Translation Deserves Better: Analyzing Translation Artifacts in Cross-lingual Visual Question Answering

Anonymous ACL submission

Abstract

Building a reliable visual question answering (VQA) system across different languages is a challenging problem, primarily due to the lack of abundant samples for training. To address this challenge, recent studies have employed machine translation systems for the cross-lingual VQA task. This involves translating the evaluation samples into a source language (usually English) and using monolingual models (i.e., *translate-test*). However, our analysis reveals that translated texts contain unique characteristics distinct from human-written ones, referred to as *translation artifacts*. We find that these artifacts can significantly affect the models, confirmed by extensive experiments across diverse models, languages, and translation processes. In light of this, we present a simple data augmentation strategy that can alleviate the adverse impacts of translation artifacts.

1 Introduction

Visual question answering (VQA) aims to answer an open-ended question by reasoning about a given image (Agrawal et al., 2015). Despite recent advances in vision-language (VL) modeling, building proficient models across various languages is still challenging. This issue primarily arises from the limited availability of annotated datasets, which are predominantly in high-resource languages such as English. Although recent efforts in developing multilingual VL models can address this issue to some extent (Zhou et al., 2021; Qiu et al., 2022; Li et al., 2023b; Geigle et al., 2023), training on datasets in the target languages is still crucial for enhanced model performance in those languages.

To mitigate the data scarcity issue, cross-lingual transfer learning focuses on utilizing extensive datasets in a *source* language (typically English) to build models effective in a *target* language (Artetxe et al., 2020; Bugliarello et al., 2022). One of the popular approaches, namely *translate-train*, translates training samples into individual target languages and uses them to train models for target languages. This approach is advantageous as it does not perform translation during inference, but it requires training individual models for each target language. Furthermore, recent VL models (Singh et al., 2022; Liu et al., 2023b; Li et al., 2023a), which are mostly tailored in English, are not suitable for the translate-train approach. Another widely adopted approach, called *translate-test*, translates test samples written in target languages into the source language and uses VL models of the source language for the inference. These translation-based approaches have shown remarkable performance in cross-lingual tasks.

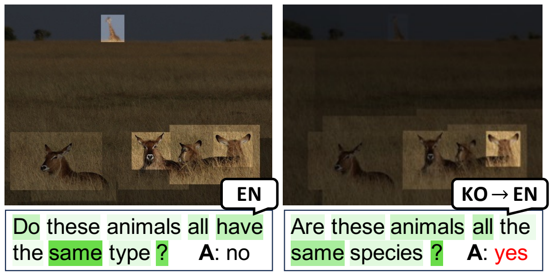


Figure 1: Predictions of LXMERT (Tan and Bansal, 2019) on the **original** (left) and **translated** (right) questions. The model is correct for the human-written question but is incorrect for the correctly translated one. The original Korean question is *"이 동물들은 모두 같은 종입니까?"*. For model visualization, we use an attention-based method by Chefer et al. (2021).

Despite the effectiveness of translation systems in cross-lingual VL tasks, using machine-translated texts as input inevitably introduces a mismatch between the training and inference phases. In the translate-test approach, models are trained on human-written texts but evaluated on machine-translated texts. This distribution shift could hurt the generalization of models to different languages (Yu et al., 2022; Wang et al., 2022). For instance, as illustrated in Fig. 1, leveraging machine-translated texts might lead to undesirable model outcomes, even when both questions convey the same meaning. In this paper, we refer to artifacts in translations that cause such unwanted behaviors as *translation artifacts*. We argue that the translation artifacts have been overlooked in previous cross-lingual VQA studies despite their significance.

To explore the effect of mismatched data distribution on cross-lingual VQA, we alleviate this mismatch in the data origins[1] by employing machine-translated texts in both training and inference. Our investigation focuses on the translate-test, which can take advantage of strong monolingual models and efficiently serve multiple target languages with a single VL model. Our results reveal that models trained on machine-translated texts generally outperform those trained on human-written texts, increasing the averaged accuracy over languages and models from 51.82 to 53.14 points. This improvement, as confirmed by our qualitative analysis, is primarily attributed to the subtle nuances in translated texts (*i.e.*, translation artifacts). Our comprehensive study covers various components in cross-lingual VQA, including 14 models, 13 languages, 5 machine translation systems, and diverse translation setups. We also observe that recent VL models (Li et al., 2023a; Dai et al., 2023) integrated with large language models also suffer from translation artifacts. Finally, we present simple data augmentation techniques, verifying their effectiveness in both human-written and translated texts.

Our contribution can be summarized as follows:

1. This is, to our knowledge, the first study to investigate translation artifacts in cross-lingual visual question answering.

2. We provide extensive analyses across a variety of languages and models, providing a foundation for future research.

3. We present simple yet effective data augmentation strategies using translated texts.

2 Related Work

2.1 Cross-lingual VQA

The study of VQA has predominantly focused on English and other high-resource languages (Zhu et al., 2015; Agrawal et al., 2015; Goyal et al., 2016; Marino et al., 2019; Schwenk et al., 2022). To extend the use of VQA to various languages, researchers have introduced cross-lingual transfer techniques (Ni et al., 2021; Zhou et al., 2021; Nooralahzadeh and Sennrich, 2022; Liu et al., 2023a). One effective approach involves pretraining VL models on multilingual image-text pairs and then fine-tuning them on English VQA, which is known as *zero-shot* transfer (Jain et al., 2021; Lee et al., 2022; Zeng et al., 2022; Chen et al., 2022, 2023; Li et al., 2023b). Another popular approach that leverages advanced machine translation shows promise in adapting to various languages. The *translate-train* involves translating the text pairs from high-resource languages to the target language for finetuning (Thapliyal and Soricut, 2020; Zeng et al., 2022; Chen et al., 2023; Li et al., 2023b). On the other hand, the *translate-test* uses machine translation to convert test data into English, allowing the use of English-only models for inference (Jain et al., 2021; Bugliarello et al., 2022; Pfeiffer et al., 2022). This latter approach is particularly beneficial, considering the strong performance of existing English-only models (Singh et al., 2022; Li et al., 2023a; Dai et al., 2023; Gao et al., 2023).

2.2 Translation Artifacts

Translated texts often exhibit unique characteristics, referred to as *translation artifacts* or *translationese* (Gellerstam, 1986; Lembersky et al., 2012; Baker, 2019; Edunov et al., 2020). These characteristics can negatively influence model outcomes due to their stylistic deviations from the original texts (Volansky et al., 2015; Bizzoni et al., 2020; Yu et al., 2022). Yang et al. (2021) examined the representation discrepancies between English and other languages in the translate-train approach for various language understanding tasks. Wang et al. (2022) explored the effects of translation artifacts on model evaluation in cross-lingual summarization. To mitigate the effects of translation artifacts, researchers have proposed various methods, such as incorporating machine-translated sentences in training (Artetxe et al., 2020; Yu et al., 2022; Wang et al., 2022) or utilizing specific tags to differentiate between original and machine-translated texts (Marie et al., 2020; Riley et al., 2020; Wang et al., 2021).

However, the effect of translation artifacts on cross-lingual VQA remains largely underexplored, leading to potential risks and unexpected outcomes. While previous research has primarily focused on the application of machine translation in VL mod-

[1]We refer to *origin* as a writer of texts (*i.e.*, human or machine translation system).

els (Thapliyal and Soricut, 2020; Zeng et al., 2022; Bugliarello et al., 2022; Pfeiffer et al., 2022; Changpinyo et al., 2023; Chen et al., 2023), our study aims to identify the presence and impact of translation artifacts within cross-lingual VQA. We find that these translation artifacts are prevalent in VL models handling both image and text modalities.

3 Translation Artifacts in Cross-lingual Visual Question Answering

In this work, we analyze the impact of machine translation on cross-lingual VQA tasks, especially on the translate-test approach. To this end, we vary the origin of training datasets into human and a machine translation (MT) system and then observe how this change affects the model behavior. We use roundtrip (RT) translation to generate machine-translated training samples from the source language- English.[2]

3.1 Experimental Setup

3.1.1 Data

We use xGQA (Pfeiffer et al., 2022), a representative benchmark for the cross-lingual VQA task. Each sample in the dataset consists of an image, a structured question related to the image, and an answer. The training set is derived from the original English GQA dataset (Hudson and Manning, 2019) and consists of 72k images and 943k samples. The evaluation sets cover seven different languages - Bengali (bn), German (de), Indonesian (id), Korean (ko), Mandarian (zh), Portuguese (pt), and Russian (ru) - and is manually translated from the balanced test-dev set of the English GQA dataset by human annotators. The evaluation set consists of 398 images and 12,578 samples, and all images in xGQA datasets are sampled from the Visual Genome dataset (Krishna et al., 2017). Further details on the dataset are described in Appendix A.

3.1.2 Models

We conduct experiments with all multilingual and monolingual VL models addressed in Bugliarello et al. (2022). Specifically, for multilingual models, MUNITER (Qiu et al., 2022), XUNITER (Qiu et al., 2022), UC^2 (Zhou et al., 2021), and M^{3}P (Ni et al., 2021) are used. For monolingual English-only models, LXMERT (Tan and Bansal, 2019), UNITER (Chen et al., 2020), VILBERT (Lu et al., 2019), VisualBERT (Li et al., 2020), and VL-BERT (Su et al., 2019) are used. All models are based on transformer (Vaswani et al., 2017) architecture, and both image and text are fed to the network simultaneously. In addition, we conduct experiments with recently proposed monolingual English VL models - BLIP-2 (Li et al., 2023a), InstructBLIP (Dai et al., 2023), and FLAVA (Singh et al., 2022). More details are in Appendix B.

For the cross-lingual transfer of multilingual models, the following approaches are considered: zero-shot, translate-train, and translate-test. The zero-shot approach trains a model on the original English training set in the GQA dataset and directly uses it to infer evaluation samples in the target language.[3] The translate-train approach trains individual models for each target language on a translated training dataset. The translate-test approach trains a single model on an English training dataset and uses it for the evaluation of target languages along with a translation system. For monolingual models, only the translate-test approach is evaluated.

3.1.3 Training Dataset from Different Origins

For the translate-test, we finetune all models described above on English GQA datasets from two different origins individually: Human and MT. For Human, we use the original xGQA training set. For MT, we use the roundtrip (RT) translation to obtain training samples that are written by an MT system. We use NLLB (Costa-jussà et al., 2022) as the MT system for RT translation.[4] The German (de) is used as a pivot language during RT translation (en→de→en). For the zero-shot and translate-train, we use the original English dataset and the dataset translated from English to individual target languages, respectively. More details about translation processes are in Appendix C.

3.1.4 Evaluation dataset

Source Language For English evaluation, we use the official evaluation set released by Pfeiffer et al. (2022) (en). Besides, we also make translated versions of English evaluation sets through RT translation (en*). This process is to understand the impact of data origins on models more comprehensively.

Target Languages For zero-shot and translate-train evaluations, the target language questions released by Pfeiffer et al. (2022) are used. For the

[2] Afterward, the *source* language refers to English.

[3] The *zero-shot* denotes that the language of evaluation samples differs from the language used in the finetuning phase.

[4] `facebook/nllb-200-3.3B` is used.

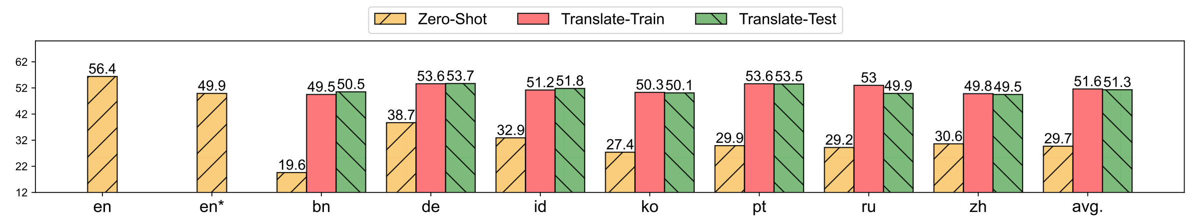


Figure 2: **Averaged multilingual models results** The *en** and *avg.* denote the RT-translated English evaluation set and the averaged cross-lingual transfer results, respectively. Full results of each multilingual model are in Fig. 9.

Models	**RT?**	en	en*	bn	de	id	ko	pt	ru	zh	**avg.**
MUNITER		57.33	50.14	50.67	54.09	52.54	50.67	54.21	49.69	49.57	51.63
	✓	55.70	52.75	52.34	55.66	53.48	53.36	54.72	53.98	52.29	53.69
XUNITER		56.98	49.90	50.76	54.63	52.37	50.52	54.24	48.91	49.94	51.62
	✓	55.22	52.45	52.10	54.97	52.66	52.51	54.18	52.85	52.23	53.07
UC^2		56.85	50.22	51.34	54.01	52.35	50.75	53.81	51.93	50.04	52.03
	✓	55.12	52.44	52.35	55.10	53.29	53.07	54.17	53.36	52.73	53.44
M^3P		54.45	49.29	49.18	52.14	49.87	48.59	51.87	49.05	48.38	49.87
	✓	52.97	51.97	50.63	53.03	51.42	50.38	52.11	51.80	50.41	51.40
LXMERT		55.40	48.42	49.64	52.83	50.80	49.17	52.49	47.54	48.02	50.07
	✓	53.44	50.51	50.20	52.93	51.34	50.41	52.47	51.44	50.25	51.29
UNITER		57.47	50.11	51.74	54.52	52.79	51.27	54.56	52.27	50.33	52.50
	✓	55.92	52.90	52.32	55.53	53.67	52.93	54.66	53.56	52.60	53.61
VILBERT		56.72	50.10	50.84	54.10	52.27	50.73	53.98	49.91	49.92	51.68
	✓	55.22	52.52	52.23	54.85	53.43	52.75	54.26	53.69	52.22	53.35
VisualBERT		55.17	48.66	49.43	52.58	50.34	48.66	52.72	50.50	48.89	50.45
	✓	53.51	50.91	50.57	53.10	51.17	50.45	52.59	51.47	50.97	51.47
VL-BERT		57.79	50.32	51.22	54.47	52.62	50.94	54.79	51.17	50.02	52.18
	✓	55.61	52.79	52.38	55.27	53.43	52.58	54.63	53.32	52.31	53.42
BLIP-2		58.05	52.10	52.03	54.70	52.99	51.57	54.91	52.36	51.22	52.83
	✓	56.11	54.76	53.18	55.70	53.98	53.51	55.11	54.25	53.31	54.15
InstructBLIP		57.85	52.26	51.80	54.91	53.01	51.29	54.85	53.16	51.34	52.91
	✓	55.84	54.62	53.04	55.06	53.82	53.17	54.32	54.08	53.18	53.81
FLAVA		**58.84**	52.91	53.47	56.26	54.11	52.85	55.84	53.64	52.18	54.05
	✓	56.87	**55.07**	**53.94**	**56.35**	**54.99**	**54.51**	**55.96**	**55.61**	**53.82**	**55.03**
avg.		56.91	50.37	51.01	54.10	52.17	50.58	54.02	50.84	49.99	51.82
	✓	55.13	52.81	52.11	54.80	53.06	52.47	54.10	53.28	52.19	53.14

Table 1: **Translate-test results with different origins of training dataset** For languages other than English, we use an evaluation set released by Bugliarello et al. (2022) translated with Google Machine Translation (GMT). Here, *en** denotes the RT-translated English evaluation set. Models finetuned on RT-translated English texts are marked with ✓. For each model within the different data origins, the higher score in each column is highlighted in underline. The highest score in each column is further highlighted in **bold**. The statistical significance analysis is in Appendix E.

translate-test evaluation, each question in the target language should be translated into English. In this work, we use an official translate-test evaluation set (Bugliarello et al., 2022) generated by the Google Machine Translation (GMT) system.

