Findings of WMT2024 English-to-Low Resource Multimodal Translation Task

Shantipriya Parida[1], **Ondřej Bojar**[2], **Idris Abdulmumin**[3], **Shamsuddeen Hassan Muhammad**[4], **Ibrahim Said Ahmad**[5]

[1]Silo AI, Finland; [2]Charles University, MFF, ÚFAL, Czech Republic;
[3]Data Science for Social Impact Research Group, University of Pretoria, South Africa;
[4]Imperial College London, UK; [5]Institute for Experimental AI, Northeastern University, USA
correspondence: shantipriya.parida@silo.ai

Abstract

This paper presents the results of the English-to-Low Resource Multimodal Translation shared tasks from the Ninth Conference on Machine Translation (WMT2024). This year, 7 teams submitted their translation results for the automatic and human evaluation.

1 Introduction

The Ninth Conference on Machine Translation (WMT24), held in conjunction with EMNLP 2024, hosted a number of shared tasks covering various aspects of machine translation (MT). This conference builds on 17 previous editions of WMT as a workshop or a conference. This year, Workshop on Asian Translation (WAT), the most recognized shared task campaign on Asian languages, merged with WMT, adding many new shared tasks to the venue.

Multi-modal translation, which involves incorporating non-text sources alongside text input for machine translation, has gained attention in the past years (Specia et al., 2016; Elliott et al., 2016). However, research in this area has focused on European languages such as English, German, French, Czech, and mainly used two datasets: Flickr30k (Young et al., 2014) and MS-COCO (Lin et al., 2014), where the text caption corresponds to the content of the associated image.

We organized the WMT2024 English-to-LowRes Multimodal Shared Task for Low-Resource Asian and African languages. One important difference is that in our setting, the text caption is attached to a rectangular region of the picture and not the picture as a whole. This approach provides an interesting opportunity to consider not only the broader image but also the localized visual context surrounding the described region, which may provide additional cues for more accurate translation.

2 Task and Datasets

In this task, participants were provided with corpora from the Visual Genome dataset in four target language: Hindi, Bengali, Malayalam, and Hausa. The specific datasets are: Hindi Visual Genome 1.1 (HVG, Parida et al., 2019)[1] for Hindi; Bengali Visual Genome (BVG, Sen et al., 2022)[2] for Bengali; Malayalam Visual Genome (MVG, Parida and Bojar, 2021)[3] for Malayalam; and Hausa Visual Genome (HaVG, Abdulmumin et al., 2022)[4] for Hausa. The datasets are split into train, test, dev and challenge test in a parallel fashion. The number of sentences in each split is provided in Table 1. Each split contains items consisting of an image, a highlighted rectangular region within the image $(x, y, width, height)$, the original English caption for this region, and the reference translation in the respective target language. These components are illustrated in Figure 1. Depending on the task track, some of these components serve as the source, while others act as references or competing candidate solutions. The specific tracks for this task are listed below.

2.1 Text-Only Translation

Labeled "TEXT" in WAT official tables, participants translate short English captions into the target language without using visual information. Additional textual resources are allowed but must be documented in the system description paper.

2.2 Captioning

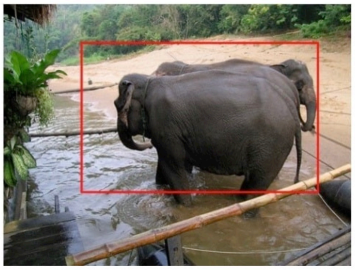
Labeled with the target language code, e.g., "HI," "BN," "ML," "HA", participants generate captions

[1]https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3267
[2]https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3722
[3]https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3533
[4]https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-4749

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 677–683
November 15–16, 2024.

data split	size
train	28,930
dev	998
test	1,595
challenge test	1,400

Table 1: Shared task dataset splits



Image_id: 2340461
x: 111 y: 54 width: 342 height: 216

English Text: Two elephants standing in the water
Hindi Text: पानी में खड़े दो हाथी
Bengali Text: পানিতে দাঁড়িয়ে দুটি হাতি
Malayalam Text: വെള്ളത്തിൽ നിൽക്കുന്ന രണ്ട് ആനകൾ
Hausa Text: giwaye biyu da ke tsaye a cikin ruwa

Figure 1: Example of a Data Point (Image Id, Region Detail, Source, and Target Languages

in the target language for the highlighted rectangular region in the input image.

