GABU@WAT 2026: Zero Shot and QLoRA Adaptation of Gemma 4 E4B for Multimodal English to Indic Translation

Shiv Ram Saud
Kathmandu University
saudshivram373@gmail.com

Aashish Mahato
Kathmandu University
aashishmtho24@gmail.com

Prajwal Ghimire
Kathmandu University
prajwalghimire22@gmail.com

Sundeep Dawadi
Kathmandu University
sundeepdwd@gmail.com

Ramesh Saud
rajusaud245@gmail.com

Sunil Regmi
Kathmandu University
sunil.regmi@ku.edu.np

Abstract

We describe *GABU*, our WAT2026 English-to-Indic multimodal translation system (WAT2026 Organizers, 2026), which translates image grounded English captions into Hindi, Bengali, Malayalam, and Odia. The task follows the WAT English-to-Indic multimodal track setup with text only, image captioning, and multimodal tracks (Parida and Bojar, 2025; Parida et al., 2024). Our system builds on the pretrained, instruction-tuned vision-language model `google/gemma-4-E4B-it`. Each input combines a cropped visual region with the English caption, rendered through a chat template with a script constrained translation instruction. We explore two approaches which includes a zero shot baseline and a fine tuned approach of the model using QLORA (rank 64, alpha 128, 4-bit NF4, double quantization), trained for 3,000 steps due to computational resource constraints, reflecting our aim to adapt the model's existing instruction-following ability rather than train it from scratch. We evaluate both systems using BLEU and CHRF3 under greedy decoding and report macro averaged scores across the target languages. The fine-tuned model substantially outperforms the zero shot baseline and achieves a BLEU score of 43.97 and 69.99 CHRF3 on the test split, and 40.37 BLEU and 68.25 CHRF3 on the challenge split. These results correspond to improvements of 24.39 BLEU and 18.72 CHRF3 points on the test split, and 16.70 BLEU and 11.90 CHRF3 points on the challenge split.

1 Introduction

The English-to-Indic multimodal machine translation (MMT) aims to improve translation of image-grounded English captions into low resource Indic languages by using paired visual context alongside source text (Parida and Bojar, 2025; Parida et al., 2024). The underlying task builds on the Visual Genome-derived Hindi and Bengali datasets (Parida et al., 2019; Sen et al., 2022) and the broader Multi30K paradigm of image-grounded translation (Elliott et al., 2016). Visual context is most valuable where the source caption is ambiguous: classic examples include polysemous words such as "court" or "dish", whose correct Indic rendering can depend on the depicted scene. Whether models genuinely exploit images rather than improving through regularization or strong textual priors remains actively debated at the same time (Wu et al., 2021; Caglayan et al., 2019; Li et al., 2021). Our work targets the WAT2026 English-to-Indic multimodal task. The system described here is implemented and evaluated for four languages: Hindi (`hi`), Bengali (`bn`), Malayalam (`ml`), and Odia (`or`). We investigate how far a modern pretrained vision-language model can be taken in this setting rather than engineering a bespoke fusion architecture: first as a zero shot image grounded translator, then adapted with parameter efficient fine-tuning under tight compute constraints. Our contributions are:

- A parameter efficient fine-tuning approach using QLoRA joint training and over the four target languages.
- A comparison between zero shot baseline and finetuning approach.
- An evaluation on BLEU and CHRF3 with greedy based decoding.

The adapted system substantially outperforms the zero shot baseline on every language

Proceedings of the 13th Workshop on Asian Translation (WAT 2026), pages 145–154
November 9, 2026 ©2026 Association for Computational Linguistics

and split (Section 6.3), with the largest gains observed where the base model could barely produce the target script.

2 Related Work

Multimodal translation architectures. Early neural MMT introduced a doubly attentive decoder that attends to source words and spatial image features independently (Calixto et al., 2017). Subsequent Transformer based MMT proposed multimodal self-attention that grounds image representations in the source text to avoid encoding irrelevant visual information (Yao and Wan, 2020), building on the Transformer backbone (Vaswani et al., 2017). The recent WAT/WMT Indic systems explore gated fusion of textual and visual encoders (Hatami et al., 2024), object tags verbalized as text for a pretrained multilingual backbone (Gupta et al., 2021), ViT visual tokens projected into a multilingual LLM with autoregressive decoding (Khan et al., 2024), and InternVL2-derived image context with knowledge distillation (Rajpoot et al., 2024).

