# Quantifying the Gap Between Machine Translation and Native Language in Training for Multimodal, Multilingual Retrieval

Anonymous ACL submission

## Abstract

There is a scarcity of multilingual vision-language models that properly account for the perceptual differences that are reflected in image captions across languages and cultures. The existing lack of model flexibility is shown in a performance gap between training on independently written English and German captions in German text-image retrieval. In this work, we first show that using off-the-shelf machine translation is ineffective at bridging this gap. Second, we propose techniques to reduce the drop off from training on native German captions. Third, we show that part of the gap remains, which identifies an open area in which we encourage future work from the community.

## 1 Introduction

Vision-language (VL) models such as CLIP (Radford et al., 2021) are predominantly limited to use in English as a result of the pretraining supervision being mostly from English captions. This trend poses a problem for non-English speakers as languages and cultures differ in concepts of interest (Liu et al., 2021) and perception of those concepts (Nisbett and Masuda, 2013). Relying on English supervision only for vision-language pretraining thus impacts downstream tasks, such as object recognition, detection, and image-text retrieval, when the text or ground-truth concepts are not in English.

Notable differences exhibited in captions across languages are in object *specificity* and *importance*. For example, Nisbett and Masuda (2013) describe differences in how cultures perceive members (e.g. penguins) of an object group (e.g. birds), indicating stronger association for *specific* than general object terms for certain groups. Experiments in Nisbett and Masuda (2013) also demonstrate differences between East Asians and Americans in terms of the *importance* of background objects and context as opposed to foreground objects. Different cultures notice different objects more; the perceptual differences may manifest in different objects being included/excluded in a caption, and different objects being relevant in downstream tasks.

There have been a few recent multilingual vision-language datasets (Elliott et al., 2016; Yoshikawa et al., 2017; Liu et al., 2021; Thapliyal et al., 2022) and models (Chen et al., 2022, 2023b; Carlsson et al., 2022; Chen et al., 2023a). The models often leverage off-the-shelf machine translation techniques to improve multilingual functionality. In contrast, we demonstrate that there are performance gaps between training with machine-translated and natively written captions for a given language. We reason that machine translation may not account for *specificity* differences in the ways cultures name and group objects (e.g. "person on bike" vs. "bicyclist" and "banjo" vs. "musical instrument"). Additionally, machine translation may not significantly add or remove supervision to account for *importance* differences. If included in the captions, certain objects can function as distractors, leading to undesirable and/or unnecessary correlations being learned for a given language/culture.

To better understand these problems, we quantify differences in non-English retrieval performance when finetuning a multilingual CLIP model (Chen et al., 2023a) with different data, either directly provided in English, a *target* language, or translated. We specifically use Multi30K (Elliott et al., 2016) with German as the target language. We find that there is a large gap depending on the data used to train the retrieval model, i.e. (1) English, (2) German translated from English, and (3) native German (i.e. captions directly written in German to describe the image).

To bridge the gap between (1-2) and (3), we further test three paraphrasing techniques, focusing on object differences in captions. First, using the observations in prior literature and our own analyses about differences in languages in terms of the specificity with which objects are mentioned, we

experiment with a *hypernymization* data augmentation technique. We specifically hypernymize object terms in the English translations, translate the result into German, and train with these hypernymized captions as additional finetuning data. Second, we use a large language model (LLM), LLaMA-3 (Touvron et al., 2023), to produce *structurally different, but semantically similar* paraphrases of each English caption before translating to German. Third, we explore LLM reasoning to produce *targeted* paraphrases that capture the perceptual properties captured in a sample set of captions, and translate these captions for finetuning. These techniques improve over default machine translation and bridge part of the gap to native German finetuning data. However, part of the gap still remains, indicating the challenge and importance of this open problem.

## 2 Background and Related Work

**Cultural differences in perception.** Prior work explores how culture affects perception and expression. For example, the dichotomy between Western individualism and East Asian collectivism manifests in perception, e.g. Americans pay more attention to foreground/objects than East Asians, but conversely for background/context (Nisbett and Masuda, 2013). Further, different cultures group objects differently (e.g. based on shape or material) and ascribe different properties to objects, because of unique grammar (e.g. gendered nouns) (Boroditsky, 2006). Further examples can be found in work on linguistic relativity (Kay and Kempton, 1984).