2.3 Multi-Modal Translation

Labeled "MM", given an image, a rectangular region within it, and an English caption for that region, participants translate the caption into the target language. Both textual and visual information are available for this task.

3 Evaluation Methods

3.1 Automatic Evaluation

We evaluated translation results by two metrics: BLEU (Papineni et al., 2002), and RIBES (Isozaki et al., 2010). BLEU scores were calculated using SacreBLEU (Post, 2018). RIBES scores were calculated using `RIBES.py` version 1.02.4.[5] All scores for each task were calculated automatically using the corresponding reference translations by the evaluation system through which the participants make their submissions.

[5] http://www.kecl.ntt.co.jp/icl/lirg/ribes/index.html

Automatic Evaluation System The automatic evaluation system receives translation results by participants and automatically gives evaluation scores to the uploaded results. As shown in Figure 2, the system requires participants to provide the following information for each submission:

- Human Evaluation: whether or not they submit the results for human evaluation;
- Publish the results of the evaluation: whether or not they permit to publish automatic evaluation scores on the WAT2024 web page;
- Task: the task to which the results belong;
- Used Other Resources: whether or not they used additional resources; and
- Method: the type of the method including SMT, RBMT, SMT and RBMT, EBMT, NMT and Other.

Evaluation scores of translation results that participants permit to be published are disclosed via the WAT2024 evaluation web page. Participants can also submit the results for human evaluation using the same web interface. This automatic evaluation system will remain available even after WMT-WAT2024.

3.2 Human Evaluation

In WMT2024, we conducted **JPO adequacy evaluation**.

JPO adequacy evaluation The evaluation was carried out by translation experts based on the JPO adequacy evaluation criterion, which was originally defined by Japan Patent Office to assess the quality of translated patent documents.

Sentence selection and evaluation For the JPO adequacy evaluation, the 200 test sentences were randomly selected from the test sentences. For each test sentence, input source sentence, translation by participants' system, and reference translation were shown to the annotators. To guarantee the quality of the evaluation, each sentence was evaluated by two annotators. Note that the selected sentences are basically the same as those used in the previous workshop.

Evaluation Criterion Table 2 shows the JPO adequacy criterion from 5 to 1. The evaluation is performed subjectively. "Important information" represents the technical factors and their relationships. The degree of importance of each element is also considered in evaluating. The percentages

