

From Text to Multi-Modal: Advancing Low-Resource-Language Translation through Synthetic Data Generation and Cross-Modal Alignments

Bushi Xiao
xiaobushi@ufl.edu

Qian Shen
qian.shen1@ufl.edu

Daisy Zhe Wang
daisyw@cise.ufl.edu

Abstract

In this study, we propose a novel paradigm for multi-modal low resource language dataset generation without relying on existing parallel multi-modal datasets. Leveraging advances in image-generation models, we introduce a systematic pipeline that transforms text-only parallel corpora into rich multimodal translation datasets. We then validate the generated content through human evaluation. We design and implement a new MMT model framework suitable for our new generated dataset. The model contains a verification mechanism with a large language model to ensure consistency between visual content and textual translations. Experimental results across four African low-resource languages with less than 10k training corpus demonstrate significant improvements over NLLB baselines, with average gains of up to 9.8% in BLEU score and 4.3% in METEOR score. Our method shows particular effectiveness in correctly translating concrete objects and contextual elements, suggesting its potential for improving low-resource machine translation through visual grounding.

1 Introduction

Large-scale pre-trained language models lead neural machine translation research into a new stage (Han et al., 2023), researchers are applying these powerful models to enhance translation for mainstream languages. However, these new models have not yet been widely adopted for translation tasks between low-resource and mainstream languages, especially some low-resource African languages. The few studies about machine translation for Hausa (Akinfaderin, 2020), Tigrigna (Hailu, 2024) and Kanuri (Tukur et al., 2024) mainly used LSTM or HMM-based models, not powerful transformer-based models. Although there are relatively more studies on machine translation for Yorùbá, there is still a huge gap in the application of multimodal machine translation models for Yorùbá,

which deserves further exploration (Adebara et al., 2022; Isaac; Timothy et al., 2024).

Currently, multimodal models have also been widely used in research works on machine translation. But in the research of multimodal low-resource language machine translation, several issues still warrant further discussion. For example, existing multimodal low-resource language parallel translation corpora are scarce, and some of them include only a limited number of low-resource languages. Most low-resource languages haven't been included in widely used multimodal low-resource parallel translation corpora. Some low-resource multimodal parallel translation corpora, such as Multi30K (Elliott et al., 2016), are created by researchers through manual annotation, which involves significant time and financial costs. While these datasets are of very high value, most researchers cannot adopt the same approach when creating more low-resource multimodal parallel translation corpora. It's hard to deny that the scarcity of corpora and the high costs of creating them pose significant obstacles to research in multimodal translation for low-resource languages. Additionally, the difficulty of adapting pre-trained multimodal models to low-resource languages also hinders the application of works such as CLIPTrans (Gupta et al., 2023) to translation tasks involving low-resource languages.

Furthermore, in the existing low-resource multimodal parallel translation datasets, the semantic consistency between text and images is still an issue worth considering. For instance, in the low-resource multimodal parallel translation dataset WIT (Srinivasan et al., 2021) created using Wikipedia, many texts consist of proper nouns, such as the names of some special buildings. If multimodal models cannot perfectly distinguish each building's unique name from a large number of similar building images, the translation results will easily be erroneous.

Our contributions

- Visual Data Generation¹: We employ state-of-the-art text-to-image generation models to create corresponding visual content for source-target text pairs, incorporating rigorous quality control mechanisms to ensure semantic consistency. This makes it possible to create corresponding multimodal low-resource translation corpora at a low cost based on existing text corpora for low-resource language translation.
- Model Architecture Development: We present an enhanced architecture which adopts a transformer-based encoder-decoder structure with cross-attention mechanisms, while introducing innovative LLM prompt engineering techniques as cross-modal alignment strategies.

This approach not only addresses the scarcity of multi-modal low-resource translation resources but also establishes a scalable framework for expanding multi-modal translation capabilities to low-resource languages. Consequently, our work can effectively avoid several major issues currently faced in multimodal low-resource language translation research, and facilitate the application of multimodal machine translation research to a greater number of low-resource languages.