English-to-Indic and low resource MMT. The WAT2025 findings define three tracks (text only, captioning, multimodal) across Hindi, Bengali, Malayalam, and Odia (Parida and Bojar, 2025); the WMT2024 low resource MMT task covers a related English-to-low resource setting including Indic languages (Parida et al., 2024). The strong text only multilingual backbones are central: IndicTrans2 covers all 22 scheduled Indian languages (Gala et al., 2023), mBART provides multilingual denoising pretraining transferable even to unseen pairs (Liu et al., 2020), and web scale NLLB-style models demonstrate scaling to 200 languages (NLLB Team, 2024). Task specific WAT2025 systems build directly on IndicTrans2 with parameter efficient adaptation and data correction (text only track) (Betala et al., 2025), or probe selective CLIP based visual grounding on top of it (Kumar et al., 2025).

Datasets. Hindi Visual Genome provides English–Hindi image-caption pairs created by post editing automatic translations with the image in view (Parida et al., 2019). Bengali Visual Genome extends the same methodology to English–Bengali for both translation and captioning (Sen et al., 2022). These follow the Multi30K precedent of pairing images with multilingual descriptions (Elliott et al., 2016).

When does vision help? Whether visual context genuinely improves MMT is actively debated. Controlled studies show models can exploit images under degraded textual context (Caglayan et al., 2019), while other analyses find gains may stem from regularization rather than visual grounding (Wu et al., 2021), and that Multi30K-style references translated without seeing images weaken visual reliance (Li et al., 2021). Stronger vision encoders help when probed carefully (Li et al., 2022), with contrastive vision-language models such as CLIP (Radford et al., 2021) widely used for selective grounding (Kumar et al., 2025). For language pairs lacking multimodal parallel data, visual prediction networks that supply pseudo-visual features enable zero shot cross-modal transfer (Hirasawa et al., 2023), which is relevant for very low resource targets.

Parameter efficient adaptation and evaluation. Low rank adaptation freezes pretrained weights and learns rank decomposed residual updates that can be merged at deployment without inference latency (Hu et al., 2022); its quantized variant combines a frozen 4-bit NormalFloat base with double quantization and trainable adapters (Dettmers et al., 2023). For evaluation, BLEU computed follows the standardized detokenized reporting practice (Post, 2018), and chrF is the character n-gram F-score whose beta parameter balances precision and recall (Popović, 2017). Concretely, BLEU combines a brevity penalty BP with the geometric mean of n-gram precisions p_n (Papineni et al., 2002),

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{4} \tfrac{1}{4} \log p_n\right),$$

computed here with 13a tokenization, case sensitively, against a single reference. chrF is the F_β score over character n-grams (here $\beta = 3$, character order 6 with word-bigram statistics),

$$\mathrm{chrF}_\beta = \frac{(1+\beta^2)\, P\, R}{\beta^2 P + R},$$

where P and R are the averaged n-gram precision and recall.

3 Dataset and Preprocessing

3.1 WAT Task Setup

The WAT English-to-Indic multimodal track setup comprises text only, image captioning, and multimodal tracks (Parida and Bojar, 2025; Parida et al., 2024). The WAT2025 instantiation covered English into Hindi, Bengali, Malayalam, and Odia (Parida and Bojar, 2025). The related WMT2024 low resource MMT task covers a comparable English-to-low resource setting including Indic languages (Parida et al., 2024). A distinguishing property of this task family is that captions are attached to rectangular image regions rather than whole images, so localized visual context is the relevant signal (Parida and Bojar, 2025). Our submission is a multimodal, multilingual system that uses the image, the annotated region, and the English caption jointly in one model covering all four languages.

