Quantifying the Gaps Between Translation and Native Perception in Training for Multimodal, Multilingual Retrieval

Kyle Buettner[1], **Adriana Kovashka**[1,2]
[1]Intelligent Systems Program, [2]Department of Computer Science, University of Pittsburgh
buettnerk@pitt.edu, kovashka@cs.pitt.edu
https://krbuettner.github.io/MultilingualRetrievalStudy

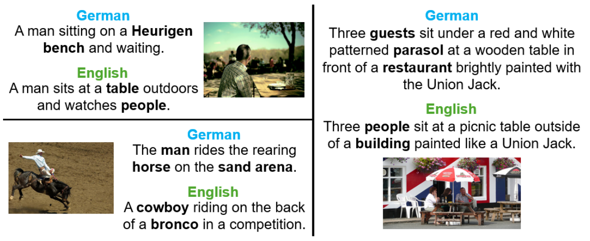
Abstract

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. In this work, through a multimodal, multilingual retrieval case study, we quantify the existing lack of model flexibility. We empirically show performance gaps between training on captions that come from native German perception and captions that have been either machine-translated or human-translated from English into German. To address these gaps, we further propose and evaluate caption augmentation strategies. While we achieve mean recall improvements (+1.3), gaps still remain, indicating an open area of future work for the community.

1 Introduction

Vision-language models (VLMs) such as CLIP (Radford et al., 2021) are predominantly limited to use in English as a result of the pretraining supervision consisting mostly of English captions. This trend naturally poses an accessibility barrier for non-English speakers. Furthermore, cultures around the world differ in their salient concepts (Liu et al., 2021) and visual perception (Nisbett and Masuda, 2013). Relying on English supervision in pretraining thus hinders consideration of cross-cultural concepts in object-based tasks such as recognition, detection, and image-text retrieval.

Example cultural differences present in language with respect to object *specificity* and *importance*. For example, past literature (Nisbett and Masuda, 2013) describes differences in how cultures perceive members of an object group (e.g. penguins within the group of birds), indicating that certain groups have stronger associations for *specific* rather than general object terms. Experiments in Nisbett and Masuda (2013) also illustrate differences between East Asians and Americans with respect to the perceived *importance* of background objects and context as opposed to foreground objects. Different cultures notice different objects more; perceptual differences may manifest in objects being included/excluded in a caption, and different objects being relevant in tasks. Fig. 1 shows examples of differences in AI datasets for English and German (Young et al., 2014; Elliott et al., 2016).



A man sitting on a **Heurigen bench** and waiting.

A man sits at a **table** outdoors and watches **people**.

The **man** rides the rearing **horse** on the **sand arena**.

A **cowboy** riding on the back of a **bronco** in a competition.

Three **guests** sit under a red and white patterned **parasol** at a wooden table in front of a **restaurant** brightly painted with the Union Jack.

Three **people** sit at a picnic table outside of a **building** painted like a Union Jack.

Figure 1: **Example perception differences between native English and German speakers.** Examples are captions from Flickr30K (Young et al., 2014) and Multi30K (Elliott et al., 2016). Note differences in mentioned objects ("sand arena", "parasol") and specificity ("Heurigen bench" vs. "table", "horse" vs. "bronco"). German captions here are translated to English.

There has been some progress in multilingual, multimodal modeling (Chen et al., 2022, 2024; Carlsson et al., 2022; Chen et al., 2023) and multilingual data creation (Elliott et al., 2016; Yoshikawa et al., 2017; Liu et al., 2021; Thapliyal et al., 2022). The models often leverage off-the-shelf machine translation techniques to improve multilingual functionality. In this work, we investigate the performance gaps between training models with translations (which reflect English speaker perception) and natively written captions (which reflect non-English speaker perception) for a task in a given language. In line with the observed differences in Nisbett and Masuda (2013), we reason that translation may not account for specificity differences and may not alter supervision to account for importance differences.

We quantify potential differences through an exploration of non-English image-text retrieval. In

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5863–5870
November 12-16, 2024 ©2024 Association for Computational Linguistics

particular, we finetune and benchmark multilingual CLIP (Chen et al., 2023) on Multi30K (Elliott et al., 2016) using *German* as the target. We explore Multi30K's native German captions (reflecting German speaker perception) and professionally translated captions (from English to German), as well as use of an external machine translation model over Flickr30K's English captions (Young et al., 2014). We find significant performance differences depending on the data used to train the model, i.e. (1) English, (2) German translated from English by a machine translation model, (3) German translated from English by humans, and (4) native German.