2 Background

2.1 Multi-modal Translation

Researchers are currently using multimodal frameworks to build multimodal low-resource machine translation models that integrate both text and images. [Kwon et al. \(2020\)](#) built a modulation network based on text information from the encoder and visual information from the pre-trained ResNet, but it is still worth considering whether the image features watched by ResNet can align with the corresponding text. Visual content was used to extend the mask language model and generate a visual mask language model for unsupervised machine translation ([Tayir and Li, 2024](#)). During this study, researchers spent three months manually creating the dataset and it’s difficult to scale for applying to a larger number of languages. As a powerful multi-modal pre-trained model, CLIP ([Radford et al., 2021](#)) has also been applied to multi-modal

translation research. For example, the multi-modal machine translation model CLIPTrans is based on multi-modal M-CLIP and multilingual mBART ([Gupta et al., 2023](#)), and a new language-driven zero-shot multi-label recognition framework by using the aligned CLIP multi-modal embedding space ([Liu et al., 2024](#)). However, CLIPTrans was not trained using low-resource language datasets, making it difficult to extend its application to translation tasks involving more low-resource languages.

2.2 LLM-assisted Machine Translation

As one of the most popular research areas in recent years, researchers have begun to use LLMs to assist machine translation tasks. [Zeng et al. \(2023\)](#) proposed Collaborative Decoding (CoDec), which considers the NMT system as a pre-translation model and the MT-oriented LLMs as a complementary solution to handle complex scenarios beyond the capabilities of NMT. The research findings of [Ki and Carpuat \(2024\)](#) claimed that prompting LLMs to perform post-editing on MT can effectively improve the accuracy of translation results. [Qian \(2023\)](#) demonstrated that human-machine collaboration (HMT) using GPT-4 LLMs instructions enhances the effectiveness of translations. In addition to assisting with the translation tasks themselves, LLMs have also been utilized to assist machine translation from other different angles, such as supervised machine translation quality assessment models ([Huang et al., 2023](#); [Wang, 2023](#)) and data management tasks in neural machine translation ([Yin et al., 2024](#)).

2.3 LLM in Low Resource Language Translation

LLMs such as Claude 3 Opus in low-resource machine translation to English show stronger machine translation capabilities than other LLMs, surpassing strong baselines such as NLLB-54B and Google Translate on specific tasks ([Merx et al., 2024](#); [Enis and Hopkins, 2024](#)). Numerous research findings indicate that LLMs can effectively help improve the performance of machine translation models for low-resource languages such as Amharic ([Andersland, 2024](#)), Faroese ([Simonsen, 2024](#)), Ge’ez ([Wassie, 2023](#)), Indic languages ([Rajpoot et al., 2024](#)), Nepali ([Rimal and Abbas, 2024](#)) and Owens Valley Paiute ([Coleman et al., 2024](#)). As mentioned above, most applications of LLMs in low-resource language translation are limited to text-to-text translation tasks. In the few studies

¹Multimodal Dataset

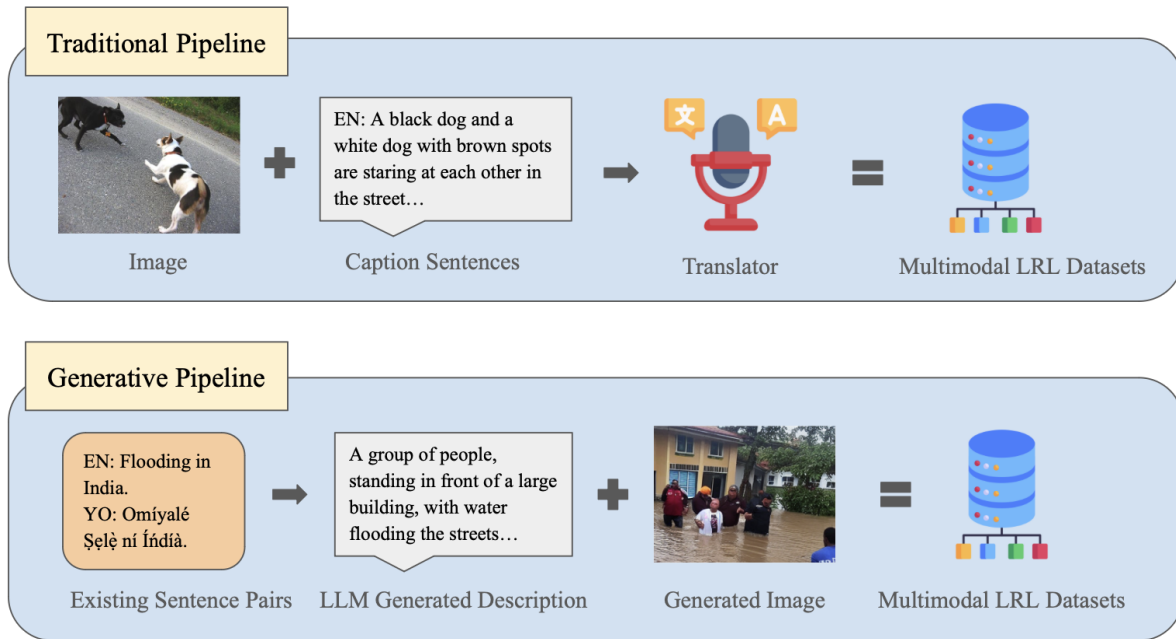


Figure 1: Comparison of dataset construction approaches: Our novel method for generating low-resource language multi-modal translation datasets versus traditional construction methods.

on multi-modal machine translation tasks, the semantic consistency between text data and images remains an unresolved issue.

