Multilingual evaluation of image captioning: How far can we get with CLIP models?

Anonymous ACL submission

Abstract

The evaluation of image captions, looking at both linguistic fluency and semantic correspondence to visual contents, has witnessed a significant effort. Still, despite advancements such as the CLIPScore metric, multilingual captioning evaluation has remained relatively unexplored. This work assesses the use of CLIPScore in multilingual captioning, evaluating different models in a variety of settings. To address the lack of multilingual test data, we consider two different strategies: (1) using machine-translated datasets with human judgements, and (2) re-purposing multilingual datasets that target inference and reasoning. Our results show that multilingual CLIP models can perform on par with their English-centric counterparts on English benchmarks while allowing for multilingual assessments. Performance increases with model finetuning and according to model size. Larger models, trained with more data, attained similar performance to more advanced methods that extended the original CLIPScore. Tests with machine-translated data show that multilingual CLIPScore can also maintain a high correlation with human judgements across different languages, and additional tests with natively multilingual and multicultural data further attest to the high-quality assessments.

1 Introduction

Computer-generated image captions are nowadays commonly used as descriptive annotations. The image captioning task has been extensively studied, including in multilingual settings, with many recent approaches combining established vision encoders with large language model decoders (Ramos et al., 2023b,a; Yang et al., 2023; Geigle et al., 2023; Ramos et al., 2024). The automatic evaluation of captions, accounting for linguistic and visual contents, has also witnessed a significant effort. Approaches such as CLIPScore (Hessel et al., 2021) have been proposed to evaluate captions through cosine similarity between image and text embeddings, leveraging large-scale pre-trained vision and language models and achieving high correlations with human judgments. Still, despite the many recent advancements, most approaches are English-centric, while multilingual image captioning evaluation has remained relatively unexplored.

This work explores the use of CLIPScore in multilingual captioning, evaluating different CLIP models under various settings. Given the lack of available benchmarks for the evaluation of multilingual captioning metrics, we propose two different evaluation strategies: (1) using Machine Translation (MT) to obtain multilingual data from English-centric benchmarks, and (2) re-purposing multilingual benchmarks originally designed for the evaluation of the semantic inference and reasoning capabilities of vision-language models.

Through extensive experiments, we show that multilingual CLIP models achieve comparable or even better performance on English benchmarks while allowing for multilingual assessments. We also propose a multilingual fine-tuning strategy for CLIPScore, that allows to account for linguistic and cultural diversity while learning from human judgements, resulting in further performance improvements. Performance generally increases according to model size. Larger models, trained on more data, attained similar or even better performance to methods that extended the original CLIPScore (Sarto et al., 2023; Kim et al., 2022; Hu et al., 2023; Kim et al., 2023; Narins et al., 2024; Wada et al., 2024). Tests with machine-translated data show that multilingual CLIPScore can also maintain a high correlation with human judgements across different languages, and additional tests with natively multilingual and multicultural datafurther attest to high-quality of assessments across languages.

Our primary contributions include ① a comprehensive evaluation of existing models on English-centric and multilingual benchmarks, assessing correlation to human judgements; ② an extension of

English-centric benchmarks to multiple languages, incorporating human evaluations and diverse linguistic phenomena while preserving the original benchmarks' quality; ③ an adaptation of existing multilingual and multicultural datasets for captioning evaluation; ④ a finetuning strategy that accounts for linguistic and cultural diversity as well as alignment with human judgements, leading to multilingual CLIPSCore models that outperform previous work across several benchmarks[1].

2 Related Work

Conventional image captioning evaluation has relied on reference-based assessments, where machine-generated captions are compared against human-generated ones (i.e., the references). Frequently used metrics such as BLEU or CIDEr (Vedantam et al., 2015) rely on lexical matches, and hence may fail to capture finer nuances and semantic overlaps in rich captions. A recent shift in the evaluation paradigm involves the use of learned vision-and-language models to enable evaluation through reference-free metrics.

The CLIPScore metric (Hessel et al., 2021) was one of the first proposals for evaluating image captions that departed from the traditional metrics. Grounded in a vision-and-language encoder, specifically the original Contrastive Language-Image Pretraining (CLIP) model (Radford et al., 2021), this strategy employs a modified cosine similarity between representations for the input image and the caption under evaluation. CLIPScore exhibits a high correlation with human judgments across various datasets, and despite being a reference-free metric, it even surpasses established reference-based metrics like BLEU and CIDEr. The authors also introduced a reference-augmented version named RefCLIPScore, which additionally uses the cosine similarity between candidate and reference captions to further improve the correlation with human assessments. CLIPScore and RefCLIPScore are currently the most widely used learned metrics for captioning evaluation. However, many studies in the area still only report results using traditional lexical-based metrics. Some previous studies have also proposed the combination of CLIPScore and CIDEr through a simple weighted average (Qiu et al., 2023), arguing that this can further boost the correlation with human assessments.

PACScore (Sarto et al., 2023) extended CLIPScore by introducing a contrastive strategy that uses curated data to further finetune the CLIP projection layers for captioning evaluation. By augmenting with generated images and paraphrased texts, PACScore can achieve better correlations with human judgements across several datasets.

InfoMetIC also builds on CLIP, aiming to provide detailed and explainable feedback in the context of captioning evaluation (Hu et al., 2023), reporting incorrect words and unmentioned image regions at a fine-grained level, while also providing a text precision score, a vision recall score, and an overall quality score at a coarse-grained level. InfoMetIC was found to outperform CLIPScore even when the latter is finetuned on captioning data.

Mutual Information Divergence (MID) is instead a unified metric for multimodal generative models (Kim et al., 2022), targeting both text-to-image and image-to-text tasks. The metric quantifies the alignment between generated visual and textual features by employing the concept of Mutual Information (MI) from information theory.

Some recent studies have noted that metrics like CLIPScore can lack rating granularity, arguing the need for better benchmark datasets for assessing image captioning evaluation metrics (Ahmadi and Agrawal, 2023). Datasets like VICR (Narins et al., 2024) or Polaris (Wada et al., 2024) have recently been proposed, relying on more rigorous procedures for collecting human ratings, although still focusing only on the English language. Together with the proposal of the new datasets, the authors also showed that models trained specifically for image captioning evaluation can slightly outperform CLIPScore. We corroborate these findings, proposing a finetuned version of CLIPScore that outperforms other variants.

One previous study has specifically looked into multilingual image captioning evaluation (Kim et al., 2023), proposing a method based on finetuning the text encoder of CLIP with a language-agnostic method to distinguish the perturbed text from the original text. The authors have also developed a novel dataset to evaluate multilingual image captioning metrics, but unfortunately, this dataset is not yet publicly available.

3 Multilingual CLIPScore

The CLIPScore metric uses an adjusted cosine similarity to compare representations for the input im-

[1]The code and adapted datasets supporting our evaluation experiments will be made publicly available upon acceptance.

age and the caption being assessed, as originally described by Hessel et al. (2021). In our work, we adopted the original formulation. A more detailed explanation of the CLIPScore and RefCLIPScore metrics can be found in Appendix A.

To boost performance, we propose a strategy to finetune multilingual CLIP models in a setting that considers both linguistic and cultural diversity, while accounting for human preference alignment.

Two distinct datasets were used for finetuning. The first, CrossModal-3600, focuses on multilingual and multicultural imagery (Thapliyal et al., 2022), whereas the second, VICR, comprises English image-caption pairs that are evaluated by humans (Narins et al., 2024) and we machine translated to different languages following a strict translation scheme to help maintain quality. We also combined different losses, tailored to the specific characteristics of each dataset.

In more detail, to enhance the model's ability to process multilingual and multicultural instances, we finetuned it on both datasets using the original CLIP contrastive loss, which can be formally described as follows:

$$\mathcal{L}_C = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{e^{s_{i,i}/\tau}}{\sum_{j=1}^{N}e^{s_{i,j}/\tau}} + \log\frac{e^{s_{i,i}/\tau}}{\sum_{j=1}^{N}e^{s_{j,i}/\tau}}\right], \quad (1)$$

where N is the number of image-text pairs in a batch, $s_{i,j}$ is the similarity score between the i-th image and the j-th text description, and τ is a temperature parameter that scales the similarity scores and helps in controlling the concentration level of the distribution.

For the second dataset, to improve the alignment of CLIPScores with human ratings, we also consider a Pearson correlation loss:

$$\mathcal{L}_P = 1 - \frac{(x-\overline{x})^T(y-\overline{y})}{||(x-\overline{x})||\cdot||(y-\overline{y})||}, \quad (2)$$

where x is the vector of CLIPScores values, y is a vector with the human rating scores, and $\overline{x}$ and $\overline{y}$ are the respective average values.

