A Multimodal Recaptioning Framework to Account for Perceptual Diversity Across Languages in Vision-Language Modeling

Kyle Buettner[1], **Jacob T. Emmerson**[2], **Adriana Kovashka**[1,2]
[1]Intelligent Systems Program, [2]Department of Computer Science, University of Pittsburgh
buettnerk@pitt.edu, jte27@pitt.edu, kovashka@cs.pitt.edu
https://krbuettner.github.io/PerceptualDiversityAcrossLanguages

Abstract

When captioning an image, people describe objects in diverse ways, such as by using different terms and/or including details that are perceptually noteworthy to them. Descriptions can be especially unique across languages and cultures. Modern vision-language models (VLMs) gain understanding of images with text in different languages often through training on machine translations of English captions. However, this process relies on input content written from the perception of English speakers, leading to a perceptual bias. In this work, we outline a framework to address this bias. We specifically use a small amount of native speaker data, nearest-neighbor example guidance, and multimodal LLM reasoning to augment captions to better reflect descriptions in a target language. When adding the resulting rewrites to multilingual CLIP finetuning, we improve on German and Japanese text-image retrieval case studies (up to +3.5 mean recall, +4.4 on native vs. translation errors). We also propose a mechanism to build understanding of object description variation across languages, and offer insights into cross-dataset and cross-language generalization.

1 Introduction

People vary in how they describe the same visual scene. They may note different foreground or background objects (*e.g. person* vs. *sky*). Objects may be described apart or grouped under umbrella terms (*e.g. sofa*, *table*, and *chair* vs. *furniture*). The same object may be noted with a base term (*e.g. dog*), hypernyms (*e.g. animal*), hyponyms (*e.g. Boston Terrier*), or synonyms (*e.g. canine*). Context, like attributes (*e.g. yellow*), may be described if noteworthy or unusual, and captions may vary in detail.

Differences are especially unique *across languages* (Nguyen et al., 2024; Ye et al., 2025), where speakers have diverse perspectives, knowledge, and experiences that contribute to language production. For instance, as shown in Fig. 1, an English speaker may describe an image as containing a *plastic container*, while a Japanese speaker may perceive a *bento box*. The diversity is sometimes subtle, and reflected in object descriptions having different frequencies, abstraction, and usage patterns. For example, we find Japanese captions to mention *sunglasses* 5.6× more often than English ones, potentially due to their relative uncommonness in Japan (thus noteworthiness). There is notably a distribution gap between English captions and captions from other languages in their *perceptual diversity*.

Figure 1: **English captions (and their translations) do not capture the perceptual diversity of object and scene descriptions in other languages.** They often fail to include cultural terms (*e.g. bento box*) and miss differences in native perspective (*e.g.* German emphasis of *American football*). More subtlely, we find that they differ from cross-language captions in the use of common nouns, for instance in Japanese STAIR (Yoshikawa et al., 2017) where *bread* is more frequently described, especially with its contents (*e.g. vegetables*). Our multimodal recaptioning method considers these differences to enhance cross-lingual training data generation.

With the rise of multilingual vision-language models (VLMs) (Zhai et al., 2023; Carlsson et al., 2022; Yue et al., 2024; Chen et al., 2023b, 2024; Geigle et al., 2023), machine translation from English has been used to generate cross-lingual data. A key observation is that *machine translation does*

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1989–2006

not significantly adapt semantic content. It relies on source object naming and context, leading to an English perceptual bias that limits understanding of native text in other languages. To reduce bias, text can be diversified with strategies like paraphrasing and general captioning. We explore such options, but find they achieve limited understanding of culture-specific perceptual differences.

As a more compelling strategy to reduce perceptual bias, we propose a multimodal recaptioning framework that alters object descriptions to reflect properties in a target language. We simply incorporate a small amount of reference data and image context in the prompting of a multimodal LLM to infer how concepts are described cross-language, and then produce rewrites. This process is guided by reference captions from similar scenes, selected as nearest neighbors in image similarity space.

Rewrites are then integrated as random augmentations in the training of mCLIP (Chen et al., 2023a), a multilingual image-text retrieval model. We compare the use of rewrites generated through this *targeted image recaptioning* to the use of captions produced from non-targeted prompts that encourage diversity (i.e. paraphrasing and general recaptioning). Performance is evaluated with Japanese STAIR (Yoshikawa et al., 2017), German Multi30k (Elliott et al., 2016), and XM3600 (Thapliyal et al., 2022), including in cross-dataset and cross-language settings. Notably, targeted image recaptioning improves default training by up to +2.4 mean recall and outperforms both diverse paraphrasing and image captioning on native vs. translation error sets which isolate perceptual differences between languages by up to +4.4. In combination, all rewrites improve mCLIP by +3.5. We share further insights by highlighting key differences in object term distributions across datasets.

In summary, along with our framework, our main contributions are insights into these questions:

- In which contexts is targeted image recaptioning beneficial?
- How do differences in object descriptions uniquely manifest across languages?
- Does understanding of perceptual diversity gained from one target language dataset generalize to other datasets and languages?

2 Related Work

Perceptual differences in object descriptions. Much progress in computer vision has been driven by models trained on datasets with mostly English text, such as the CLIP dataset (Radford et al., 2021), or on datasets with English noun hierarchies, such as ImageNet (Deng et al., 2009). The underlying concept representations are thus biased towards the details English speakers find salient, and the entry-level categories English speakers use to name objects (Ordonez et al., 2013). With English conceptual understanding lacking universality (Liu et al., 2021), models may fail to adequately capture perceptual diversity across cultures (*e.g.* representation of the Japanese *koto* instrument). There is much research on cultural differences in perception, such as on attention to foreground vs. background (Nisbett and Masuda, 2013) and on attributes prescribed to objects because of grammar (Boroditsky, 2006). Recent work has found that perceptual diversity in multilingual datasets helps English vision tasks (Nguyen et al., 2024). Our work addresses English perceptual bias to achieve greater understanding of native speaker text in other languages.

Multilingual, vision-language modeling. Recent works have moved to instill popular English-centric VLMs and multimodal LLMs with multilingual capabilities (Zhai et al., 2023; Carlsson et al., 2022; Yue et al., 2024; Chen et al., 2023b, 2024; Geigle et al., 2023). In the absence of native speaker data, it is common to machine-translate captions (Carlsson et al., 2022) and instructions (Yue et al., 2024), though the resulting text carries an English bias. We show that *targeted, image recaptioning* can address this bias, and improve multilingual CLIP retrieval (Chen et al., 2023a) on two languages (German/Japanese). Our work fits with recent VLM works that effectively leverage LLMs to generate diversity in text for retrieval/classification (Fan et al., 2023) and compositional understanding (Doveh et al., 2023b,a). It also relates to aligning multiple texts to an image (Sarto et al., 2023; Bulat et al., 2024), but past work does not consider differences across languages. We aim to align multiple, perceptually diverse views to an image through random sampling of text in training.

Machine translation and image captioning. Gaps have been shown in image-text retrieval when using translated vs. native speaker training data (Kádár et al., 2018). Buettner and Kovashka (2024) show that *text-only paraphrasing* techniques can partially address gaps (+1.3 performance). Our perspective is that text changes need to be *visually* driven as speakers across cultures may uniquely focus on different parts of the image, which para-

phrasing does not address. We thus employ new *image*-based reference sampling and targeted, *image* recaptioning techniques, and further consider a greater scope of languages and datasets. Other work (Yang et al., 2023) uses k-NN in image embedding space to help retrieve knowledge for few-shot English captioning. Our method is unique as the nearest image neighbor guides creation of *perceptually diverse*, *cross-lingual* data to be used as training augmentations in *image-text retrieval*. Ramos et al. (2024) train a model for multilingual captioning using retrieval with reference translations. We alternatively show importance of a reference set with captions directly produced by native speakers of another language, and our recaptioning does not require training (just prompting). Our work is loosely related to nearest-neighbor machine translation (Khandelwal et al., 2021) and multimodal machine translation (Yao and Wan, 2020). Unlike in machine translation, we significantly adjust caption content to address perceptual bias.

3 Experimental Methodology

We consider an image-text retrieval case study where the text queried and retrieved come from native speakers of a target language (*e.g.* Japanese, German). The goal is to design a multimodal framework that improves VLM understanding of perceptually diverse text across languages. Specifically, we aim to enhance training with machine translation of English text, which is a practical strategy when a *large* amount of target language text from native speakers is unavailable. The challenge is that caption properties, such as which objects are mentioned, the level of detail or context described, and use of certain synonyms, hypernyms, or hyponyms, are biased towards English perception.

To address this bias, for a given input caption and image, we propose to use a multimodal LLM (*Llama-3.2-11B-Vision-Instruct*) to produce rewrite(s), specifically by leveraging multimodal context and reference examples selected by image similarity. Outlined in this section are our framework's retrieval model (Sec. 3.1), mechanisms for generating data with adequate diversity (Sec. 3.2), and strategy to identify object description differences between datasets (Sec. 3.3).