3.1.5 Implementation Details

For finetuning VL models, we follow hyperparameters reported in Bugliarello et al. (2022) for a fair comparison. Specifically, all models are trained for 5 epochs, and the batch size and initial learning rate are set to 256 and 4e-5, respectively. AdamW (Loshchilov and Hutter, 2018) is used for optimization. All models are trained with a classification head on top of image-language representation. We evaluate models after every training epoch and choose the best checkpoint based on its accuracy on the original English development set. More implementation details are in Appendix D.

4 Results and Analysis

4.1 Main Results

Multilingual Models Fig. 2 presents averaged evaluation results of multilingual models with different cross-lingual transfer approaches. The models show decreased accuracy when transferred to languages other than English. For instance, the average accuracy is 56.4 for the original English dataset, but are 51.6 and 51.3 for translate-train and translate-test approaches, respectively. Among the different cross-lingual transfer approaches, translate-train and translate-test are comparable, while the zero-shot approach usually performs worse.

Misaligned Data Origins in Translate-Test Table 1 presents translate-test evaluation results of models with different training data origins. Regarding models trained on human texts, FLAVA usually performs better than other models. Regarding the effects of different training data origins, we observe that models generally show higher accuracy when the origins of training and evaluation datasets are matched. Specifically, for the original English evaluation set, models trained on human texts consistently perform better than ones trained on MT texts. On the contrary, for the translate-test, in which all questions are generated by MT systems, models trained on MT texts outperform those trained on human texts. By only aligning the data origins of training and evaluation sets, the averaged translate-test scores across models and languages are increased from 51.82 to 53.14. Note that this trend is consistent in RT-translated English evaluation set (en*), where the average score increases from 50.37 to 52.81. Based on our results, we suggest a reconsideration of factors contributing to lower scores of target languages in cross-lingual VQA, indicating that data origin misalignment, alongside translation errors, could significantly impact the success of language transfer.

4.2 Human Analysis

We next analyze translated questions in the evaluation set to examine where the increased performance of models trained on MT texts comes from. To this end, we annotate translate-test evaluation samples in which a model trained on human texts makes wrong predictions, but a model trained on MT texts makes correct ones. Note that we only consider the translate-test samples in which both models with human and MT texts correctly predicted the paired human-written English samples to avoid wrong predictions arising from sample complexity. UC^2 is selected as a VQA model, and 200 questions from the Korean (ko) evaluation set are annotated. Two native speakers annotate the MT errors in each question, and the annotation schema is based on multidimensional quality metric (MQM) ontology (Mariana, 2014) following Moghe et al. (2023). More details about human annotation and annotated examples are in Appendix G.

As shown in Fig. 3, although the model trained on human texts changes its prediction from the correct to wrong ones, a majority of translated questions (>60%) do not contain crucial translation errors. In terms of translated questions without

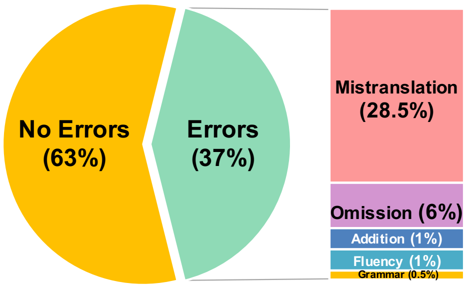


Figure 3: A distribution of different translation errors in sampled questions from Korean translate-test set.

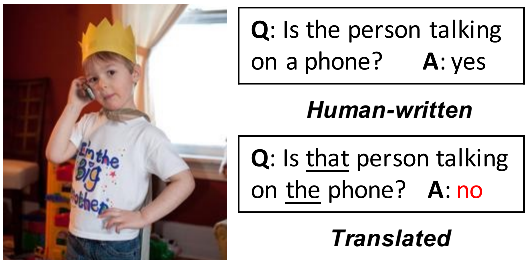


Figure 4: A model is accurate for the original human-written question, but fails for a translated one. The Original Korean question is *"그 사람이 전화 통화를 하고 있습니까?"*. Further annotation results are in Fig. 11.

translation errors, most of them can be regarded as paraphrased sentences of their paired English questions as shown in Fig. 4. Based on these results, we confirm that models trained on human texts often make wrong predictions about translations that convey similar meanings to human ones. In other words, subtle differences between human and translated texts caused by translation artifacts indeed influence model behavior.

4.3 Are Translation Artifacts Actually Presented in Translated Questions?

To enhance our comprehension of the increased translate-test accuracy of models trained on MT texts, we scrutinize model performances across samples categorized by the prevalence of translation artifacts. Specifically, we quantify the *human-likeness* of each translated question and assess its influence on VQA models. To derive the human-likeness score $p_h(x)$ for every translated question x, we train a classifier based on RoBERTa (Liu et al., 2019), designed to discern whether a given English question is written by a human or generated by round-trip (RT) translation.[5] After training, the classifier assigns a score $p_h(x)$ on how likely humans write each translate-test evaluation sample. We

[5]The accuracy of the trained classifier is 86.97 in a class-balanced validation set.

Train \ Test	NMT-like	Human-like
Human	48.60	53.56
NMT	51.98	53.85

Table 2: Averaged translate-test accuracy of VL-BERT models trained on different data origins. Each column denotes the group of translate-test evaluation samples.

Metric \ Test	NMT-like	Human-like
TTR	92.52	95.14
LD	48.44	49.76

Table 3: Lexical diversity results of translate-test evaluation samples. TTR and LD denote the token-type ratio and lexical density, respectively.

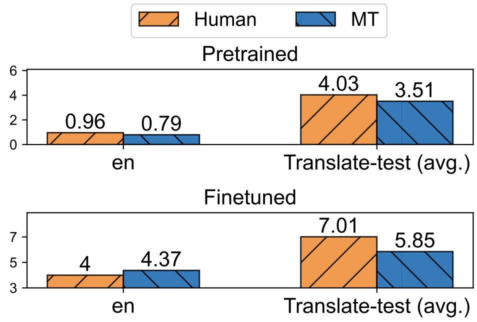


Figure 5: Representation discrepancy of translate-test evaluation samples against training samples from different data origins (Human and MT). Pretrained or finetuned VisualBERT is used to encode representation, and FID is used as a distance metric. A lower score indicates a low distance between training and evaluation samples. Full results across different languages are in Fig. 10.

then categorize the translated evaluation samples into two groups with the same size - *human-like* and *NMT-like* - based on their respective human-likeness scores $p_h(x)$. The accuracy of VL-BERT models trained on different data origins (*i.e.*, human and NMT) is compared in these groups. More experimental details are in Appendix H.

From the results in Table 2, we find that the model trained on human texts performs worse when the input questions are less likely to be written by humans. Specifically, the average accuracy across different target languages is 53.56 for human-like questions but 48.60 for NMT-like questions. Conversely, a model trained on MT texts shows less accuracy degradation on NMT-like questions compared to the one trained on human texts. These results indicate that VQA models are prone to make more errors when the given question is not likely to be written by humans, and training on (RT-) translated texts can alleviate such problems.

We next delve into the lexical diversity within each question group, inspired by previous findings that the translated texts are often simpler than human ones (Zhang and Toral, 2019; Wang et al., 2022). Specifically, we use two metrics to measure the lexical diversity of translated questions used in the translate-test approach: (1) *Token Type Ratio (TTR)* calculates the ratio of unique words over all words in the sentence, (2) *Lexical Density (LD)* calculates the ratio of content words (words that likely to convey significant meaning - nouns, verbs, adverbs, and adjectives) over all words in the sentence. Lexical diversity results in Table 3 indicate that NMT-like questions generally exhibit less variety in word usage. We suspect that such characteristics in translated questions make a difference in training and evaluation, resulting in performance degradation of models.

4.4 Representation Analysis

Our previous observations reveal that translated texts exhibit distinct impacts compared to human ones when they are used for training and evaluation. We next analyze whether these different characteristics of translated texts appear in model representation. Specifically, we compare the representations of training samples from different origins (human and MT) against evaluation samples. As evaluation samples, we use the translate-test samples from different target languages and English evaluation samples written by human or RT translations. We employ the penultimate layer output of visualBERT as the sample representation, and the Fréchet Inception Distance (FID) (Heusel et al., 2017) score is used to quantify the representation distance between training and evaluation samples. Additionally, to assess the impact of finetuning on model representation, we analyze VisualBERT at checkpoints before and after finetuning.

As shown in Fig. 5, we observe clear trends indicating that translated samples cluster more closely in the model representation space. In detail, all translate-test samples show lower FID scores with MT training samples than human ones. Note that these trends are consistent for both pretrained and finetuned models. These results indicate that characteristics shared within translated texts also affect the internal representation of VL models.

RT Pivot	en	*Translate-Test*							
		bn	de	id	ko	pt	ru	zh	**avg.**
bn	53.93	52.46	53.18	52.34	50.93	52.82	52.54	50.31	52.08
de	55.13	**52.47**	**54.77**	52.68	51.05	54.10	53.31	49.72	52.59
id	54.64	52.45	53.53	**53.57**	51.42	53.60	53.11	**50.50**	52.60
ko	53.62	51.63	52.73	51.97	**51.78**	52.72	52.02	50.41	51.89
pt	**55.64**	52.45	54.53	52.82	50.96	**55.02**	53.57	49.52	**52.69**
ru	54.94	52.42	54.10	52.98	51.15	53.87	**54.00**	50.31	**52.69**
zh	51.56	48.64	49.55	49.00	48.20	49.46	48.88	48.43	48.88

Table 4: Evaluation results of models trained on RT translation with different pivot languages. Each row indicates the pivot language used in RT translation, and scores of all models with the same pivot languages are averaged. The highest scores in each column are highlighted in **bold**. Full results of all pivot languages and models are in Table 17.

4.5 Varying NMT and Pivot Languages

Based on our previous results, we confirm that addressing the misalignment of data origins between training and evaluation is effective for the translate-test approach. We now aim to understand how these benefits vary with changes in the MT systems or translation setups. To this end, we conduct experiments by varying (1) the MT system used for translating the training and evaluation sets and (2) the pivot language during the RT translation.

Varied MT systems We use the following four MT systems in our experiments: M2M-100-418M/1.2B (Fan et al., 2021) and NLLB-200-600M/-3.3B (Costa-jussà et al., 2022). Each MT system is used to make RT-translated training and translate-test evaluation sets. In detail, we use RT translation with different MT systems to make training sets, and the pivot language is fixed to German (de). All models described in Section 3.1.2 are individually trained on these four RT-translated datasets. For the evaluation set, we translate every target language into English using different MT systems, resulting in four different evaluation sets.

Evaluation results are shown in Table 5. Notably, models trained on translated texts usually outperform those trained on human texts in translate-test sets. These results suggest that, despite a mismatch between the MT systems used for RT translation and the translate-test, leveraging RT translation for training remains advantageous for cross-lingual transfer. In terms of MT system comparison, models usually show higher accuracy when MT systems used to make training and evaluation sets are in the same model family. In the original English evaluation set, models with human texts perform best, followed by the ones with NLLB-200-3.3B texts.[6]

Varied Pivot Language in RT Translation We vary the pivot languages used in RT translation to make different versions of translated training sets. All target languages presented in xGQA datasets are selected as pivot languages. As an MT system, we use NLLB-200-3.3B to translate both training and evaluation samples. As shown in Table 4, models usually show higher accuracy when a pivot language matches its target language. This tendency is consistent with previous findings (Ni et al., 2022), where the texts within the same translation direction contain shared characteristics.

Train \ Test	GMT	M2M-small	M2M-large	NLLB-small	NLLB-large
Human	51.92	45.64	47.79	49.53	50.62
M2M-S	51.25	48.85	49.39	50.06	50.55
M2M-L	52.31	**49.37**	**50.43**	50.94	51.48
NLLB-S	52.75	48.73	50.00	51.39	52.04
NLLB-L	**53.18**	48.59	50.04	**51.65**	**52.52**

Table 5: Translate-test evaluation results with different MT systems to make RT-translated training and translate-test evaluation examples. Each row and column denote the origin of training and evaluation datasets, respectively. The best scores on each evaluation set are highlighted in **bold**. Each score denotes the averaged accuracy of models described in Section 3.1.2. Full results across languages and models are in Table 16.

4.6 Experiments with MaXM dataset

We evaluate models trained on the xGQA dataset with MaXM (Changpinyo et al., 2023), a recently proposed evaluation-only benchmark for multilingual VQA. The MaXM dataset covers seven different languages: English (en), French (fr), Hindi (hi), Hebrew (iw), Romanian (ro), Thai (th), and Chinese (zh). Each evaluation sample consists of an image, a question, and an answer. As the answers in the MaXM dataset are not exactly matched with the ones in xGQA that models are trained, we only use a question whose answer is either "yes" or "no". More details about the MaXM dataset and

[6]MT system evaluation results are in Appendix I.

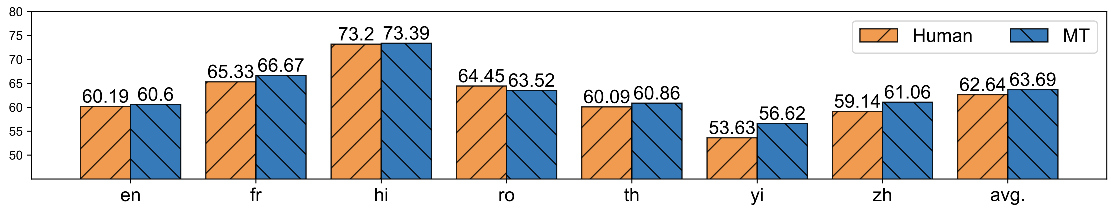


Figure 6: The averaged translate-test evaluation results of models with different training data origins on the yes/no question type in MaXM benchmark. The full results are presented in Table 18.

		Translate-Test							
	en	bn	de	id	ko	pt	ru	zh	**avg.**
Human	**56.91**	51.01	54.10	52.17	50.58	54.02	50.84	49.99	51.82
MT	55.13	52.11	54.80	53.06	52.47	54.10	53.28	52.19	53.14
MERGE	56.52	**52.80**	**55.54**	**53.75**	53.04	**55.08**	**54.10**	**52.73**	**53.86**
TAG	56.67	52.65	55.44	53.56	**53.11**	54.83	53.89	52.65	53.73

Table 6: Data augmentation results. The highest scores in each column are highlighted in **bold**. All model scores with the same data origin are averaged. Full results are in Table 19.

full evaluation results are in Appendix A and Table 18, respectively. As shown in Fig. 6, we observe results consistent with the xGQA dataset. Training on RT-translated texts increases the translate-test scores except for Romanian (ro) cases.

5 Reducing the Effect of Translation Artifact on Cross-lingual VQA

Our findings demonstrate that training VQA models on translated texts induces higher accuracy in language transfer through the translate-test approach. Despite such gains, translated texts inevitably contain wrongly translated information due to the imperfection of MT systems. Moreover, as translations are known to be different from the naturally written human texts (Volansky et al., 2015; Zhang and Toral, 2019), training models solely on the translated texts may degrade overall performance. These problems can be observed in our previous results; in Table 1, the models trained on translated texts show a relatively low average score in the English evaluation set compared to those trained on human texts (56.91→55.13).

To resolve this, we leverage a simple data augmentation technique that uses both RT-translated texts and the original human-written texts for model training (**MERGE**). Furthermore, following Marie et al. (2020), we also adopt the approach that includes special tagging tokens in front of translated texts in both training and evaluation phases (**TAG**). As MERGE and TAG double the number of training examples, we reduce the total training steps to half for a fair comparison across methods. Results with data augmentation methods are in Table 6. The accuracy of the original English evaluation set is increased in both MERGE and TAG compared to solely using translated samples. The overall scores for the translate-test are also improved with data augmentation. These results indicate that augmenting training data with both human and MT texts is helpful for cross-lingual transfer while maintaining its performance on the original English texts.

