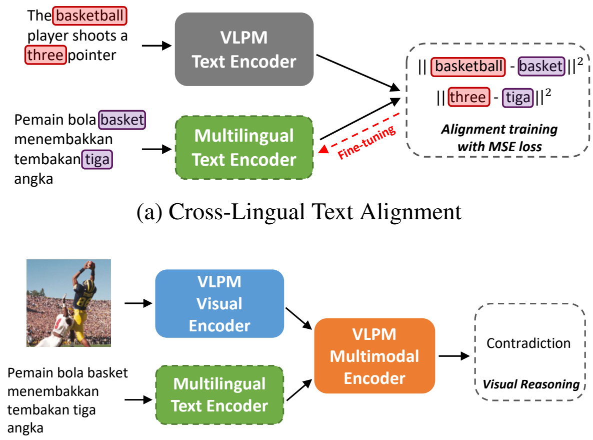
Stop Pre-Training: Adapt Visual-Language Models to Unseen Languages

Yasmine Karoui[μ*] **Rémi Lebret**[λ] **Negar Foroutan**[λ] **Karl Aberer**[λ]
[μ]Technical University of Munich, Germany
[λ]EPFL, Switzerland

Abstract

Vision-Language Pre-training (VLP) has advanced the performance of many vision-language tasks, such as image-text retrieval, visual entailment, and visual reasoning. The pre-training mostly utilizes lexical databases and image queries in English. Previous work has demonstrated that the pre-training in English does not transfer well to other languages in a zero-shot setting. However, multilingual pre-trained language models (MPLM) have excelled at a variety of single-modal language tasks. In this paper, we propose a simple yet efficient approach to adapt VLP to unseen languages using MPLM. We utilize a cross-lingual contextualized token embeddings alignment approach to train text encoders for non-English languages. Our approach does not require image input and primarily uses machine translation, eliminating the need for target language data. Our evaluation across three distinct tasks (image-text retrieval, visual entailment, and natural language visual reasoning) demonstrates that this approach outperforms the state-of-the-art multilingual vision-language models without requiring large parallel corpora. Our code is available at https://github.com/Yasminekaroui/CliCoTea.



(a) Cross-Lingual Text Alignment

(b) Zero-shot Transfer on VL downstream task

Figure 1: Overview of our approach. We adapt the text encoder of a monolingual VL model to an unseen language (a). Then we use the adapted model for a VL downstream task in a zero-shot setting (b).

1 Introduction

Inspired by the recent advancements in language model pre-training, Vision-Language Pre-trained Models (VLPMs) have demonstrated state-of-the-art performance across a wide range of vision-language (VL) tasks such as text-to-image retrieval, visual reasoning, visual entailment, and visual QA (Chen et al., 2020; Li et al., 2021, 2022).

However, extending VLPMs to multilingual scenarios is still challenging. On one hand, the majority of these models are trained on monolingual (English) corpora and thus cannot perform well for other languages. On the other hand, the multilingual pre-trained language models (Devlin et al., 2018; Conneau et al., 2019) cannot handle vision data (e.g., images or videos) directly.

Lately, there have been attempts (M^{3}P, nUNITER, UC^2) to pivot on images or English texts to align multilingual representations with vision features (Chen et al., 2020; Ni et al., 2021; Zhou et al., 2021). However, a recent benchmark on multilingual multimodal pre-training (IGLUE) (Bugliarello et al., 2022) shows that although these models achieve promising zero-shot cross-lingual transfer performance on some VL tasks, they still fall short in comparison to the "translate-test" baseline (using an English-only VLPM on the translations of the text examples).

A more recent work (CCLM) achieves promising performance on the IGLUE benchmark by exploiting massive parallel text and image-text corpora to pre-train a VL model (Zeng et al., 2022). This approach is motivated by a key observation that multilingual and multimodal pre-training essentially achieves the same goal of aligning two different views of the same object into a common semantic space. Although this framework performs well on

[*] Yasmine performed this work while interning at EPFL.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 366–375
July 9-14, 2023 ©2023 Association for Computational Linguistics

the IGLUE benchmark, it requires a large amount of parallel data. Its pre-training phase relies on 19M multilingual parallel sentence pairs extracted from WikiMatrix (Schwenk et al., 2021), jointly trained with 4 million image-text pairs in multiple languages.