3 Datasets

3.1 Text-only Datasets

The dataset we used includes parallel translation datasets for four low-resource languages from East and West Africa: Hausa, Kanuri, Tigrinya, and Yorùbá, along with English. The parallel translation datasets for Hausa, Kanuri, and Tigrinya with English are sourced from the Tatoeba corpus², while the parallel translation dataset for Yorùbá and English comes from the AI4D Yorùbá Machine Translation Challenge on the online data science platform Zindi³. In these datasets, the parallel translation datasets for Kanuri and Tigrinya to English contain 5,000 samples each, while the parallel translation datasets for Hausa and Yorùbá to English contain approximately 10,000 samples each. For each parallel translation dataset from language to English, we uniformly and randomly selected 25% of the samples to construct the test dataset, with the remaining data being used as the training dataset.

²Hausa - English, Kanuri - English, Tigrinya - English

³Yorùbá - English

3.2 Multimodal Dataset Generation

To construct a multimodal parallel translation dataset of low-resource African languages into English, we used the text-to-image model Stable Diffusion 3.5 Large Turbo to generate images aligned with the English text in the dataset. However, upon preliminary examination of the text in our parallel translation datasets, we found that a significant portion of the text such as "Is that your new friend?" doesn't contain sufficient visual information for the text-to-image model to generate images that can be aligned with the text.

As illustrated in Figure 1, we implemented a novel generative pipeline that differs significantly from traditional approaches to multimodal dataset construction. The traditional pipeline (top) typically combines existing images with caption sentences and processes them through a translator to create multimodal low-resource language datasets. This approach is limited by the availability of appropriate images and the quality of captions, often resulting in misalignments.

In contrast, our generative pipeline (bottom) starts with existing sentence pairs (English and the target low-resource language like Yoruba) and employs an LLM to generate detailed visual descriptions from these sentences. These descriptions serve as prompts for generating contextually

appropriate images using Stable Diffusion. This approach allows us to create highly aligned multi-modal content even for abstract concepts or culturally specific scenarios.

To implement this approach, we optimized the English text in the parallel translation dataset using Llama 3. We employed a few-shot learning approach, training Llama 3 with three sets of original English texts and their corresponding descriptions of images. We then used Llama 3 to transform the English texts into descriptions of images that maintain alignment with the original meaning. Finally, we used Stable Diffusion 3.5 Large Turbo to generate images based on these enhanced descriptions, creating a multimodal dataset with stronger text-image semantic alignment.

3.3 Multi-modal Dataset Evaluation

To evaluate the semantic consistency between the images generated according to the descriptions and the corresponding original texts, we randomly selected 200 original texts and their corresponding images. Then we performed a manual evaluation to determine whether the original texts contained visual information and whether the generated images were aligned with the corresponding original texts.

Figure 2 shows examples of aligned and misaligned text-image pairs in our dataset. In the aligned example (left), the English text "Can your brother drive?" corresponds appropriately with the image showing people with documents near a car, suggesting a context related to driving. In contrast, the misaligned example (right) shows a significant semantic gap between the English text "Life is hard" and an image of people looking at a large tree, where the visual content fails to represent the abstract concept expressed in the text.

The evaluation statistics in Table 1 indicate that the overall proportion of generated images consistent with the corresponding original texts reaches 80%. For original texts that contain visual information, the consistency rate with the corresponding images is 86.27%; for original texts that do not contain visual information, the consistency rate is lower, at 73.47%. Considering that only 51% of the original texts contain visual information, we believe that using Llama 3 to generate descriptions can effectively enhance the alignment between texts and corresponding images in the process of constructing parallel multimodal translation datasets.



Figure 2: Examples of aligned (left) and misaligned (right) text-image pairs in our dataset, showing how semantic consistency is evaluated across different languages.

Feature	Percentage
Overall Consistency ¹	80
Consistency with Visual Info ²	86.27
Consistency without Visual Info ³	73.47
Text with Visual Info ⁴	51
Text without Visual Info ⁵	49

Table 1: Multi-modal Dataset Evaluation Results.

