A THOROUGH INTRODUCTION TO MULTIMODAL MACHINE TRANSLATION

Kouassi Konan Jean-Claude

Faculty of Computer Science via distance learning, Bircham International University, Madrid, Spain

ABSTRACT

Five years before the release of ChatGPT, the world of Machine Translation (MT) was dominated by unimodal AI implementations, generally bilingual or multilingual AI models with only text modality. The era of Large Language Models (LLMs) led to various multimodal translation initiatives with text and image modalities, based on custom data engineering techniques that introduced expectations for improvement in the field of MT when using multimodal options. In our work, we introduced a first of its kind AI multimodal translation with four modalities (text, image, audio and video), from English towards a low resource language and vice-versa. Our results confirmed that multimodal translation generalizes better, always brings improvement to unimodal text translation, and superior performance as the number of unseen samples increases. Moreover, this initiative is a hope for worldwide low resource languages for which the use of non-text modalities is a great solution to data scarcity in the field.

KEYWORDS

Artificial Intelligence (AI), Machine Learning (ML), Multimodal Machine Translation, Data Engineering, Multimodal Dataset, Baoulé language (bci)

1. INTRODUCTION

In a previous paper, we explained that Data Engineering (obtaining, cleaning, and representing data) is at least as important as Algorithm Engineering (cf. [1], introduction of section 18.11 and section 18.11.2), and we presented the responsibility for the Data Engineer to design the right data for further ML tasks. By this means, we consequently provided the Secret Behind Data Engineering for Machine Translation Tasks [2]. The present work aims to go further.

With the recent releases of Large Language Models (LLMs) such as ChatGPT 3.5 in January 2023 and Bard in May 2023, the effectiveness of Multimodal AI is still at its early stage in 2025. Subsequently, the Multimodal models' skill share in AI job postings in the United States was of 0.15% in 2023 and 0.67% in 2024, against more than 60% for the requirement of skill in Generative artificial intelligence (cf. [3] p. 227). Indeed, the first most significant multimodal models (with only text and image modalities) had been released from major tech companies and research institutions since December 2023 (Google's TinyGPT-V and Gemini Nano, MoE-LLAVA, DeepSeek-VL, LLAVA-Gemma, etc.), cf. Figure 11.1 of [4]. In February 2025, the field of multimodal AI is led by top leading industry and academia, with Open AI and Google as top providers of multimodal models, globally covering audio, image, video, and text modalities, depending on the specific model’s name, but without a specific focus on Machine Translation. Despite these proprietary frameworks don't allow the download of fine-tuned models, they store the weights of trained models on various nodes, so creating a global environment for research and development activities.

Especially, for the field of **Multimodal Machine Translation**, the traditional LLMs also known as traditional transformers led to various multimodal translation initiatives with text and image modalities, based on custom data engineering techniques. However, the recent breakthroughs in LLMs implied the use of more modalities and introduced hopes for significant improvement in the field of multimodal translation that is at its early stage, not well defined, and so still under exploration. As good examples illustrating our analysis, two papers released in 2024 used the path for traditional transformers for their multimodal translation systems. In a low-resource setting, [5] used images annotated with solid red lines considering the coordinates of a part of the image (partial image) and full image to produce a Multimodal Machine Translation model with text and image modalities. Their experimental results demonstrated that the multimodal approach with text and image modalities consistently outperforms the text-only baseline. In a high-resource setting, [6] used full images and ambiguity scores to try to enhance hybrid multimodal translation over the text-only approach. They found that despite the benefits of incorporating visual data into Multimodal Machine Translation, its use can sometimes lead to reduced translation quality compared to text-only approaches (the multimodal approach outperformed the text-only baseline in 32% of cases). Both approaches in [5] and [6] used the gated fusion approach (Wu et al., 2021) to fuse both textual and visual information. While in January 2025, an architecture based on recent breakthroughs in LLMs allows for more modalities (text, image, video and audio) via an implementation of a multimodal LLM-based Multi-Agent System (MAS), accessible via a chat interface and providing positive results in terms of system's practicality and scalability (cf. Fig. (1) and Fig. (7) of [7]). This framework showed successful implementations in the global field of multimodal AI without a specific focus on MT. To the best of our knowledge, there are currently (in May 2025) few or no initiatives of Multimodal Machine Translation covering at least four (04) modalities at once, the MAS being a QA framework. Moreover, there is no standalone model capable of overcoming this struggle, we only find frameworks accessible via nodes that assemble the training weights.