In this work, we are proposing a simple yet efficient way to adapt VLP models to unseen languages without requiring large parallel corpora. We propose to align a VLPM monolingual text encoder (achieving start-of-the-art performance on English downstream VL tasks) with a multilingual pre-trained language model (e.g., mBERT), using only small in-domain parallel text corpus. The recent progress in Neural Machine Translation (NMT) has enabled us to create such a parallel corpus from automatically translating the data from English to any other language, even for low-resource languages (i.e., Swahili). However, since our approach relies on token alignment, it is robust to errors made by NMT. Our zero-shot evaluation across three of the four IGLUE tasks shows that the proposed method achieves state-of-the-art results while using small set of in-domain parallel sentences. The key steps of our approach are illustrated in Figure 1.

2 `CLiCoTEA` : Cross-Lingual Contextualised Token Embedding Alignment

We propose `CLiCoTEA` , an approach to transfer a monolingual vision-language (VL) pre-trained model in one language L_1 where there is an abundant number of training pairs of image and text (i.e., English) to a second language L_2. As we focus in this paper on the zero-shot setting, we do the transfer after fine-tuning the pre-trained monolingual VL model on a downstream task t, where training samples are available in language L_1.

`CLiCoTEA` consists of six steps:

1. Pre-train a monolingual VL model on a massive collection of image-text pairs, where text is written in language L_1.
2. Fine-tune the VL pre-trained model on the downstream task t in language $L1$.
3. Create a parallel text corpus by translating the training set from step 2 in the target language L_2. Note that this step can be done automatically using neural machine translation.
4. Create a list of aligned tokens for each (potentially noisy) parallel sentence using a token alignment model.
5. Cross-lingual transfer by aligning contextualised token embeddings. As illustrated in Figure 1a, it transfers the VL fine-tuned model to the new language L_2 by aligning a pre-trained multilingual LM (e.g., mBERT or XLM-R) with the text encoder of the VL pre-trained model using the list of aligned tokens created in step 4.
6. Zero-shot transfer to L_2 by swapping the monolingual text encoder from the VL pre-trained model with the aligned multilingual text encoder learned in step 5. An example of visual reasoning in Indonesian is illustrated in Figure 1b.

In practice, steps 1 and 2 are the most computationally expensive. Therefore, we propose to adapt VL fine-tuned models to new languages by only doing the steps from 3 to 5 which can be computed in a few hours on a single GPU.

We note that `CLiCoTEA` could be used with any multimodal pre-trained model where one of the modalities is a monolingual text encoder. We focus in this paper on VL models, but `CLiCoTEA` could be applied for instance to a language-knowledge model such as GreaseLM (Zhang et al., 2021) or DRAGON (Yasunaga et al., 2022).

3 Experiment

3.1 Pre-trained Models

Vision-Language Model In step 1 of `CLiCoTEA` , we use the Align BEfore Fuse (ALBEF) framework[1] (Li et al., 2021) as our Vision-Language Pre-trained Model (VLPM). ALBEF has been fine-tuned on multiple downstream VL tasks and achieves state-of-the-art performance. We use the ALBEF fine-tuned models in step 2 for the downstream tasks described in Section 3.3. Unlike other competitive VL pre-trained models (such as BLIP (Li et al., 2022)) that inject visual information by inserting cross-attention for each transformer block of the text encoder, ALBEF first encodes the image and text independently with a detector-free image encoder and a text encoder. Then it uses a multimodal encoder to fuse

[1]Code and models are available at `https://github.com/salesforce/ALBEF`.

the image features with the text features through cross-modal attention. All encoders are based on transformer networks with the text encoder being a 6-layer transformer initialised using the first 6 layers of the $\text{BERT}_{\text{base}}$. We thus extract this 6-layer text encoder for cross-lingual transfer training in step 5.