3.1 Preliminaries: Multilingual CLIP

We select the retrieval model to be multilingual CLIP or *mCLIP* (Chen et al., 2023a), due to its support for multiple languages and CLIP's success as a VLM. This method extends the mostly English CLIP by replacing the text encoder with a multilingual model, XLM-R (Conneau et al., 2020), and training lightweight projection layers to align multilingual text embeddings to CLIP image and text embeddings. We are primarily interested in *fine-tuning* to adapt mCLIP's alignment of images and captions. We specifically train the image-to-text (I2T) and text-to-image (T2I) matching losses in Chen et al. (2023a), which operate over a batch of size N where each sample k has an image i_k and text t_k. Shown in Eqs. 1 and 2 are these losses with the multilingual text encoder f, the CLIP image encoder g, a temperature τ, and a similarity function $\langle\rangle$ (cosine):

$$\mathcal{L}_{I2T} = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp(\langle f(t_k), g(i_k)\rangle/\tau)}{\sum_{n=1}^{N}\exp(\langle f(t_n), g(i_k)\rangle/\tau)} \tag{1}$$

$$\mathcal{L}_{T2I} = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp(\langle f(t_k), g(i_k)\rangle/\tau)}{\sum_{n=1}^{N}\exp(\langle f(t_k), g(i_n)\rangle/\tau)} \tag{2}$$

The overall loss is $\mathcal{L} = \frac{1}{2}(\mathcal{L}_{I2T} + \mathcal{L}_{T2I})$.

3.2 Multimodal Recaptioning

Main approach. Assume we have a dataset $\mathcal{D}_{train}$ of image-caption pairs with text in a source language $\mathcal{L}_{src}$ (i.e. English). We also have a target language $\mathcal{L}_{tgt}$ for which we wish to generate training data (*e.g.* Japanese/German). We propose to alter the object descriptions of captions in $\mathcal{D}_{train}$ before machine translation to $\mathcal{L}_{tgt}$ for enhanced diversity. A natural approach is to paraphrase English text to produce alternative descriptions of objects (*e.g.* calling a *Boston Terrier* a *dog*). However, it is difficult to infer from only the source text adequate culture-specific terms (e.g. *yakitori* vs. *chicken*). In addition, what may be deemed salient by an English speaker may exclude salient objects often written in text from a speaker of another language.

We reason that image context, along with effective guidance from native speaker data, are needed to adequately diversify captions. We thus propose to use a multimodal LLM, with strong reasoning capabilities, to rewrite each training caption in a targeted manner. The idea is to leverage a small

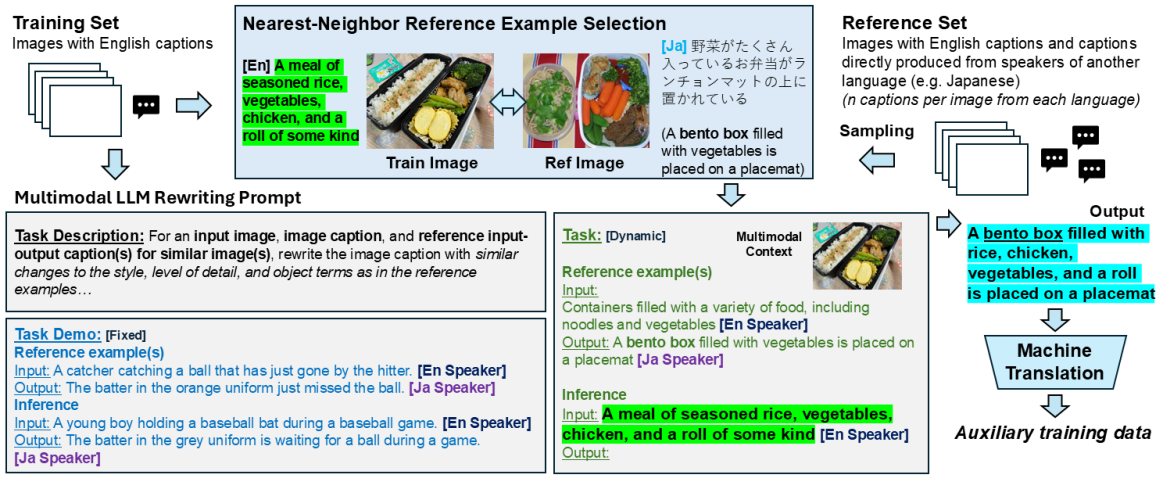

Figure 2: **Our multimodal, LLM-based recaptioning method to adapt object descriptions before translation.** For a set of images with only English captions, we generate new captions which better represent perceptual properties of a target language (*e.g.* Japanese). Each generation is guided by a reference example selected as the nearest neighbor in image similarity from a small set of native speaker data. Using the prompt shown, the multimodal LLM leverages the reference example and image context to infer targeted changes. This example shows the model adding the cultural term *bento* while also listing foods relevant to the input image. Text in brackets is not in the prompt.

reference set $\mathcal{D}_{ref}$, disjoint from $\mathcal{D}_{train}$, which consists of images and captions produced in English *and* from native speakers of $\mathcal{L}_{tgt}$ (*e.g.* German/Japanese). A single input-output *guidance example* is selected from $\mathcal{D}_{ref}$ to instruct the LLM on how an (input) English caption can be rewritten as an (output) caption from a native speaker of $\mathcal{L}_{tgt}$. The multimodal LLM then reasons how reference descriptions can apply to a new image.

We specifically use LLaMA 3.2 (Touvron et al., 2023), with the instruction in Fig. 2. Provided is a task description, demo for formatting (constant across inferences), the training image *and* caption, and the sampled reference texts for the given input. Note in the example how the reference output shows description of *bento box*, and that the LLM generalizes its use in the rewrite while also including input image-relevant ingredients (*e.g.* *rice*, *chicken*). The guidance example is translated to English (with Google Translate), and the LLM produces output in English. We recaption in English, then machine-translate, to ensure quality in a language with which we have familiarity.

Choosing guidance examples. Reference examples are selected with the idea that images with similar content may be described similarly. Thus if there is native speaker data for a similar image, object description properties can be inferred for the image to be recaptioned. We choose the nearest neighbor in image feature space as the guidance example, using image embeddings obtained from zero-shot mCLIP. This process captures culture-specific terms like in Fig. 2 and also more subtle differences across languages, as shown in Fig. 3.

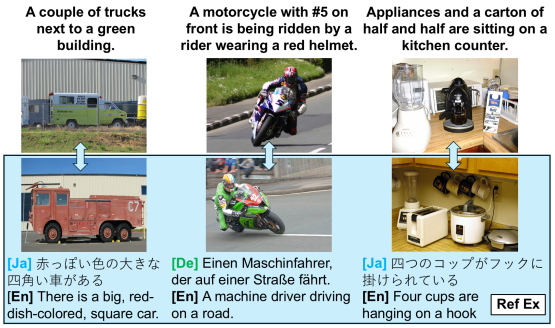

Figure 3: **Guidance from nearest neighbors can reveal subtle differences in object naming.** A reference is chosen based on image similarity to acquire diverse descriptions of related concepts and scenes. Notice how *truck* may be described loosely as *car* in Japanese. Similarly, there are differences in object grouping and objects that are deemed salient (described).

Other approaches. We also consider non-targeted methods (without use of a reference set), given that there is general variability of object descriptions across languages. We propose a multimodal captioning strategy where given an image and caption, we encourage the LLM to rewrite the caption with differences in phrases, sentence structure, semantic content, which objects are described, and/or level of detail. Otherwise the format is the same as the targeted scenario, thus ablating the impact of reference guidance. Then as a baseline similar

to prior work (Fan et al., 2023; Buettner and Kovashka, 2024), we prompt LLaMA to produce a paraphrase that reflects diversity in how speakers around the world describe objects across languages. We refer to the above two strategies as *Diverse Image Recaptioning* and *Diverse Paraphrasing*, respectively, and compare to our proposed *Targeted Image Recaptioning*, when used separately and in conjunction. The prompts are in Appendix A.1.

Incorporating rewrites into training. To use rewrites (after machine translation), we leverage a similar mechanism to Fan et al. (2023), where generated captions are treated as random augmentations in retrieval training. This simple mechanism encourages images to match to multiple, perceptually diverse views over the course of training. Referring to Eqs. 1 and 2, instead of using the original batch B of image-text pairs (i_k, t_k), a new batch B' is constructed consisting of image-text pairs (i_k, t'_k). Each t'_k is sampled from a uniform distribution of all possible positives for an image (i.e. the original caption t_k and each possible rewrite p^i_k). Formally, this process is shown in Eq. 3 for n positive rewrites:

$$t'_k \sim \text{Uniform}([t_k, p^1_k, ..., p^n_k]) \quad (3)$$

When one rewriting strategy is used, n=1, meaning that approximately 50% of training instances are rewrites. We explore up to n=3 when considering all 3 rewrite strategies in combination.