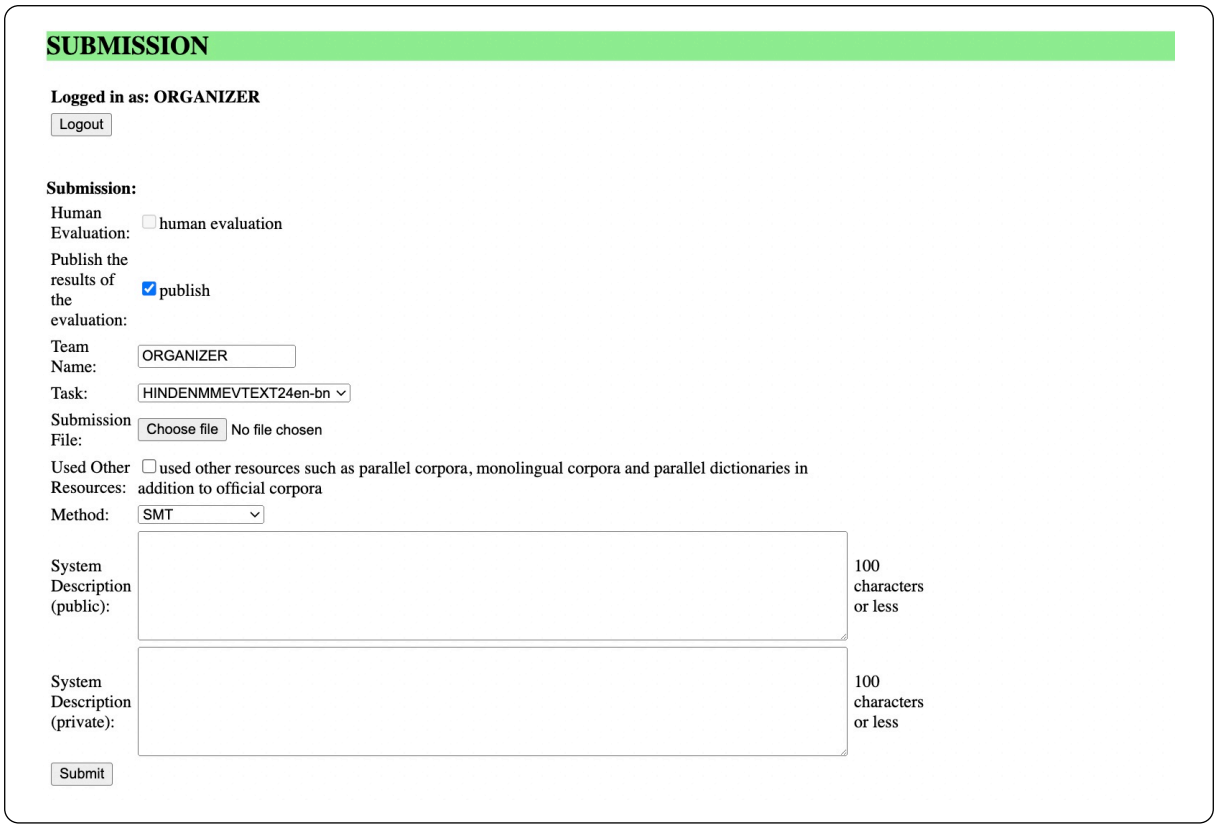


Figure 2: The interface for translation results submission

Sccore	Description
5	All important information is transmitted correctly. (100%)
4	Almost all important information is transmitted correctly. (80%–)
3	More than half of important information is transmitted correctly. (50%–)
2	Some of important information is transmitted correctly. (20%–)
1	Almost all important information is NOT transmitted correctly. (–20%)

Table 2: The JPO adequacy criterion

in each grade are rough indications for the transmission degree of the source sentence meanings. The detailed criterion is described in the JPO document (in Japanese).[6]

4 Baseline Systems

Human evaluations were conducted as pairwise comparisons between the translation results for a specific baseline system and translation results for each participant's system. That is, the specific baseline system served as the standard for human evaluation.

At WMT2024, we adopted some of neural machine translation (NMT) as baseline systems. The NMT baseline systems consisted of publicly available software, and the procedures for building the systems and for translating using the systems were published on the WAT web page.

Tokenization The shared task datasets come untokenized and we did not use or recommend any specific external tokenizer. The standard OpenNMT-py sub-word segmentation was used for pre/post-processing for the baseline system and each participant used what they wanted.

NMT Methods We used the NMT models for all tasks. For the English→Hindi, English→Malayalam, and English→Bengali Multimodal tasks we used the Transformer model (Vaswani et al., 2018) as implemented in OpenNMT-py (Klein et al., 2017) and used the "base" model with default parameters for the multimodal task baseline. We have generated the vocabulary of 32k sub-word types jointly for both the source and target languages. The vocabulary is shared between the encoder and decoder.

[6] http://www.jpo.go.jp/shiryou/toushin/chousa/tokkyohonyaku_hyouka.htm

5 Participating Teams and Results

We describe the teams' profiles and submissions as described in their respective description papers. Table 3 shows the team IDs, their respective organizations, and countries.