In this paper, we built upon our previous multilingual Machine Translation work (cf. [2]) to create a Multimodal Machine Translation Model. The adopted approach is based on our own Data Engineering techniques, coding and currently available relevant multimodal LLMs. More specifically, we designed our custom Multimodal Dataset and training strategy, and coded it to match the chosen LLM provider fine-tuning processes. In the following lines, after a comprehensive explanation of the adopted multimodal data collection process, we will provide more details about our experiments, and we will discuss the results obtained from our strategies via a thorough analysis.

2. THE DATA COLLECTION PROCESS

We have already presented in the introduction three (03) previous works related to Multimodal MT and to Multimodal Dataset creation with two (02) and four (04) modalities. In an Omniscien Technologies' webinar about AI and Language Processing Predictions for 2025 [8], Professor Kohen also presented the Audio-Visual Speech Translation (AV2AV) framework (cf. Figure 3 of [9]). It is modality agnostic and assembles three (03) research areas into a standalone framework approach (A2AVT with audio input, V2AVT with video input from lips reading videos, and AV2AVT with audio/video input). So, existing initiatives cover one to two-modality MT, modality agnostic MT, and global multimodal AI with four modalities. In this section, we will show that a successful Multimodal Dataset creation process is not obtained hazardously, but through a thorough analysis and an excellent data collection strategy.

2.1. Data Extraction Technique

We know that the quality of this delicate, complex, and hard data creation stage will highly impact the performance of related Multimodal Machine Translation models (cf. section 4.2.1.1 of [2]). Indeed, internet pages and raw documents or files cannot directly be applied to Algorithms (for AI tasks) without a watchful Data Engineering stage because they are just raw or dark data. The Multimodal Machine Translation model we want to create is intended to cover audio, image, video, and text modalities. It will translate from English (en) to Baoulé (bci) and vice-versa via a standalone multimodal model that will handle the four (04) modalities. Therefore, we will present below the specific approach we adopted for each modality.

Text Modality

We first created a **French Latin Script Converter model** for handling both the Latin-based script and French-based script of the bci language for the Text modality. The dataset used for training this model was carefully selected and restructured from the mtBCI-1.0-Corpus provided in our previous paper (cf. Appendices of [2]). This model results (TER and WER) presented better prediction quality for conversion from French towards the Latin script. This is why we chose the French Script as the format of the text modality of our multimodal model; the converter model will then be applied for getting the Latin version of translations. We could build a GUI to allow users to choose independently Latin or French-based script as output. However, we will start by providing the Text modality directly in French Script with respect to the results provided by the Converter model that is made available in the appendices with a gradio UI layer as shown below.

Figure 1. A Gradio UI for the French Latin Script Converter model

Next, we made another thorough text selection from the mtBCI-1.0-Corpus to create a dataset that will be used to train the text modality of our **Multimodal Machine Translation Model** and also train a **Multilingual (Bilingual) Model** against which compare our multimodal model.

Image Modality

For this modality, we created a dataset of 107 English annotated images labeled in Baoulé and vice-versa. So, the total number of images for the image modality is 214. As we recommended in section 6.2 of [2], we also used Google translate Baoulé translations to accelerate the translations of the image labels, including some edits in order to really match high quality labels (the Baoulé language has been added to Google translate last year in June 2024).

Audio Modality

For the audio modality, we got 347 audio files with Latin script transcriptions from the *Baule Speech Dataset For Automatic Speech Recognition Task* [10], and we completed this dataset with

our own English audio files leading the total number of files for training the audio modality to 696 (348 English files and 348 Baoulé files). The dataset was then created via manual selection and coding. The French Latin Script Converter model was also used here to convert Latin transcriptions from the ASR dataset towards the French Script.

Video Modality

For the video modality, we created small size videos from some Baoulé videos freely available through the Internet (cf. [11-19]). Then from these videos, we created 160 short Baoulé videos. But we didn't provide English videos for the training.

For each of these modalities (text, image, audio and video), we defined a coherent data structure and filenames that were then used to prepare the training sets via custom hardcoding.

2.2. Corpus Analysis and Statistics

Below we propose tables of statistics about the collected data (number of files or language pairs, depending on the data type).

Table 1. Statistics about the French Latin Script Converter Dataset.