6 Conclusion

In this work, we analyze the impacts of translation artifacts presented in machine-translated English texts for cross-lingual VQA. Through extensive experiments, we find that current VL models usually suffer from distributional shifts caused by translation artifacts during cross-lingual transfer, resulting in undesirable performance degradation. As a remedy, we conduct experiments with simple data augmentation strategies and observe consistent performance gains.

Our work focuses on translations that are semantically similar but written differently from human texts. In future work, we will explore mistranslation arising from context-free translation, where image information is not considered during a translation process. To this end, recently advanced multimodal translation systems can be utilized (Yao and Wan, 2020). Other important directions include considering a variance among different translations generated from a single text and devising an advanced training strategy to consider translation artifacts.

Limitations

Our study is mainly conducted on a translate-test approach for a cross-lingual VQA task. We recognize that some of our results may not generalize other tasks, like image captioning. Nevertheless, as reasoning over natural language and image is a crucial ability for vision-language models, we believe it is a fundamental step to comprehend the impacts of translation in the VQA task to transfer across different languages seamlessly. Besides, since we mainly consider the conventional *finetune-then-evaluate* pipelines, some experimental setups do not directly apply to recent models that do not perform parameter updates for learning (e.g., GPT-4V (OpenAI, 2023)). As discussed in Appendix K, we observe that these models also can suffer from translation artifacts to some extent when performing VL tasks. Performing extended analysis and proposals across diverse learning algorithms and models remains our future work.

Ethics Statement

Most of the models in our experiments are trained on English datasets only, so the generalizability towards other source languages is not examined. Besides, as the current MT systems are imperfect, training on translated texts may introduce unintended behaviors or favors to specific questions. Future research should investigate such undesirable bias in translations and VQA models.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. *arXiv preprint arXiv:2305.14240*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684.

Mona Baker. 2019. Corpus linguistics and translation studies*: Implications and applications. In *Researching translation in the age of technology and global conflict*, pages 9–24. Routledge.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*. Aslib.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14. International Workshop on Spoken Language Translation.

Beer Changpinyo, Linting Xue, Michal Yarom, Ashish Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2023. Maxm: Towards multilingual visual question answering. In *Findings of ACL: EMNLP*.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 387–396.

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.

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*.

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

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*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

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

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, multitask representations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3449–3463.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Youhan Lee, KyungTae Lim, Woonhyuk Baek, Byungseok Roh, and Saehoon Kim. 2022. Efficient multilingual multi-modal pre-training through triple contrastive loss. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5730–5744.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves smt. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the International Conference on Machine Learning (ICML)*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275.

Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuanjing Huang, and Zhongyu Wei. 2023b. Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5939–5958.

Chen Liu, Jonas Pfeiffer, Anna Korhonen, Ivan Vulić, and Iryna Gurevych. 2023a. Delving deeper into cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2408–2423.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Valerie R Mariana. 2014. *The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment*. Brigham Young University.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320.

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

Farhad Nooralahzadeh and Rico Sennrich. 2022. Improving the cross-lingual generalisation in visual question answering. In *AAAI Conference on Artificial Intelligence*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Chen Qiu, Dan Oneață, Emanuele Bugliarello, Stella Frank, and Desmond Elliott. 2022. Multilingual multimodal learning with machine translated text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4178–4193.

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.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in “multilingual” NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022*, pages 146–162. Springer Nature Switzerland.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* *and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Ashish V. Thapliyal and Radu Soricut. 2020. Cross-modal Language Generation using Pivot Stabilization for Web-scale Language Coverage. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 160–170.

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.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Jiaan Wang, Fandong Meng, Tingyi Zhang, Yunlong Liang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2022. Understanding translationese in cross-lingual summarization. *arXiv preprint arXiv:2212.07220*.

Shuo Wang, Zhaopeng Tu, Zhixing Tan, Shuming Shi, Maosong Sun, and Yang Liu. 2021. On the language coverage bias for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4778–4790.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2021. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

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.

Sicheng Yu, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370.

Yan Zeng, Wangchunshu Zhou, Ao Luo, and Xinsong Zhang. 2022. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. In *Annual Meeting of the Association for Computational Linguistics*.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81.

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

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2015. Visual7w: Grounded question answering in images. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004.

A Dataset Details

xGQA (Pfeiffer et al., 2022) In our study, we used the English-balanced GQA (Hudson and Manning, 2019) training set for model training, which consists of 943k training examples and 72k training images. For model validation, the English GQA validation set, containing 132k samples and 10k images, is used. For evaluation, we used the balanced test-dev subset of the xGQA dataset, which includes 12,578 systematically structured questions with an average length of 8.5 words, associated with 398 images. The xGQA dataset extends the test-dev set of GQA by translating into seven different languages, each from a unique language family. In the translate-test approach, we used the official evaluation set released by Bugliarello et al. (2022), which translates samples written in target languages into English with the Google Machine Translation system. Further details on the xGQA dataset are provided in Pfeiffer et al. (2022).

Language	# Examples
English	75
French	70
Hindi	82
Hebrew	70
Romanian	77
Thai	75
Chinese	52

Table 7: Number of selected examples for each language in MaXM dataset.

MaXM (Changpinyo et al., 2023) The MAVERICS-XM3600 (MaXM) dataset, an evaluation-only VQA benchmark, originates from the Crossmodal-3600 dataset (XM3600) (Thapliyal et al., 2022) and consists of translation-based question-answer pairs. MaXM includes 7 languages which are chosen based on their typological, genealogical, and geographical diversity. The statistics of selected evaluation samples for each language are presented in Table 7.

B Model Details

Table 8 summarizes the key characteristics of all models described in Section 3.1.2. For visual tokens, we utilize 36 image regions from a ResNet101 backbone (He et al., 2016), and 10 to 100 image regions from a ResNeXt-101 backbone (Xie et al., 2017). For BLIP-2, InstructBLIP, and FLAVA, we use the official implementations released by authors. For other models, we use the implementation released by Bugliarello et al. (2022).

C Translation Details

RT Translation We use roundtrip (RT) translation to make translated English training dataset. Unless otherwise specified, NLLB-200-3.3B is used as an MT system, and German (de) is used as a pivot language. Following Artetxe et al. (2023), we use stochastic and deterministic decoding strategies for RT translation. Specifically, for forward translation (en $\rightarrow$ de), we use nucleus sampling (Holtzman et al., 2019) with $p = 0.9$. For backward translation (de $\rightarrow$ en), we use beam search with beam size as 5. For both translation directions, the maximum number of repeated n-gram is set to 5.

Translate-Test Unless otherwise specified, we use the evaluation set released by Bugliarello et al. (2022) for a fair comparison. When constructing the translate-test evaluation set ourselves, as in Section 4.5, we use beam search with beam size 4.

Translate-Train We translate the original English training set into every target language in xGQA. NLLB-200-3.3B is used as an MT system for this process, and beam search is used with beam size 5.

D Implementation Details

For finetuning VL models, we follow hyperparameters described in Bugliarello et al. (2022) for a fair comparison. Specifically, we utilize the AdamW optimizer (Loshchilov and Hutter, 2018) with betas set at $(0.9, 0.999)$ and ϵ=1e-8. The maximum number of tokens in the input sequence is set to 40, and the batch size is set to 256. The total training epochs are set to 5. The learning rate is set to 1e-4, and a linear learning late scheduler is used with a 0.5 warm-up epoch. For training, we used cross-entropy loss for all 1,842 labels available in the GQA dataset. In overall experiments, a single NVIDIA-A100 GPU with 40GB of memory is used for BLIP-2, InstructBLIP, and FLAVA, and a single model is trained in one day. Other models are trained with a 3090 RTX GPU with 24GB of memory and are trained in 5 hours. The experiments are implemented with PyTorch (Paszke et al., 2019).

E Statistical Test Results

Based on the findings in Section 4.1, we observe an improvement in test accuracy when the data origins

Model	Language Model	Visual Tokens	# Trainable Params (M) / # Total Params (M)
MUNITER	bert-base-multilingual-cased	36 RoIs from Faster R-CNN with ResNet-101	116.46M / 116.46M
XUNITER	xlm-roberta-base	36 RoIs from Faster R-CNN with ResNet-101	116.46M / 116.46M
UC^2	xlm-roberta-base	36 RoIs from Faster R-CNN with ResNet-101	281.64M / 281.64M
M^{3}P	xlm-roberta-base	10-100 RoIs from Faster R-CNN with ResNeXt-101	376.90M / 376.90M
LxMERT	bert-base-uncased	36 RoIs from Faster R-CNN with ResNet-101	213.33M / 213.33M
UNITER	bert-base-uncased	36 RoIs from Faster R-CNN with ResNet-101	116.46M / 116.46M
VILBERT	bert-base-uncased	36 RoIs from Faster R-CNN with ResNet-101	244.04M / 244.04M
VisualBERT	bert-base-uncased	36 RoIs from Faster R-CNN with ResNet-101	116.84M / 116.84M
VL-BERT	bert-base-uncased	36 RoIs from Faster R-CNN with ResNet-101	118.03M / 118.03M
BLIP-2	opt-2.7b	-	190.29M / 3827.78M
InstructBLIP	flan-t5-xl	-	189.27M / 4024.92M
FLAVA	ViT-B/16 based text encoder	-	243.36M / 243.36M

Table 8: We report the key properties, training parameters, and total parameters for all the models.

Language	**p-value**
bn (RT >Human)	2.89e-17
de (RT >Human)	3.51e-15
id (RT >Human)	2.48e-17
ko (RT >Human)	1.70e-26
pt (RT >Human)	7.33e-05
ru (RT >Human)	2.14e-14
zh (RT >Human)	5.21e-27

Table 9: We gather all the results of each row in Table 15 to get the results for each model and performed a t-test on these aggregated results. Here $RT > Human$ means models trained with round-trip translated texts (*RT*) are better than models trained with human texts (*Human*).

Language	**p-value**
MUNITER (RT > Human)	3.78e-07
XUNITER(RT > Human)	1.26e-04
UC^2 (RT > Human)	2.06e-08
M^{3}P(RT > Human)	1.35e-07
LXMERT (RT >Human)	2.43e-06
UNITER (RT >Human)	1.05e-07
VILBERT (RT >Human)	8.29e-07
VisualBERT (RT >Human)	6.76e-07
VL-BERT (RT >Human)	1.56e-07
BLIP-2 (RT >Human)	4.82e-10
InstructBLIP (RT >Human)	5.14e-07
FLAVA (RT >Human)	1.62e-07

Table 10: We gather all the results of each column in Table 15 to get the results for each language and perform a t-test. Here $RT > Human$ means models trained with round-trip translated texts (*RT*) are better than models trained with human texts (*Human*).

for training and evaluation are aligned. To demonstrate that this improvement consistently occurs in the translate-test, we train models three times with different seeds and report the average performance of models in Table 15.

Furthermore, we perform significance tests to demonstrate that this improvement consistently occurs in the translate-test. Specifically, we conduct significant tests on all the aggregated results as well as on the language-specific and model-specific results in Table 15.[7] Our evaluation specifically compared the performance of models trained with MT texts against those trained with human texts in the translate-test.

First, we aggregate all the outcomes from models trained on machine-translated texts and compare them to those trained on human texts in Table 15. The results demonstrate a significant advantage in training with machine-translated texts over human texts, with a p-value of 6.49e-66.

[7]For significant tests, we use the paired t-test with $\alpha = 0.05$.

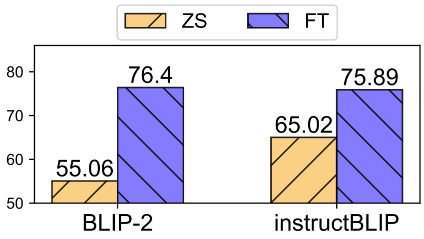


Figure 7: Comparison of zero-shot and finetuning with yes/no questions in xGQA. *ZS* and *FT* denote the accuracy of zero-shot and finetuned models, respectively.

Moving on to specific language results in Table 9, we gather all the model outcomes for each language except for English. We then compare the performance of models trained on machine-translated texts to those trained on human texts.

When considering the performance of individual models, we collect the results from all evaluated languages for each model (*i.e.*, collect all the results of each row in Table 15). We then compare the performance of models trained on machine-translated texts to those trained on human texts in Table 10.

As a result, we can observe that evaluating models trained on human-written English data with translated texts could negatively impact the generalization of models to other languages. By simply aligning the data origins for both training and evaluation sets, the overall performance in the translate-test can be improved.

F Zero-shot Evaluation of VL models

Recent VL models like BLIP-2 (Li et al., 2023a) or InstructBLIP (Dai et al., 2023) can perform target downstream tasks without task-specific finetuning by relying on its language generation ability. To compare the effectiveness of finetuning on these models, we evaluate the models with and without finetuning by using xGQA evaluation samples whose answers are either "yes" or "no". This subset contains 4,525 samples out of 12,578 total evaluation samples. For zero-shot evaluation, we prompt the model with task description as follows: '`Answer the following question in "yes" or "no".\n Question:` **<question>** `\n Answer:` '. Models generate the next token as an answer for the given question in the prompt with an image. We conduct post-processing steps including case-normalization or punctuation mark removal to derive the binary prediction of models. For model implementation, we use the models released by Wolf et al. (2020).

As shown in Fig. 7, although models exhibit competitive zero-shot scores, their performance is lower than finetuned ones. These results imply that finetuning the models on task-specific datasets is also crucial for recent VL models. In this regard, it is still essential for the VL models to consider and address data origin misalignment presented in training and evaluation.

MT	SacreBLEU	chrF	METEOR
	de → en		
M2M-Small	30.82	55.55	0.61
M2M-Large	33.68	57.86	0.63
NLLB-Small	39.34	61.48	0.68
NLLB-Large	**42.98**	**64.09**	**0.70**
	en → de		
M2M-Small	25.77	54.40	0.55
M2M-Large	30.36	58.09	0.58
NLLB-Small	32.03	58.64	0.59
NLLB-Large	**34.79**	**60.92**	**0.62**

Table 11: Evaluation results of different MT systems on IWSLT 2017 benchmarks (Cettolo et al., 2017). The best scores on each metric are highlighted in **bold**.

G Human Evaluation Details

We first identified examples where questions, initially correct in English, became incorrect in the translate-test. Among these examples, we specifically focused on cases where the UC^2 model, trained using the original English GQA dataset, provided incorrect results, but the UC^2 model trained with RT-translated data generated correct responses. From the examples that conformed to these restrictions, we analyzed a subset of 200 examples.

Following Moghe et al. (2023), we annotated any machine translation (MT) errors in these examples, utilizing the Multidimensional Quality Metrics (MQM) ontology (Burchardt, 2013). This framework categorizes errors into a hierarchical structure, allowing for the evaluation of translations based on this hierarchy. Our analysis focused on 5 error types within the MQM ontology, including *Mistranslation*, *Addition*, *Omission*, *Fluency*, and *Grammar*. Two authors with a master's degree or higher separately annotated each evaluation sample. The annotated examples from our case study are presented in Fig. 11.

H Human-likeness Analysis Details

We use a confidence score of a text classifier to analyze the prevalence of translation artifacts in every translated question. The classifier is trained to discriminate whether the data origin of a given

Method	Code	Notes
METEOR	`https://huggingface.co/spaces/evaluate-metric/meteor`	
chrF	`https://huggingface.co/spaces/evaluate-metric/chrf`	signature: "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0"
SacreBLEU	`https://huggingface.co/spaces/evaluate-metric/sacrebleu`	signature: "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0"

Table 12: Code and versions for each MT metric.

question is a human or MT system. Specifically, we finetune RoBERTa-base (Liu et al., 2019) to classify whether the given question is from the original human-written dataset or the translated dataset from another target language. The training epochs, batch size, and learning rate are set to 3, 24, and 2e-5. The finetuned classifier assigns the confidence score $p_h(x)$ about how likely a human writes the question to each translated question in translate-test evaluation sets. This confidence score is regarded as *human-likeness* of translated questions.