3.3 Identifying Object Description Differences Across Languages

It remains an open question how object description differences manifest across inter-language datasets. Specific questions include how often certain synonyms or hyponyms are used (*e.g.* in Fig. 3, *truck* vs. *car* for *vehicle*), whether grouping terms are used (*e.g. furniture*), and if certain terms are used significantly more in one language than another. To explore, we design a strategy with WordNet (Miller, 1995) to create object description distributions.

Consider a set $\mathcal{H}$ consisting of "supercategories" which cover various objects. In this study, we choose $\mathcal{H}$ to be {person, conveyance, furniture, animal, container, food, device}, representing common objects in COCO. Then given source language captions $\mathcal{D}_{src}$ and target language captions that have been translated to English $\mathcal{D}_{tgt}$, we process each caption in $\mathcal{D}_{src}$ and $\mathcal{D}_{tgt}$ with a SpaCy part of speech tagger (*en_core_web_sm*, v3.6.1) to identify nouns. Each noun is mapped to the probable WordNet synset (the first noun synset definition listed). Then if possible, we match each mapped synset to the closest synset of a top-level term in $\mathcal{H}$ by performing a closure of hypernyms. We compare term frequencies under each supercategory in Sec. 5 to learn about differences across languages.

4 Experimental Settings

Datasets. We train on English COCO and evaluate on Japanese STAIR (Yoshikawa et al., 2017), which includes Japanese captions from native speakers for COCO images. Similarly, we train on English Flickr30k and evaluate on German Multi30k (Elliott et al., 2016), which includes German captions from native speakers for Flickr images. The English captions serve as input to recaptioning and machine translation for retrieval training. Disjoint sets of native captions are used in targeted recaptioning. Specifically, in both case studies, there are 5 English caption sets (5 captions per image). STAIR contains 5 Japanese captions for each COCO image, and Multi30k contains 5 German captions for each Flickr image. We randomly split the 5 sets for each dataset/language into reference, training, and evaluation sets, so there is disjoint data for the targeted method. The image split sizes for STAIR are 9,666/73,117/10,668 and for Multi30k are 9,666/9,666/10,668. For each image, our targeted method samples 1 caption from the 5 English reference sets and 1 caption from the 5 cross-language sets. Evaluation is averaged across the 5 test sets. For cross-dataset experiments, we evaluate on XM3600 (Thapliyal et al., 2022), as it includes native speaker data for 36 languages. We do not train on XM3600 due to size (3,600 images only). We report results for languages related to the ones for which we have native speaker data.

Recaptioning. Rewrite generation is performed with the multimodal LLaMA 3.2 (*Llama-3.2-11B-Vision-Instruct*). A temperature of 0, seed of 42, and max tokens 448 are used for all experiments.

Translation. To generate cross-lingual training data, we use two models from HelsinkiNLP (Tiedemann and Thottingal, 2020), *opus-tateoba-en-ja* for Japanese and *opus-mt-en-de* for German, as these are amongst the most downloaded on HuggingFace. We also test No Language Left Behind (Costa-Jussà et al., 2022) in Appendix A.2. Translations of English captions, including rewrites, are generated with greedy decoding at a max token count of 200. We notably choose a much higher to-

Method	I2T Retrieval			T2I Retrieval			Mean Recall
	R@1	R@5	R@10	R@1	R@5	R@10	
Train: English COCO (to Japanese) / **Eval**: STAIR (Japanese)							
mCLIP	10.2	25.0	33.9	9.2	23.2	31.9	23.0
FT on Japanese Data MT from English	20.2	43.0	54.7	19.4	41.9	53.3	39.3
+ Rewrites from Diverse Paraphrasing	21.7	45.1	56.7	20.6	43.9	55.4	40.6
+ Rewrites from Diverse Image Recaptioning	22.0	45.3	57.1	20.7	44.1	55.6	40.8
+ Rewrites from Targeted Image Recaptioning	**22.5**	**46.5**	**58.1**	**21.4**	**45.0**	**56.7**	**41.7**
FT on Japanese Data from Native Speakers	24.8	50.2	62.0	24.3	49.4	61.2	45.7
Train: English Flickr30k (to German) / **Eval**: Multi30k (German)							
mCLIP	13.8	31.6	41.8	13.0	30.6	40.7	28.6
FT on German Data MT from English	22.3	45.4	56.4	21.5	44.4	55.6	40.9
+ Rewrites from Diverse Paraphrasing	22.5	46.0	57.1	21.7	44.9	56.2	41.4
+ Rewrites from Diverse Image Recaptioning	**22.7**	46.3	57.5	22.1	45.5	56.8	41.8
+ Rewrites from Targeted Image Recaptioning	**22.7**	**46.5**	**57.8**	**22.3**	**45.8**	**57.1**	**42.1**
FT on German Data from Native Speakers	22.5	46.1	57.6	22.1	45.9	57.1	41.9

Table 1: **Targeted image recaptioning is the most beneficial augmentation in text-image retrieval on native speaker data from STAIR (Japanese) and Multi30k (German).** The + symbol indicates that data is added as augmentations to the "FT on X Data MT from English" setting. FT=finetuned, MT=machine-translated.

ken count in translation than the paraphrasing work of Buettner and Kovashka (2024), as it is found to significantly improve the quality of translation.

Retrieval. For *training*, results are collected with the settings in Chen et al. (2023a) (batch size 512, learning rate 0.001, 30 epochs, LAMB optimizer, temp. 0.07) on 1 NVIDIA A100 GPU. For *evaluation*, I2T and T2I retrieval scores are calculated as recall@1/5/10. The mean of these six scores, termed mean recall (Chen et al., 2023a), is calculated and reported as averages over the 5 test sets.

Native vs. translation error sets. We construct other test sets that isolate differences from using translation (with English perceptual bias) vs. native speaker data. We train mCLIP models with captions translated from English to Japanese/German and with native Japanese/German captions. Then we collect I2T/T2I cases that the native models correctly retrieve within 10 samples, but the translation models get incorrect@10, as we reason these cases include errors that would be addressed with understanding of perceptual diversity. We refer to these collectively as *Native vs. Translation Error Sets*. Evaluation considers retrieval over the full test sets, but mean recall is only calculated across cases in these sets. The I2T/T2I sample counts are 1,409/1,391 for STAIR and 994/946 for Multi30k. This evaluation is shown in Table 2, while all other tables follow overall retrieval evaluation.

5 Results and Analysis

We evaluate *Targeted Image Recaptioning* vs. *Diverse Image Recaptioning* vs. *Diverse Paraphrasing*; the latter represents baselines (Fan et al., 2023; Buettner and Kovashka, 2024). We baseline mCLIP without finetuning, and mCLIP finetuned on data that has been machine-translated from English to the target language (without recaptioning). The datasets from rewrite strategies are incorporated into finetuning as augmentations that are randomly sampled with original machine translations. As a reference, we evaluate finetuning with captions from speakers of each language (Native Ja/De), though this is not a strict upper bound since more data is used in rewrite settings. We also test combinations of all rewrite strategies.

Method	I2T Retrieval			T2I Retrieval			Mean Recall
	R@1	R@5	R@10	R@1	R@5	R@10	
Train: English COCO / **Eval**: STAIR (Japanese)							
FT (MT Data)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
+Paraphrase	0.5	8.9	26.6	0.4	7.5	24.2	11.4
+Diverse Recap	0.7	8.8	26.4	0.9	9.5	25.0	11.9
+Tgt Recap	**1.3**	**14.1**	**32.4**	**1.4**	**13.4**	**35.2**	**16.3**
FT (Native Ja)	16.0	60.0	100.0	16.4	61.1	100.0	58.9
Train: English Flickr30k / **Eval**: Multi30k (German)							
FT (MT Data)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
+Paraphrase	0.1	6.8	20.5	0.1	5.3	22.1	9.2
+Diverse Recap	0.2	7.2	26.1	0.5	7.7	24.2	11.0
+Tgt Recap	**0.8**	**10.5**	**28.8**	**0.8**	**11.1**	**29.8**	**13.6**
FT (Native De)	10.9	52.2	100.0	10.6	56.2	100.0	55.0

Table 2: **Targeted image recaptioning is especially helpful on error cases which come from not using native speaker text.** Retrieval on STAIR (Ja)/Multi30k (De) *Native vs. Translation Error Sets*.

