Findings of WAT2025 English-to-Indic Multimodal Translation Task

Shantipriya Parida[†], **Ondřej Bojar**[‡]

[†]AMD Silo AI, Finland; [‡]Charles University, MFF, ÚFAL, Czech Republic

correspondence: shantipriya.parida@amd.com

Abstract

This paper presents the findings of the English-to-Indic Multimodal Translation shared task from the Workshop on Asian Translation (WAT2025). The task featured three tracks: text-only translation, image captioning, and multimodal translation across four low-resource Indic languages: Hindi, Bengali, Malayalam, and Odia. Three teams participated, submitting systems that achieved competitive performance, with BLEU scores ranging from 40.1 to 64.3 across different language pairs and tracks.

1 Introduction

The 12th Workshop on Machine Translation (WAT2025), held in conjunction with IJCNLP AACL 2025, hosted a number of shared tasks that covered various aspects of machine translation (MT).

Multi-modal translation, which involves incorporating non-text sources alongside text input for machine translation, has gained attention in recent 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 WAT2025 English-to-Indic Multimodal Shared Task for Low-Resource Indic 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 languages: Hindi, Bengali, Malayalam and Odia. 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 Odia Visual Genome (OVG)[4] for Odia. 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 the 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," "OD", 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] http://hdl.handle.net/11234/1-5979

Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025), pages 103–108
December 24, 2025

Split	Train	Dev	Test	Challenge
Sentences	28,930	998	1,595	1,400

Table 1: Dataset statistics across all language pairs.

Figure 1: Example of a data point showing image ID, region details, 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.

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 WAT2025 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 WAT2025 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 WAT2025.

3.2 Human Evaluation

Due to time constraints, human evaluation was not carried out in WAT2025.

4 Baseline Systems

At WAT2025, we adopted some of the 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.