Considering that both loss functions can benefit from larger batch sizes, we sample instances for training alternating between each task, without mixing instances from the different datasets in the same batch and applying the respective loss functions. We accumulate gradients for two steps before updating the network, effectively combining both loss effects while leveraging the benefits of larger batches, i.e., $\mathcal{L} \sim \mathcal{L}_C + \mathcal{L}_P$.

4 Experimental Evaluation

This section presents the datasets, the experimental setup, and the results for different CLIP models, considering English, multilingual, and multicultural scenarios for image captioning evaluation.

4.1 Datasets

For the assessment of human judgment correlations, experiments were conducted using the following well-established English-only datasets containing one or more human quality assessments for each image-caption pair.

- **Expert** (Flickr8K-Expert) contains $5,664$ pairs (Hodosh et al., 2013).
- **Crowdflower** (Flickr8K-CF) contains $47,830$ pairs (Hodosh et al., 2013).
- **Composite** contains $13,146$ pairs (Aditya et al., 2015).
- **VICR** contains $10,175$; $2,310$; and $3,161$ pairs respectively in training, validation and test splits (Narins et al., 2024).

To evaluate the robustness of our models to different linguistic phenomena, we used the VALSE(Parcalabescu et al., 2021) dataset, which contains $6,704$ correct image-caption pairs, plus their respective foil caption versions.

Comparing multilingual versus monolingual models on English data provides a limited overview of model performance. However, high-quality multilingual resources with curated data featuring human assessments of caption quality are scarce (or even unavailable), hindering multilingual evaluation. To mitigate this limitation, we proposed a translation scheme leveraging large machine translation models (Fan et al., 2021; Alves et al., 2024; Liu et al., 2020), combined with language and translation quality estimation models(Rei et al., 2022), to automatically translate English captions for which we already have human assessments, preserving the highest translation quality possible. With high-quality translations, human judgments should be valid across the different target languages. This strategy is further detailed in Appendix C.

Our language selection is in line with recent machine translation studies (Alves et al., 2024), covering high-resource languages (i.e., English, French, German, Spanish, and Chinese) and also mid- (i.e., Portuguese, Italian, and Russian) to low-resource languages (i.e., Dutch and Korean). We machine-translated both the VICR and VALSE datasets into nine languages using this technique.

In addition, to further expand our multilingual evaluation, we used naturally multilingual and multicultural datasets, i.e., XVNLI (Bugliarello et al., 2022) and MaRVL (Liu et al., 2021), re-purposing them for the evaluation of image captioning metrics. The original datasets feature differences in terms of language composition, and we thus also expanded the XVNLI dataset by translating its data from English into the languages present in the MaRVL dataset. We chose not to further extend MaRVL to other languages because the dataset features image-caption pairs that focus on culturally specific concepts in association with the target languages.

4.2 Evaluation Metrics

We evaluate the different models using correlation with human judgements, and also with classification tasks. Regarding the correlation experiments, we measure performance using three different correlation coefficients, namely Spearman ρ and Kendall τ with variations b and c. The correlation metrics are formally defined in Appendix B.

For the multilingual/multicultural experiments, we measure performance under the assumption that a caption entailed by an image should reflect a higher CLIPScore than a contradiction/foil caption.

4.3 Experiments and Results

This section presents experimental results for the different models and evaluation datasets, establishing a comparison with previously reported results and contributing to the multi-linguistic exploration of existing models and datasets. We also performed a qualitative study focusing on image-caption pairs that feature concepts that could be associated with cultural bias, which is reported in Appendix F.

4.3.1 Correlation Assessment on English Data

Table 1 displays the correlation results between CLIPScore values and human ratings, across the four English datasets and considering existing and publicly available English-only and multilingual CLIP models. The results show that the CLIPScore estimates can improve significantly with larger CLIP models trained with more data.

Comparing results against the performance of the original CLIPScore computed with a ViT-B/32 vision encoder, we can observe substantial improvements. Apple's ViT-H/14 model with 378px resolution achieves the highest correlation results among the English models, although the multilingual model of the same size from LAION is the second best-performing alternative (even outperforming the English-only model on the Composite dataset). The largest multilingual CLIP model, trained on approximately 5 billion instances, also outperforms the similarly sized English-only model trained with approximately 2 billion instances, perhaps indicating that exposure to diverse language data can enable CLIPScore values to better correlate with human judgements. Overall, the findings support the argument that CLIPScore with a multilingual model can maintain, or even improve the performance, over an English-only model.

In turn, Table 2 compares recent studies proposing other metrics against the best performing English-only and multilingual CLIPScore models[2]. We considered different evaluation settings, assessing results (a) without references, (b) using references, and (c) combining CIDER with RefCLIPScore when using references (Qiu et al., 2023).

Results confirm that the original CLIPScore outperforms standard captioning metrics in all evaluated datasets, such as BLEU or CIDEr. The correlations consistently improve with RefCLIPScore, but the combination of RefCLIPScore with CIDEr only improved the smaller CLIP model (by previously published results from Qiu et al. (2023)), instead decreasing the performance in the cases involving the other CLIPScore variants.

The multilingual model competes head-to-head with the best English-only models, outperforming the previously published results for almost all the metrics, whether human references are considered or not. This competitive performance in both reference-free and reference-aided settings further solidifies the potential of multilingual CLIPScore.

Lastly, it is also interesting to note that our fine-tuned model outperformed the original multilingual LAION ViT-H model, and also Apple's best model across all CLIPScore variants, on both the VICR and Expert datasets, even without using references. The version of our model that uses references achieved the best performance for these datasets, surpassing even specialized architectures such as VICR, InfoMetIC, and RefPACScore.

4.3.2 Correlation on Multilingual Data

We used the VICR multilingual variants to analyse the human ruman ratings, considering the best per-

[2]Note that we computed all the correlation scores for the CLIPScore variants and traditional image caption metrics based on lexical matches, whereas the results for other more recent metrics are taken from the corresponding publications.

		VICR			Expert			CrowdFlower			Composite		
		τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ
English	openai/clip-vit-base-patch32	60.7	67.1	76.9	51.1	51.5	63.1	34.4	17.8	42.4	50.6	51.4	67.9
	apple/DFN5B-CLIP-ViT-H-14-378	67.4	**74.4**	**83.1**	56.3	**56.6**	**68.4**	**38.5**	**19.9**	**47.1**	55.0	55.9	72.8
	apple/DFN5B-CLIP-ViT-H-14	66.5	73.5	82.4	55.6	56.0	67.7	38.2	19.7	46.8	54.5	55.3	72.3
	apple/DFN2B-CLIP-ViT-L-14	65.8	72.6	81.6	55.5	55.6	67.5	37.1	19.2	45.6	53.6	54.3	71.3
	laion/CLIP-ViT-g-14-laion2B-s12B-b42K	66.4	73.0	82.0	55.3	55.1	66.9	37.1	19.1	45.5	54.9	55.7	72.7
	laion/CLIP-ViT-H-14-laion2B-s32B-b79K	66.4	73.1	82.1	54.9	54.9	66.7	37.2	19.2	45.6	53.6	54.4	71.2
	laion/CLIP-ViT-L-14-laion2B-s32B-b82K	65.9	72.6	81.7	54.4	54.5	66.3	36.7	18.9	45.1	53.6	54.4	71.3
	BAAI/AltCLIP	65.1	71.9	81.1	54.1	54.4	66.2	36.2	18.7	44.6	53.8	54.6	71.6
	openai/clip-vit-large-patch14-336	62.0	68.5	78.1	52.6	53.0	64.6	35.4	18.3	43.7	52.8	53.6	70.3
	openai/clip-vit-large-patch14	62.3	68.9	78.5	52.6	53.0	64.6	35.2	18.2	43.4	52.5	53.3	70.0
Multilingual	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	**67.6**	73.0	82.4	**57.4**	54.3	67.3	38.2	19.4	46.2	**55.5**	**56.2**	**73.2**
	BAAI/AltCLIP-m18	66.6	73.4	82.3	54.8	55.0	66.8	37.1	19.2	45.6	55.2	56.0	73.0
	M-CLIP/XLM-Roberta-Large-Vit-B-16Plus	64.6	71.3	80.6	54.6	54.8	66.7	36.0	18.6	44.3	52.1	52.9	69.6
	BAAI/AltCLIP-m9	64.2	70.9	80.3	54.1	54.4	66.2	36.4	18.8	44.8	53.9	54.6	71.6
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	63.3	69.8	79.3	52.7	52.8	64.5	35.2	18.2	43.3	49.9	50.6	67.1
	M-CLIP/XLM-Roberta-Large-Vit-L-14	62.2	68.7	78.4	53.0	53.4	65.0	35.4	18.3	43.7	52.9	53.6	70.4
	M-CLIP/XLM-Roberta-Large-Vit-B-32	60.5	66.9	76.7	51.8	52.2	63.9	34.4	17.8	42.4	50.7	51.4	67.9
	sentence-transformers/clip-ViT-B-32-multilingual-v1	60.3	66.7	76.6	51.5	51.8	63.6	33.3	17.2	41.1	48.9	49.6	65.9

Table 1: Correlation between CLIPScore values and human rankings, considering a set of different English (top rows) and multilingual (bottom rows) CLIP models.

