	.00	.00	.01	.00	.02
de	.02	.13	.00	—	.01	.00	.03	.00	.08	.00	.00	.00	.00	.00	.02
el	.01	.12	-.01	-.01	—	-.01	.03	-.01	.07	.00	-.01	-.01	.00	.00	.01
en	.02	.13	.00	.00	.01	—	.04	.00	.08	.00	.00	.00	.01	.00	.02
hi	-.02	.10	-.04	-.03	-.03	-.04	—	-.03	.04	-.03	-.04	-.04	-.03	-.03	-.02
it	.02	.13	.00	.00	.01	.00	.03	—	.07	.00	.00	.00	.00	.00	.02
ja	-.06	.06	-.08	-.08	-.07	-.08	-.04	-.07	—	-.07	-.08	-.08	-.07	-.07	-.06
ro	.02	.13	.00	.00	.00	.00	.03	.00	.07	—	.00	.00	.00	.00	.02
ru	.02	.13	.00	.00	.01	.00	.04	.00	.08	.00	—	.00	.01	.01	.02
sk	.02	.13	.00	.00	.01	.00	.04	.00	.08	.00	.00	—	.01	.00	.02
uk	.01	.13	-.01	.00	.00	-.01	.03	.00	.07	.00	-.01	-.01	—	.00	.02
vi	.01	.13	.00	.00	.00	.00	.03	.00	.07	.00	-.01	.00	.00	—	.02

Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.21	.16	.05	-.02	-.05	.01	.02	.01	.04	-.01	.15	.02	.10	.05
az	-.21	—	-.05	-.18	-.23	-.25	-.21	-.18	-.21	-.17	-.20	-.06	-.19	-.13	-.17
cs	-.16	.05	—	-.12	-.18	-.20	-.17	-.12	-.16	-.12	-.17	.00	-.14	-.07	-.12
de	-.05	.18	.12	—	-.06	-.08	-.05	.00	-.04	.00	-.06	.11	-.02	.05	.01
el	.02	.23	.18	.06	—	-.03	.02	.05	.02	.07	.03	.17	.04	.12	.07
en	.05	.25	.20	.08	.03	—	.04	.07	.05	.08	.04	.18	.07	.13	.10
hi	-.01	.21	.17	.05	-.02	-.04	—	.03	.01	.04	-.01	.14	.02	.09	.05
it	-.02	.18	.12	.00	-.05	-.07	-.03	—	-.02	.01	-.03	.12	-.01	.05	.02
ja	-.01	.21	.16	.04	-.02	-.05	-.01	.02	—	.03	.00	.16	.01	.09	.05
ro	-.04	.17	.12	.00	-.07	-.08	-.04	-.01	-.03	—	-.05	.12	-.02	.04	.01
ru	.01	.20	.17	.06	-.03	-.04	.01	.03	.00	.05	—	.16	.03	.10	.06
sk	-.15	.06	.00	-.11	-.17	-.18	-.14	-.12	-.16	-.12	-.16	—	-.13	-.06	-.11
uk	-.02	.19	.14	.02	-.04	-.07	-.02	.01	-.01	.02	-.03	.13	—	.07	.03
vi	-.10	.13	.07	-.05	-.12	-.13	-.09	-.05	-.09	-.04	-.10	.06	-.07	—	-.04

Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.12	-.02	-.02	.01	-.04	.05	-.02	.07	.02	-.03	-.01	.00	.00	.01
az	-.12	—	-.13	-.14	-.11	-.16	-.07	-.13	-.04	-.09	-.14	-.12	-.12	-.11	-.11
cs	.02	.13	—	-.01	.03	-.02	.06	.00	.09	.04	-.01	.01	.02	.02	.03
de	.02	.14	.01	—	.03	-.02	.07	.00	.10	.05	.00	.02	.02	.02	.04
el	-.01	.11	-.03	-.03	—	-.05	.04	-.03	.06	.01	-.04	-.02	-.01	-.01	.00
en	.04	.16	.02	.02	.05	—	.09	.02	.11	.06	.01	.03	.04	.04	.05
hi	-.05	.07	-.06	-.07	-.04	-.09	—	-.07	.03	-.03	-.07	-.05	-.05	-.05	-.04
it	.02	.13	.00	.00	.03	-.02	.07	—	.09	.04	-.01	.01	.02	.02	.03
ja	-.07	.04	-.09	-.10	-.06	-.11	-.03	-.09	—	-.05	-.10	-.08	-.07	-.07	-.07
ro	-.02	.09	-.04	-.05	-.01	-.06	.03	.04	.05	—	-.05	-.03	-.02	.02	-.01
ru	.03	.14	.01	.00	.04	-.01	.07	.01	.10	.05	—	.02	.03	.03	.04
sk	.01	.12	-.01	-.02	.02	-.03	.05	-.01	.08	.03	-.02	—	.01	.01	.02
uk	.00	.12	-.02	-.02	.01	-.04	.05	-.02	.07	.02	-.03	-.01	—	.00	.01
vi	.00	.11	-.02	-.02	.01	-.04	.05	-.02	.07	.02	-.03	-.01	.00	—	.01