As (2) and (3) have gaps vs. (4), we also attempt to improve upon translation. We test three paraphrasing techniques to diversify object descriptions in English before translation, and use the resulting translations as additional finetuning data. *First*, we experiment with a *hypernymization* data augmentation technique, where object terms are updated before translation to represent different levels of specificity. *Second*, we use a large language model (LLM), LLaMA-3 (Touvron et al., 2023), to produce *structurally different, but semantically similar* paraphrases of English captions before translation. *Third*, we explore LLM reasoning to produce *targeted* paraphrases that capture the perceptual properties captured in a sample set of captions. These techniques outperform the baselines. **However, a gap between translation and native perception remains, indicating an open problem.** We conclude with analysis in pursuit of this direction.

2 Background and Related Work

Cultural differences in perception. Prior work considers how culture may influence perception and expression. For example, Western and East Asian cultural differences are found to manifest in visual attention, e.g. Americans appear to pay more attention to foreground/objects than East Asians, but conversely for background/context (Nisbett and Masuda, 2013). Furthermore, Boroditsky (2006) describes empirical studies that indicate that 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). Since German uses gendered nouns, this observation may manifest in native German captions (and retrieval) as objects being described with unique attributes. The work of Berthele et al. (2015) notes that Germanic language speakers describe object relationships with notably specific spatial information (e.g. posture/manner information in addition to object relationships). Hofstede (2001) conducts analysis to show that there are cultural differences between Germany and United States in terms of individualism vs. collectivism, which could impact the perception of visual content as argued by Nisbett and Masuda (2013). Further examples can be found in work on linguistic relativity (Kay and Kempton, 1984).

Multilingual multimodal modeling. Our work aligns with works that extend VLMs for use in languages besides English (Chen et al., 2022, 2024; Carlsson et al., 2022; Chen et al., 2023). Models notably often rely on translations, and works do not have analysis into performance differences between translations and captions of native perception. In contrast, 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. Our work differs from Kádár et al. (2018) as we also explore machine translation, with a more modern VLM (Chen et al., 2023). We also explicitly address the lack of techniques to overcome gaps by experimenting with paraphrasing augmentations. Our strategies are related to past paraphrasing work (Wieting and Gimpel, 2018; Hu et al., 2019), but these approaches use machine translation to generate large-scale English paraphrase datasets, while we leverage in-context learning and LLMs to generate paraphrases *for use as input to machine translation to enhance diversity*. We are inspired by Fan et al. (2024), as the work shows zero-shot image classification improvements with LLM-based caption rewriting.

Data-wise, we explore Multi30K (Elliott et al., 2016) as it contains native German captions and parallels the English Flickr30K captions (Young et al., 2014). XM3600 (Thapliyal et al., 2022) also provides natively perceived captions, in 36 languages for 3,600 images. Due to size, we do not train with this set, though we provide initial analysis on it to inspire future work. WebLI (Chen et al., 2022) is another dataset that contains crawled captions in 109 languages, though it is proprietary.

3 Experimental Methodology

We benchmark training with captions that reflect native perception by German speakers, ones that have been *machine-translated* from captions reflecting English speaker perception, and ones *human-*

translated from English speaker perception. We also test strategies to improve upon translation.

3.1 Benchmarking Details

Task. We evaluate on German image-text (I2T) and text-image (T2I) retrieval. The German captions used in eval are written directly by native speakers about images. They are *not* translated from English and represent natural non-English perception.

Data. English data is from Flickr30K (Young et al., 2014), and German data is from Multi30K (Elliott et al., 2016). Flickr30K contains 31,014 images that are annotated with 5 independently written English captions per image. Likewise, Multi30K provides 5 independently written German captions for the same images. These German captions are collected from 185 native speakers using a similar interface to Flickr (Hodosh et al., 2013). Multi30K also provides professional German translations. In particular, for each image, 1 of the 5 English captions is sampled from Flickr30K, and professional translators produce corresponding captions in German (just from source text, not using the images). We refer to the separate caption sets as *Independently Written* (5 sets for each language) and *Human-Translated* (1 set per language). For all sets, we randomly split data to create a disjoint reference set (9,666 samples) to be used with our strategies (Sec. 3.2), as well as retrieval train/val/test sets (9,666/1,014/10,668 samples respectively).