5.1 In which contexts is targeted image recaptioning beneficial?

Image-text retrieval with native speaker text. Reported are results on (1) overall STAIR and Multi30k (Tab. 1) and (2) the *Translation vs. Native Error Sets* which isolate perceptual differences

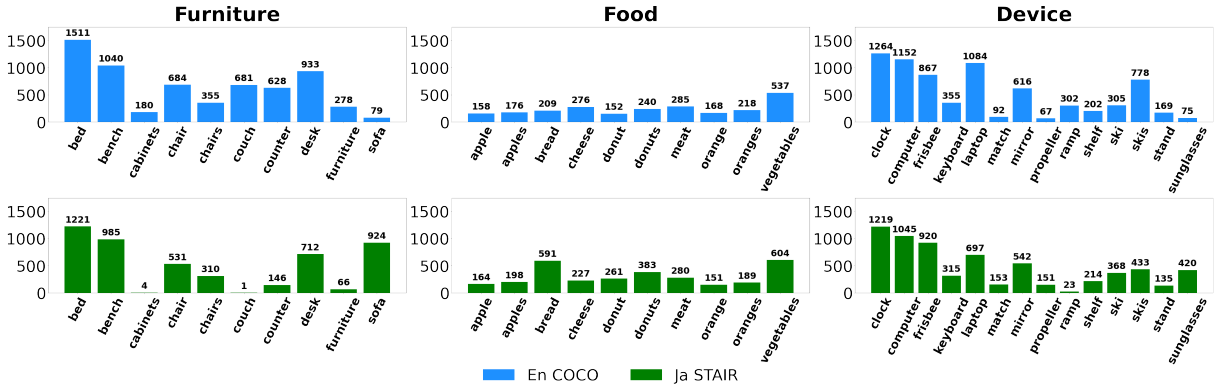

Figure 4: **When comparing English COCO vs. Japanese STAIR captions, object term distributions are found to vary across languages.** For each supercategory, any term with count > 150 is identified, and the union of terms across languages is shown. Note unique variation across common objects (*e.g.* counter, furniture, bread, sunglasses).

Method	I2T Retrieval			T2I Retrieval			Mean
	R@1	R@5	R@10	R@1	R@5	R@10	Recall
Train: English COCO / **Eval**: XM3600 (Japanese)							
FT (MT Data)	39.4	67.6	78.3	38.9	68.5	78.2	61.8
+Paraphrase	41.1	69.6	80.0	41.1	69.5	79.5	63.5
+Diverse Recap	42.6	70.0	80.6	42.1	70.6	79.6	64.2
+Tgt Recap	**42.9**	**71.6**	**80.8**	**42.9**	**72.1**	**81.7**	**65.3**
Train: English Flickr30k / **Eval**: XM3600 (German)							
FT (MT Data)	38.4	67.3	77.4	37.8	65.0	75.9	60.3
+Paraphrase	39.1	68.3	78.2	37.3	64.8	75.8	60.6
+Diverse Recap	**39.5**	**68.6**	78.8	38.0	65.2	**77.1**	61.2
+Tgt Recap	**39.5**	68.5	**78.9**	**38.7**	**66.6**	76.6	**61.5**

Table 3: **Targeted image captioning is effective cross-dataset.** Shown is retrieval on XM3600 (intra-language) for models trained on Japanese/German.

Rewrites for Image			Mean
Paraphrase	Diverse Img Recap	Targeted Img Recap	Recall
			39.3
✓			40.6
	✓		40.8
		✓	41.7
✓	✓		41.6
	✓	✓	42.5
✓		✓	42.4
✓	✓	✓	**42.8**

Table 4: **Targeted image recaptioning is complementary to other augmentation strategies.** Shown is mean recall on STAIR (Ja) when combining rewrite strategies with default translation data (30 epochs).

between English and German/Japanese (Tab. 2). The best method in both tables is *Targeted Image Recaptioning*. In Tab. 1, mean recall gains over default finetuning are +2.4 on STAIR and +1.2 on Multi30k. Gains are especially notable for Japanese, which differs much from English. In this case study, the targeted method outperforms *Diverse Paraphrasing* by +1.1 and *Diverse Image Recaptioning* by +0.9, respectively. These results illustrate benefits in using a targeted mechanism with a modest amount of native speaker data (≈10k total). We test other reference set sizes in App. A.3.

Error cases which capture perceptual differences. Tab. 2 illustrates the value of *Targeted Image Recaptioning* in addressing perceptual gaps. It outperforms all methods by +4.4 on Japanese and +2.6 on German. The method may perform well on these cases due to enhanced use of culture-specific terms, which we validate with captioning in App. A.4 and term counts in App. A.5. Describing one example, the counts of *bento* for default English captions, paraphrasing, and general captioning are 6, 4, and 5, respectively. The targeted method has 12, closer to the native Japanese 25. We also verify that LLM is not simply hallucinating terms, and show example rewrites in App. A.6.

Across datasets. To explore the generalizability of the learned perceptual understanding, we test methods cross-dataset on XM3600 image-text retrieval (Tab. 3). We find that *Targeted Image Recaptioning* is also the top method cross-dataset. The Japanese model performs especially well, improving by at least +1.1 over all methods. This experiment shows that our method results in understanding that is applicable outside of the training domain.

In combination with other augmentations. We test each method together by allowing random sampling from combined sets during training (Tab. 4). The top gains are from the combination of all methods (+3.5), and next are respective combinations of *Targeted Image Recaptioning* with *Diverse Image Recaptioning* (+3.2) and *Diverse Paraphrasing* (+3.1). In analyzing these complementary benefits, we reason that paraphrasing can address general term diversity shared across speakers of different

[Ja] 大きな緑の看板に英語で書かれている

It's written in English on a big green sign

[De] Ein Fußballspieler beim Training.

A football player during training.

[Ja] サングラスを掛けた男性が、スケートボードに乗ってジャンプしている

A man wearing sunglasses is jumping on a skateboard

Figure 5: **I2T retrievals that *Targeted Image Recaptioning* gets correct but default finetuning gets incorrect@10.** The targeted method addresses unique differences in perspective and level of detail.

languages (*e.g.* describing a *car* sometimes as a *vehicle*), and both forms of multimodal recaptioning can help address differences in object focus by incorporating new concepts not in the original caption (*e.g.* adding *clothes* if not mentioned). The targeted method has further advantages by addressing unique properties for a given language (*e.g.* often referring to a *lunch box* as a *bento box*).

Using other image neighbors. We test *Targeted Image Recaptioning* using different image neighbors (in terms of similarity) to guide rewrites. On STAIR, using k=1 results in 41.7 mean recall, k=2 results in 41.6, and k=3 results in 41.6. These results suggest that other neighbors can be effective and add further diversity.

5.2 How do differences in object descriptions uniquely manifest across languages?

There are culture-specific term frequency differences expected across languages (*e.g. futon* appearing more in Japanese). However, less obvious are differences in the distributions of common nouns. To quantify such differences, we use the method in Sec. 3.3, and calculate supercategory-grouped counts of common objects for Japanese and German vs. English. In Fig. 4, we show examples for *furniture*, *food*, and *device* (Japanese vs. English). Results for German vs. English and for other categories are in Appendix A.7.

In Fig. 4, for the supercategory *furniture*, the En set uses the term *furniture* 4.2× the Ja set while Ja describes *sofa/couch* 1.2×, showing potential differences in object grouping. For *food*, *bread* is described 2.8× more in Ja. Upon inspection, we find phrases like *bread with meat* are used synonymously with *sandwich*. For *device*, *sunglasses* is described 5.6× more in Ja, potentially a result of sunglasses culturally being less common in Japan (and more noteworthy). Cases like these point to perceptual diversity that is worthy of future study.

Method	Ja	Ko	Zh	De	Fr	Cs	Da
Train: English COCO to Japanese / **Eval**: XM3600							
FT (MT Data)	61.8	23.6	51.9	55.8	34.7	42.7	34.0
+Paraphrase	63.5	25.9	54.5	57.1	36.3	44.8	36.5
+Diverse Recap	64.2	24.8	**54.8**	58.2	36.6	**45.6**	36.3
+Tgt Recap	**65.3**	**26.0**	**54.8**	**58.3**	**37.3**	44.8	**36.8**
Train: English Flickr30k to German / **Eval**: XM3600							
FT (MT Data)	54.6	21.6	49.8	60.3	34.8	42.6	37.3
+Paraphrase	**55.5**	21.7	**50.7**	60.6	**36.1**	43.9	38.3
+Diverse Recap	55.1	**22.8**	50.3	61.2	36.0	**44.1**	**38.6**
+Tgt Recap	54.7	22.2	50.6	**61.5**	35.5	43.9	38.3

Table 5: **There is a need to learn about perceptual diversity in other languages.** We report mean recall for retrieval on different language-specific sets when performing targeted recaptioning for Japanese/German. These results show targeted recaptioning to be best intra-language, but other strategies to be similarly (or more) compelling cross-language. Perceptual diversity learned from Japanese/German may thus be unique.

Further examples show up in the retrieval error cases. Shown in Fig. 5 are some I2T retrieval cases in the *Translation vs. Native Error Sets* for Japanese/German that *Targeted Image Recaptioning* gets correct, but the finetuned baseline does not. Observe how the targeted method improves on unique cases, such as an out-group view of a New York sign, and the text describing *sunglasses* (a description difference identified in Fig. 4).

5.3 Does perceptual diversity understanding gained from one target language dataset generalize to other datasets/languages?