SigLIP2 SO400M															
Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.09	-.02	-.02	-.01	-.02	.00	-.01	.05	-.01	-.01	-.01	-.01	-.01	.00
az	-.09	—	-.11	-.11	-.10	-.12	-.09	-.10	-.04	-.10	-.10	-.10	-.10	-.10	-.10
cs	.02	.11	—	-.01	.01	-.01	.01	.00	.07	.01	.00	.01	.01	.01	.02
de	.02	.11	.01	—	.01	.00	.02	.01	.08	.01	.01	.02	.01	.02	.02
el	.01	.10	-.01	-.01	—	-.02	.01	.00	.06	.00	-.01	.00	.00	.00	.01
en	.02	.12	.01	.00	.02	—	.02	.01	.08	.01	.01	.02	.01	.02	.03
hi	.00	.09	-.01	-.02	-.01	-.02	—	-.01	.06	-.01	-.01	.00	.01	.00	.00
it	.01	.10	.00	-.01	.00	-.01	.01	—	.07	.00	.00	.01	.00	.01	.02
ja	-.05	.04	-.07	-.08	-.06	-.08	-.06	-.07	—	-.06	-.07	-.06	-.06	-.06	-.06
ro	.01	.10	-.01	-.01	.00	-.01	.01	.00	.06	—	.00	.00	.00	.00	.01
ru	.01	.10	.00	-.01	.01	-.01	.01	.00	.07	.00	—	.01	.00	.01	.02
sk	.01	.10	-.01	-.02	.00	-.02	.00	-.01	.06	.00	-.01	—	.00	.00	.01
uk	.01	.10	-.01	-.01	.00	-.01	.01	.00	.06	.00	.00	.00	—	.00	.01
vi	.01	.10	-.01	-.02	.00	-.02	.00	-.01	.06	.00	-.01	.00	.00	—	.01

Gemma 3 12B														
	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
ar	—	.70	.75	.78	.80	.78	.75	.80	.76	.78	.81	.71	.80	.81
az	.80	—	.67	.70	.69	.69	.67	.69	.69	.68	.71	.65	.70	.69
cs	.82	.76	—	.75	.74	.72	.69	.73	.71	.74	.76	.78	.77	.73
de	.89	.80	.83	—	.81	.81	.78	.80	.76	.78	.83	.71	.81	.80
el	.87	.80	.83	.89	—	.80	.78	.81	.74	.79	.82	.72	.81	.80
en	.88	.80	.82	.90	.89	—	.78	.80	.75	.77	.80	.69	.79	.80
hi	.85	.80	.81	.87	.86	.87	—	.75	.73	.73	.78	.65	.76	.76
it	.86	.80	.82	.88	.86	.88	.86	—	.75	.80	.81	.71	.79	.81
ja	.84	.81	.80	.86	.84	.86	.85	.85	—	.73	.77	.69	.75	.75
ro	.88	.79	.84	.89	.87	.90	.86	.88	.85	—	.79	.71	.78	.79
ru	.89	.81	.84	.91	.90	.91	.88	.88	.87	.89	—	.72	.83	.81
sk	.86	.80	.84	.87	.86	.87	.85	.87	.84	.86	.88	—	.76	.71
uk	.88	.80	.83	.90	.89	.89	.87	.88	.85	.89	.92	.88	—	.80
vi	.88	.78	.81	.88	.87	.88	.85	.87	.83	.88	.88	.84	.88	—
Avg.	.86	.79	.81	.85	.84	.83	.80	.82	.78	.79	.81	.73	.79	.77

Qwen 2.5 VL 7B														
	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
ar	—	.63	.79	.84	.79	.83	.78	.84	.81	.79	.85	.78	.82	.84
az	.69	—	.64	.64	.62	.61	.64	.65	.64	.63	.65	.64	.65	.63
cs	.83	.68	—	.85	.78	.81	.77	.84	.81	.80	.85	.85	.85	.80
de	.88	.70	.87	—	.81	.89	.81	.89	.84	.82	.92	.81	.88	.85
el	.86	.69	.84	.90	—	.79	.77	.82	.79	.76	.82	.75	.79	.79
en	.87	.68	.86	.94	.89	—	.80	.88	.83	.80	.89	.79	.84	.84
hi	.81	.69	.81	.83	.82	.83	—	.80	.78	.75	.81	.75	.80	.77
it	.87	.69	.87	.92	.88	.92	.83	—	.84	.84	.89	.80	.86	.85
ja	.85	.70	.85	.90	.87	.89	.85	.89	—	.78	.85	.79	.83	.81
ro	.85	.70	.86	.89	.87	.88	.82	.88	.86	—	.82	.77	.80	.79
ru	.88	.68	.87	.93	.89	.93	.84	.92	.90	.89	—	.83	.90	.86
sk	.84	.70	.90	.88	.85	.86	.82	.87	.87	.86	.88	—	.81	.77
uk	.87	.70	.88	.91	.89	.90	.84	.91	.89	.88	.91	.89	—	.83
vi	.88	.69	.85	.91	.87	.90	.82	.89	.89	.87	.90	.86	.89	—
Avg.	.84	.69	.84	.87	.83	.85	.80	.85	.83	.80	.85	.79	.82	.80

Qwen 3 VL 8B														
	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
ar	—	.73	.84	.89	.88	.89	.87	.88	.87	.85	.88	.84	.88	.88
az	.82	—	.73	.76	.74	.75	.76	.75	.74	.74	.75	.73	.75	.74
cs	.87	.79	—	.87	.86	.86	.86	.85	.84	.83	.87	.87	.87	.85
de	.91	.83	.87	—	.89	.92	.90	.90	.88	.87	.91	.86	.90	.89
el	.90	.81	.84	.90	—	.88	.89	.87	.86	.86	.89	.85	.88	.87
en	.92	.83	.88	.94	.91	—	.89	.90	.88	.86	.89	.85	.89	.89
hi	.90	.82	.86	.90	.88	.91	—	.88	.87	.86	.89	.85	.89	.88
it	.92	.83	.88	.93	.91	.94	.90	—	.87	.87	.88	.85	.88	.88
ja	.89	.82	.85	.91	.89	.91	.90	.91	—	.84	.87	.83	.86	.86
ro	.90	.82	.87	.92	.89	.92	.89	.92	.90	—	.86	.83	.86	.86
ru	.90	.82	.89	.94	.90	.93	.90	.92	.91	.92	—	.86	.91	.88
sk	.89	.82	.88	.91	.89	.91	.89	.92	.89	.90	.91	—	.86	.84
uk	.90	.81	.88	.91	.89	.92	.89	.91	.89	.91	.93	.91	—	.89
vi	.91	.82	.86	.92	.89	.92	.89	.92	.90	.91	.91	.90	.90	—
Avg.	.89	.81	.86	.90	.88	.90	.88	.89	.87	.86	.88	.85	.87	.86