	Latin Script	**French Script**
TOTAL	**6 201**	**6 198**
	12 399	

Table 2. Statistics about the English/Baoulé Multimodal Dataset.

		English				**Baoulé**			
Modalities		**Text en-bci**	**Audio en-bci**	**Image en-bci**	**Video en-bci**	**Text bci-en**	**Audio bci-en**	**Image bci-en**	**Video bci-en**
TOTAL	**13 582**	6 256	348	107	0	6 256	348	107	160
	13 582	**6 711**				**6 871**			

2.3. Coverage of Our Contribution

We underline that we created our Multimodal Dataset following the scientific method aiming to provide an easy way to improve the performance of Baoulé-related language models in a very-low resource context. We proposed a first converter model for covering the two (02) types of scripts in which the bci language is available. The French Latin Script Converter ensures that everyone will be able to use his preferred script according to his own goal. This model will follow-up new rule definitions from linguists for further improvements.

On the other side a working bci-related Multimodal Dataset will considerably reduce the difficulties about the data collection process, focusing on modalities that are easier to provide, perhaps audio or video, mostly for a very-low resource context. Moreover, our methods are adaptable and reproducible to other very low-resource and low-resource languages. In Côte d'Ivoire alone, over seventy-nine (79) very low-resource languages defined by linguists,

according to Ethnologue, we already have three (03) extincted: Esuma (esm), Gbin (xgb), and Tonjon (tjn). Additionally, fourteen (14) of them are currently endangered (cf. ethnologue.com, Côte d'Ivoire). We think initiatives like ours will contribute to overcoming the struggle of providing enough data to the NLP domain scientific community for language preservation.

3. BASELINE SYSTEM

In this section, after the collection of a representative multimodal dataset via a thorough Data Engineering Process, we will provide a **baseline multimodal model** proving that the collected data is enough and sufficient for a fair comparison of the multimodal model performance against the traditional multilingual one.

3.1. Experimental Setup

In order to provide a suitable experimentation environment for training the curated Multimodal Dataset, we should first choose a foundation or base model that can directly or indirectly handle multimodal data. A good solution for training our Multimodal MT is to use an Agent-based Multimodal Translation approach. And as currently there is no standalone approach available that covers the four modalities, an architecture similar to the multimodal LLM-based Multi-Agent System (MAS) presented in [7] is a good plausible solution. However, we could also use proprietary models that store the weights of trained models on various nodes and make them available for research and development activities. The latter solution will prevent us from having to train several agents or adapters one by one.

Therefore, after a thorough analysis matching our use case, given that existing standalone multimodal base models generally offer up to three modalities at max, we chose Gemini 2.0 Flash-Lite that is currently the cost-effective Gemini model of Google to support high throughput. It supports many data types as an input for enabling multimodal research with audio, image, video, text, code, and PDF input modalities and only text output. So, after some experimentations, we formatted our multimodal dataset to match the chosen base model requirements, and we used main hyperparameters settings for mitigating with overfitting/underfitting and providing a fair comparison of the results. We trained our baseline models with the following hyperparameters:

Table 3. Main training setup for the baseline models related to the multimodal MT.

Base LLM	Fine-Tuning Task	Adapter Size	Target Modules	Epochs	Learning Rate	Early Stopping
gemini-2.0-flash-lite-001	Supervised Fine-Tuning	1	q_proj, v_proj, gate_proj, up_proj, down_proj	60	0.00001	N/A

For the evaluation, we chose among the ROUGE, TER, WER, BLEU and SacreBLEU metrics. Indeed, the ROUGE metric computes the similarity between the machine-generated text and the human reference text. The TER metric estimates the post-editing effort required on the translation to match the reference. The WER metric indicates the percentage of words that were incorrectly predicted and the BLEU and SacreBLEU scores are based on the Longest Common Subsequence and Skip-Bigram Statistics.

Table 4. Chosen metrics.

Chosen METRICS	Low Quality range	Medium Quality range	High Quality range
ROUGE Score	[0.0 - 0.3]	[0.3 - 0.6]	[0.6 - 1.0]
TER Score	[0.7 - 1.0]	[0.4 - 0.7]	[0.0 - 0.4]
WER Score	[20 - 100]	[10 - 20]	[0 - 10]
BLEU Score	[0.0 - 0.2]	[0.2 - 0.4]	[0.4 - 1.0]
SacreBLEU Score	[0 - 20]	[20 - 40]	[40 - 60]

During our experiment, the Multimodal Dataset was split as shown in Table 5 below. Let's note that the unseen evaluation set is thoroughly chosen in order to ensure diversity assessment through the model output, in a very low-resource context (more details in section 3.3).