- ¹ The ratio of the generated image to the corresponding original text.
- ² When the corresponding original text contains visual information, the proportion of generated images is consistent with the text.
- ³ When the corresponding original text doesn't contain visual information, the proportion of generated images is consistent with the text.
- ⁴ The proportion of original text containing visual information.
- ⁵ The proportion of original text that doesn't contain visual information.

4 Methodology

We propose a multimodal translation framework built upon NLLB (Costa-jussà et al., 2022) that enhances traditional text-based translation with visual context awareness. Our approach integrates three specialized components: NLLB's powerful transformer-based translation capabilities, vision model BLIP (Li et al., 2022) for image image caption generation, and open-source large language model Llama3 (Grattafiori et al., 2024) for cross-modal alignment supervision. This architecture is specifically designed to address the challenges of low-resource language translation by leveraging visual cues while maintaining NLLB's broad multilingual support.

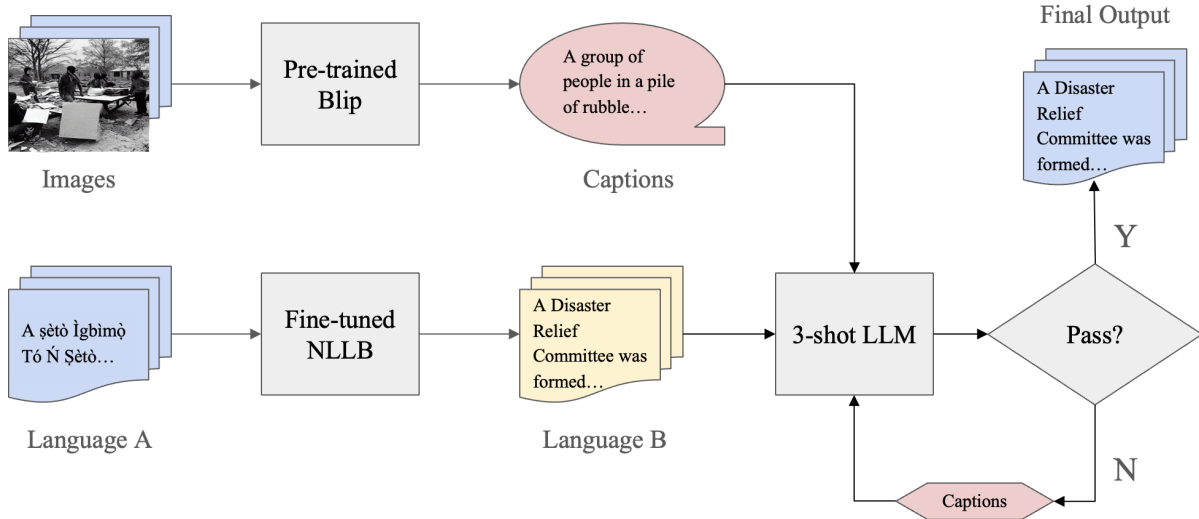


Figure 3: Architecture of our proposed translation model: A dual-path approach combining visual understanding (via BLIP) and text translation (via NLLB) with LLM-based quality supervision for low-resource language to English translation. Here, Language A represents a low-resource language sentence (Yoruba), and Language B represents English.

4.1 Baseline

The No Language Left Behind (NLLB) model is developed by Costa-jussà et al. (2022). It represents a significant breakthrough in addressing these challenges, particularly for traditionally understudied languages. By implementing a sophisticated Mixture-of-Experts (MoE) architecture within the Transformer framework, NLLB effectively manages the computational complexity while maintaining high translation quality across diverse language pairs. The model’s innovative approach to low-resource languages includes targeted data mining strategies and rigorous filtering mechanisms, ensuring robust performance despite limited training data. The model achieved a remarkable 44% improvement in BLEU scores for low-resource languages compared to previous state-of-the-art multilingual models.

4.2 New MMT Model

Figure 3 illustrates our flexible multimodal translation architecture that specifically addresses the challenges in low-resource language scenarios. The system employs a dual-path approach where visual and textual information are processed separately before integration.