Aya Vision 8B														
	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
ar	—	.40	.59	.54	.64	.48	.65	.58	.58	.63	.61	.56	.62	.62
az	.54	—	.44	.43	.45	.40	.45	.43	.45	.44	.45	.47	.44	.44
cs	.72	.53	—	.66	.69	.62	.64	.69	.66	.71	.72	.68	.72	.67
de	.77	.55	.77	—	.72	.73	.61	.75	.67	.69	.73	.57	.73	.72
el	.78	.54	.74	.82	—	.71	.71	.74	.71	.74	.77	.62	.77	.78
en	.79	.53	.76	.84	.84	—	.58	.71	.66	.65	.67	.53	.68	.72
hi	.75	.53	.73	.78	.77	.80	—	.64	.67	.67	.68	.59	.68	.68
it	.80	.55	.76	.84	.84	.85	.79	—	.69	.74	.73	.61	.73	.74
ja	.75	.52	.71	.78	.81	.81	.76	.81	—	.68	.69	.59	.70	.70
ro	.80	.55	.76	.82	.81	.81	.78	.84	.77	—	.72	.64	.73	.73
ru	.80	.55	.78	.84	.82	.84	.79	.84	.79	.82	—	.64	.79	.74
sk	.71	.55	.74	.72	.70	.70	.68	.71	.66	.73	.73	—	.63	.60
uk	.79	.55	.77	.82	.79	.78	.82	.76	.81	.86	.74	—	.74	—
vi	.79	.52	.74	.80	.82	.82	.77	.83	.79	.80	.81	.68	.79	—
Avg.	.75	.53	.71	.74	.75	.72	.69	.72	.68	.70	.71	.61	.69	.68

Jina VL														
	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
ar	—	.26	.54	.58	.43	.57	.45	.61	.48	.56	.56	.55	.55	.47
az	.46	—	.30	.24	.15	.26	.16	.23	.16	.29	.33	.34	.31	.16
cs	.65	.43	—	.53	.36	.55	.38	.53	.39	.55	.58	.68	.55	.40
de	.75	.48	.70	—	.44	.61	.48	.61	.50	.56	.62	.53	.57	.53
el	.70	.41	.64	.72	—	.38	.54	.44	.45	.39	.39	.35	.42	.48
en	.72	.41	.65	.80	.71	—	.44	.61	.49	.57	.58	.56	.55	.48
hi	.69	.46	.63	.71	.67	.69	—	.46	.52	.42	.39	.36	.43	.48
it	.76	.45	.68	.80	.71	.79	.70	—	.50	.56	.56	.53	.56	.51
ja	.71	.49	.65	.76	.67	.72	.69	.72	—	.40	.41	.38	.40	.52
ro	.76	.47	.68	.79	.71	.75	.71	.81	.72	—	.61	.59	.56	.42
ru	.75	.49	.71	.83	.71	.77	.71	.78	.78	.77	—	.62	.65	.45
sk	.71	.49	.76	.76	.69	.72	.70	.73	.72	.75	.78	—	.59	.38
uk	.72	.51	.70	.79	.72	.72	.70	.76	.73	.76	.84	.76	—	.43
vi	.76	.48	.65	.78	.68	.72	.69	.74	.74	.77	.76	.72	.74	—
Avg.	.70	.45	.64	.70	.59	.63	.57	.62	.55	.57	.57	.54	.53	.44

NLLB-SigLIP Base														
	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
ar	—	.57	.58	.70	.85	.90	.88	.68	.86	.66	.88	.61	.85	.72
az	.52	—	.51	.53	.58	.57	.57	.54	.56	.55	.57	.55	.57	.49
cs	.49	.39	—	.58	.60	.61	.60	.58	.57	.60	.59	.68	.59	.51
de	.66	.43	.40	—	.71	.73	.72	.66	.70	.62	.72	.59	.71	.58
el	.85	.50	.48	.66	—	.89	.87	.70	.84	.68	.88	.62	.84	.72
en	.88	.52	.50	.67	.88	—	.93	.71	.87	.70	.91	.64	.89	.73
hi	.84	.51	.48	.64	.84	.89	—	.69	.86	.69	.91	.63	.88	.72
it	.71	.52	.46	.57	.69	.73	.68	—	.67	.60	.70	.62	.70	.60
ja	.82	.50	.46	.64	.80	.85	.83	.69	—	.65	.85	.61	.82	.70
ro	.61	.44	.44	.54	.61	.63	.61	.57	.59	—	.67	.62	.68	.57
ru	.85	.50	.49	.68	.86	.89	.86	.70	.83	.63	—	.61	.89	.71
sk	.54	.43	.56	.48	.54	.56	.53	.50	.52	.49	.54	—	.63	.54
uk	.82	.50	.47	.65	.82	.85	.83	.69	.83	.61	.87	.54	—	.70
vi	.69	.44	.43	.54	.71	.72	.70	.58	.70	.51	.72	.46	.70	—
Avg.	.71	.48	.48	.61	.73	.76	.74	.64	.72	.61	.75	.60	.75	.64

SigLIP2 Base														
	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
ar	—	.47	.74	.78	.70	.75	.63	.76	.61	.67	.77	.72	.72	.73
az	.55	—	.45	.50	.45	.47	.42	.49	.35	.48	.49	.46	.46	.51
cs	.78	.55	—	.77	.69	.76	.60	.78	.58	.69	.78	.76	.71	.74
de	.79	.55	.83	—	.74	.85	.66	.85	.62	.72	.87	.78	.75	.80
el	.78	.57	.78	.77	—	.72	.62	.71	.56	.65	.74	.67	.69	.71
en	.81	.54	.81	.88	.77	—	.65	.84	.61	.70	.85	.77	.75	.80
hi	.73	.53	.72	.72	.70	.71	—	.66	.53	.59	.67	.63	.62	.66
it	.81	.54	.82	.87	.78	.87	.72	—	.62	.72	.85	.79	.76	.81
ja	.64	.49	.62	.63	.60	.63	.62	.64	—	.53				