3.2 Languages and Data

We work with the Visual Genome derived corpora for our four implemented languages. The hindi Visual Genome provides English–Hindi image-caption pairs created by post editing automatic translations with the image in view, together with a challenge subset concentrating ambiguous words (Parida et al., 2019). The bengali Visual Genome extends the same methodology to English–Bengali for both translation and captioning (Sen et al., 2022). Malayalam and Odia are covered by the Malayalam Visual Genome 1.0 and Odia Visual Genome releases (Parida and Bojar, 2021; Parida et al., 2025).

These follow the Multi30K precedent of pairing images with multilingual descriptions (Elliott et al., 2016). Our training uses the train and development splits only whereas test and challenge material is held out for evaluation (Table 1). For Odia, no challenge reference text was released (images only). Therefore, challenge evaluation covers Hindi, Bengali, and Malayalam.

Language	Train	Dev	Test	Challenge
Hindi	28,930	998	1,595	1,400
Bengali	28,930	998	1,595	1,400
Malayalam	28,930	998	1,595	1,400
Odia	28,930	998	1,595	–

Table 1: Dataset statistics.

3.3 Image Processing

Each annotated bounding box is cropped with 15% contextual padding (clamped to image bounds; degenerate boxes fall back to the full image) and converted to an RGB thumbnail capped at 448×448 preserving aspect ratio. Only this crop is fed to the model, with text preprocessing limited strictly to whitespace stripping.

3.4 Region Coordinate Representation

The padded box is additionally expressed on a 0–1000 scale and inserted into the prompt as `Region coordinates [x1, y1, x2, y2] on a 0-1000 scale`, so the model receives both the pixels of the region and its normalized location. We use annotated region cropping and coordinates only. There is no object detector and no learned grounding module beyond the vision-language model itself.

3.5 Prompt Construction

We render each instance as a single turn chat message: the system instruction, the region image, and the user text. The system instruction states that the model is given a cropped image region with its location and a short English caption describing that region, and must output exactly one line. The caption in the target language and script only, with no explanation, quotes, English, or romanization. Target scripts are named explicitly (Devanagari, Bengali, Malayalam, Odia). The user text gives the 0–1000 coordinates, the English caption, and requests the target language caption. The reference translation during finetuning is appended as the assistant turn; at inference the template ends with the generation prompt. Formally, with target language L_t, cropped region image C, normalized box B, and English caption x, the message sequence is

$$M = \big[\, s(L_t),\ C,\ u(B, x, L_t) \,\big],$$

where s is the system instruction, u the user text, and the rendered prompt is $P = \mathrm{ChatTemplate}(M)$ with the generation prompt appended; training appends the reference y as the assistant turn. The identical template is used in both approaches, so all measured differences are weight effects rather than prompt

effects. Table 2 summarizes the prompting and generation settings.

Setting	Value
System instruction	expert-translator framing + one-line target-script-only contract
User text	0–1000 box + English caption + target-script cue
Image turn	cropped 448-px region
Few-shot examples	none
Chat template	processor `apply_chat_template`, generation prompt appended
Training prompt	same template + reference as assistant turn
Decoding	greedy, max. 32 new tokens
Postprocessing	special tokens skipped, whitespace collapsed single line
Seed	3407 (training only)

Table 2: Prompting and generation settings.

4 Methodology

4.1 System Architecture

Figure 1 shows the implemented pipeline. Shared input processing turns each Visual Genome sample (image, bounding box, English caption) into a padded region crop plus a chat template prompt carrying the 0–1000 region coordinates. The both approaches feed this identical input to Gemma 4 E4B-it (4-bit NF4), which generates the Indic caption.

4.2 Zero Shot Baseline

The baseline loads `google/gemma-4-E4B-it` with its auto processor under 4-bit NF4 quantization with double quantization and FP16 computation, and scores it without any adapter. Evaluation covers the test split in all four languages and the challenge split where reference text is available (Hindi, Bengali, Malayalam); the Odia challenge text was not released, so Odia challenge scoring is skipped. Two isolated worker processes (one per GPU) generate per-language outputs with resumable, Hub synced prediction files.