At the core of our implementation, we utilize three key components: a pre-trained BLIP (Li et al., 2022) model for visual feature extraction, a fine-tuned NLLB⁴ model to handle low-resource

⁴NLLB-3.3B and NLLB-600M model in our experiments.

language translation, and a few shot Llama3-70b-Instruct (Grattafiori et al., 2024) serving as a quality supervisor. The NLLB component processes the low-resource language input (Language A), while BLIP generates English captions from the image. When the translation quality falls below our threshold, the system leverages these image-derived captions to supplement or replace the direct translation. Our quality control mechanism works as follows: the Llama evaluates the semantic alignment between the NLLB translation and the image context (represented by BLIP captions). Each translation receives a score on a scale of $1-10^5$. When score belows 6, the system push back the existing translation to llama3 and let it considers BLIP-generated caption to regenerate the final output⁶. This decision process ensures that the final output preserves semantic accuracy.

4.3 Evaluation Metrics

In neural machine translation evaluation, BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) serve as fundamental automatic evaluation metrics. While BLEU score primarily focuses on n-gram overlap between candidate and reference translations, METEOR enhances evaluation by incorporating stemming, synonymy, and paraphrase

Both models were further fine-tuned in pairs of languages X to English for our task.

⁵Detailed scoring criteria shown in Table 3 in Appendix D.

⁶See Appendix E.

matching, making it particularly valuable for low-resource scenarios where exact matches might be scarce (Denkowski and Lavie, 2014).

BLEU score is calculated as:

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

where:

$$BP = \min(1, \exp(1 - \frac{r}{c}))$$

r is the length of the reference translation

c is the length of the candidate translation

p_n is the modified n-gram precision

w_n is the weight for each n-gram level

METEOR score is calculated as:

$$METEOR = F_{mean} \cdot (1 - Penalty)$$

where:

$$F_{mean} = \frac{10PR}{R + 9P}$$

$$Penalty = 0.5 \cdot \left(\frac{ch}{m} \right)^3$$

P is precision

R is recall

ch is the number of chunks

m is the number of matched unigrams

5 Results

We evaluated our proposed multimodal machine translation framework with our newly constructed dataset. The experimental results demonstrate both quantitative and qualitative performance metrics. For quantitative evaluation, we measure translation quality using standard metrics including BLEU and METEOR scores. We also conduct a qualitative analysis on semantic preservation and contextual alignment between source texts, generated translations, and corresponding images. The results are compared with the baseline approaches to demonstrate the benefits of adding visual context in low-resource language translation.

5.1 Quantitative Evaluation

As shown in Table 2, while the NLLB 3.3B model generally outperforms its 600M counterpart due to its larger parameter size. Our proposed MMT

model has improvements over the 3.3B baseline across most language pairs. For Yoruba-English translation, our model achieves a BLEU score of 23.96 and a METEOR score of 50.26, significantly outperforming the NLLB 3.3B model (19.57/44.93) with relative improvements of 22.4% and 11.9%, respectively. The improvements are also evident in Hausa-English translation, where our model reaches a BLEU score of 37.21, and METEOR score of 71.84. Most notably in Kanuri-English translation, our model achieves substantial improvements with a BLEU score of 16.08, and METEOR score of 49.32, representing relative gains of 16.4% and 5.6% over the NLLB 3.3B baseline. These results demonstrate that our model architecture can effectively leverage its design advantages to outperform larger models in low-resource African language translation.

Language	Metrics	
	BLEU	METEOR
<i>Fine-tuned NLLB 600M</i>		
Yoruba - English	21.24	48.03
Hausa - English	34.24	68.22
Kanuri - English	12.46	45.77
Tigrinya - English	18.76	55.24
<i>Fine-tuned NLLB 3.3B</i>		
Yoruba - English	19.57	44.93
Hausa - English	36.75	70.35
Kanuri - English	13.81	46.69
Tigrinya - English	27.32	63.12
<i>New MMT Model</i>		
Yoruba - English	23.96	50.26
Hausa - English	37.21	71.84
Kanuri - English	16.08	49.32
Tigrinya - English	27.13	61.48

Table 2: Translation performance comparison between NLLB baseline model and our proposed model across four African low-resource languages.

The performance of our proposed model can be attributed to several key advantages of the multimodal dataset. First, by integrating visual information through the pre-trained BLIP model, our approach is particularly effective in correcting noun and subject errors in translation. This is due to the visual context providing a direct basis for these entities. Second, for extremely low-resource languages with less than 10k training sentences, the additional visual information can serve as an important supplementary for word recognition and disambiguation. Visual cues help to establish stronger se-




Image	LRL Sentence	Translation #1	Final Output	True Translation
	Àṣá ò lè balẹ kó gbéwúrẹ. <Yoruba>	The caterpillar can't sit still and dig.	The kite can't sit still the goat on the cliff.	The kite cannot swoop down and carry off a goat.
	Ngudowa kəla garbe so dəga kofi yakkinna ro lawargən. <Kanuri>	I asked the birds on the roof to share a cup of coffee.	I asked the birds on the roof while drinking a cup of coffee.	I watched the birds on the balcony while drinking my coffee.
	ደሮ ሓገይ ኩይት። <Tigrinya>	It's already winter.	It's already summer.	It's summer already.