Arabic										
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large
Gemma 3 12B	—	.68	.72	.40	.08	.64	.66	.64	.61	.64
Qwen 2.5 VL 7B	.79	—	.78	.34	.10	.67	.68	.66	.62	.64
Qwen 3 VL 8B	.82	.83	—	.40	.09	.73	.73	.73	.65	.67
Aya Vision 8B	.74	.73	.76	—	.06	.39	.39	.38	.36	.37
Jina VL	.60	.67	.62	.59	—	.07	.06	.06	.07	.07
NLLB-SigLIP Base	.72	.72	.75	.68	.57	—	.78	.75	.63	.64
NLLB-SigLIP Large	.73	.72	.76	.70	.57	.80	—	.75	.63	.65
MEXMA-SigLIP2	.73	.74	.76	.67	.59	.76	.75	—	.58	.63
SigLIP2 Base	.64	.62	.66	.59	.49	.60	.60	.61	—	.76
SigLIP2 Large	.65	.63	.67	.59	.49	.60	.61	.60	.75	—
SigLIP2 SO400M	.63	.60	.64	.58	.47	.60	.60	.59	.74	.82
SigLIP2 Giant	.62	.60	.64	.57	.47	.58	.60	.57	.71	.80
Avg.	.70	.69	.71	.56	.36	.59	.59	.58	.58	.60

Czech										
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large
Gemma 3 12B	—	.68	.69	.44	.09	.46	.57	.63	.59	.62
Qwen 2.5 VL 7B	.78	—	.76	.41	.11	.50	.62	.68	.63	.65
Qwen 3 VL 8B	.80	.81	—	.47	.09	.52	.65	.72	.68	.70
Aya Vision 8B	.72	.71	.72	—	.06	.34	.43	.47	.45	.46
Jina VL	.58	.64	.59	.55	—	.05	.06	.05	.06	.05
NLLB-SigLIP Base	.45	.48	.48	.43	.39	—	.56	.55	.47	.45
NLLB-SigLIP Large	.62	.65	.65	.58	.50	.57	—	.68	.59	.61
MEXMA-SigLIP2	.72	.75	.77	.66	.57	.49	.67	—	.62	.65
SigLIP2 Base	.66	.66	.68	.62	.51	.40	.57	.67	—	.77
SigLIP2 Large	.65	.66	.69	.62	.50	.42	.58	.65	.78	—
SigLIP2 SO400M	.67	.66	.69	.62	.50	.42	.56	.66	.77	.82
SigLIP2 Giant	.65	.65	.67	.61	.49	.42	.56	.66	.73	.79
Avg.	.66	.67	.67	.55	.35	.42	.53	.58	.58	.60

Greek										
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large
Gemma 3 12B	—	.65	.73	.59	.09	.65	.68	.66	.59	.63
Qwen 2.5 VL 7B	.80	—	.74	.46	.05	.65	.67	.65	.57	.60
Qwen 3 VL 8B	.81	.84	—	.57	.06	.72	.74	.72	.64	.67
Aya Vision 8B	.73	.75	.76	—	.10	.56	.57	.54	.49	.53
Jina VL	.58	.65	.60	.52	—	.05	.05	.05	.04	.05
NLLB-SigLIP Base	.72	.75	.77	.70	.57	—	.79	.74	.60	.62
NLLB-SigLIP Large	.74	.77	.78	.72	.57	.78	—	.76	.61	.65
MEXMA-SigLIP2	.74	.77	.80	.70	.58	.77	.79	—	.59	.60
SigLIP2 Base	.61	.62	.64	.58	.46	.59	.62	.60	—	.70
SigLIP2 Large	.64	.63	.66	.59	.46	.61	.63	.62	.73	—
SigLIP2 SO400M	.63	.63	.66	.60	.47	.62	.63	.62	.72	.78
SigLIP2 Giant	.63	.62	.64	.59	.46	.60	.63	.62	.71	.77
Avg.	.69	.70	.71	.60	.35	.60	.62	.60	.57	.60

Hindi										
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large
Gemma 3 12B	—	.59	.67	.51	.09	.61	.63	.61	.49	.58
Qwen 2.5 VL 7B	.76	—	.75	.39	.08	.64	.66	.66	.52	.58
Qwen 3 VL 8B	.80	.78	—	.47	.07	.73	.75	.73	.58	.64
Aya Vision 8B	.71	.68	.76	—	.09	.47	.47	.46	.39	.44
Jina VL	.59	.65	.58	.53	—	.05	.06	.06	.06	.07
NLLB-SigLIP Base	.72	.71	.75	.67	.55	—	.79	.75	.55	.59
NLLB-SigLIP Large	.74	.73	.77	.71	.56	.77	—	.76	.56	.60
MEXMA-SigLIP2	.75	.73	.78	.71	.57	.76	.79	—	.55	.58
SigLIP2 Base	.55	.53	.57	.52	.40	.51	.57	.55	—	.64
SigLIP2 Large	.63	.61	.67	.59	.47	.62	.65	.64	.66	—
SigLIP2 SO400M	.63	.60	.65	.59	.45	.61	.64	.63	.69	.77
SigLIP2 Giant	.57	.56	.60	.55	.42	.56	.60	.59	.67	.74
Avg.	.68	.65	.69	.57	.34	.58	.60	.59	.52	.57