Table 5. Statistics about the splits of the Multimodal Dataset.

Data Sets	English				Baoulé in French-based script			
	Text en-bci	Audio en-bci	Image en-bci	Video en-bci	Text bci-en	Audio bci-en	Image bci-en	Video bci-en
Training Data	6 230	328	97	0	6 230	328	97	150
Unseen Evaluation set	60	20	10	10	60	20	10	10
All	**6 290**	**348**	**107**	**10**	**6 290**	**348**	**107**	**160**

Therefore, for the matter of diversity coverage in the evaluation set we carefully selected 26 pairs from the provided text modality and completed them to 60 for creating the text modality of the evaluation set, adding 34 new unseen language pairs. Let's note that in order to have a fair comparison, for the unseen evaluation set we provided only text at the input of both baseline models (traditional multilingual and multimodal) so that they are evaluated exactly on the same inputs. For that we used the transcriptions of Image, Audio and Video files while keeping their labels. We think that after the training, if there is an improvement within the multimodal model, it will provide better performance on the same unseen evaluation set.

3.2. Results

In this section, we will present the results of our experiments. We will look first at the behavior of our learning curves to know if our models learn well. Then, we will provide an evaluation of each baseline model on our chosen metrics (cf. Table 4 of this paper), we retained ROUGE and BLEU among them (checking for human-like, syntactic and semantic correctness in their outputs). The multimodal model performance will be evaluated and compared to the traditional multilingual model.

3.2.1. With our French Latin Script Converter Model

The learning curves presented below show that for our French Latin Script Converter model, the training and evaluation scores gradually decrease from 3.5 towards 0.5 and lower values.

Figure 2. The French Latin Script Converter model's training curves

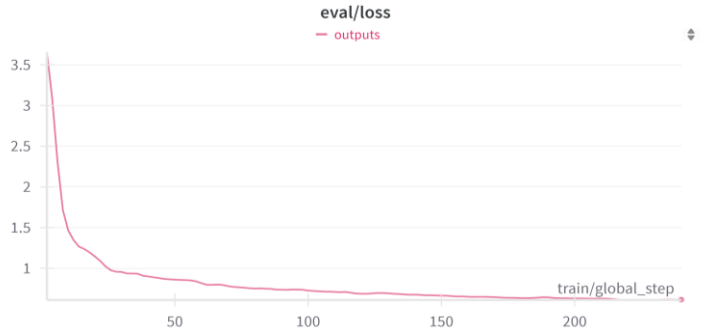


Figure 3. The French Latin Script Converter model's validation curves

Our hypothesis

As explained in section 2, this French Latin Script Converter model helped us in building our Multimodal Dataset. Then, we started our experiments with the hypothesis that a multimodal Machine Translation (MT) model with four modalities (text, image, audio and video) built upon a multilingual Machine Translation baseline model should always be beneficial to the multilingual MT (never decrease its performance) and even improve it.

Next, in the direction of our hypothesis, we built a small multimodal dataset (50 text data, 696 audio data, 32 video data and only 4 images data) selected according to our data engineering strategy explained in section 2. The results showed that this small multimodal dataset supported the only 50 text samples to yield a good model performance (80% of accuracy). It is evident that in the domain of multilingual MT, the 50-training data was not able to lead to that performance.

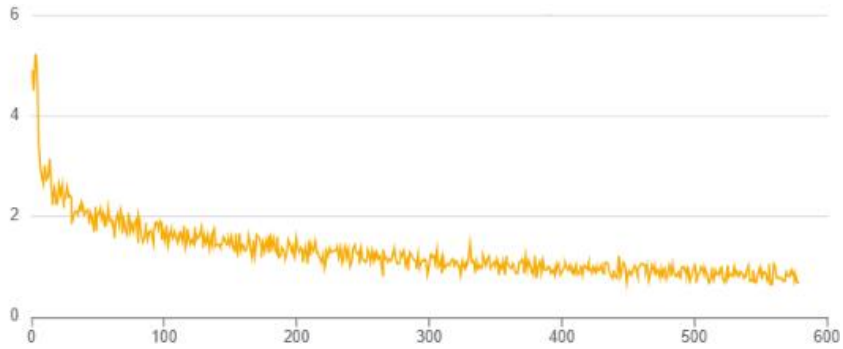


Figure 4. The training curves of the small multimodal model (hypothesis dataset)