Multilingual Language Model As a multilingual pre-trained language model, we use the multilingual BERT (mBERT)[2] (Devlin et al., 2018). It has been trained on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) objective and has demonstrated remarkable zero-shot cross-lingual transfer capabilities (Wu and Dredze, 2019; Pires et al., 2019; Hu et al., 2020; Conneau et al., 2018). We extract the first 6-layer transformer to be aligned with the text encoder of ALBEF in step 5.

3.2 Implementation Details

Word Alignment Since the parallel sentences do not contain word-level alignment information, in step 4 of `CLiCoTEA` we utilize `awesome-align`[3] (Dou and Neubig, 2021) which is a tool that automatically extracts word alignments from mBERT. The generated word pairs are then filtered for keeping only one-to-one, one-to-many or many-to-one alignments and removing many-to-many alignments. This is done for all languages except Chinese because otherwise less than 3% of the training data would remain in the set. The advantage of this filtering is twofold: a) it removes the noise from the matching word pairs; b) it reduces the training time and computation. For words that are split into sub-word tokens, we consider either the left-most token embedding alignment (i.e., the first sub-word token of a word) or, the average embedding across all sub-word tokens.

Contextualised Token Alignment Training Given a set of aligned contextual word pairs extracted from parallel sentences, we define $\{\mathrm{x}_i, \mathrm{y}_i\}_{i=1}^n$, where $\mathrm{x}_i \in \mathbb{R}^d$ is the contextualised embedding of token i in the target language (obtained from mBERT), and $\mathrm{y}_i \in \mathbb{R}^d$ is the contextualised embedding of its alignment in the source language (obtained from the fine-tuned ALBEF)[4]. In step 5 of `CLiCoTEA`, we minimise the following training objective: $\sum_{i=1}^n ||\mathrm{x}_i - \mathrm{y}_i||^2$.

The parameters of the source language encoder are frozen, while the ones of the target language encoder are fine-tuned at training time. The learning rate is set to 5.10^{-5}. The batch size is set to 128. These hyperparameters are set through the NLVR2, Flickr30k, SNLI validation sets, for each task respectively. For each target language, the training is done on a single GeForce GTX TITAN X in a few hours.

Data Augmentation As multilingual language models are generally pre-trained on the source language L_1, the contextualised token alignment can be trained not only with sentences from the target language L_2, but also with sentences from the source language L_1. This strategy doubles the training size, and consequently, the training time but it could be used with tasks where the number of available training sentences is limited.

3.3 Downstream Tasks

In step 6, we evaluate `CLiCoTEA` on three tasks from the IGLUE benchmark[5] in the zero-shot setting:

- **xFlickr&CO**: The dataset is composed of 1000 images from Flickr30K (Plummer et al., 2015) and 1000 images from MSCOCO dataset (Lin et al., 2014). These images come along with croudsourced image captions in 6 different languages. xFlickr&CO is a *retrieval* task dataset. It is composed of two subtasks: image-to-text retrieval (TR) and text-to-image retrieval (IR).

- **XVNLI**: The dataset consists in merging SNLI hypothesis with Flickr30K (Plummer et al., 2015) images and translate the test set in four languages. The task is called *visual entailment* (VE) which is a fine-grained reasoning task to determine whether a text hypothesis "contradicts", "entails", or is "neutral" with respect to an image.

- **MaRVL**: The dataset is a multilingual expansion of NLVR2 dataset (Suhr et al., 2017),

[2] Available on HuggingFace hub at `https://huggingface.co/bert-base-multilingual-cased`.

[3] `https://github.com/neulab/awesome-align`

[4] Note that the special [`CLS`] token is always included.

[5] We do not include the Cross-lingual Grounded Question Answering task (xGQA (Pfeiffer et al., 2021)) as it requires aligning the answer decoder too. We leave it as future work.

with images related to concepts of five languages and cultures. The task is called *visual reasoning* (VR) which consists in determining whether a statement is correct given a pair of images.

Step	Retrieval	VE	VR
Fine-tuning	Flickr30K	SNLI	NLVR2
Alignment	Flickr30K*	SNLI*	NLVR2*
Zero-shot Test	xFlickr&CO	XVNLI	MaRVL

Table 1: The datasets used in the different steps of CLiCoTEA . Translated train and validation captions are denoted with $*$.