4.3 Parameter Efficient Fine-Tuning

The fine-tuning adapts the same base model with LoRA (Hu et al., 2022) under 4-bit quantization (Dettmers et al., 2023). Training rows from all four languages are flattened into a single dataset with a single adapter (joint multilingual training); only train and development splits are touched. The collator masks the rendered prompt and supervises only the reference translation. The base weights stay frozen via the PEFT adapter mechanism. The adapter contributes 139,526,144 trainable parameters, i.e. 1.73% of the 8,080,626,976 total parameters, using `peft` 0.19.1 with QLoRA because `google/gemma-4-E4B-it` is already instruction tuned. Our objective was to adapt the model's existing multimodal instruction-following ability to English-to-Indic translation rather than perform extensive task-specific pretraining. The adapter was trained for 3,000 steps Section 5.

4.4 QLoRA Configuration

The base model is loaded with 4-bit NormalFloat quantization, double quantization, FP16 computation dtype, and scaled dot-product attention.

LoRA uses rank 64, alpha 128, dropout 0.05, no bias, causal LM task type, and targets the query, key, value, and output projections as well as the gate, up, and down projections of the feed-forward blocks. Each adapted weight matrix takes the form

$$W' = W_0 + \frac{\alpha}{r} BA,$$

where W_0 is the frozen base weight, A and B are the trainable low-rank factors ($r = 64$, $\alpha = 128$), A is randomly initialized, B starts at zero so that $W' = W_0$ at the beginning of training, and dropout is applied to the LoRA input (Hu et al., 2022).

4.5 Final Model for Evaluation

Training was run for 3,000 steps (Section 5), and the resulting `best_adapter` checkpoint is merged into the base model (`merge_and_unload`) and re-quantized to 4-bit NF4 at scoring time with the identical worker configuration, so the final numbers are produced by the same code path as the baseline. The evaluation script pulls the adapter from the Hub head without a pinned revision. The stored `best.json` identifies step 3000 (macro CHRF3 71.59) as the best and final checkpoint, so the evaluated adapter corresponds to that recorded selection artifact rather than to a pinned Hub revision.

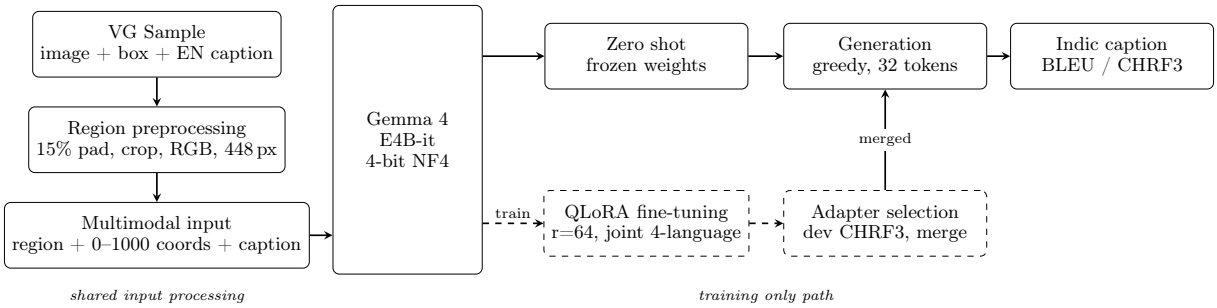


Figure 1: System architecture.

5 Training and Inference

Table 3 lists the training configuration. Optimization uses 8-bit AdamW with learning rate 10^{-4}, a cosine schedule, and a 0.05 warmup ratio. Each worker uses per device batch size 1 with gradient accumulation 8, giving an effective batch size of 16 over two-way data-parallel training; gradient checkpointing is enabled and the random seed is 3407. Maximum sequence length is 1024, with early stopping (patience 3, threshold 0.001); checkpoints are written every 500 steps (three retained) and pushed, rank-0 only, to a dedicated Hub repository together with training progress. Training is resumable from the latest local or Hub checkpoint. Training ran for 3,000 steps; the step-3,000 checkpoint is the one evaluated.