**Multilingual multimodal modeling.** Training VL models across languages has recently received interest. Some works use machine translation to enable cross-language tasks (Sharma et al., 2018; Chen et al., 2023a), but Kádár et al. (2018) and our work show differences in retrieval performance when captions are natively written in a language or translated into that language from English. WebLI (Chen et al., 2022) crawls captions in 109 languages, without a constraint that these describe the same image. While this is a realistic setting, Kádár et al. (2018) shows benefits in performance when techniques can be employed with captions in multiple languages *for the same image*. WebLI is also *not* publicly available. Elliott et al. (2016) provides both native and translated German captions for images in Flickr30K (Young et al., 2014). Thapliyal et al. (2022) provides native captions from a variety of languages on the same 100 images. On the model development side, Chen et al. (2022) achieve cross-language ability through a diverse mixture of training tasks and Chen et al. (2023a); Carlsson et al. (2022) through multilingual embeddings and machine translation. However, none investigate the reason why using native captions in a language vs. those translated in that language from English, have different statistics, nor offer techniques to cope with these differences. More distinct but motivating our work, Liu et al. (2021) examines the different concepts that are important for different languages, focusing on unique objects.

## 3 Experimental Methodology

We benchmark the use of native captions and translations when training a multilingual, multimodal model for *non-English* (i.e. German) retrieval. We also consider strategies to push model performance closer to the upper bound of native language use.

### 3.1 Benchmarking Details

**Task and data.** We focus on *native* German image-text (I2T) and text-image (T2I) retrieval. We use Multi30K (Elliott et al., 2016), which is an extension of the English-based image-caption dataset Flickr30K (Young et al., 2014), augmented with German captions. The original Flickr30K contains 5 English captions, for each of 31,014 images. Multi30K contains two different sets with German captions. In the *Human Translation* set, an English caption is sampled from Flickr30K, and professional translators produce corresponding captions in German (just from source text, not using the images). In the *Independently Written* set, 5 German captions for each image are gathered directly from the perception of native German speakers. We emphasize that these *native* captions are *not* translations as they have been produced from the image and written directly in German. We randomly create splits of Multi30K to create a disjoint reference set (9,666 samples) for our proposed strategies (Sec. 3.2) and retrieval train/val/test sets (9,666/1,014/10,668 samples respectively).

**Modeling.** We use mCLIP (Chen et al., 2023a) as a multilingual VLM. It leverages knowledge distillation, projection layers, and the multilingual text encoder XLM-R (Conneau et al., 2020) instead of CLIP's text encoder to instill multilingual capabilities. We explore *finetuning* mCLIP with German captions for German I2T and T2I retrieval. For experimentation that involves automatically trans-

lating English captions to German, we use *opus-mt-en-de* (Tiedemann and Thottingal, 2020). We use a deterministic setting (no sampling) and infer at most 40 tokens for each caption. Models are trained for 30 epochs on 1 Quadro RTX 5000 GPU with batch size 16 and learning rate 0.0005.

**Metric.** We report *mean recall* as in Chen et al. (2023a). Recall@1,5,10 is computed for both T2I and I2T retrieval on each native German test set (5 sets total). *Mean recall* is the average of these six values. We further average over each set.

### 3.2 Methods Compared

**Baselines** include:

• ENG, a lower bound: finetuning using data natively provided in English (in the *Independently Written* set). Since there are 5 sets of captions, we average over trials using each set for training.

• ENG2GER: finetuning on English sentences translated to German using a generic machine translation model (Tiedemann and Thottingal, 2020). We translate the *Human Translation* English set to German for finetuning.

• ENG2GER-TRN: same as above but the translation model is trained on images and captions from Multi30K in the disjoint reference split we create, with the intuition that translation finetuning may capture caption differences. We train for 10 epochs with learning rate 0.00001 and batch size 16, using the *Human Translation* pairs.

**Upper bounds** include:

• GER: finetuning using data natively provided in German (in the *Independently Written* set).

• GER-INDIR (German-Indirect): finetuning on Multi30K German captions translated from English by humans (in the *Human Translation* set). This training is different and expected to perform worse than native German, but better than baselines.