		VICR			Expert			CrowdFlower			Composite		
		τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ
Related Work	BLEU1	57.9	63.7	74.0	32.2	32.3	40.4	17.9	9.3	22.3	45.8	46.2	63.0
	BLEU4	54.8	60.4	70.5	30.6	30.8	38.7	16.9	8.7	21.0	46.4	46.9	63.7
	CIDEr	63.1	69.8	79.3	43.6	43.9	54.3	24.6	12.0	29.3	48.1	48.8	65.0
	CLIPScore	60.7	67.1	76.9	51.1	51.5	63.1	34.4	17.8	42.4	50.6	51.3	67.9
	RefCLIPScore	66.3	73.3	82.2	52.0	52.4	63.7	36.4	18.8	44.7	56.8	57.6	74.7
	CLIP+CIDEr	66.8	73.8	82.6	53.1	53.4	65.3	33.9	17.5	41.8	54.3	55.1	72.2
	MID (Kim et al., 2022)	–	–	–	–	54.9	–	37.3	–	–	–	–	–
	InfoMetIC (Hu et al., 2023)	–	–	–	–	54.2	–	36.3	–	–	–	59.2	–
	InfoMetIC$^+$ (Hu et al., 2023)	–	–	–	–	55.5	–	36.6	–	–	–	**59.3**	–
	PR-MCS (Kim et al., 2023)	–	–	–	–	50.6	65.6	–	–	–	–	–	–
	VICR (Narins et al., 2024)	–	75.8	–	–	53.1	–	–	–	–	–	–	–
	PACScore (Sarto et al., 2023)	–	–	–	53.9	54.3	–	36.0	18.6	–	51.5	55.7	–
	RefPACScore (Sarto et al., 2023)	–	–	–	55.5	55.9	–	37.6	19.5	–	53.0	57.3	–
CLIPScore variants	English CLIPScore	67.4	74.4	83.1	56.3	56.6	68.4	38.5	19.9	47.1	55.1	55.9	72.8
	English RefCLIPScore	68.3	75.4	83.8	56.8	57.1	68.9	38.6	**19.9**	**47.3**	56.4	57.2	74.0
	English CLIP+CIDEr	67.6	74.8	83.3	56.0	56.4	68.5	35.6	18.4	43.8	53.9	54.7	71.7
	Multilingual CLIPScore	67.6	73.0	82.4	57.4	54.3	67.3	38.2	19.4	46.2	55.5	56.2	73.2
	Multilingual RefCLIPScore	69.1	74.6	83.6	58.1	55.0	67.9	**38.8**	19.7	46.9	**57.3**	58.1	**75.0**
	Multilingual CLIP+CIDEr	67.6	74.7	83.4	55.5	55.9	67.8	36.3	18.8	44.7	55.6	56.4	73.5
	Finetuned CLIPScore	68.7	75.4	84.1	59.8	57.1	70.2	37.8	19.3	45.9	47.4	48.0	64.7
	Finetuned RefCLIPScore	**69.7**	**76.5**	**84.8**	**60.2**	57.5	**70.6**	37.6	19.1	45.6	53.3	54.0	70.8
	Finetuned CLIP+CIDEr	68.5	75.8	84.2	57.3	**57.7**	70.0	35.8	18.5	44.0	52.5	53.2	70.0

Table 2: Comparison between published results and our best English, multilingual and finetuned (also multilingual) CLIPScore models, considering settings (a) without human references, (b) using human references, and (c) combining CIDEr with CLIPScore when using human references.

forming multilingual model in English-only data, and also our finetuned model version.

Table 3 displays the correlation between multilingual CLIPScore values and human ratings, across the different languages. We observe that our finetuned version achieves a better correlation with human judgments in both the reference and reference-free settings, across all evaluated languages and for all correlation metrics. The finetuned CLIPScore is strongly correlated with human preferences in high-resource languages (i.e., English, French, German, Spanish, and Chinese), and it also exhibits excellent performance in medium- and low-resource languages. We achieve average correlations of 67.9, 74.6, and 83.5, respectively with the τ_b, τ_c, and Spearman ρ metrics in the reference-free setting, and an average correlation of 68.9, 75.7, and 84.3, respectively with the τ_b, τ_c, and Spearman ρ using references. An additional table is provided in Appendix G regarding different loss variants for model finetuning, and with different model sizes. The findings in the appendix support the idea that

	Without Finetuning									With Finetuning					
	CIDEr			Multilingual CLIPScore			Multilingual RefCLIPScore			Finetuned Multilingual CLIPScore			Finetuned Multilingual RefCLIPScore		
	τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ
English	63.1	69.8	79.3	67.6	73.0	82.4	69.1	74.6	83.6	68.7	75.4	84.1	**69.7**	**76.5**	**84.8**
German	58.3	64.4	74.4	66.1	72.3	81.5	68.0	74.4	83.2	68.0	74.8	83.6	**69.1**	**75.9**	**84.4**
French	60.0	66.4	76.3	65.9	71.8	81.3	67.6	73.6	82.7	68.0	74.7	83.6	**69.1**	**75.9**	**84.4**
Spanish	61.3	67.8	77.6	66.3	72.7	81.8	66.3	72.7	81.8	67.8	74.5	83.5	**68.9**	**75.7**	**84.3**
Chinese	58.5	64.6	74.6	65.0	71.4	80.8	66.7	73.2	82.3	67.4	74.0	83.1	**68.5**	**75.2**	**83.9**
Portuguese	61.3	67.8	77.6	66.1	72.2	81.6	67.8	74.1	83.0	67.9	74.6	83.5	**69.0**	**75.8**	**84.3**
Italian	60.7	67.1	77.0	65.8	72.1	81.4	67.5	74.0	82.9	67.9	74.6	83.5	**68.9**	**75.7**	**84.3**
Russian	54.0	59.6	69.4	65.0	71.4	80.7	66.7	73.3	82.2	67.6	74.2	83.2	**68.8**	**75.6**	**84.2**
Korean	58.1	64.3	74.1	64.3	70.5	80.0	66.3	72.7	81.8	67.4	74.1	83.1	**68.3**	**75.0**	**83.8**
Dutch	58.1	64.2	74.2	65.8	72.0	81.4	67.7	74.1	83.0	68.2	74.9	83.8	**69.0**	**75.8**	**84.3**
Average	59.3	65.6	75.5	65.8	71.9	81.3	67.4	73.7	82.7	67.9	74.6	83.5	**68.9**	**75.7**	**84.3**

Table 3: Correlation between M-CLIPScore values and human rankings, considering machine-translated versions of the validated image caption rating dataset into 9 different languages besides the original English. The last row presents macro-averaged correlation results across all the languages (including English).

smaller models can have higher human judgment correlation gains using our finetuning strategy, compared to the respective original models.

In an ideal scenario, i.e. assuming perfect machine translation results and that CLIP performs equally well across the languages, the correlations between CLIPScore values across the different languages would be equal one, signifying a perfect alignment. To explore deviations from this behaviour, we can use heatmaps to visually represent the interrelationships between CLIPScore values across languages. In Figure 1, we plot the Pearson correlations between the best multilingual CLIPScore and our finetuned version for different languages, considering either (a) the complete set of original/translated instances from the VICR dataset, (b) the subset of instances with COMETKiwi (Rei et al., 2022) translation quality scores below the 25th percentile value for each language, and (c) the subset of instances with COMETKiwi scores above the 75th percentile value for each language.

Looking at the left plot in Figure 1, which features the Pearson correlations when considering the entire VICR dataset, we observe consistently high values across all languages. As expected, the plots confirm that the CLIPScore correlations do depend on the quality of the translated captions, and that the most significant differences occur primarily in languages using non-Latin scripts. In turn, the upper diagonal of the second and third plots contains the correlations when considering only high-quality translations, corresponding to slightly higher values compared with the values of their respective lower diagonal. We also observe a significant improvement across the correlation between the languages when using our finetuned CLIPScore version. Although this improvement is somewhat expected, since the finetuned model saw "in distribution" data, it may also be an indication of the quality of our multilingual data during training, thus supporting the fact that our translation strategy maintains the quality of the data in the different languages.