The training objective is causal language model cross-entropy restricted to the reference answer tokens. With rendered prompt P and reference token sequence $y = (y_1, \ldots, y_T)$, let $\mathcal{A}$ be the positions of the answer tokens (prompt prefix and padding positions are masked to -100 by the collator). The loss over the trainable LoRA parameters θ is

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \log P_\theta(y_t \mid y_{<t}, P),$$

with no label smoothing or token reweighting. The learning rate follows linear warmup for the first $s_w = 5\%$ of steps and cosine decay afterwards:

$$\mathrm{lr}(s) = \begin{cases} \mathrm{lr}_{\max} \cdot s / s_w, & s < s_w, \\ \frac{\mathrm{lr}_{\max}}{2} \left(1 + \cos \pi \frac{s - s_w}{S - s_w}\right), & s \geq s_w, \end{cases}$$

with $\mathrm{lr}_{\max} = 10^{-4}$ and $S = 3{,}000$ training steps.

Parameter	Value
Base model	`google/gemma-4-E4B-it`
Quantization	4-bit NF4, double quantization
LoRA rank / α	64 / 128
LoRA Dropout	0.05
Target modules	q, k, v, o, gate, up, down
Optimizer	AdamW (8-bit)
Learning rate	10^{-4}
Scheduler	cosine
Warmup	0.05
Batch size	$2 \times 1 \times 8 = 16$ effective
Hardware	2× T4
Gradient checkpoint	enabled
Max. sequence length	1024
Training steps	3,000 (compute-resource constraint)
Early stopping	patience 3, threshold 0.001
Checkpoint selection	best eval. loss + best macro CHRF3 adapter (step 3,000)
Generation	greedy, max. 32 new tokens

Table 3: Fine-tuning configuration as implemented.

Development monitoring combines forward loss scoring on a 400 row subset (100 per language) with generation scoring on a 200 row subset (50 per language) at every save. Checkpoint selection is dual: the trainer keeps the best model by development loss, while a separate `best_adapter` directory (with `best.json` and metric history) is refreshed whenever macro CHRF3 improves. Inference in both approaches is greedy (`do_sample=False`, at most 32 new tokens),

$$\hat{y}_t = \arg\max_w P_\theta(w \mid \hat{y}_{<t}, P),$$

with special tokens skipped and outputs whitespace normalized.

Setting	Value
Splits scored	test n=1,595 × 4; challenge n=1,400 × 3
References	dataset target captions, single reference per row
Predictions	per row JSONL (`idx`, `ref`, `decoded`) + plain text files
Metric code	`corpus_bleu` defaults + `corpus_chrf` (β=3)
Rounding / aggregation	2 dp; unweighted macro mean over evaluated languages
Dev role (fine-tune)	loss subset 400 + generation subset 200 for selection only

Table 4: Evaluation configuration.

6 Evaluation

6.1 Metrics

We report BLEU computed following the standardized detokenized reporting practice (Post, 2018), and chrF with beta 3 (CHRF3), the character n-gram F-score variant (Popović, 2017). Scores are computed per language and aggregated as unweighted macro means across the evaluated languages.

6.2 Experimental Setup

The zero-shot baseline is evaluated on the held-out test split (all four languages) and the challenge split (Hindi, Bengali, Malayalam). The fine-tuned adapter was monitored on development generation subsets during training and then evaluated on the same held-out splits. Decoding is greedy, so reruns are comparable without sampling variance. Final fine-tuned numbers come from scoring the merged best adapter on the full test (1,595 rows per language) and challenge (1,400 rows per language) splits, using the same prompt code, generation settings, row order, and metric code as the baseline. Table 4 summarizes the evaluation configuration, and Table 5 compares the two approach.

6.3 Results

On the development trajectory Table 6 where macro CHRF3 peaks at step 3,000 (71.59), which is the last checkpoint reached and therefore also the submitted adapter. Final held-out results are in Table 7.

