

# OdiaGenAI’s Participation in WAT2024 English-to-Low Resource Multimodal Translation Task

Shantipriya Parida<sup>1</sup>, Shashikanta Sahoo<sup>2,3</sup>, Sambit Sekhar<sup>3</sup>,  
Upendra Kumar Jena<sup>4</sup>, Sushovan Jena<sup>5</sup>, Kusum Lata<sup>6</sup>

<sup>1</sup>Silo AI, Finland; <sup>2</sup>Government College of Engineering Kalahandi, India;

<sup>3</sup>Odia Generative AI, India; <sup>4</sup>Creanovation Technologies Pvt. Ltd., India;

<sup>5</sup>IIT Mandi, India; <sup>6</sup>Sharda University, India

correspondence: shantipriya.parida@siloi.ai

## Abstract

This paper covers the system description of the team “ODIAGEN’s” submission to the 11th Workshop on Asian Translation (WAT 2024). We participated in the English-to-LowRes Multimodal Translation Task, in two of the tasks, i.e. Text-only Translation and Multi-modal Translation. For Text-only Translation, we trained the Mistral-7B model for English to Multi-lingual (Hindi, Bengali, Malayalam, Hausa). For Multi-modal Translation (using both image and text), we trained the PaliGemma-3B model for English to Hindi translation.

## 1 Introduction

Machine translation (MT) is a well-established area within Natural Language Processing (NLP), focusing on the development of software that can automatically translate text or speech between languages. While substantial progress has been made in achieving human-level translation for high-resource languages, significant challenges persist for low-resource languages (Popel et al., 2020) (Parida et al., 2023). Recent research has also investigated how to effectively incorporate other modalities, such as images, into the translation process.

Since 2013, the WAT (Workshop on Asian Translation) has been an open evaluation campaign centered on Asian languages (Nakazawa et al., 2021). The multimodal translation tasks in WAT2024 involve image caption translation, where the input includes a descriptive caption in the source language paired with the image it describes, and the output is a caption in the target language. This multimodal input leverages image context to clarify source words with multiple meanings.

The evaluation of these translation tasks is conducted using established metrics such as

Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) and Rank-based Intuitive Bilingual Evaluation Score (RIBES) (Isozaki et al., 2010). In this system description paper, we (team “ODIAGEN”) outline our approach to the tasks and sub-tasks in which we participated.

- Task 1: English→Hindi (EN-HI) Multi-modal Translation
  - EN-HI text-only translation
  - EN-HI multimodal translation
- Task 2: English→Malayalam (EN-ML) Text-only Translation
- Task 3: English→Bengali (EN-BN) Text-only Translation
- Task 4: English→Hausa (EN-HA) Text-only Translation

## 2 Datasets

We used only Hindi (Parida et al., 2019), Bengali (Sen et al., 2022), Malayalam, and Hausa (Abdulmumin et al., 2022) Visual Genome datasets specified by the organizer for text-only and multi-modal translation without any additional synthetic data.

### 2.1 Pre-processing

#### 2.1.1 For Text-only

A few Hindi samples were excluded due to identical Hindi and English text in the Hindi dataset, and one Malayalam sample was removed for similar reasons. Formatting issues in the Hindi dataset were corrected, and duplicate samples were excluded from all language datasets. Image metadata (image\_id, X, Y, Width, Height) was excluded from the text-to-text translation task. The final dataset sentence/sample count is provided in Table 1.

All four different language datasets were combined to make a common translation

Language	Sentence Count
Hindi	28,927
Bengali	28,927
Malayalam	28,922
Hausa	28,927

Table 1: Training Dataset Sentence Count

dataset with a single task of translating from English to instructed Target Language like Hindi, Bengali, Malayalam, and Hausa.

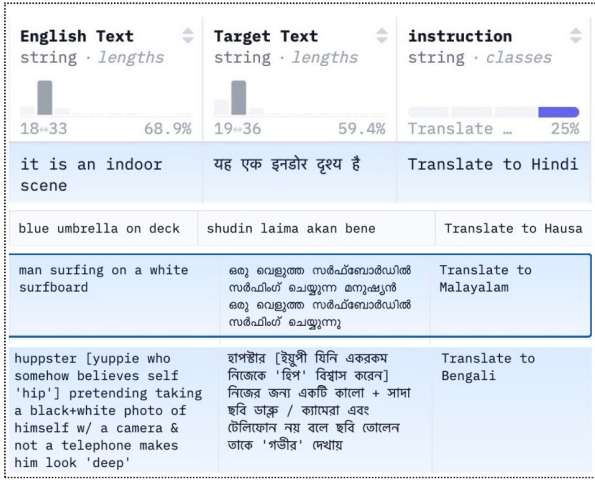


Figure 1: Instruction set in different language

### 2.1.2 For Multimodal involving both Text and Image

The multimodal dataset comprises both text and images. The text portions of the dataset (train and test sets) are organized in simple tab-delimited plain text files. Each text file contains seven columns as follows:

- Column 1: imageid,
- Column 2: X,
- Column 3: Y,
- Column 4: Width,
- Column 5: Height,
- Column 6: English text,
- Column 7: Hindi Text.