4.3.3 Multilingual Classification

This section explores the robustness of the multilingual CLIPScore assessments through different types of classification tasks. Inspired by previous work (Hessel et al., 2021; Sarto et al., 2023) which assessed accuracy in English-only benchmarks, our goal is to delve deeper into the nuanced realm of multilingual and multicultural understanding.

Robustness to linguistic phenomena: One of our experiments used machine translated versions of the VALSE dataset, explicitly designed to evaluate the robustness to particular phenomena, such as inconsistencies in numeric quantities or spatial relations (Parcalabescu et al., 2021). VALSE specifically comprises seven tests that encompass a range of linguistic structures. In each test, a model is presented with a visual input and is tasked with distinguishing genuine captions from altered versions (i.e., foils), where a word or phrase has been modified to exhibit a specific linguistic phenomenon. Additional information about the dataset is provided in Appendix D. Table 4 displays the average performance across different language variants, considering that the CLIPScore values regarding the true captions should be higher than the ones related to the corresponding foils. A more detailed breakdown of the performance across the different tasks within VALSE is available in Appendix G. The results show that our finetuned model delivers the highest average performance across nearly

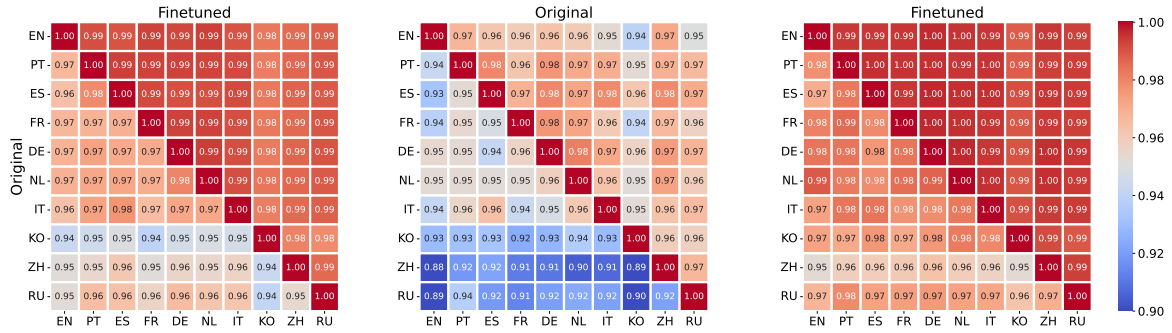

Figure 1: Pearson correlation scores between different languages, for the best multilingual CLIPScore model and our finetuned version. The first heatmap considers the complete set of instances from the VICR dataset, reporting results for both the original and finetuned model versions (lower/upper diagonal values). The second and third heatmaps consider the subset of instances with COMETKiwi scores below/above the 25th/75th percentile value for each language (lower/upper diagonal values) for the best performing multilingual CLIPScore model and our finetuned model version, respectfully.

all languages for both model size variants. Compared to the models reported in the original VALSE dataset paper, our finetuned model was only outperformed by the multi-task ViLBERT 12-in-1 model proposed by Lu et al. (2020).

Classification of multicultural instances: We also used naturally multilingual datasets (i.e., XVNLI and MaRVL) to assess the multilingual and multicultural capabilities of CLIPScore models.

Each instance in the XVNLI dataset contains an image-caption pair and a categorical label associated with the relationship between the pair. This label can be either (a) contradiction, (b) neutral, or (c) entailment. Based on these labels, we defined three multilingual classification experiments under this scenario, leveraging concordant/discordant instances as illustrated in Figure 2:

Experiment 1: This setting only considers contradiction and entailment instances, under the assumption that the order of the CLIPScore values should match the order of the labels.

Experiment 2: In a more challenging scenario, we can consider a larger set of duplets and the ordering between the three possible labels.

Experiment 3: In this case, we also consider the three possible labels, but we now assess triples of instances A, B, and C from the dataset, sharing the same image. We only assume a correct classification when we achieve a perfect match between the order of the labels and the CLIPScore values.

In the case of the MaRVL dataset, each instance consists of a caption, a pair of images, and a Boolean label with the value *true* when the caption accurately matches the images and *false* when the caption is incorrect (e.g. because its contents only

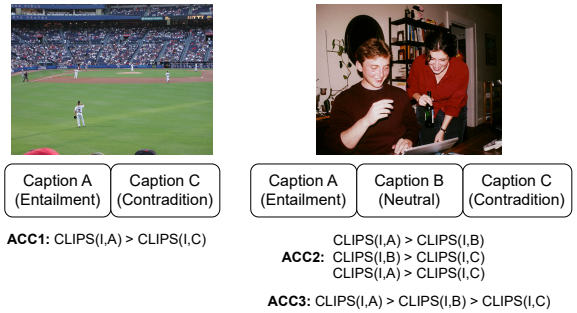

Figure 2: The three different XVNLI multilingual classification experiments, where accuracy is defined with basis on comparisons between CLIPSCore values.

describe at maximum one of the images instead of both). The data can be analyzed considering four instances simultaneously, sharing the same caption but featuring distinct pairs of images.

Experiment 1: We consider the scenario where the CLIPScore for the image that best aligns with the caption should be higher than the CLIPScore for the image that least aligns with the caption.

Experiment 2: In a more challenging scenario, we consider sets of four instances sharing the same caption and decide on a correct classification only if all best-aligning captions have a CLIPScore higher than all the least aligning captions.

A more detailed explanation for both the XVNLI and MaRVL experiments is given in Appendix E.

Table 5 contains the results for the two multilingual classification tasks. In the XVNLI experiments, as expected, we notice a consistent trend where results for Experiment 3 are worse than those for Experiment 2, which in turn are worse than those for Experiment 1. The performance is generally high in Experiments 1 and 2, for all the languages except Swahili and Tamil. The lower perfor-

Proposed Models											VALSE Models	
Finetuning method	English	German	French	Spanish	Chinese	Portuguese	Italian	Russian	Korean	Dutch	Model	English
Both	69.7	**67.3**	**67.5**	**67.9**	66.9	**69.0**	**66.0**	**67.1**	**60.6**	**66.2**	CLIP	64.0
Pearson	69.2	67.2	66.8	67.4	**67.3**	67.2	66.0	65.0	59.4	65.8	LXMERT	59.6
Contrastive	68.4	65.0	65.5	65.7	65.6	66.4	64.6	64.8	59.2	63.7	ViLBERT	46.4
None	67.6	64.8	64.2	65.6	64.0	65.4	64.4	63.4	58.4	62.5	12-in-1	**75.1**

Table 4: Average accuracy scores for the different classification tasks present in the VALSE dataset and its multilingual variants, considering different CLIP models.

	XVNLI						MaRVL			
	Accuracy 1		Accuracy 2		Accuracy 3		Accuracy 1		Accuracy 2	
	Original	Finetuned	Original	Finetuned	Original	Finetuned	Original	Finetuned	Original	Finetuned
English	92.3	91.5	80.7	81.6	47.7	51.0	91.1	91.7	81.4	80.7
Indonesian	89.9	92.0	80.5	80.8	50.0	49.0	92.4	93.4	82.3	83.3
Mandarin	88.1	90.7	78.8	79.5	46.7	46.7	89.4	91.5	80.1	81.3
Swahili	67.1	66.0	60.7	60.8	24.0	22.4	85.0	85.0	65.3	65.3
Tamil	71.6	68.7	62.8	62.7	25.3	25.7	85.7	88.0	74.7	77.5
Turkish	87.8	88.1	77.9	79.7	44.1	49.7	93.5	92.9	87.0	86.0
Arabic	84.6	85.4	76.3	76.3	44.7	41.8	-	-	-	-
French	86.7	90.5	77.9	79.9	45.4	47.7	-	-	-	-
Spanish	86.2	87.8	77.7	78.4	46.7	45.1	-	-	-	-
Russian	87.5	88.9	78.0	79.2	44.4	46.7	-	-	-	-
Overall	84.2	85.0	75.1	75.9	41.9	42.6	89.5	90.4	78.5	79.0

Table 5: Accuracy for different classification tasks defined over the datasets derived from XVNLI and MaRVL.

mance in these languages can perhaps be attributed to a lower quality in the machine translation results, and to the ability of the multilingual CLIP model in handling text in these lower resource languages. However, in the MaRVL experiments where the instances never involve machine translation, the task is performed with high accuracy across all languages. Once again, we observe that our finetuned CLIPScore version is capable of achieving moderate performance gains when compared to the original model version, across all evaluated multicultural classification tasks.