Modeling. We explore mCLIP (Chen et al., 2023), an approach which has made CLIP multilingual through knowledge distillation-based training of projector modules and replacement of CLIP's text encoder with the multilingual text encoder XLM-R (Conneau et al., 2020). We *finetune* mCLIP with images and captions for German I2T and T2I retrieval. For experimentation that involves machine-translating English captions to German, we use *opus-mt-en-de* (Tiedemann and Thottingal, 2020) from Hugging Face. With this model, we use a deterministic setting, where tokens are generated according to highest token probability, and infer at most 40 tokens for each caption. mCLIP 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. (2023). 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

Baseline finetuning strategies include:

- ENG, a "lower bound": finetuning using data natively provided in English (in the *Independently Written* sets). Since there are 5 sets of captions, we average over trials using each set for training.
- ENG2GER-MT: finetuning on German sentences that have been translated from English using an English-to-German machine translation model (Tiedemann and Thottingal, 2020). English sentences come from the *Human Translation* set.
- ENG2GER-MT (TRN): same as above, but the translation model is further trained on captions from the Multi30K 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.
- ENG2GER-HT: finetuning on German captions translated from English by professional annotators (in the *Human Translation* set). This training is different from and expected to perform worse than native German, but better than naive translation.
- GER: finetuning using data natively provided from German perception (in the *Independently Written* sets). Since there are 5 sets of captions, we average over trials using each set for training.

Strategies: We find significant gaps between these methods, notably ENG2GER-MT and GER, motivating experimentation with potential improvements. We test adding training data that has been augmented in English then translated to German. Some proposed changes involve object names, so for this purpose, we define an object vocabulary $\mathcal{V}$ including COCO object terms (Lin et al., 2014). Category detection involves consideration of these terms, synonyms (Lu et al., 2018), plurals, and word sense. For each strategy, mCLIP is trained as in ENG2GER-MT, but with an augmented dataset of captions added. Methods include:

- HYPER: After identifying each COCO class with a synset id, if available, object mentions are hypernymized to be a random 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 diversification. Our approach differs as we diversify before translation to guide translation to more generalizable descriptions.

- Para-Tgt (paraphrase-targeted): We ask LLaMA-3 to paraphrase each caption using examples of object naming "style". For each caption, a total of k=100 captions are randomly sampled from the reference split of the first native German set, such that if possible, sampled captions share at least one non-person object mention with the current caption (since most captions mention people). Translations of these are provided in the LLaMA-3 prompt as examples. Then for the input caption, LLaMA-3 is instructed to find relevant noun phrases, and to convert the noun phrases to more aligned representations based on the examples.
- Para-Cmb combines both sets above.

Please refer to the appendix for prompt details.

4 Key Findings

In the top block of Table 1, zero-shot mCLIP is shown to achieve the lowest recall (24.5). Finetuning mCLIP with English Multi30K data improves performance to 26.9 (+2.4). English data can help to a degree on German retrieval due to alignment learned in pretraining the multilingual text encoder. However, much more significant gains are achieved when the finetuning data is in German. Training with German data that has been translated from English using an off-the-shelf translation model (Eng2Ger-MT) reaches 33.4 (second block). Compared to human translation (Eng2Ger-HT - fourth block), there is a notable gap from machine translation (3.4), and finetuning the translation model only bridges this gap by 0.6. These results indicate existing challenges with off-the-shelf translation for retrieval. Then most significantly, the gap between off-the-shelf translation and native German captions (Ger) is 5.0. There is a notable gap between professional translation (Eng2Ger-HT) and Ger (1.6), which we reason is the gap due to differences in English and German perception. For example, these gaps could be due to specificity and importance differences. Expert translation does *not* address these factors.

In the third block, our methods are found to be somewhat effective for bridging the gap between Eng2Ger-MT and Ger. Hyper improves the result by 0.3, and Para-Rnd and Para-Tgt by 0.7. These models are notably more appropriate for low-resource target languages than Eng2Ger-MT (Trn) since they use no/few reference captions compared to translation finetuning. Further combining random and targeted paraphrasing results in the largest gain of 1.3. The result is still 3.7 away from Ger. Addressing differences in the perception of the visual world and the way captions are written across cultures is thus an open challenge.