Azerbaijani										
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large
Gemma 3 12B	—	.50	.58	.30	.05	.39	.44	.55	.36	.42
Qwen 2.5 VL 7B	.63	—	.61	.25	.07	.38	.43	.51	.33	.39
Qwen 3 VL 8B	.74	.65	—	.32	.04	.46	.49	.64	.40	.45
Aya Vision 8B	.49	.46	.51	—	.03	.26	.29	.30	.22	.27
Jina VL	.35	.47	.35	.29	—	.03	.02	.02	.03	.02
NLLB-SigLIP Base	.43	.43	.46	.33	.25	—	.55	.56	.32	.36
NLLB-SigLIP Large	.51	.47	.56	.39	.28	.55	—	.58	.39	.43
MEXMA-SigLIP2	.68	.58	.72	.48	.34	.51	.61	—	.43	.47
SigLIP2 Base	.40	.36	.44	.31	.22	.26	.35	.43	—	.55
SigLIP2 Large	.41	.37	.45	.33	.25	.27	.34	.43	.59	—
SigLIP2 SO400M	.44	.38	.47	.32	.24	.28	.36	.45	.56	.64
SigLIP2 Giant	.38	.31	.40	.29	.21	.26	.29	.38	.49	.60
Avg.	.50	.45	.50	.33	.18	.33	.38	.44	.37	.42

German										
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large
Gemma 3 12B	—	.69	.72	.56	.11	.55	.61	.66	.64	.67
Qwen 2.5 VL 7B	.83	—	.81	.56	.11	.61	.67	.72	.70	.72
Qwen 3 VL 8B	.84	.87	—	.59	.10	.61	.68	.73	.72	.73
Aya Vision 8B	.74	.76	.78	—	.10	.47	.53	.54	.56	.55
Jina VL	.66	.72	.68	.62	—	.08	.08	.07	.08	.08
NLLB-SigLIP Base	.60	.62	.63	.59	.53	—	.66	.64	.56	.57
NLLB-SigLIP Large	.71	.74	.74	.68	.62	.67	—	.71	.63	.62
MEXMA-SigLIP2	.77	.79	.80	.73	.65	.62	.75	—	.67	.69
SigLIP2 Base	.69	.69	.69	.63	.57	.50	.59	.65	—	.83
SigLIP2 Large	.68	.69	.70	.62	.56	.52	.60	.66	.80	—
SigLIP2 SO400M	.67	.68	.69	.63	.55	.51	.58	.66	.78	.82
SigLIP2 Giant	.66	.66	.67	.60	.54	.50	.58	.63	.75	.80
Avg.	.71	.72	.72	.62	.40	.51	.58	.61	.63	.64

English										
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large
Gemma 3 12B	—	.70	.71	.63	.07	.65	.66	.65	.65	.64
Qwen 2.5 VL 7B	.86	—	.85	.67	.08	.74	.75	.74	.73	.72
Qwen 3 VL 8B	.86	.89	—	.68	.08	.73	.75	.74	.73	.72
Aya Vision 8B	.80	.81	.81	—	.11	.63	.63	.63	.62	.63
Jina VL	.76	.81	.80	.72	—	.06	.07	.06	.06	.06
NLLB-SigLIP Base	.76	.78	.78	.74	.72	—	.79	.76	.67	.69
NLLB-SigLIP Large	.78	.79	.79	.74	.73	.81	—	.77	.69	.68
MEXMA-SigLIP2	.78	.80	.80	.75	.74	.79	.80	—	.68	.66
SigLIP2 Base	.71	.71	.72	.67	.65	.67	.69	.68	—	.83
SigLIP2 Large	.71	.70	.71	.67	.65	.67	.68	.68	.81	—
SigLIP2 SO400M	.69	.69	.70	.66	.64	.65	.67	.66	.80	.84
SigLIP2 Giant	.66	.66	.66	.62	.61	.62	.65	.63	.76	.80
Avg.	.76	.76	.76	.69	.46	.64	.65	.64	.66	.67

Italian										
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large
Gemma 3 12B	—	.68	.72	.56	.12	.52	.64	.65	.67	.67
Qwen 2.5 VL 7B	.82	—	.80	.52	.12	.58	.69	.69	.68	.69
Qwen 3 VL 8B	.83	.87	—	.57	.12	.60	.70	.72	.71	.72
Aya Vision 8B	.74	.79	.78	—	.12	.44	.52	.51	.52	.52
Jina VL	.68	.73	.69	.62	—	.08	.09	.08	.09	.09
NLLB-SigLIP Base	.60	.63	.64	.59	.55	—	.66	.64	.54	.52
NLLB-SigLIP Large	.70	.76	.77	.71	.63	.69	—	.72	.64	.65
MEXMA-SigLIP2	.75	.78	.79	.72	.67	.66	.77	—	.66	.68
SigLIP2 Base	.66	.68	.70	.63	.58	.51	.62	.66	—	.84
SigLIP2 Large	.69	.69	.72	.65	.59	.52	.63	.67	.79	—
SigLIP2 SO400M	.65	.65	.68	.62	.55	.49	.61	.65	.76	.80
SigLIP2 Giant	.64	.65	.67	.60	.55	.47	.60	.63	.75	.79
Avg.	.71	.72	.73	.62	.42	.51	.59	.60	.62	.63

Table 20: Models’ agreement as Cohen-Kappa coefficient, for each language, part 1.

Japanese											
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large	SigLIP2 SO400M
Gemma 3 12B	—	.65	.69	.51	.10	.58	.59	.60	.48	.48	.51
Qwen 2.5 VL 7B	.80	—	.77	.48	.12	.65	.67	.67	.50	.51	.56
Qwen 3 VL 8B	.80	.84	—	.54	.11	.69	.72	.72	.55	.55	.60
Aya Vision 8B	.71	.73	.76	—	.12	.49	.50	.50	.38	.38	.42
Jina VL	.59	.67	.60	.51	—	.07	.08	.08	.07	.07	.08
NLLB-SigLIP Base	.70	.73	.74	.68	.56	—	.75	.74	.49	.50	.55
NLLB-SigLIP Large	.71	.76	.76	.69	.57	.79	—	.75	.48	.50	.55
MEXMA-SigLIP2	.74	.76	.77	.68	.57	.79	.78	—	.47	.50	.53
SigLIP2 Base	.49	.50	.52	.47	.39	.47	.47	.47	—	.61	.61
SigLIP2 Large	.53	.52	.55	.49	.39	.49	.50	.49	.63	—	.66
SigLIP2 SO400M	.54	.54	.56	.51	.40	.50	.50	.51	.63	.69	—
SigLIP2 Giant	.50	.50	.53	.46	.37	.47	.47	.49	.59	.65	.70
Avg.	.65	.65	.66	.55	.34	.54	.55	.55	.48	.49	.52