5.1 Systems' Descriptions

DCU_NMT participated in the English-to-Hindi track only, developing both text-only and multimodal neural machine translation (NMT) systems. They trained the text-only models from scratch on constrained data and further enhanced them with back-translated data. For the multimodal approach, they used a context-aware transformer to integrate visual features by first encoding the image captions with a BERT model and then concatenating them with the textual input. They reported that while the back-translated text-only model achieved the best performance overall, the multimodal systems, despite lacking back-translated data, outperformed the text-only baseline, indicating the potential of visual context. However, their findings revealed that the impact of visual features was inconsistent, showing less effectiveness on the challenge set, highlighting the need for further exploration into effective multimodal integration.

ODIAGEN participated in and reported results for all the tracks, including both text-only and multimodal translation. For text-only translation, they trained the Mistral-7B (Jiang et al., 2023) model to handle English to multiple low-resource languages: Hindi, Bengali, Malayalam, and Hausa. In the multimodal English-to-Hindi task, they employed the PaliGemma-3B (Beyer et al., 2024) model, integrating both image and text inputs. However, their findings revealed that the multimodal systems were suboptimal due to improper normalization of location coordinates, which hindered the models ability to map these coordinates accurately to the provided images. While the PaliGemma-3B model demonstrated strong performance in text translation tasks, it struggled to leverage visual context effectively, underscoring the importance of refining multimodal techniques for better accuracy.

Arewa_NLP participated in the English-Hausa text-only translation task, fine-tuning the OPUS-MT-en-ha transformer model. While the system performed well on standard test set, it struggled with the more complex content in the Challenge Test, suggesting a need for further training.

v036 participated in the English-to-Indic tracks only with the help of visual context. They utilized InternVL2 (Chen et al., 2023) to extract features from the marked image region, which was then passed into a Rapid Automatic Keyword Extraction (RAKE) algorithm to generate keywords for use as hash-tags to provide context to the source text. They then used an LLM (Llama 405B) to generate chain-of-thoughts prompts, consisting the original source and target sentences, extracted keywords as hash-tags and some reasoning why that translation was generated, that serve as training data. Finally, they fine-tuned Llama 8b Instruct model, one for each language, on the generated prompts. They reported that although their predictions were mostly correct, the model failed to generate similar translations as the ground truth, indicating the need for human evaluation as the best method to assess the quality of the translations.

Brotherhood participated in all the tracks, leveraging LLMs such as GPT-4o and Claude 3.5 Sonnet to enhance cross-lingual image captioning without traditional training or fine-tuning (Betala and Chokshi, 2024). They used instruction-tuned prompting to generate contextual conversations around cropped images, incorporating the original English captions as context, and translated these conversations into target languages. They employed weighted prompting strategy to balance the original captions with the translated conversations for more descriptive outputs. They reported that their training-free approach minimizes error propagation from flawed datasets while offering flexibility in balancing source fidelity with descriptiveness, demonstrating promise for improving low-resource language datasets. However, they identified challenges such as dependence on LLM APIs, hallucination risks, computational demands, and the limitations of traditional metrics like BLEU for evaluating enriched descriptions, highlighting the need for more comprehensive evaluation methods.

UNLP participated in the English-to-Hindi, Malayalam, Bengali, and Hausa tracks. They used visual context to improve translation accuracy, employing a gated fusion mechanism to integrate visual information with textual data, combining the outputs of visual and textual encoders to create context-aware translations. For each language,

Team ID	Organization	Country
DCU_NMT	Dublin City University	Ireland
ODIAGEN	Odia Generative AI	India
Arewa_NLP	FUTB, BUK, and Arewa Data Science Academy	Nigeria
v036	SCB DataX, Walmart Global Tech	Thailand, India
Brotherhood	Indian Institute of Technology Madras	India
UNLP	University of Galway, and Lua Health, Galway	Ireland
00-7	Krutrim AI	India

Table 3: List of participants who submitted translations for the WMT2024 English-to-LowRes Multimodal Translation Task

they fine-tuned their multimodal model on this combined input, ensuring a nuanced understanding of both linguistic and visual cues. The team reported that while their multimodal model consistently outperformed text-only baselines across BLEU, ChrF2, and TER metrics, some discrepancies with the ground truth translations highlight the importance of incorporating human evaluation for a more reliable assessment of translation quality.