In addition to cross-dataset tests with XM3600, we also test models in a cross-lingual manner by performing targeted recaptioning for one language (Japanese/German) and evaluating on language-specific retrieval sets for geographically proximate languages: Korean/Chinese for Japanese and French/Czech/Danish for German. Note that high performance in each case is *not* a goal of our model, but this study provides insight into whether some perceptual understanding may be generalizable across languages. Table 5 shows the results. While the targeted method is best intra-language, the other rewrite strategies become more compelling cross-language, with smaller gaps or gains vs. the targeted method. These results indicate that other languages may not benefit much from learning German/Japanese perceptual details, and likely have their own unique perceptual diversity. This insight can inspire future work to ensure adequate consideration of native speaker data from other languages.

6 Conclusion

In this work, we provide a multimodal framework to encourage VLMs to learn diverse perceptual understanding across languages. We find that multimodal LLMs, with nearest-neighbor guidance, are effective at inferring how to change object descriptions across languages. Our targeted method improves image-text retrieval, especially on error cases that come from English perceptual bias. The targeted method is also found to be complementary to other augmentations, and gains generalize across datasets. We provide unique insights into obvious and more subtle ways that object text differences manifest across cross-language datasets. Future work needs to be dedicated to the acquisition of native speaker captions across other languages for more expansive investigation.

Acknowledgement. This work was supported by NSF Grant 2329992 and a University of Pittsburgh Intelligent Systems Provost Fellowship.

7 Limitations

First, our framework is limited by the availability of native speaker image captions across languages. This constraint is the primary reason we study Multi30k German and STAIR Japanese, since they have captions directly produced from native speakers to pair with English text of Flickr and COCO, respectively. We encourage the acquisition of native speaker data from more languages, especially low-resource ones, for study in the future.

Second, we use a single, small set of reference examples. While the mechanism is performant and data-efficient, there is intra-language diversity that remains uncaptured in such a set. Furthermore, the domain shift between the reference set and a test set of interest can be significant. Future work can expand the language and image diversity represented in the reference set to address such cases.

Third, our framework and analysis depend on machine translation quality. While we verify effectiveness across multiple machine translation techniques, improvements in machine translation are likely needed to maximize cross-lingual performance. Some object description count differences may be a result of translation artifacts (*e.g. sofa*/*couch*), though we combine cases in analysis.

Fourth, our WordNet mechanism to identify object description differences across datasets is imperfect due to the idiosyncratic synset structure of WordNet. Polysemy can affect interpretation of counts, through we try to handle some aspects of disambiguation (*e.g.* differentiating between noun and adjective forms of *orange* through part-of-speech tagging). Future methods can be developed to probe differences more precisely.

Fifth, we only explore one model each for recaptioning and retrieval training, but models of different scale may show different behaviors. We reason larger models with stronger instruction following may be more effective at altering input captions to reflect the guidance examples. They may be able to identify the most relevant differences between languages to use for adaptation, and even discover subtle differences. In addition, there may be less risk of hallucination. Conversely, smaller models with less task-following capability may be less effective at incorporating desired changes. Future work can investigate the impact of scaling on a model's ability to understand perceptual diversity. Retrieval models like SigLIP (Zhai et al., 2023) may be worth studying.

8 Ethical Considerations

While we address one form of bias, there are various other biases in the datasets we use for training and evaluation, such as racial and gender biases. An extra filtering or rewriting step could potentially help address these biases for downstream use cases. Our targeted mechanism could also help generate training data to teach models less biased descriptions of general content.

References

Lera Boroditsky. 2006. Linguistic relativity. *Encyclopedia of Cognitive Science*.

Kyle Buettner and Adriana Kovashka. 2024. Quantifying the gaps between translation and native perception in training for multimodal, multilingual retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5863–5870, Miami, Florida, USA. Association for Computational Linguistics.

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. 2024. FFF: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14172–14182.

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.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2024. On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14432–14444.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Alexander Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023b. Pali: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations (ICLR) (2023)*.

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.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. 2023a. Dense and aligned captions (DAC) promote compositional reasoning in VL models. *Advances in Neural Information Processing Systems*, 36:76137–76150.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. 2023b. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2657–2668.

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. Association for Computational Linguistics.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving CLIP training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575.

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

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, Brussels, Belgium. Association for Computational Linguistics.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. *International Conference on Learning Representations (ICLR)*.

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*.

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

Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei Koh, and Ranjay Krishna. 2024. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*.

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.

Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. From large scale image categorization to entry-level categories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2768–2775.

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.

Rita Ramos, Emanuele Bugliarello, Bruno Martins, and Desmond Elliott. 2024. PAELLA: Parameter-efficient lightweight language-agnostic captioning model. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3549–3564, Mexico City, Mexico. Association for Computational Linguistics.

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*, pages 6914–6924.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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*.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, et al. 2023. Re-ViLM: Retrieval-augmented visual language model for zero and few-shot image captioning. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, and Ranjay Krishna. 2025. Semantic and expressive variations in image captions across languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 29667–29679.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale Japanese image caption dataset. pages 417–421.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal LLM for 39 languages. *arXiv preprint arXiv:2410.16153*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision (ICCV)*, pages 11941–11952.

A Appendix

A.1 Prompts for Rewrite Strategies

Diverse Paraphrasing

Task: The objective is to paraphrase an English caption to reflect diversity in how speakers around the world describe objects, especially across languages. It is very important to strictly follow the listed requirements.

Requirements:
- Output only a single paraphrased caption which must start with <final> and end with </final>.
- Example: <final> There is a blue bicycle and red motorcycle on the street. </final>
- Do not output any additional quotes, text, comments, explanations, or details. Just the caption.

Please complete this example:
Input: {input}
Output:

Diverse Image Recaptioning

Task Description: For an input image and an input caption, produce a one-sentence image caption that differs significantly from the input caption in order of phrases, sentence structure, semantic content, which objects are described, and/or level of detail. Make sure the output differs from the input caption and use

the image for guidance. Only perform changes that are correct and semantically relevant to the given input image. After "Output: ", always output a <final> tag, followed by a rewritten caption, then </final>. Never any other text or explanation. One task demo for formatting and change instruction is provided.

Task Demo: Inference
Input: A young boy holding a baseball bat during a baseball game.
Output: <final> The batter in the grey uniform is waiting for a ball during a game. </final>

Now perform the task exactly as above:
Inference
Input: {input}
Output:

Targeted Image Recaptioning
Task Description: For an input image, image caption, and reference input-output caption(s) for similar image(s), rewrite the image caption with similar changes to the style, level of detail, and object terms as in the reference examples. Only perform changes that are correct and semantically relevant to the given input image. After "Output: ", always output a <final> tag, followed by a rewritten caption, then </final>. Never any other text or explanation. One task demo for formatting and change instruction is provided.

Task Demo:
Reference example(s)
Input: A catcher catching a ball that has just gone by the hitter.
Output: The batter in the orange uniform just missed the ball.
Inference
Input: A young boy holding a baseball bat during a baseball game.
Output: <final> The batter in the grey uniform is waiting for a ball during a game. </final>

Now perform the task exactly as above:
Reference example(s)
{reference_examples}

Inference
Input: {input}
Output:

Method	I2T Retrieval			T2I Retrieval			Mean
	R@1	R@5	R@10	R@1	R@5	R@10	Recall
Translation: HelsinkiNLP Tatoeba (to Japanese)							
FT (MT Data)	19.4	41.9	53.3	20.2	43.0	54.7	39.3
+Paraphrase	21.7	45.1	56.7	20.6	43.9	55.4	40.6
+Diverse Recap	22.0	45.3	57.1	20.7	44.1	55.6	40.8
+Tgt Recap	**22.5**	**46.5**	**58.1**	**21.4**	**45.0**	**56.7**	**41.7**
Translation: No Language Left Behind (to Japanese)							
FT (MT Data)	19.5	41.7	52.8	18.6	40.4	51.9	37.5
+Paraphrase	20.7	43.6	54.6	19.7	42.1	53.4	39.0
+Diverse Recap	21.2	44.2	55.5	20.0	42.6	54.0	39.6
+Tgt Recap	**21.6**	**45.0**	**56.5**	**20.5**	**43.6**	**55.2**	**40.4**

Table 6: **The targeted method achieves gains across translation models, and HelsinkiNLP Tatoeba outperforms NLLB**. Translation is performed on English COCO, and evaluation is on Japanese STAIR.

Method	Reference Set Size	Mean Recall
FT (MT Data)	-	39.3
+Tgt Recap	4,833	41.5
+Tgt Recap	9,666	41.7
+Tgt Recap	19,332	41.8

Table 7: **The targeted method results in retrieval gains across reference set sizes**. Recaptioning here uses varying # of English COCO images (and corresponding STAIR captions) in the reference set. We report results with a reference set of 9,666 examples throughout the paper, but 4,833 examples is shown to also be effective. Evaluation is on Japanese STAIR.