**Strategies:** We find significant gaps between these methods, notably ENG2GER and GER, motivating experimentation with potential solutions. We explore the addition of augmented data; specifically, we augment data in English before translation to German. For some experiments, we detect object mentions: we consider an object vocabulary $\mathcal{V}$ with COCO object noun terms (Lin et al., 2014) and create reference lists of nouns which correspond to each class, based on Lu et al. (2018) and accounting for plurals and word sense. For each strategy, mCLIP is trained as in the ENG2GER setting, but with an augmented dataset of captions added. These are as follows:

| Method | Mean Recall | Vs. ENG2GER |
|---|---|---|
| MCLIP | 24.5 | -8.9 |
| ENG | 26.9 | -6.5 |
| ENG2GER | 33.4 | 0.0 |
| ENG2GER-TRN | 34.0 | +0.6 |
| HYPER | 33.7 | +0.3 |
| PARA-RND | 34.1 | +0.7 |
| PARA-TGT | 34.1 | +0.7 |
| PARA-CMB | 34.7 | +1.3 |
| GER-INDIR | 36.8 | +3.4 |
| GER | 38.4 | +5.0 |

Table 1: Main results (German I2T/T2I). Mean recall values are averaged over native German caption sets.

• HYPER: After identifying each COCO class with a synset id, if available, object mentions are hypernymized to be a term above it in the WordNet hierarchy (Miller, 1995). Our goal is to improve robustness to changes in object naming to address challenges in object specificity.

• PARA-RND (paraphrase-random): Before translation, we ask LLaMA-3 (Touvron et al., 2023) to write each caption in a structurally different manner while maintaining meaning. We are motivated by Fan et al. (2024) which shows English retrieval benefits from structure and vocabulary differences. Our approach differs in its use before translation as a way to guide translation to more diverse (and potentially applicable) descriptions that may appear.

• PARA-TGT (paraphrase-targeted): We ask LLaMA-3 (Touvron et al., 2023) to paraphrase each caption using dataset examples of object naming "style". For a given caption, we ask LLaMA-3 to first find relevant noun phrases. We then sample $k$=100 captions from the reference split of the first native German set, finding captions which share any non-person mentions with the current caption (since most captions have people mentioned). These examples are used in in-context learning to convert the found noun phrases to more aligned presentations. Please refer to the appendix for the specific prompts and configuration.

• PARA-CMB combines both sets above.

## 4 Key Findings

In the top block in Table 1, we note that using the original mCLIP achieves the lowest performance (24.5). This result indicates Multi30K has characteristics that require specialized knowledge. Finetuning with English Multi30K data improves by

2.4 to 26.9. However, much more significant gains are achieved when the finetuning data is in German. Training with English data translated to German using an off-the-shelf translation model reaches 33.4 (second block). However, when compared to using human-translated German captions, there is a gap of 3.4. Finetuning the translation model only helps by 0.6. Most significantly, the performance gap between off-the-shelf translation and native German captions is 5.0 (fourth block). Thus there is nuance to the perception of the world, as captured in captions, that is not produced with machine translation. Even human translation has a gap with native German, demonstrating nuance in native language understanding that can only be acquired through native (direct from the image) captioning.

Among our strategies (third block), we observe that our methods are somewhat effective for bridging the gap between ENG2GER and GER. HYPER improves the result by 0.3, and PARA-RND and PARA-TGT by 0.7. These models are more appropriate for low-resource target languages than ENG2GER-TRN since they use no/few reference captions compared to what is required for translation finetuning. Finally, combining random and targeted paraphrasing results in the biggest gain of 1.3. Yet these gains are still relatively small, indicating that bridging gaps in the perception of the visual world and in the way captions are written across cultures, remains challenging.

## 5 Further Analysis

**Object statistics in English/German captions.** Multi30K (Elliott et al., 2016) does not provide statistics of object mentions. To analyze, we translate German captions to English and extract nouns in both the (original) English and (translated to English) native German captions. The ratio of English/German mentions is about 1.5, i.e. English mentions object nouns 50% more frequently than German. However, this observation varies by type of object. For example, English consistently mentions clothing more often than German (pants-143% more, shirt-112%, hat-60%, jacket-43%). However, German mentions furniture more frequently (table-37% more, bed-20%, bench-15%). These languages also vary in granularity: English captions often say "people", while native German ones describe "workers", "athletes", etc.