Method	Mean Recall	Vs. Eng2Ger-MT
mCLIP	24.5	-8.9
Eng	26.9	-6.5
Eng2Ger-MT	33.4	0.0
Eng2Ger-MT (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
Eng2Ger-HT	36.8	+3.4
Ger	38.4	+5.0

Table 1: **German I2T/T2I retrieval results.** Mean recall values are averaged over native German cap sets.

5 Further Analysis

Object mentions in English/German captions. To analyze possible differences in perception, we analyze object mention frequency in Flickr30K/Multi30K. We specifically translate German captions to English and extract nouns in both (original) English and (translated to English) German captions. The ratio of English and German mentions is about 1.5, i.e. English mentions object nouns 50% more often than German. However, counts vary by object type. For example, English mentions clothing more often (pants-143% more, shirt-112%, hat-60%, jacket-43%), and German mentions furniture more often (table-37% more, bed-20%, bench-15%). These languages also vary in granularity: English captions often say "people", while German ones say "workers", "athletes", etc.
Analysis of other languages. We conduct initial analysis of the languages and captions in XM3600 (Thapliyal et al., 2022). We group XM3600 languages into European, Arabic/Farsi, Hindi/Bengali, Indonesian/Thai, East Asian, and Swahili categories. After translating each language to English, we report average mention counts and standard deviations per group for various common objects in Table 2. Language groups show large differences in terms of how commonly they mention elements of nature (e.g. mountains, trees), scenery (streets, buildings), household objects (table, plate, box, bottle), and the gender of portrayed people. It is also

	eu-mean	eu-stdev	ar-mean	ar-stdev	hi-mean	hi-stdev	id-mean	id-stdev	easia-mean	easia-stdev	sw
tree(s)	270.5$^-$	92.9	349	19.8	**581.5**$^+$	214.3	286	49.5	274.7	63.5	**383**
mountain(s)	171.1$^-$	47.8	183	24.0	185.5	31.8	173	42.4	**218**$^+$	16.5	**208**
street	**100.9**	30.2	**124**$^+$	50.9	61	7.1	38.5$^-$	10.6	76.7	19.0	82
car(s)	207.3	20.0	235	24.0	**239**	50.9	204$^-$	11.3	220	17.8	**270**$^+$
building(s)	244.8$^-$	69.3	281.5	40.3	329	108.9	**383.5**	84.2	253.3	49.9	**502**$^+$
restaurant	45.8	13.7	**54**	7.0	19$^-$	5.7	**50.5**$^+$	13.4	42.7	6.1	21
table	156.7	52.8	162.5	58.7	**240**$^+$	12.7	**228**	93.3	185.3	43.7	121$^-$
plate	112.5	25.9	90$^-$	12.7	105.5	10.6	109.5	33.2	**119.3**$^+$	5.1	**113**
box	18.1	4.5	15.5$^-$	0.7	15.5	2.1	**28**$^+$	4.2	**24**	2.6	18
bottle	10.2$^-$	2.7	12	0	10.5	2.1	11	4.2	**14.7**$^+$	0.6	**18**
dog	26.2	5.1	28	1.4	**31.5**$^+$	5.0	29.5	0.7	20.7$^-$	5.5	**34**
woman	135.5	23.7	127	5.7	114$^-$	31.1	**164.5**$^+$	20.5	133.3	27.7	**160**

Table 2: **Language shifts in terms of concept mentions in different languages.** We group XM3600 European languages (eu), Arabic/Farsi (ar), Hindi/Bengali (hi), Indonesian/Thai (id), East Asian languages (easia), and report Swahili on its own (sw). The largest two numbers per row are bolded. Observe the differences between the language with highest (+) and lowest (−) counts, which are significantly larger than the within-group standard deviations.

found that the difference between objects counts across languages is much greater than within-group standard deviations. Such results suggest differences in supervision worthy of exploration.

Paraphrasing. LLaMA picks up on granularity differences. For example, PARA-TGT changes "Man in a red shirt riding his bicycle" to "A bicyclist in a red shirt is riding". Further, LLaMA transforms "man on skis" into "skier", "person in blue and red ice climbing" into "ice climber", and "men with children" into "family". The model tends to simplify, irrespective of the reference. 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". Paraphrasing could thus result in over-simplification.