Table 1 shows the datasets used for a) fine-tuning the monolingual VL pre-trained model in step 2, b) training the alignment of contextualised token embeddings in step 5, and c) testing the zero-shot cross-lingual transfer in step 6. For creating the parallel corpus in step 3, all datasets used for fine-tuning the monolingual pre-trained VL model are translated to the corresponding test dataset languages from the IGLUE benchmark using GoogleTrans Python API[6]. Statistics about the translation datasets can be found in Section A.1. MaRVL being the smallest dataset, the data augmentation strategy described in Section 3.2 is applied only for this task. Detailed results on data augmentation can be found in Section 3.2.

3.4 Experimental Results

Results reported in Table 2 shows that CLiCoTEA outperforms the state-of-the-art CCLM models for all downstream tasks except retrieval. The larger improvement against CCLM models is obtained in visual entailment with an increase of almost 5%. The superiority of CLiCoTEA is especially high for Spanish (+7.68%), as can be seen from Table 10 in Section A.4. The average performance on visual reasoning is similar to CCLM, but CLiCoTEA significantly outperforms CCLM by $\pm 4\%$ on the low-resource languages such as Tamil and Swahili (results per language can be seen in Table 8 in Section A.3). For retrieval, CLiCoTEA outperforms all models except $\text{CCLM}_{4\text{M}}$. It is worth mentioning that, unlike the other models, $\text{CCLM}_{4\text{M}}$ has been pre-trained on COCO which could explain its superiority on Flickr&CO dataset. More details about the results on retrieval can be found in Section A.2.

Model	VE	VR	Retrieval	
	XVNLI	MaRVL	xFlickr&CO	
			IR	TR
mUNITER	53.69	53.72	8.06	8.86
xUNITER	58.48	54.59	14.04	13.51
UC^2	62.05	57.28	20.31	17.89
M^3P	58.25	56.00	12.91	11.90
$\text{CCLM}_{3\text{M}}$	74.64	65.91	67.35	65.37
$\text{CCLM}_{4\text{M}}$	73.32	67.17	**76.56**	**73.46**
CLiCoTEA	**78.15**	**68.09**	67.45	65.07

Table 2: Zero-shot performance on IGLUE benchmark. Recall@1 and Accuracy are reported for retrieval tasks (xFlickr&CO) and understanding tasks (XVNLI, MaRVL) respectively. Results of compared models are directly copied from Zeng et al. (2022).

4 Conclusion

In this paper, we present CLiCoTEA an approach for adapting Vision-Language pre-trained models to unseen languages. Unlike other approaches that rely on an expensive pre-training phase (both in terms of data and computation), our approach adapts the contextualised token embeddings of a multilingual pre-trained language model by aligning them with the contextualised token embeddings of the VLPM text encoder. By aligning ALBEF text encoder with mBERT, we show that CLiCoTEA outperforms CCLM, which exploits massive parallel text and image-text corpora. CLiCoTEA achieves start-of-the-art performance on visual entailment and visual reasoning, with an increase of almost 5% on visual entailment. It also demonstrates its effectiveness, especially for low-resource languages, as it does not require large corpora to do the adaptation.

5 Limitations

The general performance of CLiCoTEA could be improved with a better MPLM than mBERT, such as XLM-R which has a larger token vocabulary and has been pre-trained on a much larger dataset. Our approach is currently not applicable to generation tasks where a multilingual text decoder is needed to generate text in unseen languages. We leave this

[6]https://pypi.org/project/googletrans/

adaptation for future work. Unlike the statement made in Zeng et al. (2022), current multilingual VL models still do not surpass the *Translate-Test* baseline of the tasks from IGLUE benchmark. The performance of `CLiCoTEA` is promising but the best scores are still obtained when translating everything to English and using the (English-only) ALBEF model. The smallest difference in accuracy on MaRVL dataset between `CLiCoTEA` and ALBEF with *Translate-Test* is obtained in Swahili (-2%), while the gap is much larger (around -6%) for the other languages. Outperforming the *Translate-Test* achieved by ALBEF still remains an open challenge, especially for high-resource languages.

References

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. *arXiv preprint arXiv:2201.11732*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3977–3986.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2021. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *arXiv preprint arXiv:2210.09338*.