7 Error Analysis

The both approach uses the same input construction, prompt template, decoding procedure, and evaluation code, the comparison isolates the effect of model adaptation under this experimental setup. However, without text only, image mismatch, or region/full image ablations, the contribution of visual information itself cannot be isolated. The test gains are uneven across languages (Table 7): Odia shows the largest absolute BLEU gain (+44.65, from 5.84 to 50.49) starting from the weakest zero shot baseline (CHRF3 37.25 to 72.82), while Hindi shows the smallest test BLEU gain (+10.10) starting from the strongest zero shot BLEU (31.99). Bengali (+18.93) and Malayalam (+23.85) fall in between. On CHRF3 the same ordering holds (Odia +35.57, Malayalam +16.26, Bengali +15.22, Hindi +7.85). These are observed patterns, not proof of why the model behaves this way. On challenge, Hindi reaches the best metric in our evaluation (50.58 BLEU / 69.68 CHRF3), while Bengali and Malayalam gains (+20.47/+14.10), (+15.55/+13.03) respectively.

8 Limitations

Our system adaptation uses a single rank-64 adapter, which bounds expressivity relative to full fine-tuning. Training was limited to 3,000 steps on two T4 GPUs (effective batch size 16, sequence length 1024) due to compute constraints; we do not claim this is an optimal budget or that longer training would not help. There is no dedicated fusion architecture, so architectural ablations are limited. The evaluated adapter is the recorded step-3,000 checkpoint, not a pinned Hub revision. Optimizer hyperparameters beyond learning rate, scheduler, and warmup use library defaults. Checkpoint selection reuses one 200-row development subset, and the same image pool underlies all splits by dataset design. Ablations, multiple seeds, and human evaluation remain future work.

9 Conclusion

We presented a Gemma 4 E4B-it based English-to-Indic multimodal translation system with region aware visual input, a zero shot baseline, and a quantized LoRA fine-tuning under joint multilingual training, evaluated with BLEU and CHRF3 under greedy based decoding because the base model is already instruction tuned. We adapted it for 3,000 steps, a prac-

Property	Zero shot baseline	Fine-tuned system
Model	Gemma 4 E4B-it, NF4, no adapter	same base + LoRA r64/α128 (139.5M, 1.73%)
Training data	none (eval only)	train 115,715 + dev 3,992, joint 4-language
Training	N/A	3,000 steps, eff. batch 16, 10^{-4} cosine
Prompt	Sec. 5 template, greedy 32 tokens	byte-identical template and decoding
Test macro	19.58 BLEU / 51.27 CHRF3	43.97 BLEU / 69.99 CHRF3
Challenge macro	23.67 BLEU / 56.35 CHRF3	40.37 BLEU / 68.25 CHRF3

Table 5: Baseline versus fine-tuned approach.

Step	Macro BLEU	Macro CHRF3	hi	bn	ml	or
500	43.70	67.20	36.01/59.23	36.81/66.88	49.04/73.38	52.95/69.29
1000	45.82	69.31	34.14/59.31	43.84/70.65	48.04/74.41	57.25/72.87
1500	41.70	65.43	33.60/58.14	45.02/68.44	32.73/63.70	55.47/71.44
2000	45.74	69.09	34.49/58.93	43.33/72.99	50.99/70.60	54.16/73.86
2500	41.67	67.03	39.29/61.54	42.39/74.40	32.09/60.29	52.90/71.88
3000	**49.76**	**71.59**	37.78/59.79	49.04/75.36	48.75/72.73	63.47/78.49

Table 6: Development trajectory up to 3,000 steps.

tical training budget under the computational resources available to us. The adapted system improves over the zero shot baseline on every language and split, with test macro gains of +24.39 BLEU / +18.72 CHRF3 and challenge macro gains of +16.70 / +11.90, confirming that parameter efficient adaptation of a frozen vision language backbone is effective for this task within this constrained training budget. Whether further training would yield additional gains remains an open question.

Acknowledgments

We thank the creators and annotators of the Visual Genome derived Hindi, Bengali, Malayalam, and Odia datasets (Parida et al., 2019; Sen et al., 2022; Parida and Bojar, 2021; Parida et al., 2025), and the WAT2026 organizers (WAT2026 Organizers, 2026) multimodal translation shared tasks for making the task data and evaluation protocols available.

References

Siddharth Betala, Kushan Raj, Vipul Betala, and Rohan Saswade. 2025. A picture is worth a thousand (correct) captions: A vision-guided judge-corrector system for multimodal machine translation. In *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 124–137, Mumbai, India. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *Transactions on Machine Learning Research*. ArXiv:2305.16307.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. ViTA: Visual-linguistic translation by aligning object tags. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 166–173, Online. Association for Computational Linguistics.