A.2 Evaluation of NLLB Translation Model

We consider another translation model, No Language Left Behind (Costa-Jussà et al., 2022). Results after training with translation to Japanese are shown in Tab. 6. The results trail those of the HelsinkiNLP model, potentially a result of the HelsinkiNLP model being language-dedicated, while No Language Left Behind is focused on multilinguality. These results nonetheless show that the targeted method works well across translators.

A.3 Impact of Reference Set Size

The reference set size is a key parameter that could affect the quality of rewrites and thus retrieval performance. Our Japanese STAIR experiments in Table 1 demonstrate that a training set of size ~73k images can benefit from a smaller reference set of size ~10k images. To provide sensitivity analysis, we additionally test a smaller ~5k reference set and a larger ~20k reference set for the same training set size (~73k), sampling extra reference images and captions from previously unused COCO examples. The results are shown in Table 7. We find a smaller reference set (5k) to be effective vs. the baseline

Method	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
En Caps	0.332	0.095	0.031	0.011	0.293
+Paraphrase	0.321	0.085	0.029	0.010	0.278
+Diverse Recap	0.363	0.110	0.042	0.018	0.295
+Tgt Recap	**0.369**	**0.128**	**0.052**	**0.024**	**0.323**

Table 8: **Targeted image recaptioning is the most effective strategy in terms of captioning metrics, ROUGE-1/2/3/4/L (avg. F1 scores)**. The reference set here is Japanese STAIR train (translated to English). The rewrites, before translation to Japanese, are evaluated versus the references.

(41.5 vs. 39.3), indicating further opportunity for data efficiency with our method. We find a larger reference set (20k) to be effective vs. the baseline (41.8 vs. 39.3), though perform just slightly better than the 10k set (41.8 vs. 41.7), implying diminishing returns.

A.4 Captioning Evaluation

While our focus is on text-image retrieval, we provide an alternative evaluation of targeted image recaptioning in terms of captioning metrics. We particularly provide a small-scale captioning comparison of the original English training captions and the output captions from each rewrite strategy versus a set of references consisting of captions from the Japanese STAIR training set (which is unseen in the recaptioning process). This comparison is designed to gauge if the targeted rewrites more closely align the native Japanese captions vs. the other baselines in terms of the words and descriptions used.

Of note, the rewrites are first output from the multimodal LLM in English (see Fig. 2). With focus on measuring syntactic overlap, which can be done in English, we simply translate Japanese references to English (Google Translate). Then we score in terms of ROUGE-1/2/3/4 and ROUGE-L (avg. F1 scores). Results are shown in Table 8. We find our targeted recaptioning to be the most effective strategy across all metrics, with results indicating that the targeted rewrites most closely align the syntactic structure of the native Japanese captions. These results can inspire more directed captioning work in the future.

A.5 Term Count Comparisons

To validate that the targeted recaptioning is capturing object description properties of a target language, we compare term counts of training rewrites to native English and the native target language. Tab. 9 shows examples for Japanese, and Tab. 10 shows examples for German. The targeted method successfully accounts for terms which the other augmentation strategies and original English captions do not account for as well. For why these differences exist, future study is needed. We hypothesize that certain terms may be more salient and/or noteworthy. For Japanese, *sunglasses* are uncommon and perhaps noteworthy. For German, *formula* encompasses "Formula 1" racing, which is popular in Europe. Additionally, *zone* encompasses "pedestrian zones", which are popular in Germany.

A.6 Consideration of Hallucination

One potential concern about recaptioning is hallucination of concepts from the LLM. To mitigate potential hallucination, we encourage the LLM to reason about correct changes, where our prompt states, "Only perform changes that are correct and semantically relevant to the given input image". To conduct quality evaluation, we perform analysis for targeted image recaptioning. We examine a random sample of 200 generated captions, and find that 94.5% of the altered captions are entirely correct given the image. These results demonstrate that our method has limited negative impact from hallucination. Instead, the targeted process generally respects the details of the new image, while generalizing how objects should be described using reference captions from native speakers. Some examples are shown in Fig. 6 and Fig 7.

In caption generations with some degree of hallucination/incorrectness, the errors typically involve the addition of one object which does not exist in the given image. This happens when the object to be captioned is really small (*e.g.* "bird" is hallucinated when there is a "person" in the distance), or if there is fine-grained understanding required (*e.g.* understanding the difference between "third base stands" and "first base stands" at a baseball game). We expect that future improvements to multimodal LLMs will overcome these issues.

A.7 More Object Description Distributions

With respect to Japanese vs. English, we produce more distributions like Fig. 4 (using one of our

Method	platform	bento	futon	sunglasses	car	western	jumper	ramen
Native En	147	6	5	77	847	3	1	0
Paraphrase	251	4	0	74	370	3	2	1
Diverse Recap	162	5	5	62	818	1	0	3
Tgt Recap	309	12	15	145	1077	52	6	6
Native Ja	491	25	87	422	1379	181	36	9

Table 9: **With the targeted method, term counts get closer to the distribution of native Japanese.** Shown are counts in the training sets for each of the methods. Notice that the terms cover culture-specific naming, unique salient content, and different perspectives.

Method	rugby	motorcyclist	lectern	football	zone	blonde	shepherd	formula
Native En	7	5	1	63	1	0	10	0
Paraphrase	6	9	0	84	2	16	9	0
Diverse Recap	6	8	0	53	1	20	5	0
Tgt Recap	13	20	3	82	10	33	11	3
Native De	19	20	4	94	14	52	15	2

Table 10: **With the targeted method, term counts get closer to the distribution of native German.** Shown are counts in the training sets for each of the methods. Notice that the terms cover culture-specific naming, object grouping, and unique salient content.

Japanese STAIR and English COCO train sets). Fig. 8 shows *conveyance* and *animal*, while Fig. 9 shows results for *person* and *container*. There are some similarities in counts (*e.g.* the distributions for *animal* are similar in Figure 8). There are also notable differences. In Fig. 8, *jet* is used more often in the En set, while *plane* is used more frequently in the Ja set. In Fig. 9, *locomotive* and *car* are used much more frequently in the Ja set, while *truck* is used more in the En set. Also in Fig. 9, *man* and *woman* are described more in the Ja set, while the En set uses *people* more.

Fig. 10-12 also show comparisons for German vs. English, using one of the original speaker train sets from each Multi30k and Flickr30k. In general, there is more similarity between English and German, which is intuitive considering the proximity of the languages. However, some uniqueness exists. For example for the supercategory *person*, *cyclist* occurs ~4× more often in German than English, through *cowboy* occurs ~3× more often in English. For *device*, *stand* is ~1.5× more frequent in German. Then for the supercategory *container*, *bike* is mentioned ~2.5× less often in German.