00-7 competed in three tracks—Image Captioning, Text-only, and Multimodal Translation—for Indic languages, developing a multimodal model that integrates a multilingual LLM with a vision module for improved translation. Their method employs a ViT image encoder to extract visual token embeddings, which are projected into the LLM space through an adapter layer, generating translations autoregressively. They achieved state-of-the-art results for Hindi on the Challenge set, while remaining competitive for other languages. Despite the models success, they observed limited impact of the vision modality on translation quality.

5.2 Results

Automatic evaluation results Tables 4 to 8 present the automatic evaluation results of the submitted systems, indicating that the systems performed competitively against each other. Despite these promising results, participants expressed a need for human evaluations, as shown in subsequent tables. This reflects a common concern among participants who suspect that their systems may outperform the scores they received, underscoring the importance of qualitative assessments in conjunction with automatic metrics.

Lang.	System	ID	Type	RSRC	BLEU	RIBES
en-hi	00-7	7190	NMT	Yes	53.40	0.842400
en-hi	v036	7406	NMT	No	43.20	0.812507
en-hi	Brotherhood	7378	NMT	Yes	37.90	0.795538
en-hi	DCU_NMT	7372	NMT	No	30.30	0.710342
en-ml	00-7	7195	NMT	Yes	39.80	0.739973
en-ml	v036	7395	NMT	No	33.30	0.606598
en-ml	Brotherhood	7377	NMT	Yes	13.60	0.428194
en-bn	00-7	7192	NMT	Yes	45.30	0.796451
en-bn	v036	7414	NMT	No	33.90	0.736029
en-bn	Brotherhood	7375	NMT	Yes	21.70	0.644341
en-ha	Brotherhood	7376	NMT	Yes	21.10	0.636818

Table 4: MMCHMM24 submissions

Lang.	System	ID	Type	RSRC	BLEU	RIBES
en-hi	00-7	7313	NMT	No	54.10	0.858322
en-hi	ODIAGEN	7358	Other	No	44.10	0.815457
en-hi	DCU_NMT	7349	NMT	No	35.90	0.762839
en-ml	00-7	7327	NMT	Yes	34.00	0.651880
en-ml	ODIAGEN	7343	Other	No	18.10	0.505942
en-bn	00-7	7321	NMT	Yes	44.20	0.789032
en-bn	ODIAGEN	7336	Other	No	35.60	0.735341
en-ha	ODIAGEN	7366	Other	No	24.40	0.663630

Table 5: MMCHTEXT24 submissions

Lang.	System	ID	Type	RSRC	BLEU	RIBES
en-hi	v036	7411	NMT	No	44.60	0.833853
en-hi	00-7	7325	NMT	No	43.70	0.813357
en-hi	DCU_NMT	7351	NMT	No	40.60	0.806358
en-hi	UNLP	7392	NMT	No	40.30	0.800532
en-hi	Brotherhood	7379	NMT	Yes	29.70	0.725450
en-ml	00-7	7194	NMT	Yes	51.40	0.780907
en-ml	v036	7396	NMT	No	42.70	0.700828
en-ml	UNLP	7393	NMT	No	32.20	0.626281
en-ml	Brotherhood	7382	NMT	Yes	15.10	0.410674
en-bn	00-7	7191	NMT	Yes	46.40	0.775597
en-bn	v036	7418	NMT	No	44.10	0.737924
en-bn	UNLP	7391	NMT	No	42.10	0.766589
en-bn	Brotherhood	7381	NMT	Yes	22.10	0.575370
en-ha	UNLP	7394	NMT	No	41.80	0.723997
en-ha	Brotherhood	7380	NMT	Yes	17.70	0.580239

Table 6: MMEVMM24 submissions

Human evaluation results Tables 10 and 11 present the adequacy scores after human evaluation. The scores reinforce the need for human evaluations to actually determine the quality of multi-