The X, Y, Width, and Height columns define the rectangular region in the image described by the caption.

The Mistral-7B model (Beyer et al., 2024) requires data in the format  $[x_{min}, y_{min}, x_{max}, y_{max}]$ . We interpreted the provided X and Y coordinates as the

center coordinates of the rectangular region and calculated  $[x_{min}, y_{min}, x_{max}, y_{max}]$  as the coordinates of the bottom-left and top-right corners of the rectangular box.

## 2.2 Instruction Dataset

### 2.2.1 For Text-only

Alpaca prompt format was used to prepare instruction data sets for text-to-text translation for all languages. Sample prompt format is given below.

""" Below is an instruction that describes a translation task, paired with an input that provides context in Source Language. Write a response that appropriately completes translation to desired Target Language.

### Instruction:  
{ }

### Input:  
{ }

### Response:  
{ }"""

A raw training sample data for Hindi translation after prompt formatting is shown below, similar method was used for other language translations.

Below is an instruction that describes a translation task, paired with an input that provides context in the Source Language. Write a response that appropriately completes translation to desired Target Language.

Instruction: Translate to Hindi

Input: it is an indoor scene

Response: यह एक इनडोर दृश्य है

### 2.2.2 For Multi-modal involving both Image and Text

We passed the prompts in a CSV file with fields 'image id', and 'message'. The prompt in the "message" field is in the below format: 'message': [{'content': 'describe the image in Hindi <loc ymin><loc xmin><loc ymax><loc xmax>', 'role': 'user'}, {'content': English text, 'role': 'assistant'}]

## 2.3 Tokenization

Both model unsloth/mistral-7b-v0.3 and tokenizer were used from unsloth library, tok-

Set	Sentences	Tokens				
		English	Hindi	Malayalam	Bengali	Hausa
Train	28,927	143,164	145,448	107,126	113,978	113,978
D-Test	998	4,922	4,978	3,619	3,936	3,936
E-Test	1,595	7,853	7,852	5,689	6,408	6,408
C-Test	1,400	8,186	8,639	6,044	6,657	6,657

Table 2: Statistics of our data used in the English→Hindi, English→Malayalam, English→Bengali, and English→Hausa tasks: the number of sentences and tokens in text-text translation.

Set	Images	English Words	Hindi Words
Train	28,927	143,164	145,448
D-Test	998	4,922	4,978
E-Test	1,595	7,853	7,852
C-Test	1,400	8,186	8,639

Table 3: Statistics of our data used in the English→Hindi multi-modal translation.

enizer is based on SentencePiece with Byte-Pair Encoding (BPE). This is the standard approach for tokenization in many modern transformer-based language models, including those similar to Mistral.



Figure 2: Instruction set in English-Hindi for multi-modal translation

### 3 Experimental Details

This section describes the complete pipeline used to produce the translation systems for the WAT English-to-Low Resource Multimodal shared task submission.

#### 3.1 EN-HI, EN-ML, EN-BN, EN-HA Text-only Translation

For EN-HI, EN-BN, EN-ML, and EN-HA text-only (E-Test and C-Test) translation, the study fine-tunes the pre-trained Mistral-7B model (Jiang et al., 2023), which has been fine-tuned utilizing only HVG, BVG, MVG, and HaVG Datasets; aiming to develop a high-quality machine translation system.

The Mistral-7B model is a cutting-edge language model that has been fine-tuned specifically for developing high-quality machine

translation systems. Leveraging its 7 billion parameters, Mistral-7B (Jiang et al., 2023) excels in capturing linguistic nuances and context, making it exceptionally adept at translating between languages with high accuracy. The fine-tuning process involves training the model on extensive and diverse datasets, allowing it to understand and generate translations that are not only precise but also contextually relevant.

#### 3.2 EN-HI Multimodal Translation

This section discusses the multimodal translation pipeline for EN-HI. For EN-HI multimodal (E-Test and C-Test) translation, we used the object tags extracted from the HVG dataset images for image features and concatenated them with the text. The PaliGemma-3B model (Beyer et al., 2024) is finetuned on the Hindi-Visual-Genome dataset for English to Hindi Translation when a specific location is given in the input prompt as explained in Section 2.2.2. We used the script from LLaMa Factory (Zheng et al., 2024) with our configuration to fine-tune this model. During fine-tuning, we froze the vision tower and adjusted the parameters in the language model and projector layer. The hyperparameters are shown in Table 4.