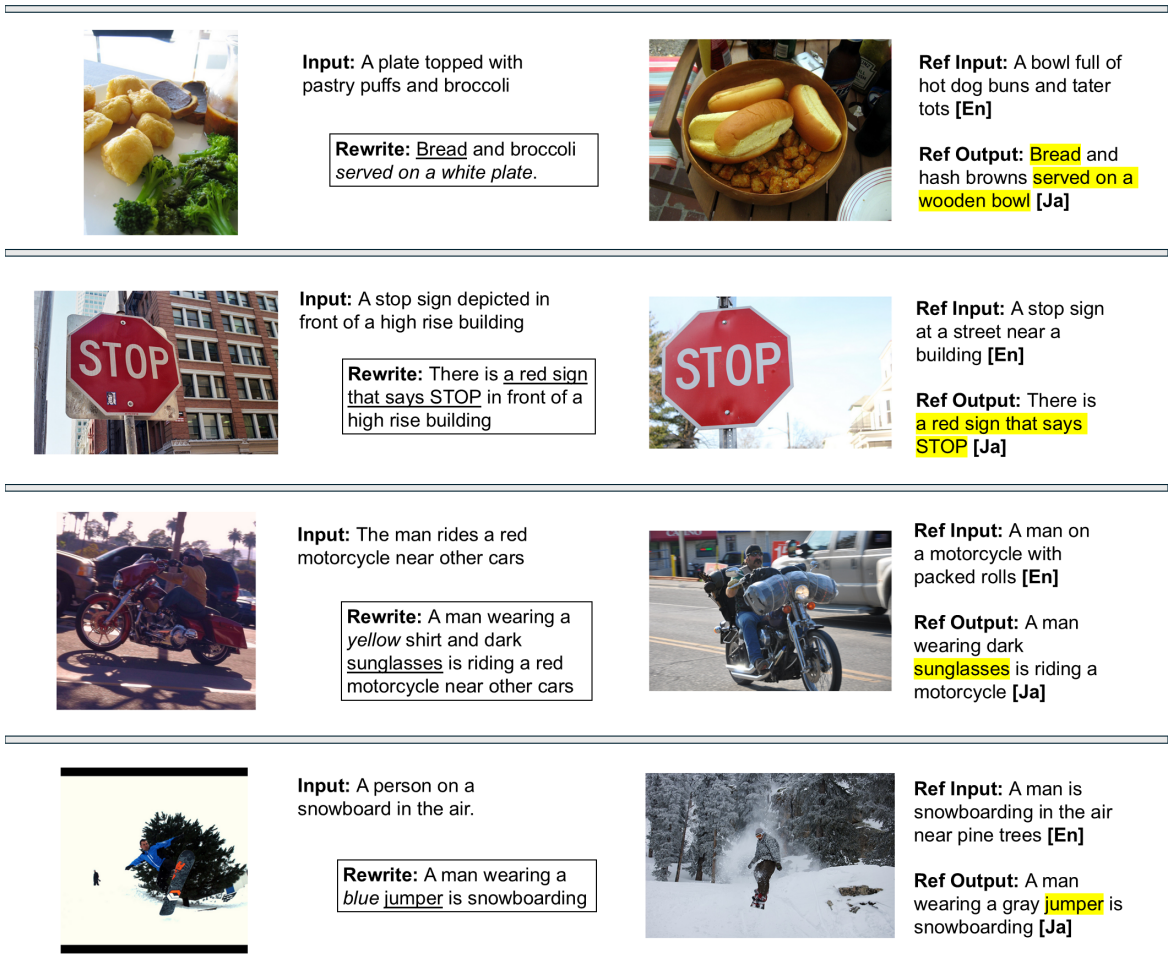

Figure 6: **Example inputs, nearest-neighbor reference images/captions, and rewrites produced with our targeted image recaptioning (Japanese).** Observe how the model can leverage text in the outputs of the reference (*e.g.* *bread*, *red sign that says STOP*, *sunglasses*, and *jumper*), while inferring relevant details for the input images (*e.g.* the fact that the *jumper* is *blue*). The language in brackets is the language in which the caption was produced.

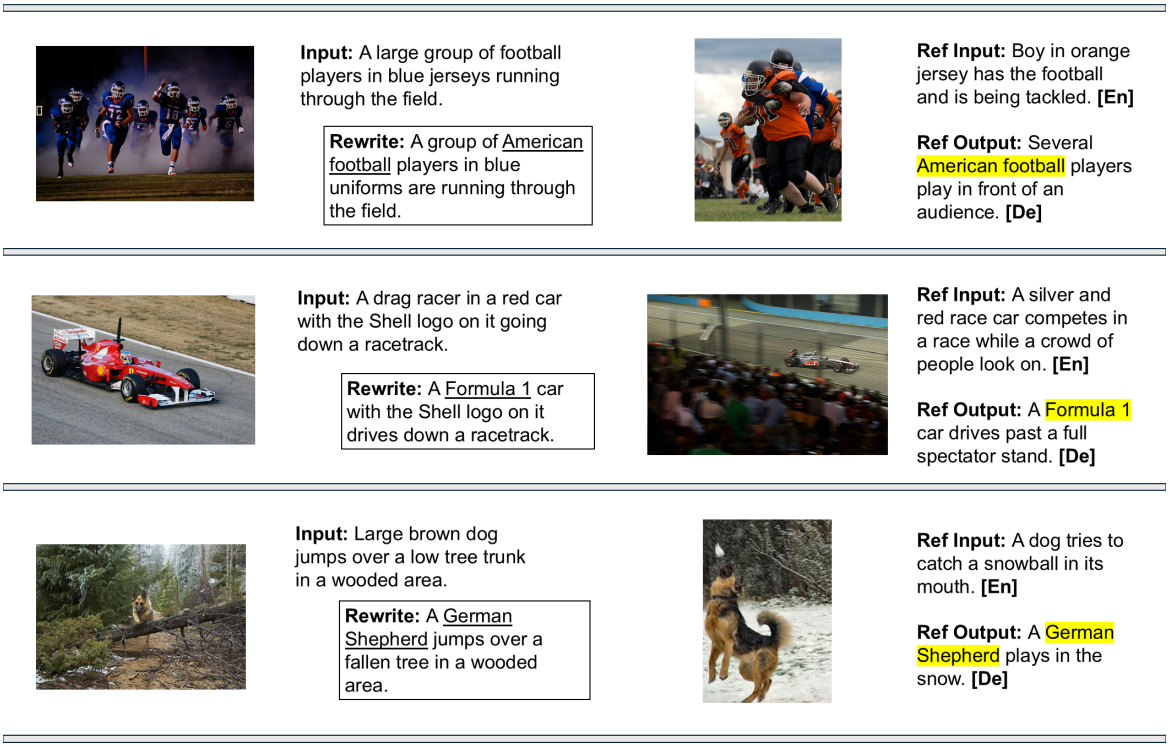

Figure 7: **Example inputs, nearest-neighbor reference images/captions, and rewrites produced with our targeted image recaptioning (German).** Observe how the model can leverage text in the outputs of the reference (*e.g. American football*, *Formula 1*, and *German Shepherd*), while inferring relevant details for the input images (*e.g.* jumping over a *fallen tree*). The language in brackets is the language in which the caption was produced.

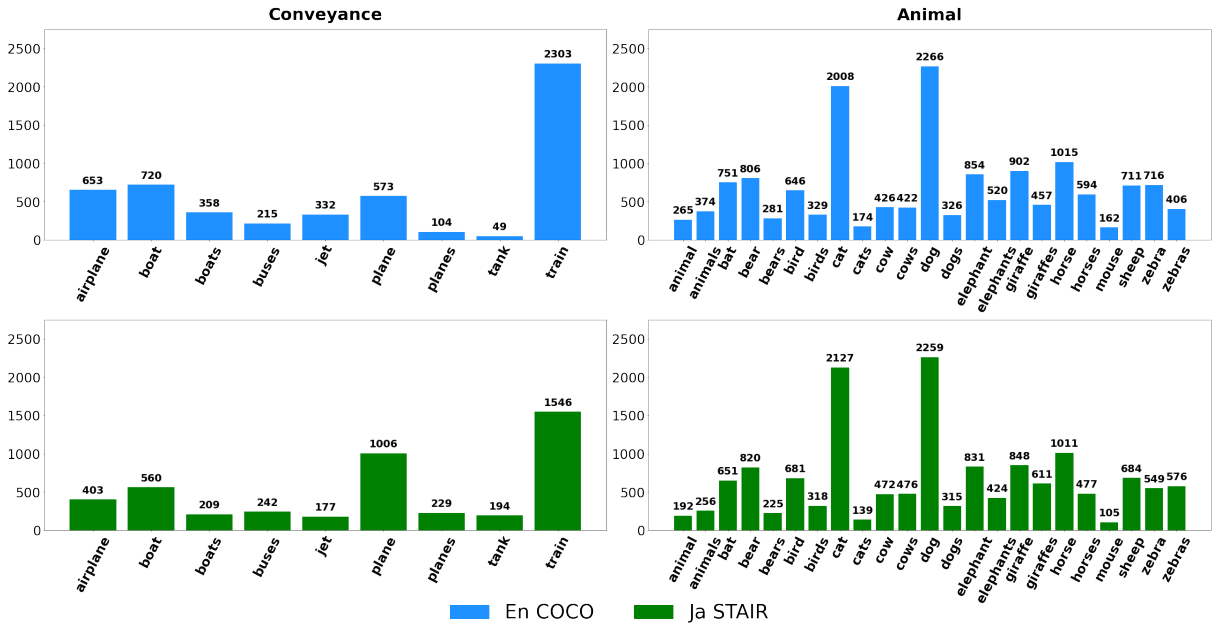

Figure 8: **Term distributions for *conveyance* and *animal*, English COCO vs. Japanese STAIR.** For each supercategory, any term with count > 150 is identified, and the union of terms across languages is shown.