I Translation Quality of MT systems

From results in Section 4.5, we observe that the accuracy ranking in the original English set and the GMT translate-test set is NLLB-200-3.3B > NLLB-200-600M > M2M-100-1.2B > M2M-100-418M. We suspect that this trend reflects the translation quality of the training data produced by each MT system. To corroborate this, we assessed these MT systems using an IWSLT 2017 (Cettolo et al., 2017) benchmark, while maintaining the same translation direction as in RT translation (i.e., en→bn and vice versa). The IWSLT2017 dataset contains 8079 parallel sentences in these language directions, which involves multilingual text translation of TED talks. For evaluation, we utilized METEOR (Banerjee and Lavie, 2005), chrF (Popović, 2015), and SacreBLEU (Post, 2018) as evaluation metrics. As shown in Table 11, we observe the results which are clearly aligned with the previously observed trends. Further details of each metric are in Table 12.

J Experiments with LLaMA-Adapter-V2

We examine whether LLaMA-Adapter-V2 (Gao et al., 2023), a recently proposed powerful VL model with a large language model, also suffers from translation artifacts for cross-lingual VQA tasks. To this end, we finetune LLaMA-Adapter-V2 with different training options (Human, MT, MERGE, and TAG) and observe their results.

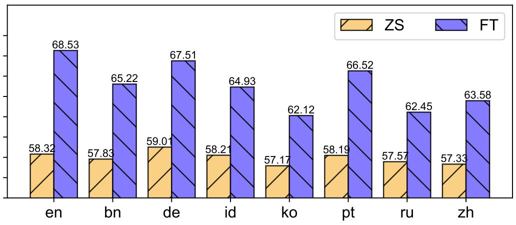


Figure 8: Accuracy of LLaMA-Adapter-V2 zero-shot and fine-tuned performance on yes/no questions on xGQA. The finetuned model is trained on all human training samples in the xGQA dataset.

Specifically, we add a classification head on top of end-of-sequence (eos) token representation in LLaMA and finetune it along with the unfrozen weights.[8] The model is parameter-efficiently finetuned, where only a small portion of the total parameters are updated (15M). We use the official codes released by the authors[9] for implementation, and LLaMA-7b (Touvron et al., 2023) with CLIP visual encoder (Radford et al., 2021) is used. The overall finetuning setups follow previously mentioned ones in Section 3.1.5. Note that we also finetune and evaluate models to directly generate the answer text, but the scores are usually lower compared to using the classification head.

As shown in Table 14, leveraging translated texts for training is beneficial to the translate-test approach of LLaMA-Adapter-V2, where the models trained on translated texts show higher accuracy compared to human texts. MERGE and TAG further improve accuracy in English and other target languages.

Besides, we also evaluate the model without finetuning on xGQA to probe its zero-shot ability. Since zero-shot classification with generation models requires roughly the number of forward passes with answer candidates, we choose evaluation samples whose label is either "yes" or "no", and measure the probability of both tokens. Regard-

[8] `LORA-BIAS-7B` is used.

[9] `https://github.com/OpenGVLab/LLaMA-Adapter/tree/main/llama_adapter_v2_multimodal7b`

	en	bn	de	id	ko	pt	ru	zh	**avg.**
Zero-Shot	60.67	64.67	65.33	61.67	66.00	64.33	62.33	59.67	63.43
Translate-Test	-	57.33	60.67	59.00	56.33	59.00	55.33	60.33	58.28

Table 13: Evaluation results of `gpt-4-1106-vision-preview` on xGQA datasets. All experiments are conducted based on 300 yes/no type questions. *Zero-Shot* denotes that the input question is written in the target language.

		Translate-Test							
	en	bn	de	id	ko	pt	ru	zh	**avg.**
Human	53.03	47.72	50.40	48.10	46.60	50.08	48.01	46.90	48.26
MT	51.41	48.76	51.11	49.36	48.93	50.64	49.98	48.94	49.67
MERGE	53.15	**49.52**	**51.79**	**50.25**	**49.73**	51.64	**50.57**	**49.42**	**50.42**
TAG	**53.22**	49.21	51.74	50.23	49.54	**52.04**	50.38	49.32	50.35

Table 14: Evaluation results of LLaMA-Adapter-V2 models parameter-efficiently finetuned with different data origins. The highest scores in each column are highlighted in **bold**.

ing the comparison with zero-shot and finetuning for yes/no question types in Fig. 8, the finetuned model scores better than the zero-shot approach. This result implies that although recent VL models show impressive zero-shot capability, finetuning on task-specific datasets is still required for better performance.

K Experiments with GPT-4-Vision

In this study, we present experimental results of GPT-4-Vision (OpenAI, 2023), a cutting-edge VL model. We use 300 evaluation samples of yes/no questions described in Appendix F. We include all target languages and their corresponding original English questions. For evaluations in the target languages, inputs consist of questions either originally written in the target language or translated into English via GMT. The prompt format and the evaluation outcomes are presented in Fig. 15, and Table 13, respectively.

Our findings indicate that GPT-4 can serve as an effective multilingual VL model. Remarkably, its performance in all languages except Chinese exceeds that of English. Directly using the target language proves more efficient than relying on the translated source language, primarily due to the inherent errors in translation processes.

However, GPT-4 falls short of the finetuned monolingual models detailed in Appendix F. The direct comparison between GPT-4 and these models is nuanced, largely because of differences in evaluation settings.[10] Despite these challenges, the translate-test with strong VL models yielded more favorable outcomes than using GPT-4, with scores of 63.43 compared to 76.4 and 75.89 for finetuned BLIP-2 and InstructBLIP models, respectively, Additionally, our qualitative analysis indicates that GPT-4 is also susceptible to translation artifacts, which can cause differences in predictions between human and MT texts. We present the qualitative results of GPT-4 on xGQA in Fig. 15 and 16.

L Additional Results

Full results of Fig. 2 Fig. 9 presents full results of Fig. 2.

Full results of Fig. 5 Fig. 10 presents full results of Fig. 5 across different target languages.

Full results of Table 5 Table 16 presents full results with varying MT systems for RT translation and a translate-test approach.

Full results of Table 4 Table 17 presents full results with varying pivot languages used in RT translation. NLLB-200-3.3B is used as an MT system.

Full results of Fig. 6 Table 18 presents the full results of different models on the MaXM dataset. NLLB-200-3.3B is used as an MT system for translate-test evaluation.

Full results of Table 6 Table 19 presents the full results of models with different data sources.

[10]This complexity arises from the differences in the number of questions asked and the categorization of any response from GPT-4 other than "yes" or "no" as incorrect.

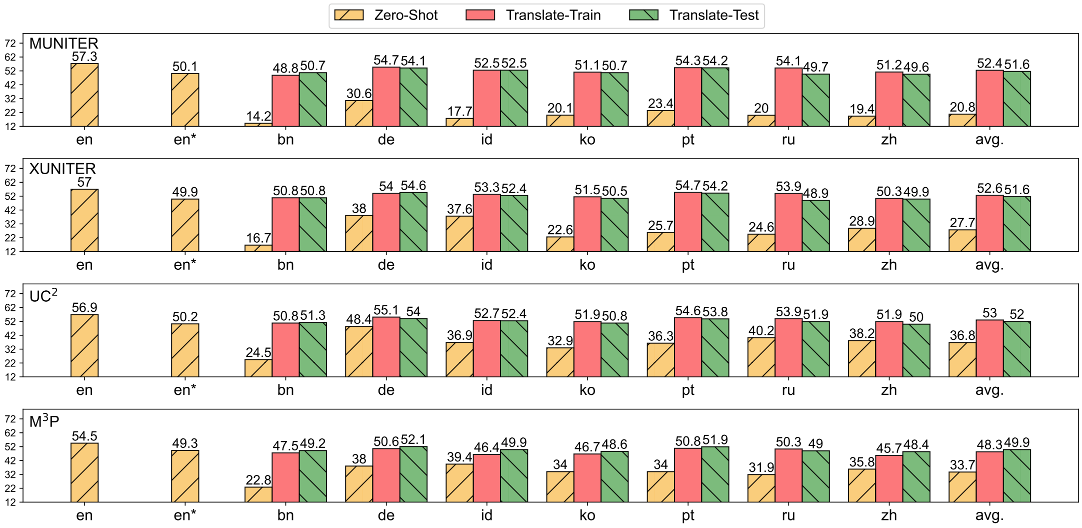


Figure 9: **Multilingual models results** The indicators are the same as Fig. 2.

Models	**RT?**	en	bn	de	id	ko	pt	ru	zh	**avg.**
MUNITER		57.20	50.79	53.95	52.40	50.60	54.30	49.85	49.63	51.65
	✓	55.57	52.14	55.36	53.35	52.92	54.54	53.68	52.43	53.49
XUNITER		56.96	51.20	54.46	52.39	50.70	54.19	49.81	49.92	51.81
	✓	55.27	52.06	54.88	52.76	52.32	54.01	53.01	52.20	53.03
UC^2		56.92	51.34	54.27	52.43	51.20	54.25	52.49	50.16	52.30
	✓	55.50	52.60	55.29	53.54	53.27	54.52	53.96	52.82	53.71
M^{3}P		54.70	48.65	51.43	49.59	47.94	51.37	48.35	47.99	49.33
	✓	53.51	49.93	52.51	50.72	49.97	51.77	50.97	49.97	50.84
LXMERT		54.89	48.94	52.39	50.36	48.58	52.20	47.59	47.76	49.69
	✓	53.66	50.65	53.09	51.59	50.83	52.59	51.65	50.48	51.55
UNITER		57.52	51.44	54.37	52.64	51.21	54.45	51.86	50.22	52.32
	✓	55.92	52.34	55.45	53.60	53.13	54.75	53.83	52.67	53.68
VILBERT		57.08	51.05	54.37	52.68	51.04	54.20	50.14	50.01	51.93
	✓	54.84	52.58	55.19	53.79	52.97	54.53	53.92	52.72	53.67
VisualBERT		55.26	49.51	52.43	50.24	48.80	52.54	50.62	48.62	50.40
	✓	53.59	50.53	53.10	51.07	50.49	52.44	51.41	50.63	51.38
VL-BERT		57.66	51.03	53.95	52.39	50.78	54.58	50.49	49.73	51.85
	✓	55.67	52.37	55.27	53.55	52.81	54.76	53.54	52.32	53.52
BLIP-2		57.84	51.59	54.52	52.73	51.26	54.61	52.02	51.06	52.54
	✓	56.35	53.36	55.86	54.13	53.54	55.12	54.42	53.40	54.26
InstructBLIP		57.76	51.65	54.81	53.04	51.08	54.79	52.92	51.53	52.83
	✓	56.15	53.20	55.54	53.90	53.36	54.74	54.45	53.19	54.05
FLAVA		**58.43**	53.27	55.90	54.00	52.66	55.72	53.32	51.73	53.80
	✓	57.20	**54.09**	**56.72**	**55.23**	**54.55**	**56.14**	**55.58**	**54.07**	**55.20**
avg.		56.76	50.85	53.88	52.05	50.49	53.89	50.81	49.87	51.69
	✓	56.19	51.22	54.11	52.35	51.11	53.93	51.61	50.60	52.13

Table 15: **Averaged translate-test results with different origins of training dataset** Each accuracy represents the average of three training with different random seeds. The indicators are the same as Table 1.

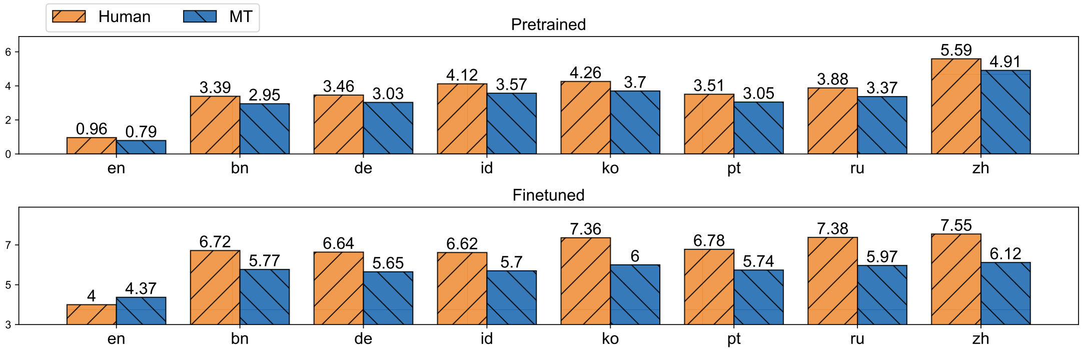


Figure 10: Representation discrepancy of translate-test evaluation samples against training samples from different data origins (Human and MT). Pretrained or finetuned VisualBERT is used to encode representation, and FID score is used as a distance metric. A lower score indicates a low distance between training and evaluation samples.