## 4 Results

### 4.1 Text-only Translation

We present the official automatic evaluation results of our models for all the tasks we participated in Table 2, along with sample outputs in Table 3. After the fine-tuning process, these

Hyperparameter	Value
Train Batch Size	2
Eval Batch Size	8
Learning Rate	$3 \times 10^{-6}$
Epochs	10
Warm-up Steps	50
LR Scheduler	Cosine
Gradient Accumulation Steps	8
Optimizer	“Adam”

Table 4: Training Hyperparameters.

models were used to generate translations for two distinct sets in each language: the evaluation set and the challenge set. The translation quality was assessed using the BLEU (Bilingual Evaluation Understudy) score and the RIBES (Ranking by Incremental Bilingual Evaluation System) score.

The English-to-Hindi model achieved a BLEU score of 41.60 on the evaluation set and 44.10 on the challenge set. Similarly, it attained a RIBES score of 0.82115 on the evaluation set and 0.8154 on the challenge set. These results underscore the model’s robust performance and its ability to manage more complex or less typical translation tasks.

In the case of the English-to-Bengali model, a BLEU score of 43.70 was achieved on the evaluation set, with a slightly lower score of 35.60 on the challenge set. Similarly, it attained a RIBES score of 0.78975 on the evaluation set and 0.73534 on the challenge set. This indicates a robust overall performance and a commendable capability to handle nuanced translations specific to the Bengali language.

For the English-to-Malayalam model, the system achieved a BLEU score of 33.10 on the evaluation set and 18.10 on the challenge set. Similarly, it attained a RIBES score of 0.66837 on the evaluation set and 0.50594 on the challenge set. Despite a slightly lower score on the challenge set, the model still demonstrates a respectable performance in translating English to Malayalam.

Lastly, for the English-to-Hausa model, the system achieved a BLEU score of 49.80 on the evaluation set and 24.40 on the challenge set. Similarly, it attained a RIBES score of 0.81289 on the evaluation set and 0.66363 on the challenge set. This indicates a robust overall performance and a commendable capability

to handle nuanced translations specific to the Hausa language.

## 4.2 Multi-modal Translation Involving both Image and Text

Contrary to our expectations, the PaliGemma-3B model showed very poor results on the mentioned dataset and we tried to investigate the factors behind it. By qualitative analysis, we figured out that the location coordinates that we normalized during pre-processing may not be the right approach required for PaliGemma-3B. We found that the normalized [xmin, ymin, xmax, ymax] coordinates provided in the input prompt did not perfectly align with the model-generated captions. Instead, they pointed to a neighboring location in the image with a significant overlap. However, this mismatch in location led to a very poor BLEU score for the predicted captions.

## 5 Availability

The text-to-text and multimodal datasets, as well as the models, are freely available for research and non-commercial use under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License via Hugging Face.

We have also released our experimental code on GitHub.<sup>1</sup>

### 5.1 EN-HI/ML/BN/HA Text-only Translation

Dataset: [https://huggingface.co/datasets/OdiaGenAIdata/wat24\\_text\\_to\\_text\\_translation](https://huggingface.co/datasets/OdiaGenAIdata/wat24_text_to_text_translation)

Model: [https://huggingface.co/OdiaGenAI-LLM/wat\\_mistral\\_7b\\_translate](https://huggingface.co/OdiaGenAI-LLM/wat_mistral_7b_translate)

### 5.2 EN-HI Multimodal Translation

Dataset: [https://huggingface.co/datasets/sahoosk/Hindi-visual-genome\\_Train](https://huggingface.co/datasets/sahoosk/Hindi-visual-genome_Train)

Model: <https://huggingface.co/sam2ai/odia-paligemma-2b-9900-v1.1>

## 6 Conclusion

In this system description paper, we presented our approach for four tasks in WAT2024: (a) English→Hindi text-only and multimodal translation, (b) English→Malayalam text-only

<sup>1</sup>[https://github.com/shantipriyap/ODIAGEN\\_WAT2024](https://github.com/shantipriyap/ODIAGEN_WAT2024)

Translation Model	Translation Type	BLEU Score (Evaluation Set)	BLEU Score (Challenge Set)
English to Hindi	Text-to-Text	41.60	44.10
	Multimodal	0.50	-
English to Bengali	Text-to-Text	43.70	35.60
English to Malayalam	Text-to-Text	33.10	18.10
English to Hausa	Text-to-Text	49.80	24.40

Table 5: Comparison of BLEU Scores for Different Translation Models and Types

Translation Model	Translation Type	RIBES Score (Evaluation Set)	RIBES Score (Challenge Set)
English to Hindi	Text-to-Text	0.8212	0.8155
	Multimodal	0.1512	-
English to Bengali	Text-to-Text	0.7898	0.7353
English to Malayalam	Text-to-Text	0.6684	0.5059
English to Hausa	Text-to-Text	0.8129	0.6636