While the experiments with the XVNLI and MaRVL datasets provide interesting insights into the effectiveness of multilingual CLIPScore models, they also involve several important limitations. For instance, considering the XVNLI experiments, previous studies have reported good results in multimodal inference leveraging CLIP (Song et al., 2022). However, the authors of SNLI-VE (Xie et al., 2019), from which XVNLI is derived, noted that good performance (i.e., an accuracy up to 67%) can be achieved when looking only at the information in the textual hypothesis, without the visual premise. This points to significant biases in the XVNLI data. In the case of the MaRVL experiments, given that the captions refer to a pair instead of individual images, the CLIPScore values can be unreliable when attempting to match images to textual sentences. Previous studies have noted that CLIP models can treat inputs as a bag-of-words and suffer from a concept association bias (Yamada et al., 2022), e.g. ignoring the missing information when two concepts are present in one of the inputs while the other only contains a single concept. Hence, while already useful, there is room for future work and analysis of the robustness of caption evaluation metrics, accounting for the discussed limitations.

5 Conclusions

We studied the potential of CLIPScore variants in multilingual image captioning evaluation, considering a variety of settings and demonstrating that multilingual CLIP maintains or even surpasses performance on English benchmarks, while enabling versatile multilingual assessments.

Performance improves with larger model sizes and increased training data, aligning with the scalability of CLIP. On English data, large multilingual CLIP models can even outperform more advanced methods that extend the original CLIPScore metric or involve specific training for captioning evaluation. Experiments with machine-translated data reveal that a finetuned version of multilingual CLIPScore strongly correlates with human judgments across languages of varying complexity, and enhances the robustness of CLIP to different linguistic phenomena across languages. Additionally, assessments with natively multilingual and multicultural datasets, specifically with data from the XVNLI and MaRVL benchmarks repurposed for evaluating captioning metrics, reaffirm the ability of multilingual CLIPScore models to consistently provide high-quality assessments in varied settings.

Limitations and Ethical Considerations

Although our work does not raise new ethical issues within the domain of vision-language models (e.g., we conducted our experiments on public datasets, carefully designed for academic research and extensively used in previous studies), there are some general important concerns.

Models like CLIP are, for instance, notorious for their internal biases, e.g. inherited from the training data itself. We therefore recommend caution in the use of the approach proposed in this paper, and anticipate further research into the specific issue of model biases, before relying on our work beyond research environments. Another important limitation in the work reported on this paper concerns the use of machine translated data in some of the evaluation experiments, which despite our best efforts to avoid translation errors can still lead to different types of biases and to the reliance on artificially impoverished language. The development of manually curated benchmarks, specifically designed for the assessment of multilingual metrics for image captioning evaluation, is left as an important challenge for future work.

We also note that we used Github Copilot[3] during the development of our research work, and we used ChatGPT[4] for minor verifications during the preparation of this manuscript.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Saba Ahmadi and Aishwarya Agrawal. 2023. An examination of the robustness of reference-free image captioning evaluation metrics. *arXiv preprint arXiv:2305.14998*.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of the International Conference on Machine Learning*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107).

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv preprint arXiv:2307.06930*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. InfoMetIC: An informative metric for reference-free image caption evaluation. *arXiv preprint arXiv:2305.06002*.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. 2022. Mutual information divergence: A unified metric for multimodal generative models. In *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Yongil Kim, Yerin Hwang, Hyeongu Yun, Seunghyun Yoon, Trung Bui, and Kyomin Jung. 2023. PR-MCS: Perturbation robust metric for multilingual image captioning. *arXiv preprint arXiv:2303.08389*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task

[3] https://github.com/features/copilot
[4] https://openai.com/chatgpt/

vision and language representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10437–10446.

Lothar D Narins, Andrew Scott, Aakash Gautam, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta Ihorn, Yue-Ting Siu, James M Mason, Alexander Blum, et al. 2024. Validated image caption rating dataset. In *Proceedings of the Annual Meeting on Neural Information Processing Systems*.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2021. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*.

Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. 2023. Gender biases in automatic evaluation metrics for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International conference on machine learning*.

Rita Ramos, Emanuele Bugliarello, Bruno Martins, and Desmond Elliott. 2024. PAELLA: Parameter-efficient lightweight language-agnostic captioning model. In *Findings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023a. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of the Association for Computational Linguistics*.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023b. SmallCap: Lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ricardo Rei, Nuno M Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José GC de Souza, and André FT Martins. 2023. Scaling up COMETKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. *arXiv preprint arXiv:2309.11925*.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *arXiv preprint arXiv:2209.06243*.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022. CLIP models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. *arXiv preprint arXiv:2402.18091*.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Yutaro Yamada, Yingtian Tang, and Ilker Yildirim. 2022. When are lemons purple? the concept association bias of CLIP. *arXiv preprint arXiv:2212.12043*.

Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun, and Yuexian Zou. 2023. MultiCapCLIP: Auto-encoding prompts for zero-shot multilingual visual captioning. *arXiv preprint arXiv:2308.13218*.

A The CLIPScore Metric

We now formally describe the CLIPScore and RefCLIPScore metrics (Hessel et al., 2021), which in our study are assessed in multilingual image captioning scenarios. In brief, we have that CLIPScore is based on a modified cosine similarity between representations for the input image and the caption under evaluation. The image and the caption are both passed through the respective feature extractors from a given CLIP model. Then, we compute the cosine similarity of the resultant embeddings, adjusting the resulting value through a re-scaling operation. For an image with visual CLIP embedding $\mathbf{v}$ and a candidate caption with textual CLIP embedding $\mathbf{c}$, a re-scaling parameter is set as $w = 2.5$ and we compute the corresponding CLIPScore as follows:

$$\text{CLIPScore}(\mathbf{c}, \mathbf{v}) = w \times \max(\cos(\mathbf{c}, \mathbf{v}), 0). \quad (3)$$

To compute a corpus-level CLIPScore, e.g. for evaluating the overall quality of a captioning method over a given dataset of images, we can simply average over all the image-candidate pairs.

Note that CLIPScore does not depend on the availability of underlying references for each of the images in an evaluation dataset. However, an extension named RefCLIPScore was also proposed, which additionally extracts the vector representations $\mathbf{R}$ of each available reference with the CLIP text encoder, and computes the harmonic mean of the CLIPScore value from Equation 3, and the maximal reference cosine similarity:

$$\begin{aligned}\text{RefCLIPScore}(\mathbf{c}, \mathbf{R}, \mathbf{v}) = \\ \text{H-Mean}(\text{CLIPScore}(\mathbf{c}, \mathbf{v}), \\ \max(\max_{\mathbf{r} \in \mathbf{R}} \cos(\mathbf{c}, \mathbf{r}), 0)).\end{aligned} \quad (4)$$

B The Correlation Metrics

This appendix presents a formal definition of the metrics used in the correlation experiments.

Seeing each of our evaluation datasets as sets of n observations with the form $(\hat{y}_1, y_1), \ldots, (\hat{y}_n, y_n)$, for CLIPScore values $\hat{y}_i$ and reference ratings y_i, the Spearman correlation coefficient ρ is defined as the Pearson correlation between the results of converting the scores $\hat{y}_i$ and y_i to ranks.

Instead of using ranks, we can also define any pair of observations $(\hat{y}_i, y_i)$ and $(\hat{y}_j, y_j)$, where $i < j$, as concordant (or otherwise discordant) if the sort order of the instances agrees (i.e. if either both $\hat{y}_i > \hat{y}_j$ and $y_i > y_j$ holds, or both $\hat{y}_i < \hat{y}_j$ and $y_i < y_j$). Based on pairs, the Kendall τ correlation coefficient assesses the strength of association between the CLIPScore values and the reference ratings, with the τ_b variant accounting for ties and being defined as:

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \quad (5)$$

where n_c is the number of concordant pairs, n_d the number of discordant pairs, $n_0 = n(n-1)/2$, $n_1 = \sum_i t_i(t_i - 1)/2$, $n_2 = \sum_j u_j(u_j - 1)/2$, t_i is the number of tied values in the i^{th} group of ties for the CLIPScore, and u_j is the number of tied values in the j^{th} group of ties for the reference ratings.

In turn, τ_c accounts with the fact that the underlying scales of the scores are different for CLIPScore and the reference ratings, being defined as:

$$\tau_c = \frac{n_c - n_d}{n_0} \times \frac{n-1}{n} \times \frac{m}{m-1}, \quad (6)$$

where m is the number of values in the ranking scale for the reference ratings.

C The Machine Translation Scheme

This appendix describes the translation scheme that was used to machine translate the datasets used in our experiments. This scheme is design to mitigate low-quality translations, or hallucinations generated by the machine translation model, thus providing a reliable dataset at the end. We specifically used a large (i.e., 1.2 billion parameters) open-access multilingual machine translation model named M2M100 (Fan et al., 2021), available on the HuggingFace[5] model hub. M2M100 was trained on a range of high and low-resource languages from different families and using different scripts, achieving state-of-the-art performance across a diverse set of 100 languages.