**Example paraphrasing.** LLaMA picks up on the granularity pattern. For example, PARA-TGT modifies the original caption "Man in a red shirt riding his bicycle" into "A bicyclist in a red shirt is riding" LLaMA's reasoning states, "Combine 'man' and 'bicycle' into 'bicyclist' based on the reference captions." It further transforms "man on skis" into "skier', "person in blue and red ice climbing" into "ice climber", "man with backpack" into "backpacker", "men with children" into "family". LLaMA also tends to simplify captions, based on what is or is not common in the reference list, for example, "Two young people are approached by a flamboyant young woman dressed in a red bikini and a red feathered headress." becomes "Two young people are approached by a bikini-clad woman." While helpful overall (Table 1), paraphrasing could result in over-simplification.

| Supercat | Vehicle | Animal | Sports | Furniture | Electronic |
|---|---|---|---|---|---|
| Ger (#men) | 2604 | 2836 | 2101 | **1488** | 510 |
| Ger-Indir (#men) | **2724** | **2918** | **2127** | 1191 | **554** |
| Ger (prec) | 0.42 | 0.41 | 0.16 | 0.26 | 0.25 |
| Ger-Indir (prec) | **0.47** | **0.51** | **0.17** | **0.29** | **0.27** |
| Ger (rec) | **0.52** | **0.55** | **0.61** | **0.20** | 0.28 |
| Ger-Indir (rec) | 0.46 | 0.44 | 0.56 | 0.16 | **0.30** |

Table 2: Recognition stats by supercategory; top two rows: mention counts, middle: precision, bottom: recall

**Recognition.** We also compare objects mentioned in a target German caption, and ones predicted by a model trained with GER and GER-IND. We take object predictions to be ones with CLIP scores (i.e. similarity between the image and the text "A photo of [object]") greater than a threshold. Ground-truth is true only if the object is mentioned in German. In Table 2, we show results for the best-performing five COCO supercategories. We observe large differences in the number of mentions for the two types of data. Also GER achieves better recall (only slightly correlated with differences in mention counts), but GER-INDIR better precision. These results show that the variance in supervision from translated and native German captions is significant, and care must be taken to mimic native German content to ensure utility for German users.

## 6 Conclusion and Future Work

We show large differences in using native and translated German captions to train a retrieval model, and experiment with three strategies to reduce the gaps. We will next extend the analysis to more languages and experiment with heuristics-based data augmentation inspired by the psychology literature (Nisbett and Masuda, 2013; Boroditsky, 2006).

## 7 Limitations

We only experiment with one translation model, one language (German), and limited LLaMA-3 runs. Further, we note that image-caption datasets and LLaMA-3, used in our experiments, have been noted to contain biases that warrant consideration. We note that a future extension of our paraphrasing strategies could be to reduce in-group vs. out-of-group bias in inference.

## References


Lera Boroditsky. 2006. Linguistic relativity. *Encyclopedia of cognitive science*.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual clip. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023a. mCLIP: Multilingual CLIP via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, Toronto, Canada. Association for Computational Linguistics.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. 2023b. Pali-X: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

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

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2024. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36.

Ákos Kádár, Desmond Elliott, Marc-Alexandre Côté, Grzegorz Chrupała, and Afra Alishahi. 2018. Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 402–412.

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

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *Empirical Methods In Natural Language Processing*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Richard E Nisbett and Takahiko Masuda. 2013. Culture and point of view. In *Biological and cultural bases of human inference*, pages 49–70. Psychology Press.

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 *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Ashish Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In *EMNLP*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

# A Appendix

Shown are the prompt templates used for querying LLaMA-3 (meta-llama/Meta-Llama-3-8B-Instruct). We do not experiment with LLaMA sampling settings and generate outputs with default parameters.

*Para-Rnd Prompt Template*
Rewrite captions in a structurally different manner, while closely maintaining semantic meaning. Return as Python string. Return no other text.

*Para-Tgt Prompt Template*
1) Given a caption, 1st decompose into noun phrases, keeping all phrase content (e.g. adjectives) aside from articles. EX: "A person is riding a blue bicycle down the street on a sunny day." Noun Phrases: ["person", "blue bicycle", "street", "sunny day"]

2) Based on a provided reference list of related captions, construct a new set of noun phrases that alters the original noun phrases to be in the common styles/forms shown in the reference list. EX: If many captions say "bicyclist", combine "person" and "blue bicycle" into "bicyclist". Do not infer unnecessary information.

3) Finally, combine the new noun phrases back into a sentence, keeping the same semantics as the original caption. EX: "A bicyclist is traveling down the road on a sunny day."

Here is your reference caption list: {ref$_{caps}$}

Now run each steps 1-3 for the example: "{example}" Enclose the final output caption in <final></final> tags for easy parsing.

*System Prompt for Experiments*
I'm a researcher using LLMs for NLP tasks. Behave like an automatic processing agent for the user.