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

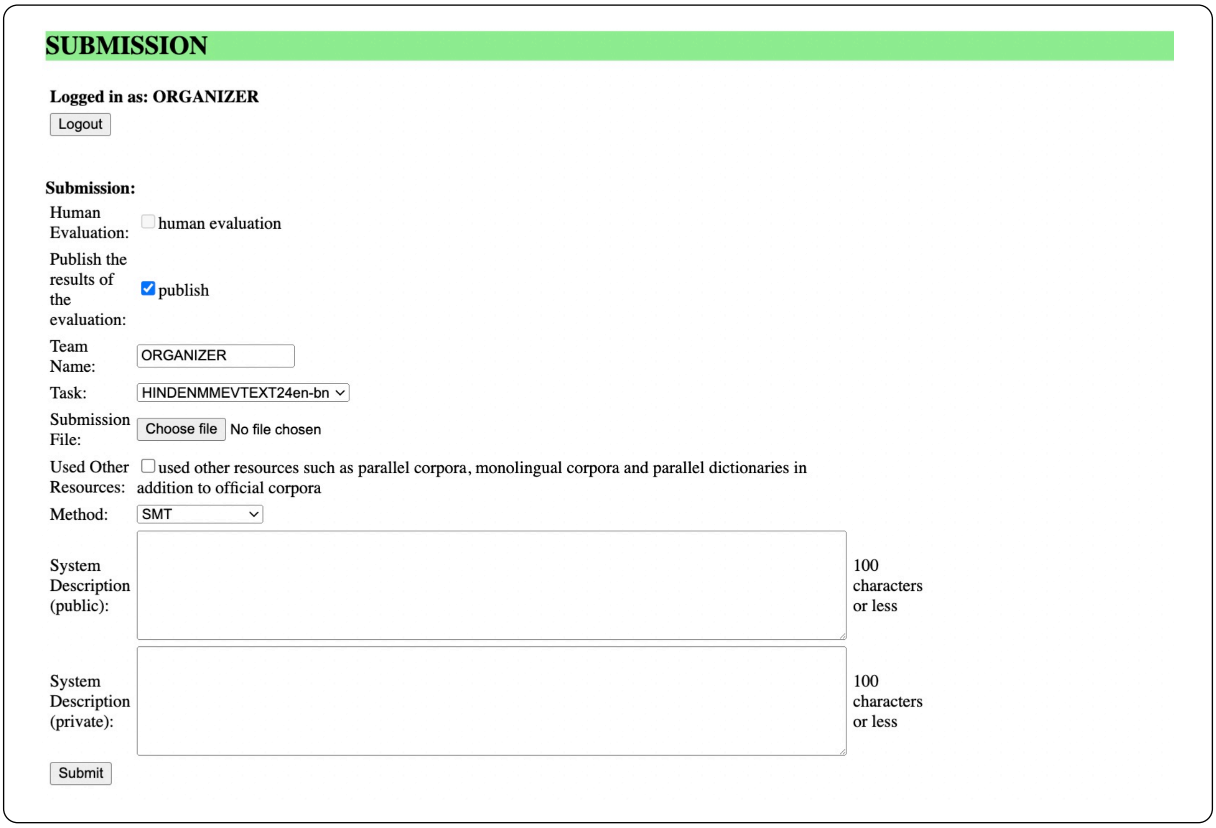

Figure 2: The interface for translation results submission

5 Participating Teams and Results

We describe the teams' profiles and submissions as described in their respective description papers. Table 2 shows the team IDs, their respective organizations, and countries.

5.1 Systems' Descriptions

IITP-AI-NLP-ML The IITP-AI-NLP-ML team participated in and reported results for both text-only and multimodal translation tracks. For text-only translation, they fine-tuned the IndicTrans model (Bhat et al., 2015) jointly on all four target languages. In the multimodal track, they enhanced IndicTrans with a CLIP-based visual grounding mechanism that selects the most semantically relevant image regions. By computing cosine similarities between text and full or cropped image embeddings, the system automatically integrates the most aligned visual features into the translation pipeline.

OdiaGenAI team participated in and reported results for all text-only translation tracks. They fine-tuned the NLLB-200 3.3B model (NLLB et al., 2022) to support English-to-multilingual translation, specifically targeting low-resource languages: Hindi, Bengali, Malayalam, and Odia. To enhance training, they applied data augmentation using 100K samples from the Samanantar dataset (Ramesh et al., 2022) provided by AI4Bharat.

BLEU Monday team participated in and reported results for the text-only translation for three language pairs: English-Hindi, English-Bengali, and English-Odia. The proposed system uses a two-stage approach: automated training data correction through a vision-augmented judge-corrector pipeline, followed by LoRA-based fine-tuning. The pipeline employs multimodal models to detect and correct translation errors, replacing ambiguous or mistranslated captions using GPT-4o-mini and IndicTrans2.

5.2 Results and Analysis

Automatic evaluation results Tables 3 to 6 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

Team ID	Organization	Country
OdiaGenAI	Odia Generative AI	India
BLEU Monday	Indian Institute of Technology Madras	India
IITP-AI-NLP-ML	Indian Institute of Technology Patna	India

Table 2: List of participants who submitted translations for the WAT2025 English-to-Indic Multimodal Translation Task.

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	IITP-AI-NLP-ML	7461	NMT	Yes	56.60	0.872157
en-ml	IITP-AI-NLP-ML	7463	NMT	Yes	38.90	0.749429
en-bn	IITP-AI-NLP-ML	7462	NMT	Yes	47.00	0.815367
en-od	IITP-AI-NLP-ML	7464	NMT	Yes	55.20	0.915999

Table 3: MMCHMM25 submissions.