Yan Zeng, Wangchunshu Zhou, Ao Luo, and Xinsong Zhang. 2022. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. *arXiv preprint arXiv:2206.00621*.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2021. Greaselm: Graph reasoning enhanced language models. In *International Conference on Learning Representations*.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165.

A Appendix

A.1 Details of Alignment Datasets

Tables 3, 4, and 5 show the average number of aligned tokens extracted from the translated sentences of Flickr30k, SNLI, and NLVR2, respectively.

Language	Total number of sentences	Avg. number of aligned tokens
German	144935	8.74
Spanish	144990	10.04
Indonesian	144858	7.46
Russian	144526	6.44
Turkish	143664	4.83

Table 3: Statistics about Flickr30k translation set.

Language	Total number of sentences	Avg. number of aligned tokens
Arabic	513683	2.95
Spanish	549785	6.31
French	549260	5.78
Russian	524308	3.60

Table 4: Statistics about SNLI translation set.

A.2 Results on Retrieval

Zero-shot performance on the Flickr&CO dataset, the image-text and text-image retrieval tasks from the IGLUE benchmark, for four available languages (DE: German, ES: Spanish, ID: Indonesian, RU: Russian, TR: Turkish) are reported in Table 6 and Table 7, respectively. CLiCoTEA outperforms all models except $\text{CCLM}_{4\text{M}}$. Compared

Language	Total number of sentences	Avg. number of aligned tokens
Indonesian	86325	8.27
Swahili	85415	5.46
Tamil	85241	4.53
Turkish	85050	5.42
Chinese	86373	10.76

Table 5: Statistics about NLVR2 translation set.

with $\text{CCLM}_{3\text{M}}$, $\text{CCLM}_{4\text{M}}$ has been trained with 1M additional image-text pairs from Visual Genome and COCO datasets. The gap in performance between the two models on retrieval tasks suggests that pre-training with COCO text-image pairs gives a clear advantage to $\text{CCLM}_{4\text{M}}$ as Flickr&CO contains 1000 images from COCO, while all other models have been fine-tuned only on Flickr30K.

Model	**Language**				
	DE	ES	ID	RU	TR
mUNITER	12.05	13.15	5.95	5.85	1.75
xUNITER	14.55	16.10	16.50	15.90	9.05
UC^2	28.60	15.95	14.60	20.00	7.15
M^3P	13.35	13.40	13.20	15.95	7.75
$\text{CCLM}_{3\text{M}}$	67.67	71.23	62.38	72.83	55.15
$\text{CCLM}_{4\text{M}}$	**73.65**	**79.62**	**69.50**	**80.65**	**65.08**
CLiCoTEA	61.48	74.50	64.98	73.50	62.80

Table 6: Zero-shot performance on multi-lingual image-text retrieval with Flickr&CO dataset. Recall@1 is reported.

A.3 Results on Natural Language Visual Reasoning

Table 8 shows the zero-shot performance on the MaRVL dataset, and the natural language visual reasoning task from the IGLUE benchmark, for all available languages (ID: Indonesian, SW: Swahili, TA: Tamil, TR: Turkish, ZH: Chinese).

As MaRVL is the smallest dataset among the three tasks from IGLUE, we apply the data augmentation for training the alignment as described in Section 3.2. Results reported in Table 9 show that there is drop of 3.35% for Turkish, and 9.99% for Chinese when training only using the target language L_2, while there is no significant difference for the three other languages (Indonesian, Swahili,

Model	Language				
	DE	ES	ID	RU	TR
mUNITER	11.85	13.05	7.55	6.80	3.25
xUNITER	13.25	15.10	16.75	14.80	10.05
UC^2	23.90	15.30	13.60	16.75	6.95
M^{3}P	11.85	12.15	12.10	14.45	8.35
CCLM$_{3M}$	66.88	68.58	60.33	69.90	54.22
CCLM$_{4M}$	**73.60**	**78.38**	**67.67**	**80.35**	**63.22**
`CLiCoTEA`	70.34	71.42	57.77	69.80	56.00

Table 7: Zero-shot performance on multi-lingual text-image retrieval with Flickr&CO dataset. Recall@1 is reported.