		en	bn	de	id	ko	pt	ru	zh	avg.	bn	de	id	ko	pt	ru	zh	avg.	bn	de	id	ko	pt	ru	zh	avg.	bn	de	id	ko	pt	ru	zh	avg.	bn	de	id	ko	pt	ru	zh	avg.
Models			**GMT**								**NLLB**								**NLLB_3.3B**								**M2M-418M**								**M2M-1.2B**							
MUNITER	Human	57.33	50.67	54.09	52.54	50.67	54.21	49.69	49.57	51.63	49.65	53.8	49.90	47.40	52.22	46.96	45.1	49.29	50.87	54.48	51.76	49.17	51.86	47.17	45.76	50.15	44.19	48.84	47.13	43.85	47.50	42.43	43.57	45.36	44.74	51.88	49.55	45.72	51.62	45.39	46.04	47.85
	NLLB-small	55.38	51.85	55.15	53.36	52.7	54.7	52.77	52.27	53.26	51.27	54.82	51.97	50.17	54.43	52.65	48.59	51.99	52.23	54.67	53.09	50.99	54.12	53.21	49.44	52.54	46.09	51.71	49.56	47.37	51.44	49.36	48	49.08	46.84	53.36	51.41	48.47	52.91	51.22	49.24	50.49
	NLLB-large	55.70	52.34	55.66	53.48	53.36	54.72	53.98	52.29	53.69	51.25	55.16	52.07	50.31	54.48	52.88	48.43	52.08	52.70	55.58	53.23	51.37	54.89	53.94	50.07	53.11	46.37	51.47	49.31	47.06	51.03	49.01	47.48	48.82	46.90	53.59	51.27	48.22	53.04	51.42	48.56	50.43
	M2M-small	53.96	50.84	53.1	51.77	51.48	52.85	51.65	50.54	51.75	50.37	52.5	50.52	48.47	52.55	50.97	47.23	50.37	50.65	52.89	51.23	49.47	52.81	51.23	48.58	50.98	47.24	51.37	49.40	47.54	51.76	49.46	47.99	49.25	47.49	52.53	50.32	47.88	51.95	50.20	48.12	49.78
	M2M-large	**54.87**	**51.47**	**54.33**	**52.22**	**52.23**	**53.35**	**52.36**	**50.87**	**52.40**	**51.03**	54.15	51.68	49.59	53.42	52.15	48.51	51.50	51.56	54.13	52.50	50.63	53.74	52.44	49.54	52.08	47.07	51.85	49.79	48.00	51.61	50.07	48.37	49.54	47.45	53.63	51.80	48.79	52.81	51.90	49.63	50.86
XUNITER	Human	56.98	50.76	54.63	52.37	50.52	54.24	48.91	49.94	51.62	49.60	53.59	49.57	47.11	51.24	45.33	45.29	48.82	50.69	54.09	51.36	48.86	51.20	45.77	46.06	49.72	43.82	48.79	46.80	43.16	44.58	37.05	43.66	43.98	44.08	51.61	48.76	45.44	50.52	41.49	46.02	46.85
	NLLB-small	54.48	51.14	54.27	52.32	51.49	53.59	52.47	51.27	52.36	50.47	53.92	51.16	49.01	53.30	51.73	47.67	51.04	51.45	53.82	52.06	50.21	53.38	51.91	48.88	51.67	46.44	51.28	49.09	47.24	50.25	48.40	47.58	48.61	46.45	52.7	50.92	47.79	52.07	50.13	48.28	49.76
	NLLB-large	55.22	52.10	54.97	52.66	52.51	54.18	52.85	52.23	53.07	51.00	54.64	51.42	49.78	53.94	52.00	48.22	51.57	51.95	54.74	52.62	51.01	54.25	52.90	49.68	52.45	46.33	51.11	49.10	47.69	50.37	48.29	47.71	48.66	46.70	53.26	51.03	48.43	52.68	50.30	48.87	50.18
	M2M-small	53.40	50.93	52.67	50.87	50.6	52.33	51.33	50.12	51.26	49.97	52.27	50.12	47.65	51.92	50.22	47.42	49.94	50.49	52.23	50.64	49.02	52.02	50.51	48.12	50.43	47.02	51.07	48.95	46.82	50.88	49.11	47.73	48.80	46.95	51.9	49.92	47.54	51.34	49.55	47.79	49.28
	M2M-large	**54.42**	**52.05**	**54.26**	**52.31**	**51.86**	**53.71**	**53.19**	**50.75**	**52.59**	**50.68**	53.95	51.32	48.71	52.95	51.85	47.8	51.04	51.03	53.74	52.00	49.71	53.32	51.98	48.94	51.53	47.31	51.92	50.10	48.26	51.56	50.26	48.51	49.70	46.97	53.45	51.19	48.87	52.92	51.76	49.32	50.64
UC^2	Human	56.85	51.34	54.01	52.35	50.75	53.81	51.93	50.04	52.03	49.73	53.08	49.75	47.51	52.77	49.90	45.76	49.79	51.42	54.13	51.64	49.32	53.35	51.08	47.48	51.20	44.17	49.17	47.31	44.78	49.69	47.17	45.59	46.84	44.44	51.37	49.17	46.16	51.63	49.25	47.06	48.44
	NLLB-small	54.87	51.80	54.55	53.09	52.27	54.18	53.13	51.95	53.00	50.71	54.23	51.84	49.63	53.76	52.50	48.23	51.56	51.88	53.99	52.84	51.14	53.89	52.35	49.2	52.18	46.30	51.41	49.62	47.14	50.93	49.69	48.06	49.02	46.36	52.9	51.45	48.22	52.61	51.26	49.34	50.31
	NLLB-large	55.12	52.35	55.1	53.29	53.07	54.17	53.36	52.73	53.44	51.07	54.44	51.77	50.04	53.98	52.75	48.58	51.80	52.42	54.7	53.10	51.14	54.05	53.59	49.71	52.67	46.14	51.38	49.40	47.86	50.48	49.32	47.87	48.92	47.07	53.2	51.11	48.82	52.82	51.13	49.28	50.49
	M2M-small	53.50	50.86	53.04	51.32	50.99	52.61	51.78	50.88	51.64	50.18	52.73	50.50	48.67	52.19	50.89	48.37	50.50	50.35	52.5	51.02	49.43	52.12	50.77	49.13	50.76	47.51	51.3	49.58	47.26	51.19	49.44	48.01	49.18	47.45	51.94	50.41	47.83	51.77	50.75	48.67	49.83
	M2M-large	54.21	52.07	53.76	52.23	52.29	53.25	52.62	50.62	52.41	50.79	53.29	51.45	49.68	52.71	51.87	48.31	51.16	51.07	53.24	51.78	50.23	52.62	52.08	49.26	51.47	47.10	51.3	50.52	48.44	51.43	50.56	48.39	49.68	47.35	52.81	51.34	48.82	52.69	51.74	49.6	50.62
M^{3}P	Human	54.45	49.18	52.14	49.87	48.59	51.87	49.05	48.38	49.87	47.88	51.5	47.59	45.62	50.37	46.63	44.53	47.73	49.13	52.07	49.10	46.90	50.45	47.61	45.95	48.74	43.08	47.5	45.61	43.20	46.88	43.15	43.55	44.71	43.66	50.21	47.23	44.42	49.66	45.25	45.42	46.55
	NLLB-small	51.88	49.65	52.07	50.32	49.25	51.19	50.63	49.48	50.37	48.65	51.77	48.87	47.08	50.96	49.73	46.07	49.02	49.44	51.3	49.64	48.23	51.05	49.86	46.8	49.47	44.51	49.63	47.25	44.73	48.53	47.19	45.48	46.76	44.91	50.85	49.05	45.59	50.11	48.90	46.61	48.00
	NLLB-large	52.97	50.63	53.03	51.42	50.38	52.11	51.8	50.41	51.40	49.38	52.37	49.82	47.94	51.92	50.68	47.44	49.94	50.86	52.9	50.52	49.29	52.27	51.43	48.2	50.78	45.36	49.63	47.99	45.77	48.86	47.60	46.1	47.33	46.13	51.22	49.51	46.67	50.81	49.45	47.27	48.72
	M2M-small	50.63	48.43	50.2	48.65	48.43	49.64	49.37	47.92	48.95	47.31	49.62	47.71	46.18	49.45	48.21	45.68	47.74	48.04	49.49	48.31	47.05	49.47	48.33	46.41	48.16	45.26	48.89	47.01	45.00	48.79	46.92	45.73	46.80	45.10	49.64	47.74	45.41	48.90	47.91	45.68	47.20
	M2M-large	51.07	49.07	50.84	49.51	49.33	50.33	49.88	48.94	49.70	48.41	50.25	48.55	46.87	49.79	49.08	46.47	48.49	48.63	50.37	49.03	47.68	50.11	49.17	46.96	48.85	44.85	49.02	47.71	45.98	48.71	47.53	45.9	47.10	45.40	50.08	48.60	46.42	50.02	49.13	47.03	48.10
LXMERT	Human	55.40	49.64	52.83	50.80	49.17	52.49	47.54	48.02	50.07	47.92	52.21	48.81	45.79	50.94	44.65	44.55	47.84	49.27	52.76	50.13	47.30	50.85	45.20	45.39	48.70	42.78	48.2	46.31	42.98	46.02	39.04	43.19	44.07	43.11	50.58	48.00	44.28	49.98	42.29	45.07	46.19
	NLLB-small	52.95	50.00	52.85	50.91	50.14	52.08	50.97	50.11	51.01	49.32	52.86	49.81	47.59	51.95	50.46	46.74	49.82	50.38	52.58	50.70	49.01	51.88	50.54	47.36	50.35	44.55	50.06	47.56	45.71	49.07	47.35	46.14	47.21	45.46	51.53	49.07	46.47	50.75	48.39	47.04	48.39
	NLLB-large	53.44	50.20	52.93	51.34	50.41	52.47	51.44	50.25	51.29	49.29	52.61	49.94	47.87	52.15	51.05	46.84	49.96	50.65	52.83	50.97	49.50	52.46	52.05	48.19	50.95	44.74	49.48	47.30	45.91	48.86	47.37	45.79	47.06	45.31	51.37	49.34	46.94	50.84	49.25	47.42	48.64
	M2M-small	51.56	48.50	50.61	48.88	48.55	50.37	49.5	48.19	49.23	48.08	50.15	48.19	46.47	50.33	48.62	45.68	48.22	48.02	50.15	48.54	47.20	50.19	49.04	46.5	48.52	45.36	49.32	47.43	45.77	49.30	47.89	46.26	47.33	45.60	50.02	48.17	45.76	49.83	48.49	46.4	47.75
	M2M-large	52.80	49.98	52.45	50.89	50.46	51.89	51.47	49.21	50.91	48.93	51.92	49.72	47.11	51.20	50.37	46.31	49.37	49.66	52.27	50.46	48.22	51.49	50.63	47.41	50.02	45.59	50.19	48.23	46.57	49.89	48.65	46.4	47.93	45.32	51.59	49.52	46.65	50.81	49.94	47.49	48.76
UNITER	Human	57.47	51.74	54.52	52.79	51.27	54.56	52.27	50.33	52.50	50.14	54.01	50.25	47.60	53.33	50.53	45.79	50.24	51.58	54.76	52.29	49.26	54.03	51.62	47.27	51.54	44.36	49.39	47.51	45.67	48.19	43.31	45.27	46.24	44.54	52.33	49.73	46.40	51.59	46.97	46.77	48.33
	NLLB-small	55.49	52.03	54.76	53.28	52.66	54.44	53.18	52.02	53.20	51.54	54.98	51.93	49.64	54.53	53.03	48.86	52.07	52.45	54.9	52.87	51.26	54.38	53.36	49.52	52.68	46.73	51.98	49.99	47.92	51.29	49.21	48.36	49.35	47.17	53.58	51.52	48.48	53.29	51.36	49.6	50.71
	NLLB-large	55.92	52.32	55.53	53.67	52.93	54.66	53.56	52.6	53.61	51.66	55.26	52.54	50.06	54.85	52.82	49.53	52.39	53.36	55.59	53.67	51.66	54.91	53.78	50.39	53.34	46.93	52.14	49.82	47.97	51.12	48.85	48	49.26	47.49	53.75	51.76	48.82	53.19	51.26	49.44	50.82
	M2M-small	54.13	50.56	52.99	51.15	50.85	52.65	51.34	49.95	51.36	49.94	52.56	50.41	48.33	52.74	50.86	47.34	50.31	50.56	52.74	51.32	49.24	52.58	51.46	48.24	50.88	46.74	51.37	49.28	47.14	51.22	49.75	47.46	48.99	46.61	52.15	50.38	47.75	51.98	50.45	48.07	49.63
	M2M-large	55.35	52.27	54.64	52.97	52.79	53.88	53.24	51.44	53.03	51.25	54.09	51.96	49.55	53.66	52.39	48.67	51.65	52.26	54.46	52.74	50.80	53.83	53.00	49.52	52.37	47.42	52.46	50.68	48.84	51.92	50.78	48.86	50.14	47.87	53.97	51.97	49.52	53.51	52.33	49.85	51.29
VILBERT	Human	56.72	50.84	54.1	52.27	50.73	53.98	49.91	49.92	51.68	49.24	53.64	49.83	47.19	51.88	47.11	45.2	49.16	50.64	54.09	51.53	48.86	52.42	48.21	46.3	50.29	44.08	49.41	47.00	43.97	46.84	39.92	44.43	45.09	44.46	51.76	48.88	45.68	50.68	43.38	46.16	47.29
	NLLB-small	54.86	51.40	54.32	52.73	51.88	54.21	52.57	51.66	52.68	50.95	54.22	51.73	49.24	53.82	52.08	48.04	51.44	51.91	54.33	52.39	51.28	53.85	52.23	48.86	52.12	46.76	51.6	49.24	47.30	50.56	48.74	48.01	48.89	47.17	52.85	51.20	47.86	52.40	50.51	48.77	50.11
	NLLB-large	55.22	52.23	54.85	53.43	52.75	54.26	53.69	52.22	53.35	51.22	54.48	52.21	49.82	54.01	52.69	48.96	51.91	52.58	54.8	52.84	51.40	54.13	53.36	49.73	52.69	46.37	51.08	49.53	47.18	50.62	48.67	47.96	48.77	47.07	53.12	51.27	48.26	52.48	51.49	49.27	50.42
	M2M-small	53.52	51.07	52.9	51.37	50.97	52.36	51.57	50.43	51.52	50.41	52.27	50.53	48.52	52.44	50.56	47.61	50.33	50.77	52.81	51.35	49.70	52.28	51.26	48.69	50.98	47.35	50.94	49.63	47.11	50.87	49.37	47.97	49.03	47.34	52.08	50.79	47.35	51.50	50.33	48.31	49.67
	M2M-large	55.12	52.15	54.32	52.66	52.21	53.6	53.43	51.44	52.83	51.11	53.66	51.72	49.56	52.81	52.25	48.12	51.32	51.25	53.94	52.10	50.45	53.20	52.27	49.12	51.76	47.27	52.39	49.97	48.33	51.53	50.55	48.71	49.82	47.48	53.44	51.83	48.74	52.67	52.22	49.46	50.83
VisualBERT	Human	55.17	49.43	52.58	50.34	48.66	52.72	50.5	48.89	50.45	48.08	51.92	48.19	45.84	51.87	48.96	44.56	48.49	49.64	52.42	49.65	47.51	52.51	49.75	45.48	49.57	43.08	47.92	45.36	43.13	45.67	40.62	43.43	44.17	43.34	50.29	47.21	44.39	49.70	44.47	44.86	46.32
	NLLB-small	53.59	50.60	52.87	51.00	50.68	52.77	51.11	50.52	51.36	48.99	52.68	49.69	47.30	52.35	50.65	46.72	49.77	50.47	52.89	50.53	49.11	52.81	51.30	47.58	50.67	44.54	50.09	47.42	45.55	49.16	47.25	46.18	47.17	45.50	51.32	48.77	46.33	51.08	48.82	47.27	48.44
	NLLB-large	53.51	50.57	53.1	51.17	50.45	52.59	51.47	50.97	51.47	49.44	52.73	49.52	47.87	52.55	50.87	46.94	49.99	50.92	53.29	50.90	49.10	52.42	51.60	48.25	50.93	45.08	49.3	47.41	45.92	48.47	46.70	45.95	46.98	45.17	51.22	49.04	46.58	50.56	49.04	47.53	48.45
	M2M-small	52.26	49.36	51.47	49.81	49.35	51.19	50.31	49.35	50.12	48.61	51.05	48.76	47.42	50.86	48.94	46.42	48.87	49.16	51.15	49.26	48.00	50.72	49.59	47.27	49.31	45.46	49.86	47.94	46.03	49.84	48.22	46.64	47.71	45.79	50.29	48.59	45.88	50.45	48.66	46.62	48.04
	M2M-large	52.71	50.10	52.56	50.35	49.9	52.01	51.25	49.95	50.87	48.83	52.15	49.44	47.61	51.52	50.24	47.05	49.55	50.02	52.42	49.98	48.63	51.60	50.62	47.8	50.15	46.02	50.26	48.09	46.60	49.76	48.07	46.85	47.95	46.00	51.73	49.48	46.98	51.32	49.86	48.12	49.07
VL-BERT	Human	57.79	51.22	54.47	52.62	50.94	54.79	51.17	50.02	52.18	50.14	53.67	50.60	47.82	53.55	48.78	45.4	49.99	51.38	54.78	52.22	49.61	53.51	49.37	46.68	51.08	44.20	48.8	46.92	44.08	47.59	41.04	44.01	45.23	44.35	51.89	49.71	46.31	51.69	45.85	46.57	48.05
	NLLB-small	55.91	52.41	55.76	53.84	52.9	55.26	53.56	52.5	53.75	51.79	55.47	52.19	50.12	54.90	52.98	48.81	52.32	52.91	55.4	53.35	51.68	55.07	53.40	49.83	53.09	47.15	52.08	50.01	47.80	51.26	49.39	48.37	49.44	47.78	53.78	51.99	48.54	53.08	50.91	49.15	50.75
	NLLB-large	55.61	52.38	55.27	53.43	52.58	54.63	53.32	52.31	53.42	51.20	54.78	51.92	49.79	54.55	52.92	48.78	51.99	52.84	55.06	53.13	51.47	54.55	53.78	49.58	52.92	46.63	51.04	49.55	47.64	50.65	48.58	47.84	48.85	47.06	53.04	51.07	48.67	52.67	51.16	49.4	50.44
	M2M-small	54.67	51.43	53.79	52.04	51.45	53.3	52.05	50.92	52.14	50.48	53.09	51.14	49.13	53.03	51.25	48.08	50.89	51.16	53.28	51.38	49.91	53.10	52.03	49.06	51.42	47.92	51.5	49.77	47.77	52.04	50.17	48.31	49.64	48.23	52.53	51.26	48.59	52.58	51.14	48.64	50.42
	M2M-large	55.64	52.23	54.71	53.09	52.3	54.64	53.2	51.14	53.04	51.17	54.61	52.22	49.55	53.75	52.47	48.31	51.73	51.64	54.78	52.46	50.59	53.74	52.89	49.37	52.21	47.35	52.46	50.59	48.31	52.02	50.72	48.62	50.01	47.31	54.28	51.80	49.25	53.30	52.50	49.3	51.11
BLIP-2	Human	58.05	52.03	54.7	52.99	51.57	54.91	52.36	51.22	53.48	50.13	54.4	50.82	47.94	53.94	50.43	46.78	50.63	52.32	55.14	52.39	49.75	54.68	51.12	47.98	51.91	44.81	49.45	47.61	45.10	48.96	46.27	45.81	46.86	45.36	52.64	49.54	46.61	51.98	48.70	47.62	48.92
	NLLB-small	55.69	52.86	55.15	53.58	52.66	54.69	53.93	52.85	53.93	52.18	55.41	52.25	49.83	54.79	53.59	49.64	52.53	53.16	55.29	53.16	51.38	54.82	53.57	50	53.05	47.18	52.11	50.32	47.85	51.62	50.12	48.38	49.65	47.29	53.94	51.63	48.76	53.26	51.46	49.78	50.87
	NLLB-large	56.11	53.18	55.7	53.98	53.51	55.11	54.25	53.31	54.39	52.21	55.35	52.73	50.29	54.96	53.24	49.21	52.57	53.74	56.07	53.70	51.96	55.15	54.37	50.8	53.68	46.98	52.03	49.97	48.28	51.52	49.40	48.52	49.53	47.81	54.08	51.62	48.76	53.32	51.66	49.9	51.02
	M2M-small	53.97	50.29	52.7	50.97	50.37	52.54	51.15	50.31	51.54	50.09	52.77	50.33	48.72	52.57	51.10	48.08	50.52	50.76	53.07	51.08	49.48	52.85	51.49	49.17	51.13	47.23	51.39	49.60	47.23	51.42	49.59	48.5	49.28	47.30	52.24	50.67	47.62	51.84	50.50	48.47	49.81
	M2M-large	55.41	52.03	54.52	52.77	52.71	54.04	53.32	51.95	53.34	51.07	54.41	51.80	49.86	54.19	52.99	49.34	51.95	52.27	54.44	52.79	51.25	54.07	53.00	50.41	52.60	47.68	52.77	50.49	48.47	51.76	51.31	49.52	50.29	48.27	53.99	52.05	49.46	53.10	52.40	50.58	51.41
InstructBLIP	Human	57.85	51.80	54.91	53.01	51.29	54.85	53.16	51.34	52.91	50.23	54.27	50.48	48.04	54.33	51.49	46.63	50.78	51.81	54.77	52.08	49.44	54.76	52.27	48.23	51.91	44.48	49.89	47.65	44.91	49.72	48.33	46.3	47.33	44.43	52.73	49.61	46.09	52.24	49.71	47.81	48.95
	NLLB-small	55.61	52.27	54.71	53.57	52.55	54.57	53.67	52.52	53.41	51.77	55.14	52.27	49.89	54.56	53.34	49.29	52.32	52.76	54.99	53.27	51.50	54.79	53.65	50.29	53.04	46.76	52.2	50.50	47.42	51.26	50.49	48.34	49.57	47.30	53.59	51.49	48.18	53.04	51.43	49.63	50.67
	NLLB-large	55.84	53.04	55.06	53.82	53.17	54.32	54.08	53.18	53.81	51.66	54.87	52.35	50.71	54.40	53.43	49.68	52.44	53.50	55.42	53.51	51.93	54.48	53.78	50.74	53.34	47.08	51.53	49.98	48.44	51.06	49.29	48.83	49.46	47.27	53.42	51.46	48.78	52.73	51.50	49.75	50.70
	M2M-small	54.13	50.70	53.03	51.56	50.99	52.44	51.57	50.97	51.61	50.14	52.73	51.14	49.52	52.41	50.94	48.39	50.75	51.01	52.97	51.34	49.82	52.69	51.29	49.55	51.24	47.59	51.36	50.21	47.67	51.88	49.95	48.7	49.62	47.84	52.34	51.07	47.57	52.15	50.76	48.99	50.10
	M2M-large	54.86	52.00	54.21	52.81	52.33	53.78	53.36	51.82	52.90	51.15	53.84	52.05	49.87	53.49	52.66	48.86	51.70	52.14	54.05	52.42	50.83	53.70	52.78	49.75	52.24	47.27	52.34	50.69	48.70	51.73	50.87	48.76	50.05	47.77	53.7	51.70	49.26	52.93	52.68	49.81	51.12
FLAVA	Human	58.84	53.47	56.26	54.11	52.85	55.84	53.64	52.18	54.65	51.74	54.93	51.69	49.13	54.91	51.26	47.23	51.56	52.94	55.76	53.13	51.03	55.13	52.17	48.53	52.67	45.37	50.16	48.86	46.87	49.90	46.76	46.48	47.77	46.14	52.99	50.72	47.81	52.74	49.17	48.4	49.71
	NLLB-small	56.50	53.64	55.84	54.42	53.66	55.17	54.67	53.18	54.64	52.31	55.71	53.08	50.82	55.06	53.43	49.05	52.78	53.39	55.76	53.92	52.14	55.36	54.25	50.38	53.60	46.80	52.8	50.82	48.99	51.72	50.14	48.84	50.02	47.77	54.24	52.57	49.59	53.97	52.08	50.18	51.49
	NLLB-large	56.87	53.94	56.35	54.99	54.51	55.96	55.61	53.82	55.26	52.55	55.95	52.86	50.92	55.45	54.19	49.82	53.11	53.50	55.42	53.51	51.93	54.48	53.78	50.74	53.34	47.08	51.53	49.98	48.44	51.06	49.29	48.83	49.46	47.53	52.27	50.75	49.13	52.19	50.41	49.13	50.20
	M2M-small	55.80	53.16	54.83	53.66	53.11	54.51	53.59	52.66	53.92	51.94	54.43	52.68	50.37	54.11	52.66	49.65	52.26	52.39	54.55	53.22	51.33	54.44	52.85	50.37	52.74	48.57	52.96	50.83	48.95	52.53	50.63	49.17	50.52	48.74	53.45	52.13	49.26	53.20	51.90	49.63	51.19
	M2M-large	55.35	52.97	54.68	53.39	53.37	54.29	53.96	51.7	53.71	51.79	54	52.13	50.21	53.47	52.56	48.86	51.86	52.05	54.31	52.76	51.51	53.86	53.15	50.06	52.53	47.64	52.5	50.78	49.20	51.84	50.88	48.7	50.22	47.75	53.77	52.12	49.51	53.12	52.70	50.26	51.32

