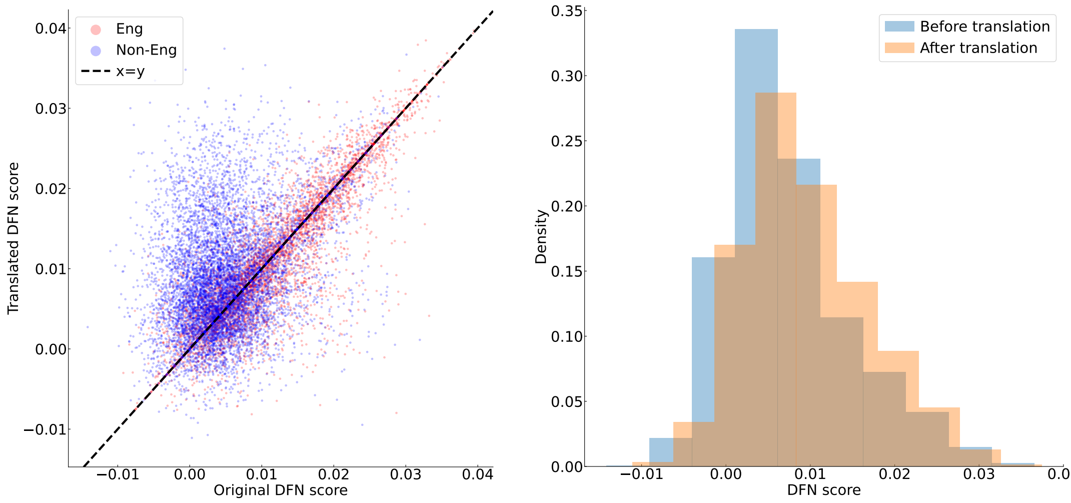

Figure 1: **Translation increases the image-text similarity scores of non-English samples, thus also improving the average scores of the data pool.** We randomly sample 10K images with English captions and 10K with non-English captions from the initial data pool, and compare how the image-text DFN scores change with translation. (Left) Unsurprisingly, DFN scores for non-English samples generally increase after the captions are translated into English. (Right) The overall score distribution observes a right shift as a result of translation. Since image-text alignment (measured by DFN score) tends to correlate with empirical performance, this shift suggests that translation helps increase the availability of beneficial training data.

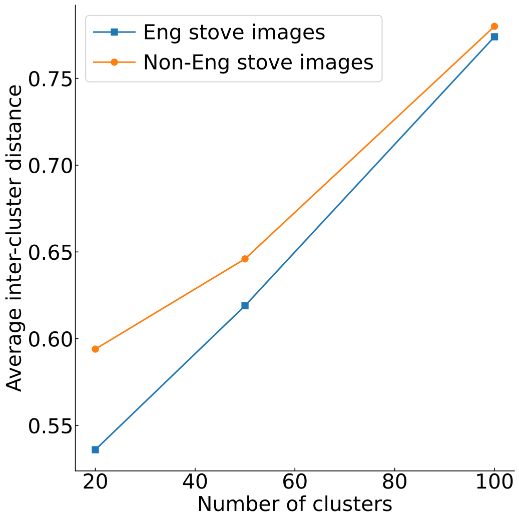

Figure 2: **For the same concept (e.g. "stove"), images from the non-English data distribution show greater visual distinction compared to those from the English distribution.** We look for raw English captions and English-translated captions that mention "stove", and randomly pick 1K corresponding images from each distribution. We then embed the images with the DINOv2 model and cluster the embeddings. Across different number of clusters uncovered, we observe that non-English data generally yields higher inter-cluster distance, suggesting that the "stove" images with multilingual captions are more heterogeneous than those with English captions.