Russian											
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large	SigLIP2 SO400M
Gemma 3 12B	—	.71	.74	.58	.09	.68	.67	.66	.66	.68	.69
Qwen 2.5 VL 7B	.84	—	.79	.51	.09	.72	.72	.70	.70	.71	.72
Qwen 3 VL 8B	.83	.87	—	.56	.07	.73	.73	.73	.72	.73	.74
Aya Vision 8B	.75	.79	.79	—	.09	.53	.53	.51	.52	.53	.54
Jina VL	.63	.68	.65	.61	—	.06	.06	.05	.06	.06	.05
NLLB-SigLIP Base	.74	.78	.78	.71	.62	—	.79	.76	.69	.68	.69
NLLB-SigLIP Large	.75	.78	.78	.71	.62	.80	—	.78	.68	.68	.70
MEXMA-SigLIP2	.78	.80	.80	.73	.63	.79	.80	—	.66	.66	.68
SigLIP2 Base	.68	.69	.70	.63	.55	.65	.64	.68	—	.82	.84
SigLIP2 Large	.70	.70	.72	.64	.55	.65	.66	.66	.80	—	.85
SigLIP2 SO400M	.68	.67	.68	.62	.52	.64	.63	.65	.79	.81	—
SigLIP2 Giant	.67	.66	.68	.63	.52	.63	.62	.65	.76	.82	.81
Avg.	.73	.74	.74	.63	.40	.63	.62	.62	.64	.65	.67

Ukrainian											
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large	SigLIP2 SO400M
Gemma 3 12B	—	.71	.74	.54	.12	.66	.66	.66	.63	.64	.66
Qwen 2.5 VL 7B	.81	—	.78	.48	.09	.68	.70	.70	.64	.66	.66
Qwen 3 VL 8B	.82	.84	—	.54	.08	.72	.74	.74	.67	.68	.69
Aya Vision 8B	.72	.75	.76	—	.07	.51	.53	.53	.49	.50	.51
Jina VL	.60	.66	.60	.57	—	.06	.06	.06	.06	.06	.05
NLLB-SigLIP Base	.72	.76	.75	.68	.57	—	.78	.74	.62	.63	.63
NLLB-SigLIP Large	.73	.77	.76	.68	.57	.79	—	.77	.64	.65	.65
MEXMA-SigLIP2	.75	.78	.79	.69	.59	.76	.78	—	.61	.64	.63
SigLIP2 Base	.64	.65	.67	.59	.50	.60	.61	.63	—	.76	.76
SigLIP2 Large	.66	.67	.68	.60	.51	.62	.62	.64	.75	—	.80
SigLIP2 SO400M	.65	.66	.68	.60	.50	.62	.62	.63	.73	.78	—
SigLIP2 Giant	.65	.66	.66	.60	.49	.62	.62	.63	.73	.78	.80
Avg.	.70	.72	.72	.60	.37	.60	.61	.61	.60	.62	.63

Romanian											
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large	SigLIP2 SO400M
Gemma 3 12B	—	.64	.70	.52	.09	.53	.62	.65	.57	.62	.64
Qwen 2.5 VL 7B	.80	—	.76	.42	.10	.52	.62	.66	.58	.62	.63
Qwen 3 VL 8B	.83	.83	—	.50	.09	.57	.67	.72	.62	.67	.69
Aya Vision 8B	.74	.74	.76	—	.06	.42	.50	.50	.45	.50	.49
Jina VL	.64	.71	.65	.61	—	.04	.06	.06	.05	.06	.06
NLLB-SigLIP Base	.56	.57	.57	.55	.47	—	.64	.61	.50	.51	.53
NLLB-SigLIP Large	.68	.70	.71	.64	.56	.63	—	.73	.58	.61	.61
MEXMA-SigLIP2	.76	.77	.78	.72	.62	.57	.70	—	.60	.65	.65
SigLIP2 Base	.61	.60	.62	.57	.48	.46	.58	.60	—	.70	.72
SigLIP2 Large	.67	.65	.68	.62	.52	.51	.60	.64	.71	—	.76
SigLIP2 SO400M	.66	.64	.66	.61	.51	.48	.59	.62	.69	.78	—
SigLIP2 Giant	.65	.63	.65	.59	.51	.49	.58	.62	.67	.77	.80
Avg.	.69	.68	.69	.58	.36	.48	.56	.58	.55	.59	.60

Slovak											
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large	SigLIP2 SO400M
Gemma 3 12B	—	.64	.65	.37	.08	.48	.55	.59	.55	.57	.59
Qwen 2.5 VL 7B	.77	—	.71	.35	.09	.51	.59	.65	.60	.61	.63
Qwen 3 VL 8B	.78	.82	—	.41	.08	.54	.66	.70	.65	.65	.68
Aya Vision 8B	.63	.68	.67	—	.06	.32	.37	.40	.38	.39	.41
Jina VL	.59	.67	.60	.56	—	.04	.05	.05	.05	.06	.05
NLLB-SigLIP Base	.48	.52	.52	.45	.43	—	.61	.56	.49	.48	.52
NLLB-SigLIP Large	.63	.65	.66	.55	.53	.55	—	.69	.58	.57	.60
MEXMA-SigLIP2	.73	.75	.77	.62	.59	.51	.69	—	.62	.62	.64
SigLIP2 Base	.64	.64	.66	.55	.50	.43	.54	.63	—	.77	.77
SigLIP2 Large	.65	.63	.66	.54	.48	.43	.54	.61	.75	—	.82
SigLIP2 SO400M	.63	.62	.65	.53	.48	.43	.53	.62	.74	.78	—
SigLIP2 Giant	.64	.63	.65	.54	.48	.42	.54	.62	.71	.76	.80
Avg.	.65	.66	.65	.50	.35	.42	.52	.55	.56	.57	.59