Table 16: Full results of translate-test evaluation with different MT systems for RT-translation and translating evaluation samples in target languages into English. The averaged results are in Table 5.

Models	RT Pivot	en	bn	de	id	ko	pt	ru	zh	avg.	bn	de	id	ko	pt	ru	zh	avg.
			GMT								**NLLB-3.3B**							
MUNITER	bn	54.08	51.89	53.96	52.46	52.35	53.01	52.60	52.30	52.65	52.54	53.84	52.85	51.20	52.97	52.85	50.30	52.36
	de	55.70	52.34	55.66	53.48	53.36	54.72	53.98	52.29	53.69	52.70	55.58	53.23	51.37	54.89	53.94	50.07	53.11
	id	55.13	52.24	54.44	54.02	53.47	54.31	53.93	52.94	53.62	53.08	54.55	54.66	51.85	54.35	53.84	51.14	53.35
	ko	53.90	51.30	53.27	52.31	52.81	53.07	52.79	52.11	52.52	51.65	53.19	52.54	52.12	53.24	52.59	50.68	52.29
	pt	56.23	52.43	55.49	53.67	52.72	55.72	54.00	52.97	53.86	53.08	55.52	53.74	51.59	55.97	54.44	50.39	53.53
	ru	55.29	51.53	54.33	52.82	53.21	53.88	53.52	52.56	53.12	52.39	54.48	53.47	51.59	54.23	54.60	50.52	53.04
	zh	53.06	50.99	52.77	51.96	52.07	52.30	51.92	52.05	52.01	49.53	50.75	49.88	48.88	50.49	49.67	49.58	49.83
XUNITER	bn	54.23	52.39	53.68	52.70	52.22	53.45	53.18	52.07	52.81	52.66	53.30	52.52	51.14	53.01	52.43	50.42	52.21
	de	55.22	52.10	54.97	52.66	52.51	54.18	52.85	52.23	53.07	51.95	54.74	52.62	51.01	54.25	52.90	49.68	52.45
	id	54.83	52.47	54.05	54.09	53.21	54.06	53.42	53.24	53.51	52.82	53.54	53.99	51.30	53.84	52.89	50.62	52.71
	ko	53.55	51.35	52.79	51.74	52.15	52.63	52.18	51.70	52.08	51.38	52.69	51.77	51.34	52.74	51.65	49.98	51.65
	pt	55.51	51.82	54.55	53.05	52.15	54.88	53.74	52.37	53.22	52.29	54.45	52.50	50.87	54.78	53.47	48.85	52.46
	ru	54.77	51.96	54.05	53.25	52.70	53.81	53.72	52.50	53.14	52.00	53.91	53.02	50.91	53.86	53.92	50.11	52.53
	zh	52.31	50.99	52.32	51.22	51.16	51.84	51.50	51.68	51.53	48.20	48.81	48.83	47.79	48.88	48.24	47.96	48.30
UC^2	bn	53.95	52.51	53.51	52.73	52.41	53.19	52.78	52.20	52.76	52.62	53.24	52.42	51.27	52.91	52.63	50.73	52.26
	de	55.12	52.35	55.10	53.29	53.07	54.17	53.36	52.73	53.44	52.42	54.70	53.10	51.14	54.05	53.59	49.71	52.67
	id	54.89	53.12	54.31	54.29	53.69	54.14	53.90	53.38	53.83	52.77	53.76	53.77	51.89	53.85	53.33	51.04	52.92
	ko	53.54	51.69	53.01	52.50	52.71	52.88	52.34	52.02	52.45	51.83	52.62	52.13	51.82	52.62	52.01	50.33	51.91
	pt	55.31	52.16	54.64	52.89	52.57	54.65	53.49	52.02	53.20	52.33	54.27	53.01	50.83	54.85	53.57	49.00	52.55
	ru	55.17	52.67	54.79	53.22	53.47	54.31	54.48	52.70	53.66	52.71	54.33	53.55	51.57	54.09	54.42	51.07	53.11
	zh	52.72	51.04	52.86	51.84	52.00	52.39	52.18	51.88	52.03	48.62	49.43	49.07	48.24	49.48	48.79	48.80	48.92
M^{3}P	bn	51.64	50.13	51.03	49.83	49.67	50.47	50.33	49.73	50.17	50.25	50.78	49.56	48.67	50.21	50.14	48.20	49.69
	de	52.97	50.63	53.03	51.42	50.38	52.11	51.80	50.41	51.40	50.86	52.90	50.52	49.29	52.27	51.43	48.20	50.78
	id	51.03	49.10	50.00	50.29	49.79	50.40	49.79	49.55	49.85	49.16	49.67	49.86	48.06	50.19	49.66	47.54	49.16
	ko	51.30	49.23	50.70	49.56	50.14	50.35	49.70	49.79	49.92	49.15	50.25	49.22	49.27	50.06	49.71	48.35	49.43
	pt	53.69	50.60	52.95	51.34	50.44	52.98	51.49	50.55	51.48	50.72	52.47	50.59	48.79	52.73	51.19	48.21	50.67
	ru	52.72	50.14	52.26	50.45	50.14	51.61	51.65	50.38	50.95	50.81	52.15	50.52	48.89	51.57	51.84	48.44	50.60
	zh	49.70	48.16	49.23	48.78	48.49	48.86	48.68	48.81	48.72	48.00	49.23	48.33	47.64	48.56	48.46	47.92	48.31
LXMERT	bn	51.94	50.42	51.33	50.47	50.36	51.16	50.58	50.10	50.63	50.72	51.20	50.76	49.33	51.09	50.99	48.76	50.41
	de	53.44	50.20	52.93	51.34	50.41	52.47	51.44	50.25	51.29	50.65	52.83	50.97	49.50	52.46	52.05	48.19	50.95
	id	53.01	50.44	51.76	52.16	51.29	52.11	51.46	50.99	51.46	50.61	51.78	51.77	49.86	52.19	51.47	49.47	51.02
	ko	51.84	49.61	51.08	50.63	50.72	51.00	50.39	50.20	50.52	50.09	51.01	50.63	50.33	50.87	50.01	48.77	50.24
	pt	53.82	50.42	53.02	51.64	50.53	53.07	51.69	50.56	51.56	50.65	52.76	51.01	49.43	53.32	51.61	48.02	50.97
	ru	53.03	49.72	52.08	50.99	50.75	51.84	51.22	50.42	51.00	50.54	51.99	51.11	49.75	51.91	52.14	48.79	50.89
	zh	51.02	49.09	50.09	49.90	49.76	50.12	49.85	49.93	49.82	46.67	47.45	46.98	46.57	47.54	46.75	46.84	46.97
UNITER	bn	54.15	52.09	53.26	52.75	52.07	52.66	52.68	51.96	52.50	52.62	52.90	52.89	50.98	52.99	52.74	50.65	52.25
	de	55.92	52.32	55.53	53.67	52.93	54.66	53.56	52.60	53.61	53.36	55.59	53.67	51.66	54.91	53.78	50.39	53.34
	id	55.61	52.66	54.33	54.78	54.00	54.38	53.76	53.47	53.91	52.92	54.42	54.33	52.16	54.26	54.23	50.74	53.29
	ko	53.90	51.46	52.91	51.90	52.54	52.81	52.06	51.99	52.24	51.85	52.98	52.15	52.18	53.10	52.38	50.77	52.20
	pt	56.38	52.16	55.22	53.61	52.58	55.29	53.94	52.73	53.65	53.03	55.44	53.30	51.47	55.86	54.36	49.72	53.31
	ru	55.94	51.55	54.60	52.97	52.92	54.05	53.87	52.55	53.22	52.42	54.62	53.68	51.76	54.76	54.76	50.60	53.23
	zh	53.12	50.72	52.19	51.90	52.04	52.30	51.60	52.02	51.82	48.72	49.71	49.26	48.36	49.71	48.94	48.09	48.97
VILBERT	bn	54.16	52.34	53.32	52.42	52.16	53.12	52.96	51.59	52.56	52.42	53.06	52.39	50.84	52.80	52.57	50.16	52.03
	de	55.22	52.23	54.85	53.43	52.75	54.26	53.69	52.22	53.35	52.58	54.80	52.84	51.40	54.13	53.36	49.73	52.69
	id	55.33	52.56	54.52	54.38	53.46	54.22	54.08	53.12	53.76	52.85	54.13	54.25	51.87	54.31	53.57	50.99	53.14
	ko	53.62	51.34	52.93	52.27	52.35	52.78	52.45	51.76	52.27	51.81	52.76	51.86	51.81	52.70	52.12	50.21	51.90
	pt	55.88	52.33	54.77	53.41	52.35	54.87	53.93	52.57	53.46	52.21	54.33	53.45	51.36	55.06	53.42	49.41	52.75
	ru	55.12	52.19	53.94	52.76	52.59	53.84	54.17	52.67	53.17	52.44	54.18	53.14	51.04	54.11	54.01	50.10	52.72
	zh	52.87	50.53	52.35	51.71	51.54	52.10	51.61	51.49	51.62	48.56	49.26	48.78	47.77	49.03	48.64	48.13	48.60
VisualBERT	bn	52.20	50.13	51.50	50.58	49.71	50.98	50.25	50.02	50.45	50.84	51.80	50.65	49.28	51.47	50.86	48.46	50.48
	de	53.51	50.57	53.10	51.17	50.45	52.59	51.47	50.97	51.47	50.92	53.29	50.90	49.10	52.42	51.60	48.25	50.93
	id	52.82	50.38	51.71	52.01	50.95	52.16	51.69	51.04	51.42	51.03	51.55	51.67	49.86	51.97	51.05	49.07	50.89
	ko	52.53	50.06	51.50	50.68	51.07	51.70	51.04	50.75	50.97	50.44	51.91	50.63	50.82	51.83	50.91	49.59	50.88
	pt	54.41	50.72	53.50	51.86	50.79	53.72	52.19	51.45	52.03	51.17	53.22	51.32	49.51	53.74	52.19	48.66	51.40
	ru	53.51	50.14	52.73	51.34	50.94	52.35	51.96	51.04	51.50	51.30	52.97	51.73	49.83	52.93	52.80	49.47	51.58
	zh	49.04	47.69	48.54	48.02	48.03	48.48	48.12	47.96	48.08	47.44	48.36	48.16	47.64	48.64	47.66	46.96	47.84
VL-BERT	bn	54.79	52.91	54.39	52.70	53.37	53.72	53.36	52.81	53.32	53.30	54.29	52.91	51.76	53.80	53.70	51.35	53.02
	de	55.61	52.38	55.27	53.43	52.58	54.63	53.32	52.31	53.42	52.84	55.06	53.13	51.47	54.55	53.78	49.58	52.92
	id	55.43	52.40	54.52	54.16	53.74	54.14	53.80	53.08	53.69	53.04	54.37	54.09	52.00	54.30	53.92	50.95	53.24
	ko	54.04	51.79	53.59	52.75	52.93	53.40	52.73	52.02	52.74	52.00	53.29	52.62	52.06	53.17	52.77	50.45	52.34
	pt	56.54	52.56	55.52	53.77	52.87	55.62	54.29	52.41	53.86	53.10	55.49	53.55	51.60	55.81	54.33	49.91	53.40
	ru	55.84	52.11	54.83	53.67	52.73	54.65	54.08	52.71	53.54	53.27	55.01	53.71	51.97	54.47	54.67	50.87	53.42
	zh	50.85	48.76	50.12	49.52	49.25	49.99	49.45	49.37	49.49	48.82	49.73	49.32	48.49	49.83	49.23	48.62	49.15
BLIP-2	bn	55.06	52.97	53.86	53.57	52.89	53.60	53.32	52.60	53.26	53.78	54.26	53.63	52.28	53.99	53.67	51.06	53.24
	de	56.11	53.18	55.70	53.98	53.51	55.11	54.25	53.31	54.15	53.74	56.07	53.70	51.96	55.15	54.37	50.80	53.68
	id	55.34	52.54	54.17	54.29	53.78	53.76	53.41	52.98	53.56	53.44	54.36	54.46	52.36	54.06	54.13	50.82	53.38
	ko	55.14	52.51	54.25	53.18	53.63	53.73	53.33	53.09	53.39	53.17	53.83	53.37	53.04	54.21	53.31	52.12	53.29
	pt	56.07	52.78	55.22	53.93	52.86	55.29	54.53	53.39	54.00	52.97	55.07	53.42	51.61	55.72	54.60	50.44	53.40
	ru	56.01	52.66	54.78	53.58	53.59	54.52	54.24	53.29	53.81	54.05	55.45	53.83	52.18	54.94	55.22	51.68	53.91
	zh	50.48	48.63	49.94	49.74	49.63	49.55	49.68	49.89	49.58	49.25	50.05	49.40	48.40	49.90	49.94	49.22	49.45
InstructBLIP	bn	55.21	52.92	54.01	53.38	53.17	53.93	53.23	53.00	53.38	53.68	54.77	53.75	51.69	54.38	53.80	51.65	53.39
	de	55.84	53.04	55.06	53.82	53.17	54.32	54.08	53.18	53.81	53.50	55.42	53.51	51.93	54.48	53.78	50.74	53.34
	id	55.97	52.96	54.61	54.46	54.13	54.64	54.56	53.67	54.15	53.67	54.85	54.86	52.85	54.73	54.73	51.67	53.91
	ko	54.62	51.93	53.67	52.91	53.24	53.27	53.29	53.16	53.07	52.76	53.82	53.08	52.90	53.39	52.87	51.63	52.92
	pt	56.64	52.85	55.54	54.07	52.98	55.61	54.75	53.45	54.18	53.77	55.39	53.78	51.90	56.09	54.87	50.67	53.78
	ru	56.27	52.39	54.85	53.50	53.39	54.34	54.00	52.94	53.63	53.81	55.29	54.13	52.08	54.87	54.77	51.15	53.73
	zh	51.66	50.22	51.06	50.57	50.46	50.54	50.33	50.40	50.51	50.02	51.03	50.51	49.43	50.83	50.21	49.40	50.20
FLAVA	bn	55.70	54.08	55.14	54.20	54.09	54.63	54.60	54.00	54.39	54.13	54.70	53.77	52.67	54.25	54.06	51.94	53.65
	de	56.94	53.97	56.41	54.99	54.52	55.95	55.70	53.87	55.06	54.08	56.28	54.01	52.77	55.66	55.14	51.30	54.18
	id	56.26	53.82	55.36	55.27	54.63	55.29	54.86	54.01	54.75	54.05	55.40	55.14	52.97	55.14	54.54	51.92	54.17
	ko	55.42	53.54	54.42	54.17	54.43	54.64	54.25	53.70	54.16	53.39	54.40	53.67	53.65	54.66	53.91	52.03	53.67
	pt	57.17	53.79	56.07	54.73	53.67	56.38	55.10	53.75	54.78	54.07	55.91	54.16	52.50	56.33	54.76	50.90	54.09
	ru	55.57	53.26	54.95	53.90	53.61	54.87	54.79	53.50	54.13	53.28	54.79	53.88	52.19	54.68	54.88	50.91	53.52
	zh	51.93	49.79	51.40	50.85	50.22	51.11	50.40	50.76	50.65	49.89	50.78	50.12	49.19	50.64	50.02	49.59	50.03