Table 6: Comparison of RIBES Scores for Different Translation Models and Types

	MALAYALAM	HINDI	BENGLI	HAUSA
English-Sentence-1	silver car is parked	fine thin red hair	A stop light	A stop light
Target-Original	സിൽവർ കാർ പാർക്ക് ചെയ്തു	सूक्ष्म पतले लाल बाल	একটি স্টপ লাইট	Hasken tasha
Target-Translated	വെള്ളി കാർ പാർക്ക് ചെയ്തിരിക്കുന്നു	ठीक पतले लाल बाल	একটি স্টপ আলো	Hasken tasha
Gloss	Silver car has been parked	Correct thin red hair	A stop light	A stop light
Remarks (Comparison)	Translated version is more formal	Original version is better; "Fine" mistranslated by our model.	Original version is more colloquial	Both are identical
English-Sentence-2	eye of the pumpkin	the cross is black	This is a person	three zebras in the wild
Target-Original	മത്തങ്ങയുടെ കണ്ണ്	क्रॉस काला है	এটি একজন ব্যক্তি	alfadarai uku a cikin daji
Target-Translated	പമ്മിക്കിന്റെ കണ്ണ്	क्रॉस काला है	এটি একজন ব্যক্তি	alfadarai uku a cikin daji
Gloss	Pumpkin's eyes	The cross is black	This is a person	Three zebras in the wild
Remarks (Comparison)	Model doesn't translate "pumpkin," which is colloquial	Both are identical	Both are identical	Both are identical
English-Sentence-3	pen on the paper	date and time of photo	the bird is black	a girl is standing.
Target-Original	പേപ്പറിൽ പേന	फोटो की तारीख और समय	পাখিটি কালো	yariyana tana tsaye
Target-Translated	പേപ്പറിൽ പേന	फोटो की तारीख और समय	পাখিটি কালো	yariyana tana tsaye
Gloss	Pen on the paper	Date and time of photo	The bird is black	A girl is standing
Remarks (Comparison)	Both are identical	Both are identical	Both are identical	Both are identical

Table 7: Comparison between original translations and our model's translations for English-Malayalam, English-Hindi, and English-Bengali language pairs.

translation, (c) English→Bengali text-only translation, and (d) English→Hausa text-only translation. The results for the multimodal English→Hindi translation, which involves both image and text, were suboptimal due to improper normalization of the location coordinates for the PaliGemma-3B model. As a result, the model was unable to accurately map the provided coordinates in the prompt to the original image. We utilized the PaliGemma-3B model with a resolution of 448, which performed well in the translation tasks but failed to generate results relevant to the precise coordinates. Due to limitations in time and computing resources, addressing this issue has been deferred to future work. The code has been released on GitHub for use by other researchers.

## Acknowledgments

We would like to express our gratitude to Silo AI in Helsinki, Finland, and Odia Generative AI in Bhubaneswar, India, for their support.

## References

- Idris Abdulmumin, Satya Ranjan Dash, Musa Abdul-lahi Dawud, Shantipriya Parida, Shamsuddeen Hassan 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. arXiv preprint arXiv:2205.01133.
- 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.
- 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, pages 944–952.
- 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.
- Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao




Image	Prompt	Predicted Answer (Hindi)	Gloss
	describe the image in Hindi <loc166><loc177><loc263> <loc298>.	एक महिला एक पुस्तक की दुकान से गुजर रही है.	A woman is passing by a bookstore.
	describe the image in Hindi <loc6><loc120><loc31> <loc136>.	एक बड़ा सफेद भवन जिसके शीर्ष पर एक घंटाघर है.	A large white building with a clock tower at the top.
	describe the image in Hindi <loc95><loc284><loc214> <loc300>.	एक सफेद भोजन कक्ष में एक कांच की टेबल सेट होती है.	A glass table set in a white dining room.

Table 8: Comparison of user prompts, predicted answers in Hindi, and their English translations with corresponding images.

Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, et al. 2021. Overview of the 8th workshop on asian translation. In Proceedings of the 8th Workshop on Asian Translation (WAT2021), pages 1–45.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Shantipriya Parida, Ondřej Bojar, and Satya Ranjan Dash. 2019. Hindi visual genome: A dataset for multi-modal english to hindi machine translation. *Computación y Sistemas*, 23(4):1499–1505.

Shantipriya Parida, Alakananda Tripathy, Satya Ranjan Dash, and Shashikanta Sahoo. 2023. Mdoic: Multi dialect odia song lyric corpus.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022. Bengali visual genome: A multi-modal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: The-*

*ory and Applications (FICTA 2021)*, pages 63–70. Springer.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. arXiv preprint arXiv:2403.13372.