Figure 4: Examples demonstrating how our model progressively improves translations by leveraging visual information and LLM verification, comparing the initial translation result (Translation 1) with our model’s final output and the ground truth.

mantic connections between source language terms and English terms, leading to more accurate lexical choices. The improvement in vocabulary accuracy is better reflected in the BLEU metric, which is particularly sensitive to exact vocabulary matches. Furthermore, our LLM verification mechanism acts as a quality control gate to help filter out translations that are semantically inconsistent with the visual and textual context. But at the same time, it reduces making changes on passed sentences, this can avoid decreasing BLEU&METEOR scores since LLMs can also make mistakes.

5.2 Qualitative Evaluation

Figure 4 shows three examples across different low-resource language datasets, revealing that the multimodal architecture significantly improves semantic accuracy, particularly in noun phrase translation. For instance, as shown in Figure X, when translating from Yoruba, our model correctly identifies "kite" from the visual context, while the baseline NLLB model incorrectly translates it as "caterpillar". Similarly, in the Kanuri example, the visual information helps specify the correct location as "balcony" rather than the more generic "roof". This pattern of improvement in entity and object recognition is consistent across our test set, where the visual context effectively grounds abstract lingu-

tic tokens to concrete visual representations.

Moreover, the LLM assistance step maintains semantic consistency between the visual scene and the generated translation. As demonstrated in the Tigrinya example, the model correctly identifies the seasonal context as "summer" based on visual cues.

However, there are also many error cases in the final outputs. For example, in the Yoruba test set, the correct translation of a sentence 'No description' is predicted as 'There is no image', while the visual information can be understood as both the former sentence and the latter sentence. The limitation of LLM’s verification method in context makes it not possible to recognize that an error has occurred here, thus making this sentence pass.

These limitations stem partly from the inherent challenge of aligning abstract sentences with visual content. Many sentences in everyday communication are conceptual or abstract in nature and cannot be perfectly represented visually. In such cases, the visual information offers limited assistance in correcting translation errors, preventing our system from achieving higher performance scores.

Our approach represents a fundamental shift from traditional multimodal translation research. While existing systems typically rely on datasets with perfectly aligned images and captions, this sce-

nario rarely occurs in real-world communication where daily language is conversational, abstract, or conceptual with only partial visual correspondence. This limitation has created a problematic cycle where computational resources remain disproportionately allocated to high-resource languages, perpetuating linguistic inequalities in NLP development. Our model addresses this inequity by working with realistic, imperfectly aligned multimodal content rather than artificially constrained datasets, though this approach introduces additional evaluation challenges and unique error patterns.

6 Conclusion

Low-resource languages have been particularly neglected in multimodal translation research precisely because they lack perfectly aligned multimodal datasets. We presented a comprehensive approach to advance low-resource machine translation. Our method transforms text-only parallel corpora into visually enriched datasets using LLM-guided image generation, ensuring semantic alignment between the generated visual content and original sentences. Building on this multimodal data set, we propose a novel translation architecture that takes advantage of visual context and LLM assistance to improve translation quality. Experimental results in four low-resource African languages demonstrate significant improvements over NLLB baselines in BLEU Scores and METEOR Scores. Beyond these immediate results, our approach offers broader implications for language preservation and accessibility. The framework’s adaptability makes it promising for any other low-resource languages, potentially aiding in the digital preservation of endangered languages and facilitating cross-cultural communication. This flexible solution could contribute significantly to reducing language barriers in digital spaces while helping preserve linguistic diversity for future generations.

7 Limitations

LLM as a Black Box

Although the LLM component enhances translation supervision, it lacks the ability to truly understand low-resource language semantics or learn their intrinsic grammatical structures. This suggests that while multimodal integration can improve evaluation metrics, truly capturing the underlying linguistic structures remains an open challenge.

Misalignment

The absence of a generalized assessment method for image-text alignment presents a significant challenge in our approach. This issue is particularly acute for low-resource languages where our used parallel corpora originate from diverse, often inconsistent sources. While data filtering could potentially improve alignment quality, it would further reduce the already limited training sets, exacerbating the data scarcity problem. A potential solution could be implementing back translation techniques to augment the dataset while maintaining semantic consistency, but this approach would need careful validation to ensure it doesn’t propagate or amplify existing misalignment.