Table 17: Translate-test results with varied pivot languages during RT translation. The averaged results are in Table 4.

Models	RT?	en	Translate-Test						
			fr	hi	ro	th	yi	zh	avg.
MUNITER		61.73	60	72.73	63.33	65.79	55.13	53.85	61.81
	✓	58.02	64	76.14	62.22	61.84	46.15	65.38	62.62
XUNITER		60.49	61.33	79.55	65.56	57.89	44.87	63.46	62.11
	✓	54.32	62.67	79.55	62.22	56.58	48.72	61.54	61.88
UC^2		60.49	60.00	68.18	58.89	56.58	51.28	59.62	59.09
	✓	51.85	64.00	75.00	60.00	56.58	57.69	57.69	61.83
M^3P		59.26	62.67	76.14	66.67	53.95	43.59	61.54	60.76
	✓	64.20	66.67	75.00	70.00	61.84	53.85	59.62	64.50
LXMERT		64.2	72.00	75.00	61.11	52.63	46.15	55.77	60.44
	✓	64.20	72.00	79.55	65.56	61.84	55.13	63.46	66.26
UNITER		61.73	70.67	76.14	67.78	59.21	46.15	63.46	63.90
	✓	61.73	68	76.14	63.33	56.58	58.97	57.69	63.45
VILBERT		60.49	66.67	75.00	66.67	63.16	48.72	61.54	63.63
	✓	62.96	66.67	76.14	63.33	60.53	53.85	59.62	63.36
VisualBERT		69.14	70.67	71.59	60.00	56.58	51.28	59.62	61.62
	✓	70.37	70.67	76.14	60.00	68.42	57.69	65.38	66.38
VL-BERT		35.80	44.00	56.82	37.78	48.68	47.44	48.08	47.13
	✓	50.62	48.00	52.27	40.00	44.74	46.15	42.31	45.58
BLIP-2		60.49	72.00	72.73	76.67	65.79	71.79	57.69	69.45
	✓	60.49	72.00	69.32	70.00	64.47	67.95	67.31	68.51
InstructBLIP		64.20	73.33	73.86	72.22	65.79	71.79	57.69	69.11
	✓	62.96	69.33	69.32	70.00	61.84	65.38	61.54	66.24
FLAVA		64.20	70.67	80.68	76.67	75.00	65.38	67.31	72.62
	✓	65.43	76.00	76.14	75.56	75.00	67.95	71.15	73.63

Table 18: Full results on MaXM dataset. The averaged results across different models are in Table 6.

Models		en	Translate-Test bn	de	id	ko	pt	ru	zh	**avg.**
MUNITER	Human	57.33	50.67	54.09	52.54	50.67	54.21	49.69	49.57	51.63
	MT	55.70	52.34	55.66	53.48	53.36	54.72	53.98	52.29	53.69
	MERGE	57.12	52.79	55.99	53.78	53.7	55.15	53.94	52.97	54.05
	TAG	57.08	52.92	56.21	54.68	53.48	55.72	54.85	53.18	54.43
XUNITER	Human	56.98	50.76	54.63	52.37	50.52	54.24	48.91	49.94	51.62
	MT	55.22	52.10	54.97	52.66	52.51	54.18	52.85	52.23	53.07
	MERGE	56.69	52.5	55.45	53.55	53.07	54.83	53.71	52.61	53.67
	TAG	56	52.2	55.00	53.12	52.62	54.61	53.54	52.04	53.30
UC^2	Human	56.85	51.34	54.01	52.35	50.75	53.81	51.93	50.04	52.03
	MT	55.12	52.35	55.10	53.29	53.07	54.17	53.36	52.73	53.44
	MERGE	57.67	53.84	56.59	54.87	54.48	56.11	55.08	53.45	54.92
	TAG	56.7	53.24	55.95	54.01	53.59	55.48	54.95	53.11	54.33
M^3P	Human	54.45	49.18	52.14	49.87	48.59	51.87	49.05	48.38	49.87
	MT	52.97	50.63	53.03	51.42	50.38	52.11	51.80	50.41	51.40
	MERGE	53.7	50.37	52.85	50.89	50.36	51.88	51.22	50.43	51.14
	TAG	54.66	51.11	53.71	51.66	50.78	53.12	52.38	51.27	52.00
LXMERT	Human	55.40	49.64	52.83	50.80	49.17	52.49	47.54	48.02	50.07
	MT	53.44	50.20	52.93	51.34	50.41	52.47	51.44	50.25	51.29
	MERGE	54.88	50.78	53.59	52.01	51.28	53.04	52.31	50.68	51.96
	TAG	54.75	51.03	53.82	52.17	51.26	53.24	52.15	51.2	52.12
UNITER	Human	57.47	51.74	54.52	52.79	51.27	54.56	52.27	50.33	52.50
	MT	55.92	52.32	55.53	53.67	52.93	54.66	53.56	52.6	53.61
	MERGE	57.26	52.97	56.19	54.05	53.65	55.53	54.44	53.1	54.28
	TAG	57.03	52.71	55.96	54.21	53.16	55.45	54.48	52.74	54.10
VILBERT	Human	56.72	50.84	54.10	52.27	50.73	53.98	49.91	49.92	51.68
	MT	55.22	52.23	54.85	53.43	52.75	54.26	53.69	52.22	53.35
	MERGE	56.97	53.01	55.46	53.73	53.54	55.05	54.33	53.05	54.02
	TAG	56.67	53.04	55.72	54.21	53.73	55.42	54.65	52.84	54.23
VisualBERT	Human	55.17	49.43	52.58	50.34	48.66	52.72	50.50	48.89	50.45
	MT	53.51	50.57	53.10	51.17	50.45	52.59	51.47	50.97	51.47
	MERGE	54.79	51.07	53.43	51.91	51.36	53.15	52.19	51.49	52.09
	TAG	54.92	51.28	53.91	52.15	51.07	53.70	51.2	51.33	52.09
VL-BERT	Human	57.79	51.22	54.47	52.62	50.94	54.79	51.17	50.02	52.18
	MT	55.61	52.38	55.27	53.43	52.58	54.63	53.32	52.31	53.42
	MERGE	57.45	52.71	55.80	53.49	53.62	54.88	54.09	52.17	53.82
	TAG	57.49	53.63	56.16	54.25	53.8	55.82	54.65	53.24	54.51
BLIP	Human	58.05	52.03	54.70	52.99	51.57	54.91	52.36	51.22	52.83
	MT	56.11	53.18	55.70	53.98	53.51	55.11	54.25	53.31	54.15
	MERGE	57.41	53.6	56.26	54.33	53.83	55.94	54.52	53.76	54.61
	TAG	57.31	53.62	56.23	54.33	53.98	55.72	55.14	53.78	54.69
InstructBLIP	Human	57.85	51.80	54.91	53.01	51.29	54.85	53.16	51.34	52.91
	MT	55.84	53.04	55.06	53.82	53.17	54.32	54.08	53.18	53.81
	MERGE	58.1	54.26	57.08	55.16	54.15	56.27	55.59	54.18	55.24
	TAG	58.24	54.65	57.20	55.06	54.52	56.69	55.79	54.32	55.46
FLAVA	Human	58.84	53.47	56.26	54.11	52.85	55.84	53.64	52.18	54.05
	MT	56.87	53.94	56.35	54.99	54.51	55.96	55.61	53.82	55.03
	MERGE	57.95	53.95	56.61	54.99	54.33	56.08	55.22	53.91	55.01
	TAG	57.44	54.21	56.56	55.18	54.51	55.95	55.42	53.73	55.08

Table 19: Full results of data augmentation experiments. The averaged results across different models are in Table 6.

Image	Question	Predictions
	**KO:** 소파 오른쪽에 있는 장치는 무엇입니까?	**Answer: speaker**
	EN: What is the device to the right of the couch?	speaker (0.32) || printer (0.3) || computer (0.06)
	KO→EN: What's the device on the right side of the sofa?	printer (0.3) || speaker (0.22) || computer (0.08)
	**KO:** 셔츠와 배가 같은 색인가요?	**Answer: yes**
	EN: Are both the shirts and the boats the same color?	yes (0.6) || no (0.4) || gray (0.0)
	KO→EN: Are the shirt and belly the same color?	no (0.53) || yes (0.47) || gray (0.0)
	**KO:** 비어 있지 않은 가방이 침대 위에 놓여 있습니까?	**Answer: no**
	EN: Is the bag that is not empty sitting on top of a bed?	no (0.99) ||| yes (0.01) ||| couch (0.0)
	KO→EN: Is there a non-empty bag lying on the bed?	yes (0.99) || no (0.01) ||| hat (0.0)
	**KO:** 이 사진의 울타리 근처에 얼룩말이 보이십니까?	**Answer: no**
	EN: Do you see a zebra near the fence in this photo?	no (1.0) || yes (0.0) || lady (0.0)
	KO→EN: See the zebra near the fence in this photo?	yes (0.71) || no (0.29) || hat (0.0)
	**KO:** 스케이트보드와 지붕의 재질이 동일합니까?	**Answer: no**
	EN: Do the skateboard and the rooftop have the same material?	no (0.57) || yes (0.43) || chairs (0.0)
	KO→EN: Are skateboards and roofs the same material?	yes (0.8) || no (0.2) || chairs (0.0)
	**KO:** 사람이 타고 있습니까?	**Answer: yes**
	EN: Is the person riding?	yes (0.99) || no (0.01) || couch (0.0)
	KO→EN: Is anyone riding?	girl (0.32) || woman (0.22) || man (0.13)
	**KO:** 어두운 차량 뒤에 출입구가 있습니까?	**Answer: yes**
	EN: Is the doorway behind the dark vehicle?	yes (0.9) || no (0.1) || chairs (0.0)
	KO→EN: Is there a doorway behind a dark vehicle?	no (0.59) || yes (0.41) || couch (0.0)

Figure 11: We present a randomly selected example, which includes the original English text (**EN**), its target language translation by a human annotator (e.g., **KO**), and translation from the target language to English (e.g., **KO** → **EN**) for translate-test. For each example, we provide the correct English answer, the top three English predictions, and the top three predictions from the translate-test, along with their respective probabilities of UC^2. In the translate-test, the examples with translation errors are specifically identified, with the type of error highlighted in **red**.