Ali Hatami, Shubhanker Banerjee, Mihael Arcan, Paul Buitelaar, and John Philip McCrae. 2024. English-to-low-resource translation: A multimodal approach for Hindi, Malayalam, Bengali, and Hausa. In *Proceedings of the Ninth Conference on Machine Translation*, pages 815–822, Miami, Florida, USA. Association for Computational Linguistics.

Tosho Hirasawa, Emanuele Bugliarello, Desmond Elliott, and Mamoru Komachi. 2023. Visual prediction improves zero-shot cross-modal machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 522–535, Singapore. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022)*. OpenReview nZeVKeeFYf9.

Shaharukh Khan, Ayush Tarun, Ali Faraz, Palash Kamble, Vivek Dahiya, Praveen Pokala, Ashish Kulkarni, Chandra Khatri, Abhinav Ravi, and Shubham Agarwal. 2024. Chitranuvad: Adapting multi-lingual LLMs for multimodal translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 839–851, Miami, Florida, USA. Association for Computational Linguistics.

Deepak Kumar, Baban Gain, Kshetrimayum Boynao Singh, and Asif Ekbal. 2025. Does vision still help? multimodal translation with CLIP-based image selection. In *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 115–123, Mumbai, India. Association for Computational Linguistics.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

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. Association for Computational Linguistics.

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).

Shantipriya Parida and Ondřej Bojar. 2025. Findings of WAT2025 English-to-Indic multimodal translation task. In *Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025)*, pages 103–108, Mumbai, India. Association for Computational Linguistics.

Shantipriya Parida, Ondřej Bojar, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, and Ibrahim Said Ahmad. 2024. Findings of WMT2024 English-to-low resource multimodal translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 677–683, Miami, Florida, USA. Association for Computational Linguistics.

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. ArXiv:1907.08948. Presented at CICLing 2019.

Shantipriya Parida, Shashikanta Sahoo, Kalyanamalini Sahoo, Ondřej Bojar, and Satya Ranjan Dash. 2025. Odia visual genome. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR. ArXiv:2103.00020.

Pawan Rajpoot, Nagaraj Bhat, and Ashish Shrivastava. 2024. Multimodal machine translation for low-resource Indic languages: A chain-of-thought approach using large language models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 833–838, Miami, Florida, USA. Association for Computational Linguistics.

Arghyadeep Sen, Shantipriya Parida, Ketan Kotwal, Subhadarshi Panda, Ondřej Bojar, and Satya Ranjan Dash. 2022. Bengali visual genome: A multimodal dataset for machine translation and image captioning. In *Intelligent Data Engineering and Analytics*, pages 63–70. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. ArXiv:1706.03762.

WAT2026 Organizers. 2026. WAT2026 English-to-Indic Multimodal Translation Task. https://ufal.mff.cuni.cz/wat2026english-indicmultimodaltranslation. Workshop on Asian Translation.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Appendix

	Zero Shot		Fine-Tuned		Improvement (Δ)	
Language	BLEU	CHRF3	BLEU	CHRF3	BLEU	CHRF3
Test Split						
Hindi	31.99	57.44	42.09	65.29	+10.10	+7.85
Bengali	22.67	53.82	41.60	69.04	+18.93	+15.22
Malayalam	17.83	56.55	41.68	72.81	+23.85	+16.26
Odia	5.84	37.25	50.49	72.82	+44.65	+35.57
Macro Avg	**19.58**	**51.27**	**43.97**	**69.99**	**+24.39**	**+18.72**
Challenge Split						
Hindi	36.51	61.10	50.58	69.68	+14.07	+8.58
Bengali	17.61	54.36	38.08	68.46	+20.47	+14.10
Malayalam	16.89	53.59	32.44	66.62	+15.55	+13.03
Macro Avg	**23.67**	**56.35**	**40.37**	**68.25**	**+16.70**	**+11.90**

Table 7: Test and Challenge results.