While machine translated data allows us to assess multilingual captioning metrics, the results will depend not only on the performance of the metrics but also on the quality of the translations. Low-quality translations, or hallucinations generated by the translation model, will impact the caption and break our assumption that human ratings for the English data can be transferred across languages. To address this issue, we propose to use the COMETKiwi (Rei et al., 2023) machine translation quality estimation metric to control for translation

[5] https://huggingface.co/facebook/m2m100_1.2B

quality, assessing the impact of low quality translations on the observed results.

We specifically began by translating the VICR dataset, featuring English captions with human ratings and also reference captions originally from the MSCOCO and Flickr8K datasets. For each caption, whether a candidate or a reference, we return 25 translations using a beam search technique with 100 beams. Subsequently, we filtered the candidates with a language checker, to ensure proper translation into the intended language. After the language check, we selected for each instance the translation that scored higher based on a large COMETKiwi model[6].

D Description of the Datasets

The following datasets were used in the tests that assessed correlation with human judgment.

- **Flickr8K-Expert (Hodosh et al., 2013):** This dataset comprises $16,992$ expert human judgments for $5,664$ image-caption pairs from Flickr8K. Human assessors graded captions on a scale of 1 to 4, where 4 indicates a caption that accurately describes the image without errors, and 1 signifies a caption unrelated to the image.
- **Flickr8K-CF (Hodosh et al., 2013):** This dataset consists of $145,000$ binary quality judgments collected with CrowdFlower, involving $47,830$ image-caption pairs with $1,000$ unique Flickr8K images. Each pair received at least three binary judgments, and we use the proportion of *yes* annotations as the score for each pair.
- **Composite (Aditya et al., 2015):** This dataset contains $13,146$ image-caption pairs taken from MSCOCO (2007 images), Flickr8K (997 images), and Flickr30K (991 images). Each image originally had five reference captions. One of these references was chosen for human rating and subsequently removed from the reference set that is to be used when assessing evaluation metrics.
- **VICR (Narins et al., 2024):** The Validated Image Caption Rating (VICR) dataset features 68,217 ratings, collected through a gamified approach, for 15,646 image-caption pairs involving 9,990 distinct images. The authors of the dataset demonstrated that it exhibits a superior inter-rater agreement compared to other alternatives (e.g., an improvement of 19% in Fleiss' κ when compared to the agreement for the Flickr8K-Expert dataset), and it features a more balanced distribution across various levels of caption quality. In our tests, we used the test split of the VICR dataset, which includes 3,161 image-caption pairs, with 2,000 images from the MSCOCO 2014 validation dataset and 1,161 images from the Flickr8K dataset. When using VICR to finetune CLIP models with a contrastive loss, we used the original image captions from MSCOCO or Flickr8K.

All the previous datasets are originally available only for English, but we translated them to nine other different languages using the approach described in Appendix C.

For the experiments that assessed accuracy in terms of distinguishing correct vs incorrect captions, we used the following datasets.

- **VALSE (Parcalabescu et al., 2021)**: VALSE is designed to test visio-linguistic grounding capabilities on specific linguistic phenomena. It is composed by seven tasks, each with the same structure: given a visual input, a model is asked to distinguish real captions from foils, where a foil is constructed from a caption by altering a word or phrase that realizes a specific linguistic phenomenon. The tests include: (a) existential quantifiers, where models need to differentiate between examples (i) where there is no entity of a certain type or (ii) where one or more of these entities are visible in an image; (b) plurality, where models need to distinguish between noun phrases denoting a single entity in an image (*exactly one flower*), versus multiple entities (*some flowers*); (c) counting, where models needs to differentiate between examples where the specific number of entities in the associated image is correct or incorrect, given the statement; (d) spatial relations, where Models need to distinguish between different spatial relations, with foils differing from the original caption only by the replacement of a spatial preposition; (e) actions, particularly (i) action replacement and (ii) actant swaping, where models need to (i) identify whether an action mentioned in the text matches the action seen in the image (e.g., *a man shouts* versus *smiles*

[6] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl

at a woman), and (ii) correctly identify the participants of an action and the roles they play (e.g., is it the man who is shouting or is it the woman); (f) coreference, where models need to perform pronominal coreference resolution, encompassing cases where (i) the pronoun has a noun (phrase) antecedent and pronoun and (noun) phrase are both grounded in the visual modality (e.g., in *a woman is driving a motorcycle*, is she wearing a helmet?), and cases where (ii) the pronoun refers to a region in the image or even to the entire image (e.g., *is this outside?*); (g) foil-it cases, in which the foil minimally differs from the original caption, only by swapping a important noun.

- **XVNLI (Bugliarello et al., 2022):** XVNLI is a multilingual dataset for evaluating vision-language inference, challenging models to predict entailment relationships between a textual hypothesis and an image premise. XVNLI includes high/mid-resource languages like Arabic, French, Spanish, Russian, and English. This dataset includes 1,164 instances per language, each featuring an image and two captions in different languages. There are 357 unique images in total.
- **MaRVL (Liu et al., 2021):** MaRVL follows a format similar to the English NLVR2 dataset (Suhr et al., 2018) and is designed as a multicultural vision-language reasoning dataset, where the goal is to determine the correctness of a sentence about a pair of images. MaRVL predominantly comprises very low-resource languages: Indonesian, Chinese, Swahili, Tamil, and Turkish. This dataset includes around one thousand instances per language, each featuring two image and one caption. There are 1,411 unique captions in total. The English content is composed by collecting the reverse translations from the low-resource languages into English, as provided by the authors on the original GitHub[7].

For model training, besides instances in the training split from the aforementioned VICR dataset, we also used data from the natively multilingual CrossModal-3600 dataset (i.e., XM3600, in short).

- **XM3600 (Thapliyal et al., 2022):** This is a geographically-diverse set of 3600 images annotated with human-generated reference captions in 36 languages. The images were selected from all across the world, covering regions where the 36 languages are spoken, and consistently annotating captions in terms of style across all languages, while avoiding annotation artifacts due to direct translation.

[7] https://github.com/marvl-challenge/marvl-code/tree/master/data

E Multicultural Experiments

This appendix details the datasets and experimental settings that were considered for the tests including natively multilingual and multicultural data.

E.1 Settings for the XVNLI Experiments

Each instance in the XVNLI dataset contains an image-caption pair and a categorical label associated with the relationship between the pair. This label can be either (a) contradiction, (b) neutral, or (c) entailment. With basis on the labels, we defined three multilingual classification experiments under this scenario, leveraging concordant/discordant instances as illustrated in Figure 2:

Experiment 1: This setting only considers instances with the extreme label classes (i.e., contradiction and entailment), noting that some previous studies have pointed to the fact that SNLI-VE, from which XVNLI is derived, has some problems in the annotations for the neutral class (Kayser et al., 2021). We compare pairs of instances A and B with the same image, in which the label associated with A differs from the label associated with B. When computing CLIPScore values individually for the instances A and B, the order of the CLIPScore values should match the order of the labels (i.e., contradiction $<$ entailment).

Experiment 2: In a more challenging scenario, we can consider a larger set of instances and the ordering between the three possible labels (i.e. entailment $>$ neutral $>$ contradiction), i.e. including also the neutral class. Similarly to the previous case, by fixing an image and comparing pairs of captions associated with that image with different labels, we assess the matching of the order between the labels against the CLIPScore values.

Experiment 3: In this case, we also consider the three possible labels, but we now assess triples of instances A, B, and C from the dataset, sharing the same image. We only assume a correct classification when we achieve a perfect match between the order of the labels and the CLIPScore values.

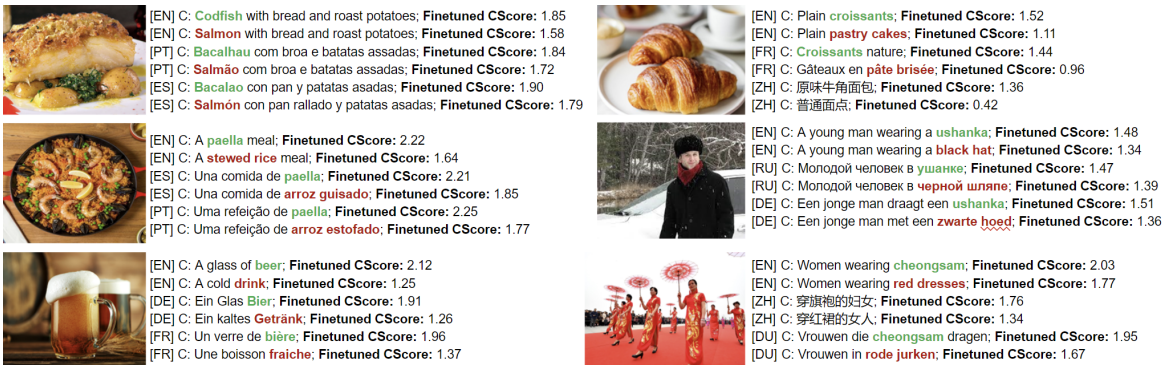

Figure 3: Multilingual CLIPScore values for image-caption pairs featuring concepts biased to particular languages.