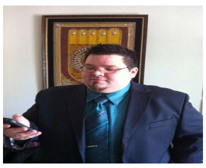
<table>
<tr><th>Image</th><th>Question</th><th>Predictions</th></tr>
<tr><td rowspan="3"></td><td>KO: 금속 울타리 뒤에 키 큰 나무가 자라고 있습니까?</td><td>Answer: yes</td></tr>
<tr><td>EN: Are the tall trees growing behind the metal fence?</td><td>yes (0.97) || no (0.03) || chairs (0.0)</td></tr>
<tr><td>KO→EN: Are there tall trees growing behind metal fences?</td><td>no (0.67) || yes (0.33) || couch (0.0)</td></tr>
<tr><td rowspan="3"></td><td>KO: 접시는 소녀의 왼쪽에 있습니까?</td><td>Answer: no</td></tr>
<tr><td>EN: Is the plate to the left of a girl?</td><td>no (1.0) || yes (0.0) || bananas(0.0)</td></tr>
<tr><td>KO→EN: Is the plate on the girl's left?</td><td>yes (0.94) || no (0.06) || couch (0.0)</td></tr>
<tr><td rowspan="3"></td><td>KO: 어떤 장치가 켜져 있습니까?</td><td>Answer: laptop</td></tr>
<tr><td>EN: What device is on?</td><td>laptop (0.39) || monitor (0.21) || screen (0.16)</td></tr>
<tr><td>KO→EN: Which device is turned on?</td><td>Keyboard (0.54) || laptop (0.12) || computer (0.1)</td></tr>
<tr><td rowspan="3"></td><td>KO: 바닥에 붉게 보이는 책이 있습니까?</td><td>Answer: no</td></tr>
<tr><td>EN: Are there books on the floor that looks red?</td><td>no (0.81) || yes (0.19) || bananas(0.0)</td></tr>
<tr><td>KO→EN: Are there any books that look red on the floor?</td><td>yes (0.82) || no (0.18) || hat (0.0)</td></tr>
<tr><td rowspan="3"></td><td>KO: 쿠키 뒤에 테이프가 있습니까?</td><td>Answer: yes</td></tr>
<tr><td>EN: Is the tape behind the cookie?</td><td>yes (1.0) || no (0.0) || train (0.0)</td></tr>
<tr><td>KO→EN: Is there a tape behind the cookie?</td><td>no (0.59) || yes (0.41) || gray (0.0)</td></tr>
<tr><td rowspan="3"></td><td>KO: 양복이 검고 더럽습니까?</td><td>Answer: no</td></tr>
<tr><td>EN: Is the suit both black and dirty?</td><td>no (0.55) ||| yes (0.45) ||| couch (0.0)</td></tr>
<tr><td>KO→EN: Is your suit black and dirty? (Mistranslation)</td><td>yes (0.62) || no (0.38) ||| couch (0.0)</td></tr>
<tr><td rowspan="3"></td><td>KO: 땅 위에 어떤 동물이 있습니까?</td><td>Answer: elephant</td></tr>
<tr><td>EN: What animal is above the ground?</td><td>elephant (1.0) || elephants (0.0) || rhino (0.0)</td></tr>
<tr><td>KO→EN: What animals are on the ground? (Mistranslation)</td><td>elephants (0.97) || birds (0.01) || bears (0.0)</td></tr>
</table>

Figure 12: *(cont'd)* We present a randomly selected example, which includes the original English text (**EN**), its target language translation by a human annotator (e.g., **KO**), and translation from the target language to English (e.g., **KO**) for translate-test. For each example, we provide the correct English answer, the top three English predictions, and the top three predictions from the translate-test, along with their respective probabilities of UC^2. In translate-test, examples with translation errors are specifically identified, with the type of error highlighted in **red**.

<table>
<tr><th>Image</th><th>Question</th><th>Predictions</th></tr>
<tr><td rowspan="3"></td><td>KO: 어떤 가구 항목이 흰색입니까?</td><td>Answer: chair</td></tr>
<tr><td>EN: What item of furniture is white?</td><td>chair (0.91) || couch (0.05) || armchair (0.01)</td></tr>
<tr><td>KO→EN: Which furniture items are white? (Mistranslation)</td><td>chairs (0.97) || tables (0.01) || couches (0.0)</td></tr>
<tr><td rowspan="3"></td><td>KO: 싱크대는 무엇입니까?</td><td>Answer: porcelain</td></tr>
<tr><td>EN: What's the sink made of?</td><td>porcelain (0.97) || glass (0.01) || plastic (0.01)</td></tr>
<tr><td>KO→EN: What is a sink? (Omission)</td><td>bathroom (0.22) ||| bathtub (0.21) ||| shower (0.19)</td></tr>
<tr><td rowspan="3"></td><td>KO: 바닥이 변기 아래에 있습니까?</td><td>Answer: no</td></tr>
<tr><td>EN: Is the floor below a toilet?</td><td>no (1.0) || yes (0.0) || cloudless (0.0)</td></tr>
<tr><td>KO→EN: Is the floor under the toilet bowl? (Addition)</td><td>yes (1.0) || no (0.0) || left (0.0)</td></tr>
<tr><td rowspan="3"></td><td>KO: 이미지의 어느 부분에 가죽 소파가 있습니까?</td><td>Answer: right</td></tr>
<tr><td>EN: In which part of the image is the leather couch?</td><td>right (1.0) || left (0.0) || bottom (0.0)</td></tr>
<tr><td>KO→EN: Where in the image is the leather sofa? (Mistranslation)</td><td>living room (0.74) || floor (0.23) || bedroom (0.01)</td></tr>
<tr><td rowspan="3"></td><td>KO: 작은 깃발이나 연이 있습니까?</td><td>Answer: no</td></tr>
<tr><td>EN: Are there any small flags or kites?</td><td>no (0.79) || yes (0.21) || hat (0.0)</td></tr>
<tr><td>KO→EN: Where in the image is the leather sofa? (Mistranslation)</td><td>yes (0.7) || no (0.3) || hat (0.0)</td></tr>
<tr><td rowspan="3"></td><td>KO: 냉동고가 있는 바닥 위에 캐비닛이 보이십니까?</td><td>Answer: yes</td></tr>
<tr><td>EN: Do you see a cabinet above the floor the freezer is on?</td><td>yes (0.82) || no (0.18) || gray (0.0)</td></tr>
<tr><td>KO→EN: See the cabinet above the floor where the freezer is? (Fluency)</td><td>no (0.52) || yes (0.48) || hat (0.0)</td></tr>
<tr><td rowspan="3"></td><td>KO: 어떤 종류의 조리 도구가 구부러져 있습니까?</td><td>Answer: cutting board</td></tr>
<tr><td>EN: Which kind of cooking utensil is curved?</td><td>cutting board (0.38) || coffee pot (0.07) || pan (0.06)</td></tr>
<tr><td>KO→EN: What kind of cookware are bent? (Grammar)</td><td>tongs (0.54) || burger (0.04) || potatoes (0.01)</td></tr>
</table>

Figure 13: *(cont'd)* We present a randomly selected example, which includes the original English text (**EN**), its target language translation by a human annotator (e.g., **KO**), and translation from the target language to English (e.g., **KO** $\rightarrow$ **EN**) for translate-test. For each example, we provide the correct English answer, the top three English predictions, and the top three predictions from the translate-test, along with their respective probabilities of UC^2. In translate-test, examples with translation errors are specifically identified, with the type of error highlighted in **red**.

Image	Question	Predictions
	**DE:** Gibt es rechts neben dem gelben Getränk einen Mixer?	**Answer: yes**
	EN: Is there a blender to the right of the yellow drink?	yes (0.52) || no (0.48) || hat (0.0)
	DE→EN: Is there a mixer to the right of the yellow drink?	no (0.74) || yes (0.26) || bananas (0.0)
	**DE:** Sieht das Fahrzeug hinter den Zebras schwarz aus?	**Answer: no**
	EN: Does the vehicle behind the zebras look black?	no (0.66) || yes (0.34) || couch (0.0)
	DE→EN: Does the vehicle look black behind the zebras?	yes (0.57) ||| no (0.43) ||| hat (0.0)
	**DE:** Scheint das Hemd ärmellos oder langärmelig zu sein?	**Answer: long sleeved**
	EN: Does the shirt seem to be sleeveless or long sleeved?	long sleeved (0.60) || sleeveless (0.33) || short sleeved (0.04)
	DE→EN: Does the shirt appear sleeveless or long sleeved?	sleeveless (0.48) || long sleeved (0.45) || short sleeved (0.03)
	**DE:** Sind der Pullover und das schwarze Hemd beide langärmlig?	**Answer: yes**
	EN: Are the sweater and the black dress shirt both long sleeved?	yes (1.0) || no (0.0) || airplanes (0.0)
	DE→EN: Are the sweater and black shirt both long-sleeved? ***(Omission)***	no (0.53) || yes (0.47) || couch (0.00)
	**DE:** Scheint der Mann links neben dem anderen Mann zu stehen?	**Answer: no**
	EN: Does the man that is to the left of the other man seem to be stan ding?	no (0.79) || yes (0.21) || hat (0.0)
	DE→EN: Does the man appear to be to the left of the other man? ***(Omission)***	yes (0.7) || no (0.3) || hat (0.0)
	**DE:** Was macht er da?	**Answer: sleeping**
	EN: What is he doing?	sleeping (0.47) || lying (0.43) || resting (0.04)
	DE→EN: What is he doing there? ***(Addition)***	lying (0.46) || sleeping (0.43) || resting (0.04)
	**DE:** Was macht der Mann?	**Answer: jumping**
	EN: What's the man doing?	jumping (0.63) || playing (0.10) || skating(0.05)
	DE→EN: What does the man? ***(Grammar)***	skateboard (0.05) || swimming pool (0.04) || water (0.03)

Figure 14: *(cont'd)* We present a randomly selected example, which includes the original English text (**EN**), its target language translation by a human annotator (e.g., **DE**), and translation from the target language to English (e.g., **DE** $\rightarrow$ **EN**) for translate-test. For each example, we provide the correct English answer, the top three English predictions, and the top three predictions from the translate-test, along with their respective probabilities of UC^2. In translate-test, examples with translation errors are specifically identified, with the type of error highlighted in **red**.

Image	Question	Predictions
	**FR: Un homme porte quelque chose.**	**Answer: Yes (entailment)**
	EN: Based on the provided image, evaluate whether the follow ing statement is true or false: {**A man is wearing something.**} Respond with 'Yes' if the statement is true, 'No' if it is false, or 'Maybe' if it is uncertain. The answer is:	Yes (entailment)
	FR →EN: Based on the provided image, evaluate whether the following statement is true or false: {**A man is carrying something.**} *(Mistranslation)* Respond with 'Yes' if the statement is true, 'No' if it is false, or 'Maybe' if it is uncertain. The answer is:	No (contradiction)
	**RU: Ребенок играет.**	**Answer: Yes (entailment)**
	EN: Based on the provided image, evaluate whether the following statement is true or false: {**A child is playing.**} Respond with 'Yes' if the statement is true, 'No' if it is false, or 'Maybe' if it is uncertain. The answer is:	Yes (entailment)
	RU →EN: Based on the provided image, evaluate whether the following statement is true or false: {**The child is playing.**} Respond with 'Yes' if the statement is true, 'No' if it is false, or 'Maybe' if it is uncertain. The answer is:	Maybe (neutral)
	**RU: Эти два мальчика плавают со своими досками для серфинга.**	**Answer: Yes (entailment)**
	EN: Based on the provided image, evaluate whether the following statement is true or false: {**The two boys are swimming with their floats.**} Respond with 'Yes' if the statement is true, 'No' if it is false, or 'Maybe' if it is uncertain. The answer is:	Yes (entailment)
	RU →EN: Based on the provided image, evaluate whether the following statement is true or false: {**The two boys are swimming with their surfboards.**} *(Mistranslation)* Respond with 'Yes' if the statement is true, 'No' if it is false, or 'Maybe' if it is uncertain. The answer is:	No (contradiction)
	**FR: Deux hommes tristes montant sur un échafaudage en bois.**	**Answer: Maybe (neutral)**
	EN: Based on the provided image, evaluate whether the following statement is true or false: {**Two sad men climbing on a wooden scaffold.**} Respond with 'Yes' if the statement is true, 'No' if it is false, or 'Maybe' if it is uncertain. The answer is:	Maybe (neutral)
	FR →EN: Based on the provided image, evaluate whether the following statement is true or false: {**Two sad men climbing on a wooden scaffolding.**} Respond with 'Yes' if the statement is true, 'No' if it is false, or 'Maybe' if it is uncertain. The answer is:	No (contradiction)

Figure 15: Sample results with `gpt-4-1106-vision-preview`. For each example, we present the original question written in the target language along with its answer (e.g., **FR**), the original question written in English and corresponding model prediction (i.e., **EN**), and the translated question from the target language and model prediction (e.g., **FR** → **EN**). Each question is given with a task description and is highlighted in **bold**. Any translation errors in translated questions are further highlighted in red.

Image	Question	Predictions
	**BN: টেবিলের উপরের বাসনপত্র কি পরিষ্কার দেখাচ্ছে এবং কালো?**	**Answer: No**
	EN: Based on the provided image, evaluate whether the following statement is true or false: **{Does the utensil on top of the table look clean and black?}** Respond with 'Yes' if the statement is true or 'No' if it is false. The answer is:	No
	BN→EN: Based on the provided image, evaluate whether the following statement is true or false: **{Does the tableware look clean and black?}** ***(Mistranslation)*** Respond with 'Yes' if the statement is true or 'No' if it is false. The answer is:	Yes
	**DE: Gibt es rechts neben dem gelben Getränk einen Mixer?**	**Answer: Yes**
	EN: Based on the provided image, evaluate whether the following statement is true or false: **{Is there a blender to the right of the yellow drink?}** Respond with 'Yes' if the statement is true or 'No' if it is false. The answer is:	No
	DE→EN: Based on the provided image, evaluate whether the following statement is true or false: **{Is there a mixer to the right of the yellow drink?}** Respond with 'Yes' if the statement is true or 'No' if it is false. The answer is:	Yes
	**ID: Apakah terdapat sikat gigi dan keset di gambar ini?**	**Answer: No**
	EN: Based on the provided image, evaluate whether the following statement is true or false: **{Are there both toothbrushes and mats in this picture?}** Respond with 'Yes' if the statement is true or 'No' if it is false. The answer is:	Yes
	ID→EN: Based on the provided image, evaluate whether the following statement is true or false: **{Is there a toothbrush and mat in this picture?}** ***(Grammar)*** Respond with 'Yes' if the statement is true or 'No' if it is false. The answer is:	No
	**KO: 하늘이 비행기 위에 있습니까?**	**Answer: No**
	EN: Based on the provided image, evaluate whether the following statement is true or false: **{Is the sky above an airplane?}** Respond with 'Yes' if the statement is true or 'No' if it is false. The answer is:	No
	KO→EN: Based on the provided image, evaluate whether the following statement is true or false: **{Is the sky above the plane?}** Respond with 'Yes' if the statement is true or 'No' if it is false. The answer is:	Yes

Figure 16: *(cont'd)* Sample results with `gpt-4-1106-vision-preview`. For each example, we present the original question written in the target language along with its answer (e.g., **BN**), the original question written in English and corresponding model prediction (i.e., **EN**), and the translated question from the target language and model prediction (e.g., **BN** → **EN**). Each question is given with a task description and is highlighted in **bold**. Any translation errors in translated questions are further highlighted in red.