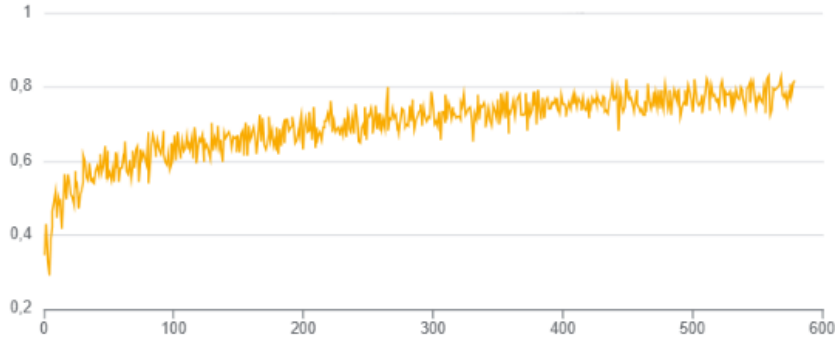
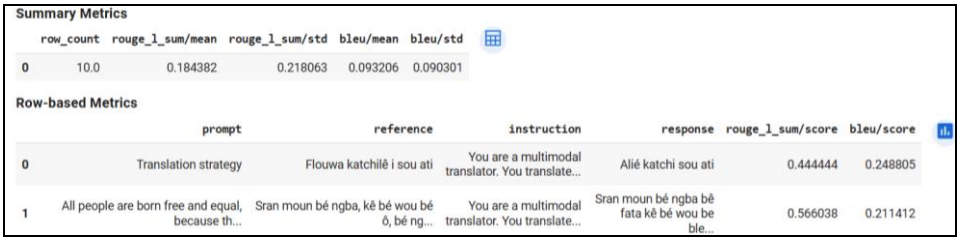


Figure 5. The small multimodal model's performance (accuracy)



Summary Metrics

	row_count	rouge_l_sum/mean	rouge_l_sum/std	bleu/mean	bleu/std
0	10.0	0.184382	0.218063	0.093206	0.090301

Row-based Metrics

	prompt	reference	instruction	response	rouge_l_sum/score	bleu/score
0	Translation strategy	Flouwa katchilê i sou ati	You are a multimodal translator. You translate...	Alié katchi sou ati	0.444444	0.248805
1	All people are born free and equal, because th...	Sran moun bé ngba, kê bé wou bé ô, bé ng...	You are a multimodal translator. You translate...	Sran moun bé ngba bê fata kê bé wou be ble...	0.566038	0.211412

Figure 6. An illustration of the evaluation of the small multimodal model

After these experiments, we were confident enough to go further in order to demonstrate that there exists a minimal number of non-text multimodal data (image, audio and video), a threshold that should not decrease the performance of the multilingual model while improving generalization during inference time, even when both (multilingual and multimodal) model performances are closely the same. So, based on the insights provided by our first experiments, we moved further with the multimodal dataset described above in table 5.

3.2.2. With our Multilingual Machine Translation baseline model

The learning curves presented below show that for the traditional Multilingual Machine Translation baseline model, the training scores gradually decrease from 8 towards 1, and the performance of the model itself (model accuracy) progressively increases from 0.1 to 0.7 and slightly above.

Figure 7. The Multilingual Machine Translation baseline model's training curves

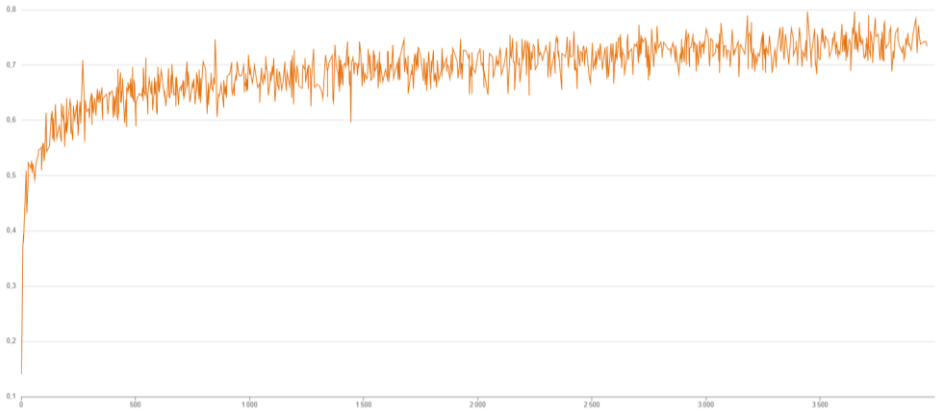


Figure 8. The Multilingual Machine Translation baseline model's performance (accuracy)