E.2 Settings for the MaRVL Experiments

In the case of the MaRVL dataset, each instance consists of a caption, a pair of images, and a Boolean label with the value *true* when the caption accurately matches the images, and *false* when the caption is incorrect (e.g., because its contents only describe at maximum one of the images instead of both). The data can be analyzed considering four instances at a time, sharing the same caption but featuring distinct pairs of images. MaRVL was designed in such a way that, within these four instances, two of them are labeled as *true* while the remaining two are labeled as *false*. We consider two multilingual classification experiments under this scenario, defined as follows:

Experiment 1: We draw comparisons between pairs of instances with distinct labels. For the instance labeled as *true*, we compute the CLIPScore values for both images associated with the caption and select the maximum, obtaining the score for the image that best aligns with the caption. Conversely, we perform a similar computation for the instance labeled as *false*, this time choosing the minimum CLIPScore value, which results in the score for the image that least aligns with the caption, presumably the incorrect image. The maximum CLIPScore value in an instance labeled as *true* should be higher than the minimum CLIPScore value of an instance labeled as *false*.

Experiment 2: In a more challenging scenario, we consider sets of four instances sharing the same caption, and decide on a correct classification only if all maximum CLIPScore values of the *true* instances are higher than all the minimum CLIPScore values of the *false* instances.

F A Qualitative Study with Captions Featuring Cluturally Related Concepts

We performed a small qualitative study on image-caption pairs that feature concepts where some languages should exhibit a particular bias (e.g., *codfish* in the case of Portugal, *paella* for Spain, *beer* for Germany, *croissant* for France, *ushanka* for Russia, and *cheongsam* for China). We attempted to see if the multilingual CLIPScore could distinguish between two plausible captions, where one mentions a specific concept that should better match the image. Figure 3 shows that the multilingual CLIPScore is indeed capable of distinguishing nuanced multicultural concepts and favouring culturally specific captions over generic ones.

G Additional Classification Results

Table 6 presents classification results on the different tasks from the VALSE dataset, separately for each of the considered languages and comparing CLIP models of different sizes under different finetuning strategies (i.e., without model finetuning, considering only the contrastive loss, only the Pearson correlation loss, or the combined loss function). The results show that the smaller CLIP model can achieve significantly higher gains from finetuning, approaching the performance of the larger CLIP model. Better results are, in general, obtained when considering the combined loss function.

In turn, Table 7 presents correlation results on the VICR dataset, separately for each language and comparing the same two CLIP models under the different finetuning strategies. Results again show that the smaller CLIP model approach the performance of the larger model, with better results consistently obtained when considering the combined loss function.