Vietnamese											
	Gemma 3 12B	Qwen 2.5 VL 7B	Qwen 3 VL 8B	Aya Vision 8B	Jina VL	NLLB-SigLIP Base	NLLB-SigLIP Large	MEXMA-SigLIP2	SigLIP2 Base	SigLIP2 Large	SigLIP2 SO400M
Gemma 3 12B	—	.66	.72	.59	.11	.54	.58	.63	.63	.65	.66
Qwen 2.5 VL 7B	.83	—	.78	.53	.10	.59	.61	.66	.63	.65	.66
Qwen 3 VL 8B	.84	.85	—	.58	.09	.60	.63	.70	.67	.69	.70
Aya Vision 8B	.74	.74	.77	—	.10	.46	.49	.54	.54	.55	.56
Jina VL	.61	.69	.61	.54	—	.06	.07	.07	.07	.08	.08
NLLB-SigLIP Base	.63	.64	.65	.59	.50	—	.67	.64	.53	.53	.54
NLLB-SigLIP Large	.65	.66	.68	.62	.52	.68	—	.67	.55	.56	.59
MEXMA-SigLIP2	.76	.77	.80	.72	.60	.68	.71	—	.61	.61	.61
SigLIP2 Base	.63	.63	.65	.59	.49	.52	.54	.61	—	.78	.77
SigLIP2 Large	.67	.66	.67	.62	.50	.54	.54	.62	.75	—	.79
SigLIP2 SO400M	.64	.63	.66	.60	.48	.52	.52	.61	.76	.79	—
SigLIP2 Giant	.65	.64	.65	.60	.49	.54	.55	.62	.73	.79	.79
Avg.	.70	.69	.69	.60	.36	.52	.54	.58	.59	.61	.63

Table 21: Models’ agreement as Cohen-Kappa coefficient, for each language, part 2.

Model	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi
Gemma 3 12B	.581 ±.031	.425 ±.032	.507 ±.032	.581 ±.031	.587 ±.031	.569 ±.031	.523 ±.031	.581 ±.032	.510 ±.032	.569 ±.031	.610 ±.030	.472 ±.031	.593 ±.031	.600 ±.031
Qwen 2.5 VL 7B	.595 ±.031	.335 ±.030	.586 ±.031	.689 ±.029	.571 ±.031	.713 ±.029	.529 ±.031	.663 ±.030	.607 ±.031	.571 ±.031	.686 ±.030	.544 ±.032	.630 ±.031	.622 ±.031
Qwen 3 VL 8B	.676 ±.030	.504 ±.031	.653 ±.030	.700 ±.029	.704 ±.029	.722 ±.028	.695 ±.029	.705 ±.029	.686 ±.030	.678 ±.029	.698 ±.029	.623 ±.030	.676 ±.030	.685 ±.029
Aya Vision 8B	.245 ±.028	.127 ±.021	.296 ±.029	.410 ±.031	.408 ±.031	.497 ±.032	.302 ±.029	.367 ±.031	.335 ±.030	.360 ±.030	.379 ±.031	.225 ±.027	.347 ±.030	.424 ±.032
Jina VL	.018 ±.008	.004 ±.004	.007 ±.005	.035 ±.012	.016 ±.008	.020 ±.009	.009 ±.006	.023 ±.009	.023 ±.009	.013 ±.007	.012 ±.007	.009 ±.006	.016 ±.008	.034 ±.011
NLLB-SigLIP Base	.665 ±.029	.339 ±.030	.382 ±.031	.488 ±.031	.661 ±.030	.677 ±.030	.651 ±.031	.478 ±.031	.635 ±.030	.450 ±.031	.682 ±.029	.407 ±.031	.654 ±.030	.517 ±.032
NLLB-SigLIP Large	.680 ±.030	.443 ±.031	.571 ±.031	.599 ±.031	.689 ±.029	.698 ±.029	.698 ±.029	.643 ±.030	.656 ±.030	.608 ±.030	.685 ±.029	.573 ±.031	.671 ±.029	.551 ±.031
MEXMA-SigLIP2	.652 ±.030	.657 ±.030	.668 ±.030	.687 ±.030	.653 ±.030	.677 ±.029	.684 ±.030	.676 ±.030	.641 ±.030	.676 ±.030	.678 ±.030	.670 ±.030	.673 ±.030	.653 ±.030
SigLIP2 Base	.630 ±.030	.339 ±.030	.645 ±.030	.686 ±.029	.587 ±.031	.719 ±.028	.501 ±.032	.674 ±.029	.450 ±.032	.591 ±.031	.687 ±.029	.620 ±.031	.624 ±.031	.642 ±.031
SigLIP2 Large	.638 ±.030	.397 ±.031	.671 ±.029	.709 ±.029	.633 ±.031	.721 ±.028	.617 ±.031	.695 ±.029	.460 ±.031	.639 ±.030	.697 ±.029	.644 ±.030	.654 ±.030	.668 ±.030
SigLIP2 SO400M	.633 ±.030	.428 ±.033	.664 ±.030	.705 ±.029	.659 ±.030	.705 ±.029	.610 ±.031	.689 ±.030	.510 ±.032	.661 ±.030	.698 ±.029	.658 ±.030	.660 ±.030	.652 ±.030
SigLIP2 Giant	.627 ±.031	.380 ±.031	.674 ±.030	.678 ±.030	.641 ±.031	.684 ±.029	.573 ±.031	.672 ±.030	.490 ±.032	.655 ±.030	.677 ±.029	.658 ±.030	.650 ±.030	.660 ±.030

Table 22: Confidence intervals for *strict accuracy* obtained with bootstrap.