Lang.	System	ID	Type	RSRC	BLEU	RIBES
en-hi	OdiaGenAI	7485	NMT	Yes	56.90	0.870254
en-hi	IITP-AI-NLP-ML	7471	NMT	Yes	56.10	0.870914
en-hi	BLEU Monday	7500	Other	Yes	54.00	0.864790
en-ml	OdiaGenAI	7483	NMT	Yes	44.20	0.775824
en-ml	IITP-AI-NLP-ML	7473	NMT	Yes	40.30	0.757277
en-bn	OdiaGenAI	7481	NMT	Yes	50.10	0.830882
en-bn	IITP-AI-NLP-ML	7472	NMT	Yes	47.50	0.819714
en-bn	BLEU Monday	7503	Other	Yes	45.60	0.808860
en-od	OdiaGenAI	7487	NMT	Yes	56.40	0.916177
en-od	IITP-AI-NLP-ML	7474	NMT	Yes	55.40	0.916776
en-od	BLEU Monday	7498	Other	Yes	40.10	0.872698

Table 4: MMCHTEXT25 submissions.

Lang.	System	ID	Type	RSRC	BLEU	RIBES
en-hi	IITP-AI-NLP-ML	7456	NMT	No	44.90	0.765514
en-ml	IITP-AI-NLP-ML	7460	NMT	Yes	50.70	0.780907
en-bn	IITP-AI-NLP-ML	7457	NMT	Yes	48.70	0.799718
en-od	IITP-AI-NLP-ML	7459	NMT	Yes	63.50	0.903624

Table 5: MMEVMM25 submissions.

5.3 Key Findings

The results show that:

- Text-only translation generally outperformed multimodal approaches
- Odia achieved the highest BLEU scores (62.9-64.3)
- Malayalam proved most challenging with lower scores (38.9-51.2)
- Data augmentation strategies proved effective across teams

Lang.	System	ID	Type	RSRC	BLEU	RIBES
en-hi	IITP-AI-NLP-ML	7467	NMT	Yes	45.40	0.834985
en-hi	OdiaGenAI	7484	NMT	Yes	45.10	0.831282
en-hi	BLEU Monday	7494	Other	Yes	42.10	0.814804
en-ml	IITP-AI-NLP-ML	7469	NMT	Yes	51.20	0.760801
en-ml	OdiaGenAI	7482	NMT	Yes	43.20	0.708217
en-bn	OdiaGenAI	7480	NMT	Yes	49.50	0.804158
en-bn	IITP-AI-NLP-ML	7468	NMT	Yes	49.50	0.801714
en-bn	BLEU Monday	7496	NMT	Yes	42.00	0.770437
en-od	IITP-AI-NLP-ML	7470	NMT	Yes	64.30	0.906478
en-od	OdiaGenAI	7486	NMT	Yes	62.90	0.903659
en-od	BLEU Monday	7504	Other	Yes	41.60	0.845874

Table 6: MMEVTEXT25 submissions.

5.4 Cross-Track Performance Comparison

Comparing performance across different tracks reveals interesting patterns:

- **Text-only vs. Multimodal**: Text-only systems achieved comparable or better performance than multimodal systems, indicating room for improvement in visual-textual integration methods
- **Language-specific trends**: Odia consistently performed best across all tracks, while Malayalam showed the most variation between different approaches
- **Team strategies**: Teams employing data augmentation and fine-tuning of large pretrained models (NLLB, IndicTrans) achieved the most competitive results

6 Conclusion and Future Directions

This paper presents an overview of the English-to-Indic Resource Multimodal Translation shared tasks at WAT2025. The task attracted strong participation from numerous teams. Out of these, three 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.

Acknowledgements

The shared tasks were supported by AMD Silo AI, Finland; and Charles University, MFF, ÚFAL, Czech Republic;

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

Irshad Ahmad Bhat, Vandan Mujadia, Aniruddha Tammewar, Riyaz Ahmad Bhat, and Manish Shrivastava. 2015. Iiit-h system submission for fire2014 shared task on transliterated search. In *Proceedings of the Forum for Information Retrieval Evaluation*, FIRE '14, pages 48–53, New York, NY, USA. ACM.

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.

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.

Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan Ak, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

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.