	Model	Finetuning	Existence quantifiers	Plurality number	Counting balanced	Counting sns.	Counting adv.	Sp.Rel relations	Action repl.	Action actant swap	Coreference standard	Coreference clean	Foil-it!	Avg.
English	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	**81.6**	**76.3**	**68.3**	**75.8**	66.1	**66.5**	86.0	**62.1**	46.9	42.3	95.0	**69.7**
		Pearson	80.8	75.9	67.1	74.7	65.8	64.7	**87.5**	62.0	46.9	41.3	94.8	69.2
		Contrastive	75.4	75.3	66.7	74.4	**66.9**	61.5	87.0	58.5	47.6	44.2	**95.2**	68.4
		None	74.7	69.6	66.9	72.4	65.3	61.7	86.7	56.1	51.3	45.2	94.2	67.6
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	74.7	70.0	65.2	70.9	68.7	59.8	80.2	61.9	48.6	49.0	90.6	67.2
		Pearson	80.4	68.7	64.3	71.2	63.7	55.7	80.1	58.8	49.4	52.9	89.2	66.8
		Contrastive	72.3	70.7	63.5	70.7	66.6	60.2	80.6	60.5	45.5	40.4	92.0	65.7
		None	70.7	64.7	62.2	67.3	63.4	56.8	81.6	53.2	**52.3**	**56.7**	89.7	65.3
German	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	78.8	**71.7**	68.4	**74.6**	58.5	**57.2**	**77.6**	62.1	53.0	43.3	**94.9**	**67.3**
		Pearson	**79.8**	71.1	**68.9**	73.7	55.3	56.8	77.2	**62.5**	**54.7**	45.2	94.2	67.2
		Contrastive	64.8	71.0	66.2	72.7	61.9	54.6	76.1	59.2	48.9	45.2	94.3	65.0
		None	69.3	69.2	68.8	72.6	56.6	53.3	75.9	57.4	53.4	44.2	91.9	64.8
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	68.9	67.2	65.8	70.8	60.8	56.3	71.3	57.3	50.7	40.4	89.9	63.6
		Pearson	73.1	64.3	67.4	69.4	53.5	55.5	70.5	54.6	52.5	**47.1**	88.2	63.3
		Contrastive	60.0	68.2	63.8	68.0	**62.8**	57.0	72.5	59.7	46.5	**47.1**	90.2	63.3
		None	64.8	65.5	65.7	67.1	54.0	55.7	69.4	57.4	50.3	44.2	86.2	61.8
French	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	80.2	**75.1**	**69.7**	72.4	61.4	**56.1**	75.5	**61.4**	48.6	48.1	**94.5**	**67.5**
		Pearson	**82.2**	73.4	69.0	**72.8**	56.9	55.9	75.8	61.3	46.6	47.1	94.0	66.8
		Contrastive	69.9	73.1	66.6	70.7	**63.5**	55.5	73.6	57.3	49.6	46.2	**94.5**	65.5
		None	70.7	65.7	66.0	69.4	53.8	53.3	**75.9**	53.6	53.8	50.0	93.4	64.2
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	70.7	67.8	66.5	68.6	60.5	53.8	71.0	60.6	54.5	42.3	89.0	64.1
		Pearson	74.5	66.2	64.7	68.7	51.8	53.5	71.1	56.9	**55.4**	49.0	88.0	63.6
		Contrastive	61.0	68.6	63.8	69.6	58.0	54.6	70.4	60.9	49.9	46.2	90.6	63.0
		None	66.9	62.2	61.5	64.8	50.1	54.0	72.2	57.2	52.8	**60.6**	87.9	62.7
Spanish	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	84.2	**73.0**	69.2	**75.1**	64.4	58.9	**80.2**	**64.6**	44.8	40.4	91.9	**67.9**
		Pearson	**85.1**	72.3	**69.6**	73.2	61.2	58.3	78.9	63.6	**47.0**	40.4	91.2	67.4
		Contrastive	74.7	72.7	66.8	73.2	**65.4**	56.6	78.7	61.7	41.9	38.5	**92.9**	65.7
		None	75.8	66.5	68.2	67.2	61.8	**61.3**	79.0	61.5	42.9	**46.2**	91.1	65.6
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	82.2	66.2	65.8	67.8	60.9	53.8	72.8	61.2	46.3	40.4	88.3	64.2
		Pearson	80.8	62.9	65.7	66.8	52.1	52.5	74.2	60.4	45.9	38.5	86.9	62.4
		Contrastive	69.9	67.8	61.5	66.6	57.3	53.6	72.7	60.6	45.6	39.4	89.3	62.2
		None	73.9	61.5	63.8	66.2	54.0	52.9	73.5	60.8	43.5	36.5	86.5	61.2
Chinese	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	78.0	**69.3**	**69.0**	**72.2**	65.4	**55.7**	74.1	63.8	48.0	49.0	91.6	66.9
		Pearson	**82.0**	68.5	68.3	**72.2**	63.4	52.7	**75.2**	**64.7**	49.2	**53.8**	90.8	**67.3**
		Contrastive	69.1	**69.3**	68.4	70.2	**66.4**	53.8	73.6	59.5	48.4	50.0	**92.2**	65.6
		None	69.1	66.4	66.8	69.3	63.0	55.1	75.9	62.2	46.3	41.3	88.3	64.0
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	76.0	63.3	62.2	67.9	62.7	50.1	71.0	59.1	48.7	49.0	86.0	63.3
		Pearson	79.0	62.3	62.9	68.2	54.1	52.1	72.1	57.0	**50.4**	44.2	85.7	62.6
		Contrastive	71.9	64.9	60.7	66.1	64.0	52.1	70.4	59.4	50.0	47.1	87.6	63.1
		None	69.3	59.8	63.8	66.2	60.1	54.4	71.6	60.8	49.6	38.5	83.5	61.6
Portuguese	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	78.8	**74.3**	**70.2**	**74.2**	63.1	54.4	**79.0**	64.1	50.7	56.7	**93.7**	**69.0**
		Pearson	**80.4**	71.0	68.0	72.1	55.9	51.8	78.5	**64.5**	47.2	**57.7**	92.7	67.2
		Contrastive	68.3	72.6	65.7	71.7	62.5	53.3	76.2	60.5	50.8	54.8	93.6	66.4
		None	68.5	67.6	68.3	71.2	62.1	56.1	77.2	58.3	50.6	48.1	91.9	65.4
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	75.2	69.3	68.3	67.0	50.8	**76.5**	59.5	54.2	**54.2**	51.0	89.2	66.0
		Pearson	78.0	65.2	64.9	67.2	60.6	50.1	77.2	58.4	50.3	53.8	87.5	64.8
		Contrastive	65.7	68.6	61.5	67.6	**64.3**	52.1	75.0	58.3	53.0	51.0	90.0	64.3
		None	65.0	65.2	63.2	65.9	61.1	52.1	75.6	56.1	47.5	51.0	86.3	62.6
Italian	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	77.2	72.3	**68.5**	**75.0**	57.9	56.6	77.5	64.0	43.5	39.4	93.5	66.0
		Pearson	**77.8**	71.2	68.2	73.1	52.7	56.4	77.8	**64.8**	45.5	45.2	93.1	**66.0**
		Contrastive	61.4	**72.6**	64.5	71.7	60.1	56.8	77.5	61.2	46.9	44.2	**94.2**	64.6
		None	67.3	67.5	65.9	68.2	52.1	**58.7**	**78.4**	62.7	46.5	50.0	91.3	64.4
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	66.1	68.6	65.9	70.2	60.6	56.8	73.1	58.7	51.8	48.1	88.4	64.4
		Pearson	73.3	66.3	65.7	69.2	51.7	55.5	72.4	58.3	52.4	48.1	87.5	63.7
		Contrastive	55.6	67.6	63.0	69.3	**60.9**	54.8	74.2	59.9	51.8	47.1	89.7	63.1
		None	65.9	63.2	63.4	64.9	56.0	58.1	73.5	63.0	**53.5**	**51.9**	87.8	63.8
Russian	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	**79.2**	**74.4**	**67.6**	**73.3**	**62.4**	**57.0**	**74.8**	62.7	47.9	**47.1**	91.3	**67.1**
		Pearson	78.6	72.3	65.4	71.3	57.5	55.0	74.4	**63.4**	45.8	40.4	91.2	65.0
		Contrastive	65.5	74.1	65.6	70.3	61.8	55.3	74.1	59.6	**50.4**	42.3	**93.1**	64.8
		None	70.5	65.8	66.0	68.0	56.2	53.6	72.2	58.3	49.9	49.0	88.3	63.4
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	69.3	68.6	65.7	68.8	60.1	53.1	71.9	61.2	45.3	42.3	88.2	63.1
		Pearson	72.5	64.3	66.8	68.2	54.6	53.3	72.5	58.5	44.8	44.2	86.1	62.3
		Contrastive	60.2	70.2	62.4	66.7	58.9	53.5	71.5	61.0	46.3	44.2	89.5	62.2
		None	65.1	64.5	63.9	64.9	52.5	50.7	72.5	55.7	46.3	39.4	85.2	60.1
Korean	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	61.0	**69.4**	64.9	**70.9**	59.9	54.4	69.0	57.5	36.4	32.7	90.9	**60.6**
		Pearson	63.2	67.1	**65.3**	**70.9**	54.4	52.9	67.9	58.4	37.4	26.0	90.0	59.4
		Contrastive	48.5	**69.4**	63.8	68.2	**61.9**	52.1	**69.1**	55.5	35.6	35.6	**91.5**	59.2
		None	56.4	61.1	63.9	66.8	54.4	**57.4**	**69.1**	49.2	**39.3**	38.5	86.3	58.4
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	59.0	63.3	63.5	67.2	58.9	52.3	64.2	**60.0**	35.0	40.4	85.6	59.0
		Pearson	**66.3**	59.7	63.0	65.2	51.4	52.1	64.0	54.8	36.7	33.7	84.2	57.4
		Contrastive	50.1	63.2	62.3	65.8	58.0	54.0	64.6	59.1	36.4	43.3	87.1	58.5
		None	63.2	59.3	61.9	64.0	58.9	52.9	66.0	55.1	37.6	**45.2**	80.6	58.6
Dutch	laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k	Both	78.2	**72.3**	**68.5**	**74.6**	61.2	53.3	75.5	60.4	48.3	42.3	**93.6**	**66.2**
		Pearson	**79.0**	70.3	66.6	72.7	56.0	53.3	**75.9**	**61.7**	49.3	46.2	93.1	65.8
		Contrastive	63.8	72.2	67.6	72.8	**61.9**	52.9	75.0	58.3	44.6	38.5	**93.6**	63.7
		None	64.4	66.5	62.8	67.2	49.2	55.1	**75.9**	59.2	46.6	47.1	92.9	62.5
	laion/CLIP-ViT-B-32-xlm-roberta-base-laion5B-s13B-b90k	Both	74.9	67.3	65.3	69.1	57.7	55.3	71.5	56.6	48.6	54.8	89.3	64.6
		Pearson	75.4	64.5	65.3	69.0	49.3	53.5	71.9	54.1	46.8	52.9	88.3	62.8
		Contrastive	62.0	68.0	61.6	67.4	59.6	**57.9**	74.1	58.5	**50.1**	**55.8**	89.7	64.1
		None	63.2	60.4	64.9	67.0	47.2	52.7	72.1	58.2	44.9	49.0	86.5	60.5

Table 6: Performance of multimodal models on the multilingual VALSE benchmark according to different metrics. **sns.** Counting small numbers. **adv.** Counting adversarial. **repl.** Action replacement. **Sp.rel.** Spatial relations.

	Multilingual CLIP ViT-B									Multilingual CLIP ViT-H								
	Contrastive			Pearson			Combined			Contrastive			Pearson			Combined		
	τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ	τ_b	τ_c	ρ
ENG	65.8	72.2	81.5	66.7	73.2	82.3	67.2	73.5	82.6	68.6	75.0	83.8	64.5	70.6	80.6	**68.7**	**75.4**	**84.1**
GER	65.1	71.4	80.8	65.6	72.1	81.4	66.5	72.6	81.9	**68.0**	74.3	83.2	62.9	69.1	79.3	**68.0**	**74.8**	**83.6**
FRE	65.4	71.7	81.1	66.0	72.4	81.7	66.7	72.9	82.1	**68.0**	74.4	83.3	62.8	69.0	79.3	**68.0**	**74.7**	**83.6**
SPA	65.1	71.3	80.8	65.9	72.6	81.7	66.4	72.7	81.9	**67.8**	74.3	83.2	62.4	68.6	78.9	**67.8**	**74.5**	**83.5**
CHI	64.4	70.6	80.2	64.7	71.3	80.7	65.7	71.9	81.3	**67.4**	73.8	82.8	62.1	68.3	78.6	**67.4**	**74.0**	**83.1**
POR	65.2	71.5	80.9	65.7	72.1	81.4	66.5	72.7	81.9	**67.9**	74.3	83.2	62.7	68.9	79.2	**67.9**	**74.6**	**83.5**
ITA	65.0	71.3	80.7	65.6	72.2	81.4	66.3	72.6	81.8	**67.9**	74.3	83.2	62.5	68.7	79.0	**67.9**	**74.6**	**83.5**
RUS	64.6	70.8	80.3	65.1	71.6	80.9	65.9	71.9	81.4	67.4	73.8	82.8	62.7	68.9	79.2	**67.6**	**74.2**	**83.2**
KOR	63.8	70.0	79.6	64.3	70.8	80.2	65.1	71.2	80.7	67.3	73.7	82.7	62.1	68.3	78.6	**67.4**	**74.1**	**83.1**
DUT	65.2	71.4	80.8	65.8	72.4	81.6	66.6	72.8	82.0	68.1	74.5	83.4	62.9	69.1	79.3	**68.2**	**74.9**	**83.8**
AVG	65.0	71.2	80.7	65.5	72.1	81.3	66.2	72.5	81.8	67.8	74.2	83.2	62.8	68.9	79.2	**67.9**	**74.6**	**83.5**

Table 7: Correlation between CLIPScore values and human rankings, using multimodal models finetuned with different loss functions. The last row presents macro-averaged correlation results across all the languages (including English). Non-English results above the English baseline are shown in bold.