Model	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
Gemma 3 12B	.476	.308	.352	.517	.494	.542	.499	.467	.470	.462	.535	.337	.506	.508	.462
Qwen 2.5 VL 7B	.525	.326	.462	.252	.507	.654	.471	.577	.510	.532	.590	.439	.527	.572	.496
Qwen 3 VL 8B	.584	.407	.546	.625	.630	.674	.619	.616	.598	.519	.646	.485	.595	.592	.581
Aya Vision 8B	.268	.101	.284	.352	.371	.299	.226	.328	.226	.328	.356	.203	.357	.327	.288
Jina VL	.010	.000	.006	.006	.001	.019	.004	.007	.007	.019	.001	.005	.002	.018	.008
Avg.	.373	.229	.330	.350	.401	.438	.364	.399	.362	.372	.426	.294	.397	.403	.367

Table 23: *Strict accuracy* for bidirectional retrieval with the instruction prompt in French, only for LLMs, i.e. c_{strict} from Equation (3).

Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.17	.05	.01	.00	-.08	.03	.01	.02	.00	-.01	.07	.01	.01	.02
az	-.17	—	-.14	-.17	-.18	-.24	-.16	-.17	-.17	-.19	-.18	-.11	-.16	-.18	-.17
cs	-.05	.14	—	-.05	-.06	-.14	-.02	-.05	-.03	-.06	-.06	.03	-.04	-.06	-.03
de	-.01	.17	.05	—	-.02	-.08	.02	.01	.01	.00	-.02	.08	.00	-.01	.01
el	.00	.18	.06	.02	—	-.07	.03	.02	.03	.01	.00	.09	.02	.01	.03
en	.08	.24	.14	.08	.07	—	.10	.09	.10	.09	.05	.16	.08	.08	.10
hi	-.03	.16	.02	-.02	-.03	-.10	—	-.02	.00	-.04	-.03	.05	-.01	-.02	-.01
it	-.01	.17	.05	-.01	-.02	-.09	.02	—	.01	-.01	-.03	.08	.00	-.01	.01
ja	-.02	.17	.03	-.01	-.03	-.10	.00	-.01	—	-.03	-.03	.06	-.01	-.02	.00
ro	.00	.19	.06	.00	-.01	-.09	.04	.01	.03	—	-.02	.09	.01	-.01	.02
ru	.01	.18	.06	.02	.00	-.05	.03	.03	.03	.02	—	.09	.02	.02	.04
sk	-.07	.11	-.03	-.08	-.09	-.16	-.05	-.08	-.06	-.09	-.09	—	-.07	-.09	-.06
uk	-.01	.16	.04	.00	-.02	-.08	.01	.00	.01	-.01	-.02	.07	—	-.01	.01
vi	-.01	.18	.06	.01	-.01	-.08	.02	.01	.02	.01	-.02	.09	.01	—	.02

Table 24: Difference between language-specific mistakes with the instruction prompt in French.

	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.949	.985	.987	.988	.982	.980	.989	.981	.986	.989	.981	.987	.989	.984
az	.967	—	.964	.969	.962	.963	.948	.969	.955	.962	.969	.962	.965	.965	.966
cs	.984	.949	—	.990	.986	.987	.976	.990	.979	.985	.990	.985	.987	.988	.984
de	.988	.952	.988	—	.985	.989	.974	.990	.979	.987	.990	.987	.988	.987	.985
el	.988	.948	.985	.986	—	.986	.983	.990	.980	.986	.990	.983	.988	.991	.985
en	.979	.946	.983	.988	.983	—	.976	.990	.980	.986	.986	.978	.984	.986	.982
hi	.984	.941	.980	.979	.986	.979	—	.983	.979	.983	.985	.976	.983	.989	.980
it	.989	.952	.989	.991	.988	.990	.981	—	.982	.989	.992	.988	.991	.991	.987
ja	.985	.946	.982	.983	.984	.984	.979	.985	—	.981	.987	.979	.983	.985	.982
ro	.986	.949	.984	.987	.987	.986	.977	.991	.978	—	.990	.983	.988	.989	.984
ru	.989	.952	.989	.992	.988	.988	.981	.992	.981	.989	—	.987	.989	.991	.986
sk	.979	.944	.981	.986	.980	.980	.969	.986	.974	.980	.986	—	.984	.982	.979
uk	.986	.951	.986	.988	.987	.985	.978	.990	.980	.986	.990	.984	—	.989	.984
vi	.990	.952	.987	.987	.991	.992	.985	.993	.983	.990	.992	.985	.990	—	.987
Avg.	.985	.952	.985	.987	.985	.985	.978	.988	.979	.985	.988	.983	.986	.987	.982

Table 25: Accuracy of the bilingual retrieval for in the random pairs scenario, *2txt* only.

Lng.	ar	az	cs	de	el	en	hi	it	ja	ro	ru	sk	uk	vi	Avg.
ar	—	.03	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
az	-.03	—	-.03	-.02	-.03	-.03	-.03	-.03	-.02	-.03	-.03	-.02	-.02	-.03	-.03
cs	.00	.03	—	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
de	.00	.02	.00	—	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
el	.00	.03	.00	.00	—	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
en	.00	.03	.00	.00	.00	—	.00	.00	.01	.00	.00	.01	.00	.00	.00
hi	.00	.03	.00	.00	.00	.00	—	.00	.00	.00	.00	.01	.00	.00	.00
it	.00	.03	.00	.00	.00	.00	.00	—	.01	.00	.00	.00	.00	.00	.00
ja	.00	.02	.00	.00	.00	-.01	.00	-.01	—	-.01	-.01	.00	.00	-.01	.00
ro	.00	.03	.00	.00	.00	.00	.00	.00	.01	—	.00	.00	.00	.00	.00
ru	.00	.03	.00	.00	.00	.00	.00	.00	.01	.00	—	.00	.00	.00	.00
sk	.00	.02	.00	.00	.00	-.01	-.01	.00	.00	.00	.00	—	.00	-.01	.00
uk	.00	.02	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	—	.00	.00
vi	.00	.03	.00	.00	.00	.00	.00	.00	.01	.00	.00	.01	.00	—	.01

Table 26: Difference between language-specific mistakes in the random pairs scenario.