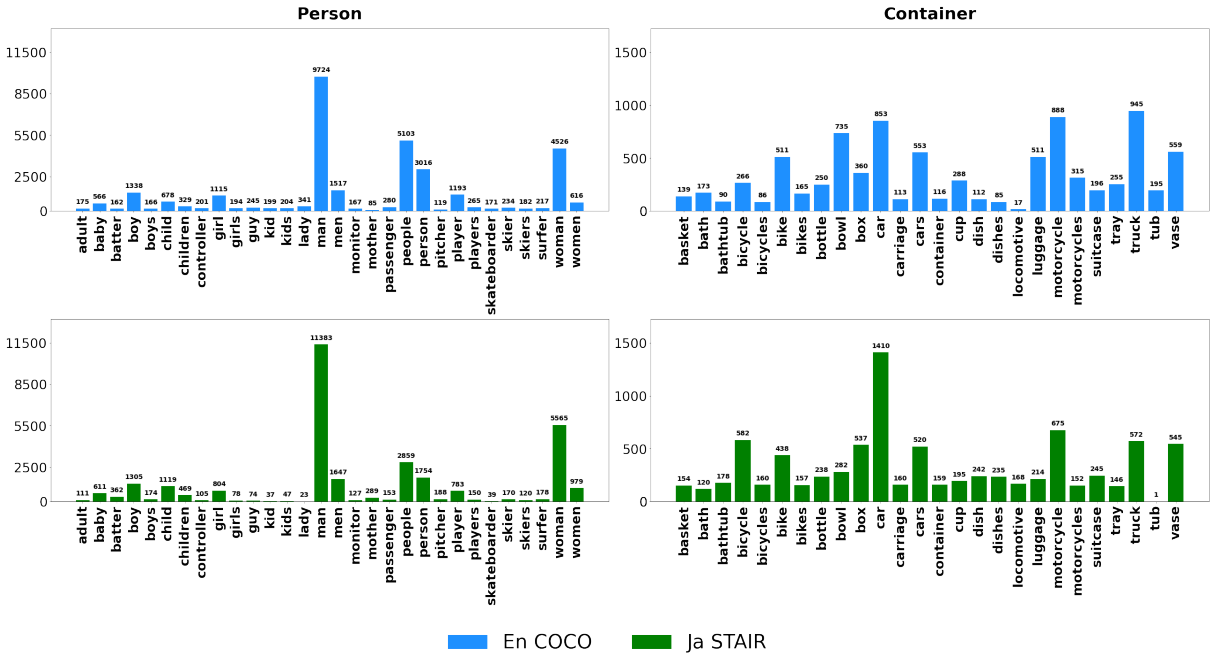

Figure 9: **Term distributions for *person* and *container*, English COCO vs. Japanese STAIR.** For each supercategory, any term with count $>$ 150 is identified, and the union of terms across languages is shown.

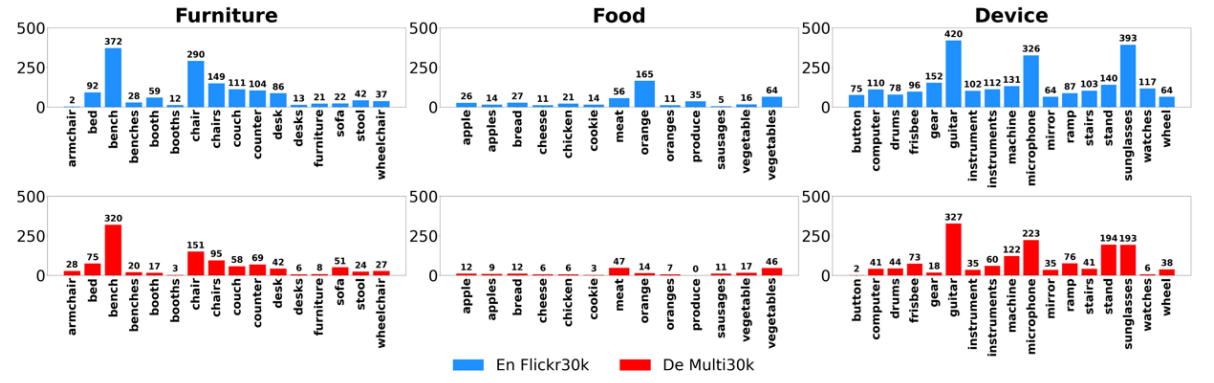

Figure 10: **Term distributions for *furniture*, *food*, and *animal*, English Flickr30k vs. German Multi30k.** For *furniture*, any term with count $>$ 10 is identified. For *food*, any term with count $>$ 10 is identified. For *device*, any term with count $>$ 60 is identified. Then the union of terms across languages is shown.

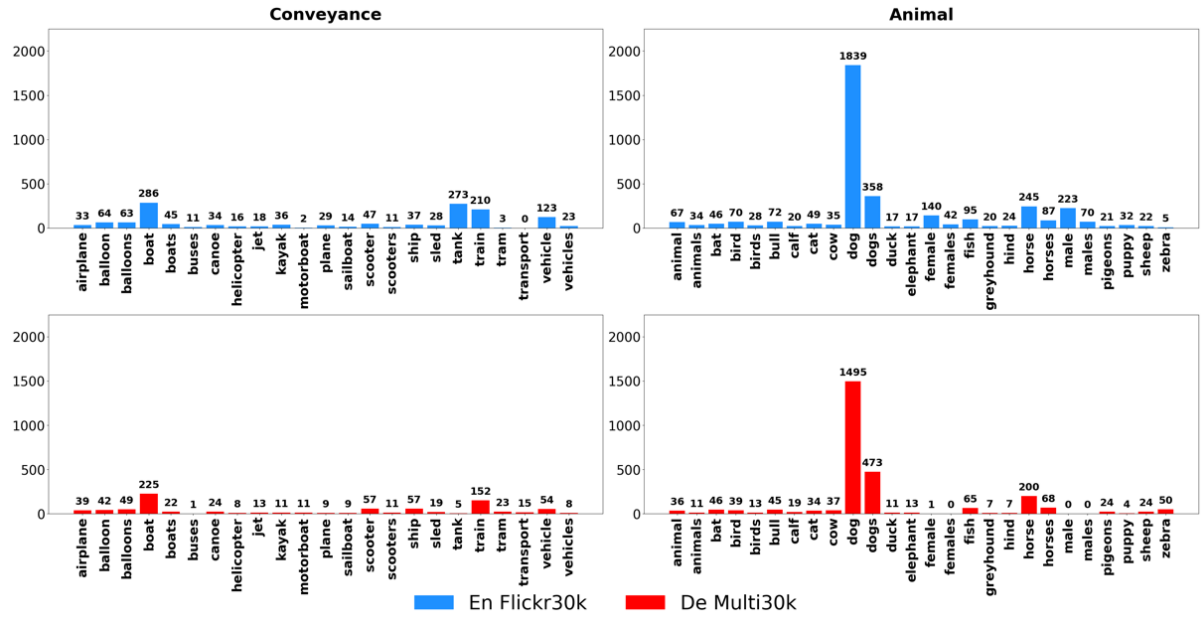

Figure 11: **Term distributions for *conveyance* and *animal*, English Flickr30k vs. German Multi30k.** For *conveyance*, any term with count > 10 is identified. For *animal*, any term with count > 15 is identified. Then the union of terms across languages is shown.

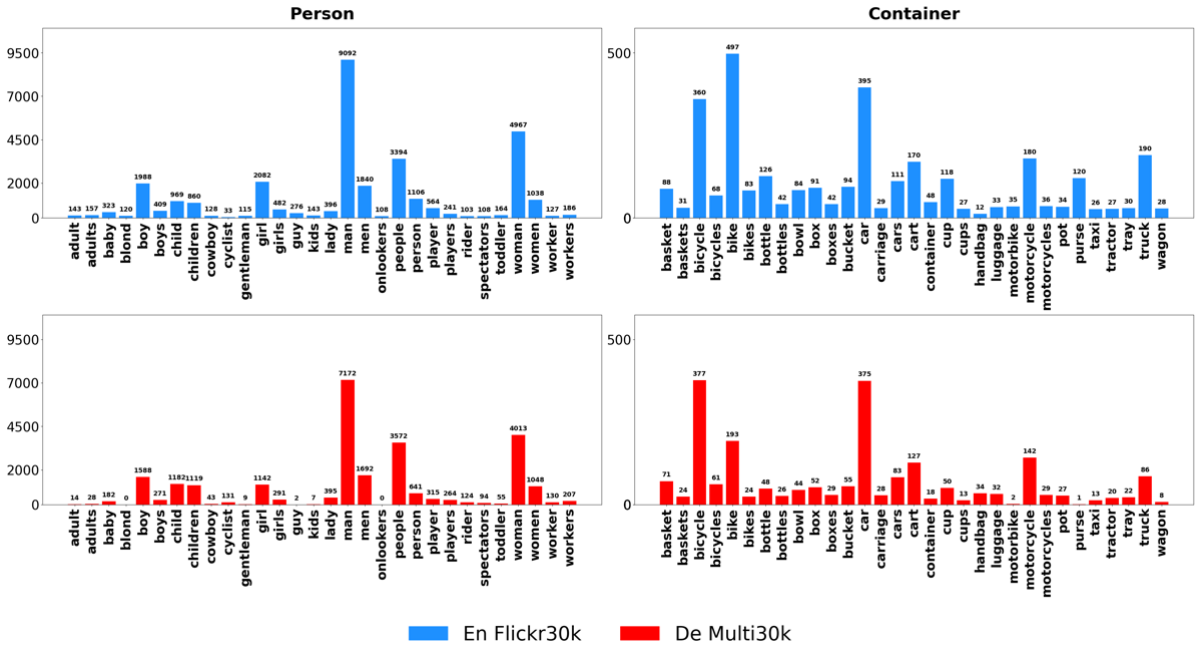

Figure 12: **Term distributions for *person* and *container*, English Flickr30k vs. German Multi30k.** For *person*, any term with count > 100 is identified. For *container*, any term with count > 25 is identified. Then the union of terms across languages is shown.