Baseline Comparison

Our baseline comparison is constrained as there are no multimodal translation models for low-resource languages that would serve as more appropriate benchmarks. Existing multimodal models like LLaVA (Liu et al., 2023) primarily support English, while CLIP’s capabilities are also limited to widely-spoken languages such as English and German (Gupta et al., 2023) on translation tasks. So we used the text-only NLLB as our baseline, despite its inability to leverage visual information. This limitation makes it difficult to fairly evaluate whether our specific architecture is optimal or if similar results could be achieved through more efficient approaches that incorporate visual modalities.

References

- Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2022. Linguistically-motivated yorùbá-english machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5066–5075.
- Adewale Akinfaderin. 2020. Hausamt v1. 0: Towards english-hausa neural machine translation. *arXiv preprint arXiv:2006.05014*.
- Michael Andersland. 2024. Amharic llama and llava: Multimodal llms for low resource languages. *arXiv preprint arXiv:2403.06354*.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

- Jared Coleman, Bhaskar Krishnamachari, Khalil Iskarous, and Ruben Rosales. 2024. Llm-assisted rule based machine translation for low/no-resource languages. *arXiv preprint arXiv:2405.08997*.
- 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](#). *Preprint*, arXiv:2207.04672.
- Michael Denkowski and Alon Lavie. 2014. [Meteor universal: Language specific translation evaluation for any target language](#). In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.
- Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.
- Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *arXiv preprint arXiv:2404.13813*.
- Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Devaansh Gupta, Siddhant Kharbada, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. Cliptrans: transferring visual knowledge with pre-trained models for multimodal machine translation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2875–2886.
- Fitiwi Hailu. 2024. [Tigrigna-English Bidirectional Machine Translation using Deep Learning](#). Ph.D. thesis, St. Mary’s University.
- Xue Han, Yi-Tong Wang, Jun-Lan Feng, Chao Deng, Zhan-Heng Chen, Yu-An Huang, Hui Su, Lun Hu, and Peng-Wei Hu. 2023. A survey of transformer-based multimodal pre-trained models. *Neurocomputing*, 515:89–106.
- Hui Huang, Shuangzhi Wu, Xinnian Liang, Bing Wang, Yanrui Shi, Peihao Wu, Muyun Yang, and Tiejun Zhao. 2023. Towards making the most of llm for translation quality estimation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 375–386. Springer.
- O Isaac. Machine translation system for numeral in english text to yorùbá language.
- Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. *arXiv preprint arXiv:2404.07851*.
- Soonmo Kwon, Byung-Hyun Go, and Jong-Hyeok Lee. 2020. A text-based visual context modulation neural model for multimodal machine translation. *Pattern Recognition Letters*, 136:212–218.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. [Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation](#). *Preprint*, arXiv:2201.12086.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.
- Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. 2024. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *International Conference on Machine Learning*, pages 32173–32183. PMLR.
- Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented llm prompting: A study on the mambai language. *arXiv preprint arXiv:2404.04809*.
- 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.
- Ming Qian. 2023. Performance evaluation on human-machine teaming augmented machine translation enabled by gpt-4. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 20–31.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- 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.
- Kshitiz Rimal and Noorhan Abbas. 2024. Enhancing nepali text understanding with machine translation and lora fine-tuning of open-source llm. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 313–319. Springer.

Annika Simonsen. 2024. *Improving Machine Translation for Faroese using ChatGPT-Generated Parallel Data*. Ph.D. thesis.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449.

Turghun Tayir and Lin Li. 2024. Unsupervised multimodal machine translation for low-resource distant language pairs. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(4):1–22.

Adeboje Olawale Timothy, Olusola Adebayo Adetunmbi, Arome Gabriel Junior, and Akinyede Raphael Olufemi. 2024. [Bilingual neural machine translation from english to yoruba using a transformer model](#). *International Journal of Innovative Science and Research Technology (IJISRT)*.

A Tukur, A Jibrin, and U Inuwa. 2024. Towards efficient part-of-speech tagging for the kanuri language: A hidden markov model-based solution. *Nigerian Journal of Engineering Science and Technology Research*, 10(2):115–123.

Yiheng Wang. 2023. Large language models evaluate machine translation via polishing. In *Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence*, pages 158–163.

Aman Kassahun Wassie. 2023. Machine translation for ge’ez language. *arXiv preprint arXiv:2311.14530*.