Lang.	System	ID	Type	RSRC	BLEU	RIBES
en-hi	00-7	7322	NMT	Yes	43.30	0.812578
en-hi	DCU_NMT	7348	NMT	Yes	42.70	0.817949
en-hi	ODIAGEN	7335	Other	No	41.60	0.821154
en-ml	00-7	7326	NMT	Yes	37.80	0.633752
en-ml	ODIAGEN	7365	Other	No	33.10	0.668374
en-bn	00-7	7320	NMT	No	45.10	0.766452
en-bn	ODIAGEN	7363	Other	No	43.70	0.789757
en-ha	ODIAGEN	7344	Other	No	49.80	0.812898
en-ha	Arewa_NLP	7314	SMT	No	40.70	0.755910

Table 7: MMEVTEXT24 submissions

Lang.	System	ID	Type	RSRC	BLEU	RIBES
en-hi	00-7	7385	NMT	Yes	2.80	0.183643
en-ml	00-7	7389	NMT	Yes	0.90	0.064375
en-bn	00-7	7386	NMT	No	1.80	0.105044

Table 8: MMEVHI24 submissions

Lang.	System	ID	Type	RSRC	BLEU	RIBES
en-hi	00-7	7346	NMT	No	1.30	0.125551
en-ml	00-7	7390	NMT	Yes	0.30	0.039097
en-bn	00-7	7387	NMT	Yes	0.40	0.041301

Table 9: MMCHHI24 submissions

modal generations. The number of sentences that were marked 4 and 5 (almost all or all information transmitted) in system 7375 Brotherhood in Table 10 indicates a higher performance than what the automatic metrics suggest for the same system in Table 4.

Lang.	System	ID	JPO adequacy scores					
			#	1	2	3	4	5
en-bn	v036	7414	**1**	2	6	29	84	79
			2	7	23	47	85	38
en-bn	Brotherhood	7375	**1**	0	1	16	71	112
			2	1	10	11	46	132
en-ha	Brotherhood	7376	**1**	11	21	40	48	80
			2	16	29	50	68	37

Table 10: MMCHMM24 Human Evaluations on random 200 Test Sentences

6 Conclusion and Future Directions

This paper presents an overview of the English-to-Low Resource Multimodal Translation shared tasks at WMT2024. The task attracted strong participation from numerous teams. Out of these, 7 teams submitted system description papers detailing their approaches and results. In the future, we aim to expand the range of low-resource languages, with a particular focus on multimodal translation, and encourage greater participation from more teams.

Lang.	System	ID	JPO adequacy scores					
			#	1	2	3	4	5
en-bn	ODIAGEN	7336	**1**	15	18	55	66	46
			2	46	43	48	40	23
en-ha	ODIAGEN	7366	**1**	18	29	62	61	30
			2	26	58	66	36	14

Table 11: MMCHTEXT24 Human Evaluations on random 200 Test Sentences

Acknowledgements

The shared tasks were supported by Silo AI, Finland; Charles University, MFF, ÚFAL, Czech Republic; the Data Science for Social Impact Research Group at the University of Pretoria; and HausaNLP Research Group. Shamsuddeen acknowledges the support received from the Google DeepMind Academic Fellowship.

Ethical Considerations

The authors do not see ethical or privacy concerns that would prevent the use of the data used in the study. The datasets do not contain personal data. Personal data of annotators needed when the datasets were prepared and when the outputs were evaluated were processed in compliance with the GDPR and national law.

References

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022. Hausa Visual Genome: A Dataset for Multi-Modal English to Hausa Machine Translation. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.

Siddharth Betala and Ishan Chokshi. 2024. Brotherhood at wmt 2024: Leveraging llm-generated contextual conversations for cross-lingual image captioning. *arXiv preprint arXiv:2409.15052*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Desmond Elliott, Douwe Kiela, and Angeliki Lazaridou. 2016. Multimodal learning and reasoning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Berlin, Germany. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 944–952, Stroudsburg, PA, USA. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Shantipriya Parida and Ondřej Bojar. 2021. Malayalam visual genome 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation. *Computación y Sistemas*, 23(4):1499–1505.

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, Belgium, Brussels. Association for Computational Linguistics.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022. Bengali visual genome 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 193–199. Association for Machine Translation in the Americas.

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