Human evaluation. We extend quantification past retrieval by asking two German speakers to gauge the likelihood that captions are made by a German speaker and their naturalness. We provide 50 random captions for each of 3 sets (ENG2GER-MT, ENG2GER-HT, PARA-TGT). Speakers do not know each set's identity and are tasked with scoring captions as 3=great, 2=good, 1=bad. On average, the speakers rate ENG2GER-HT the highest with a mean ternary score of 2.73 and mean binary score (great/good=1, bad=0) of 0.97. For PARA-TGT, the ternary score is 2.19 and binary score is 0.79. For ENG2GER-MT, the ternary score is 2.16 and binary score is 0.77. These differences approximately reflect the recall results in Table 1.

Recognition. To evaluate object recognition, we compare objects mentioned in a native German caption to ones predicted by the models GER and ENG2GER-HT. We take predictions to be ones with CLIP scores greater than a threshold (the one in range 10:5:50 that maximizes val F1). A prediction is correct only if the object is mentioned in native German. Table 3 shows train-set mentions and performance for the best-performing COCO supercategories. We observe large differences in the number of mentions, precision, and recall for several supercategories. GER achieves better recall (slightly correlated with mention count differences), but ENG2GER-HT better precision. These results suggest potential recognition differences when using translated and native captions.

Supercategory	Vehicle	Animal	Sports	Furniture	Electronic
GER (#men)	2604	2836	2101	**1488**	510
ENG2GER-HT (#men)	**2724**	**2918**	**2127**	1191	**554**
GER (prec)	0.42	0.41	0.16	0.26	0.25
ENG2GER-HT (prec)	**0.47**	**0.51**	**0.17**	**0.29**	**0.27**
GER (rec)	**0.52**	**0.55**	**0.61**	**0.20**	0.28
ENG2GER-HT (rec)	0.46	0.44	0.56	0.16	**0.30**

Table 3: **Recognition stats by supercategory**. Top rows: mention counts, middle: precision, bottom: recall.

6 Conclusion

We show notable differences in using native vs. translated German captions to train a retrieval model, and experiment with three strategies to reduce the gaps. We plan to extend investigation to more languages. Future work can also involve creation of data augmentation strategies that take inspiration from psychology literature (Nisbett and Masuda, 2013; Boroditsky, 2006) and solutions for the ambiguity challenges of machine translation, such as by using images (Futeral et al., 2023).

Acknowledgement. This work was supported by NSF Grants No. 2006885 and 2329992.

Limitations and Ethical Considerations

We only experiment with one translation model, one non-English language (German), and a small amount of runs of LLaMA-3. To ensure that insights generalize, various models, and languages (especially low-resource ones), should be analyzed. There may be intra-language variance amongst native speakers that should also be considered.

We rely on the use of image-caption datasets like Flickr30K and Multi30K. These datasets are relatively small (about 30k samples), so the coverage of concepts may not be fully representative of spoken language. Such datasets have also been noted to contain harmful biases with respect to attributes like race and gender (Van Miltenburg, 2016). The use of models like LLaMA-3 carries similar biases. There should be careful consideration regarding downstream usage of these sets and models. We note that a future extension of our paraphrasing strategies could be to mitigate the impact of in-group perspectives in the captions used for pretraining models.

Our analysis of differences in languages is limited by the fact that languages are machine-translated to English. It is possible that some differences are amplified and/or missed due to machine translation artifacts.

Finally, while we conduct initial human evaluation, we encourage larger-scale human evaluation that expands past our limited evaluation. This can be done to ensure that methods are applicable for a greater amount of people.

References

Raphael Berthele, Matthew Whelpton, Åshild Næss, and Pieter Duijff. 2015. Static spatial descriptions in five Germanic languages. *Language Sciences*, 49:82–101.

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. 2023. 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. 2024. On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14432–14444.

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.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

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.

Geert Hofstede. 2001. Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. *Thousand Oaks*.

J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6521–6528.

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.

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*.

Emiel Van Miltenburg. 2016. Stereotyping and bias in the Flickr30k dataset. *Proceedings of the Workshop on Multimodal Corpora (MMC)*.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

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.

Appendix

Shown are the prompt templates used for querying LLaMA-3 (meta-llama/Meta-Llama-3-8B-Instruct on Hugging Face). 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: $\{\text{ref}_{caps}\}$

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.