Yongjing Yin, Jiali Zeng, Yafu Li, Fandong Meng, and Yue Zhang. 2024. Lexmatcher: Dictionary-centric data curation for llm-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14767–14779.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Improving machine translation with large language models: A preliminary study with cooperative decoding. *arXiv preprint arXiv:2311.02851*.

Appendix A

Examples of transforming abstract sentences into concrete visual scenes.

Prompt Structure:

Input: [input sentence]

Scene Description: [detailed scene description]

Training Examples:

1. Example 1

Input: “He was only allowed to play for an hour a day.”

Scene Description: “A boy was sitting at the table with a disappointed look on his face. Next to him stood his parents, lecturing him. Toys were scattered on the floor.”

2. Example 2

Input: “There is some hope.”

Scene Description: “A group of people stand on the parched earth, surrounding a flower that sprouts from the earth.”

Appendix B

When evaluating whether the original English text in our dataset contains visual information, we looked for words that describe tangible objects. For example, the sentence “I was suspicious” only conveys an abstract attitude, so we consider this sample to not include visual information. Conversely, a sentence like “The bridge is open to traffic” includes at least two elements, a bridge, and vehicles, thus we consider this sample to contain visual information. Similarly, we manually evaluated whether the generated images matched the original English by checking if the key elements in these original English texts aligned with the scenes in the corresponding images.

Appendix C

We ran Stable Diffusion 3.5 Large Turbo on an NVIDIA A100 GPU for image generation, with each image taking approximately one second to generate. Since the English texts in the parallel translation datasets from Kanuri and Tigrinya to English are identical, we effectively generated images corresponding to around 25,000 English originals. In total, this process took approximately seven hours on the NVIDIA A100 GPU.

We used three NVIDIA A100 GPUs for the multimodal translation task. The image recognition and caption generation steps took about eight hours in total. Fine-tuning the 3.3B NLLB took three hours, and fine-tuning the 600M NLLB took about one hour. LLM validation task takes about two seconds per sentence, for a total of about three hours on all test sets.

Appendix D

We designed a comprehensive scoring framework for evaluating translation quality using LLMs as shown in Table 3. To assess how well translations align with their visual contexts, we created a 0-10 scale with specific criteria for each range.

We provided these scoring criteria along with example pairs to Llama in a few-shot learning approach, teaching the model how to evaluate semantic alignment between images and translations. The examples demonstrate various degrees of alignment - from perfect matches (graduation ceremony with "They graduated today") to completely contradictory pairings (birthday party with "Funeral was yesterday").

Recognizing that low-resource language datasets often have imperfect alignment with images, we intentionally designed our scoring criteria to be somewhat flexible. For instance, translations scoring in the 5-6 range may have only indirect connections to the visual scene, like "The course starts from 9 o'clock" paired with a classroom teaching scene. This flexibility acknowledges the real-world challenges of multimodal translation while still maintaining meaningful quality standards.

Appendix E

Below is the prompt template used to guide the LLM in generating alternative translations when the initial translation fails to align with the image content:

You are an expert in multimodal translation who specializes in ensuring semantic alignment between images and text in multiple languages. I'll provide you with:

1. An English translation of a sentence from a low-resource language
2. A caption describing an image that was paired with the original sentence

The current translation does not align well with the image content. Your task is to generate a new English translation that

- Maintains the core meaning of the original translation when possible
- Better aligns with the visual content described in the image caption
- Reads naturally in English

Original Sentence: {original_sentence}

Original Translation: {original_translation}

Image Caption: {image_caption}

Instructions: Generate a new translation that better preserves the meaning while aligning with the

visual content. Provide only the new translation without explanation

New Translation: {llama_output}

This prompt is used when the LLM evaluation score falls below our threshold of 6, indicating insufficient semantic alignment between the translated text and the corresponding image.

Score	Description	Example Match
9-10	Perfect semantic alignment Most key elements present	Scene: Graduation ceremony. Sentence: "They graduated today."
7-8	Strong correlation Core meaning preserved	Scene: A group of doctors and nurses surround a man in a hospital bed. Sentence: "He had to undergo surgery."
5-6	Partial alignment Indirect connection	Scene: There is a teacher teaching student in the classroom. Sentence: "The course starts from 9 o'clock."
3-4	Limited connection Major elements missing	Scene: A girl is playing skateboard in the park. Sentence: "Summer starts soon."
1-2	Minimal alignment Very loose connection	Scene: A boy sat on bed with his toys. Sentence: "Communicating with Others."
0	No connection Contradictory elements	Scene: Children at birthday party. Sentence: "Funeral was yesterday"

Table 3: Scene-Sentence Scoring Criteria