Model	Language				
	ID	SW	TA	TR	ZH
mUNITER	54.79	51.17	52.66	54.66	55.34
xUNITER	55.14	55.51	53.06	56.19	53.06
UC^2	56.74	52.62	60.47	56.70	59.88
M^{3}P	56.47	55.69	56.04	56.78	55.04
CCLM$_{3M}$	67.81	61.55	60.28	69.60	**70.52**
CCLM$_{4M}$	**71.66**	67.21	60.36	66.75	69.86
`CLiCoTEA`	69.55	**71.30**	**63.93**	**70.72**	64.93

Table 8: Zero-shot performance on visual reasoning with MaRVL dataset. Accuracy is reported.

and Tamil). As explained in Section 3.2, our noise filtering technique does not work well with Chinese. Aligning the English sentences with half of the original training set helped the model infer knowledge from English and reduced the number of wrong matching words. For Turkish, the increase in performance could be explained by the similarity between the two alphabets.

Training Set	Language				
	ID	SW	TA	TR	ZH
L_1	69.55	71.30	63.45	67.37	54.94
$L_1 + L_2$	68.53	70.31	63.93	70.72	64.93

Table 9: Zero-shot performance of `CLiCoTEA` on visual reasoning with MaRVL dataset using monolingual (L_1) or bilingual ($L_1 + L_2$) alignment training. Accuracy is reported.

A.4 Results on Visual Entailment

Zero-shot performance on the XVNLI dataset, the visual entailment task from the IGLUE benchmark, for all available languages (AR: Arabic, ES: Spanish, FR: French, RU: Russian) are reported in Table 10. `CLiCoTEA` outperforms other models by a significant margin for all languages, except Russian where CCLM$_{3M}$ achieves comparable performance.

Model	Language			
	AR	ES	FR	RU
mUNITER	46.73	56.96	59.36	51.72
xUNITER	51.98	58.94	63.32	59.71
UC^2	56.19	57.47	69.67	64.86
M^{3}P	55.24	58.85	56.36	62.54
CCLM$_{3M}$	71.04	75.80	78.14	**73.56**
CCLM$_{4M}$	69.68	73.65	77.54	72.40
`CLiCoTEA`	**75.83**	**83.48**	**80.17**	73.13

Table 10: Zero-shot performance on visual entailment with XVNLI dataset. Accuracy is reported.

A.5 In-domain vs Open-domain Data

Language	Total number of sentences	Accuracy in %
Swahili	50400	63.27
Turkish	50418	66.61
Chinese	51159	59.09

Table 11: Zero-shot performance on visual reasoning with MaRVL dataset. Alignment is done with a subset from XNLI dataset.

In order to eliminate the need for machine translations from `CLiCoTEA` in step 3, we created a parallel text corpus with sentences obtained from XNLI (Conneau et al., 2018) which is publicly available and covers 15 languages. A subset of XNLI has been used for training the alignment by considering only the sentences that were semantically close to the captions in NLVR2. To do so, we used the `Sentence-Transformers` framework[7] to compute sentence embeddings sim-

[7]Available at `https://www.sbert.net`.

ilarities between NLVR2 captions and XNLI English sentences and kept only the sentences with a cosine similarity higher than 0.5. About 50k English sentences from XNLI are semantically close to NLVR2 captions, we thus selected their parallel sentences in Swahili, Turkish and Chinese to perform an evaluation on MaRVL dataset. After the contextualised token alignment training on XNLI-based datasets, our results in Table 11 suggest that a multilingual open-domain dataset gives better results than mUNITER and xUNITER but underperforms the results obtained by translating in-domain training sets. This could be explained by the fact that although these datasets are multilingual, the sentences are not semantically close enough to NLVR2 captions.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 5

☒ A2. Did you discuss any potential risks of your work?
We could not think of any risk as we do not introduce any model or dataset.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☑ B1. Did you cite the creators of artifacts you used?
Section 3

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We cited the datasets website that includes the licenses.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We use the datasets only for evaluation.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No, because we have employed widely used public datasets and have not collected any data ourselves.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3 and Appendix A

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sections 2 and 3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.