Summary Metrics

	row_count	rouge_l_sum/mean	rouge_l_sum/std	bleu/mean	bleu/std
0	200.0	0.349537	0.308683	0.172692	0.245144

Row-based Metrics

	prompt	reference	instruction	response	rouge_l_sum/score	bleu/score
0	Translation strategy	Flouwa katchilê i sou ati	You are a multimodal translator. You translate...	Liké katchilè i si'n nzounnzoun	0.363636	0.065673
1	All people are born free and equal, because th...	Sran moun bé ngba, kê bé wou bé ô, bé ng...	You are a multimodal translator. You translate...	Nvlé koun sou sran moun bé wou ngounmin li n...	0.312500	0.091272
2	No person may be tortured, or treated in a cru...	Bé kwla'a bé ye'a sran fi liké têtê moun,...	You are a multimodal translator. You translate...	Sran koun bé kwla-man bé-man-moun-ne nzoué ...	0.225806	0.066451

Figure 9. An illustration of the evaluation of the Multilingual Machine Translation model

3.2.3. With our Multimodal Machine Translation Baseline Model

Below, for the Multimodal MT baseline model, the training scores gradually decrease from 8 towards 1, and the performance of the model itself (model accuracy) progressively increases from 0.1 to 0.7 and slightly above too.

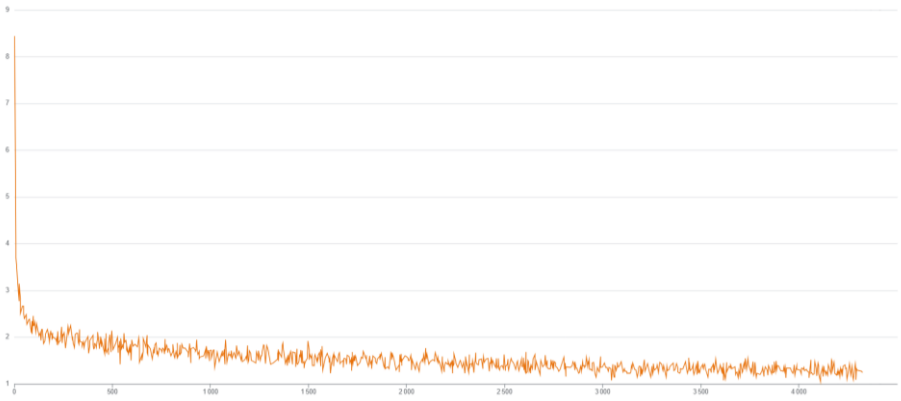


Figure 10. The Multimodal MT model's training curves

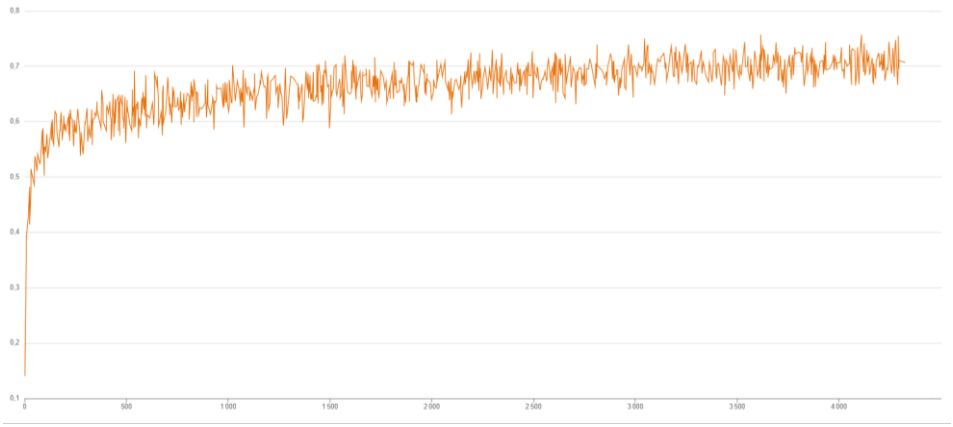
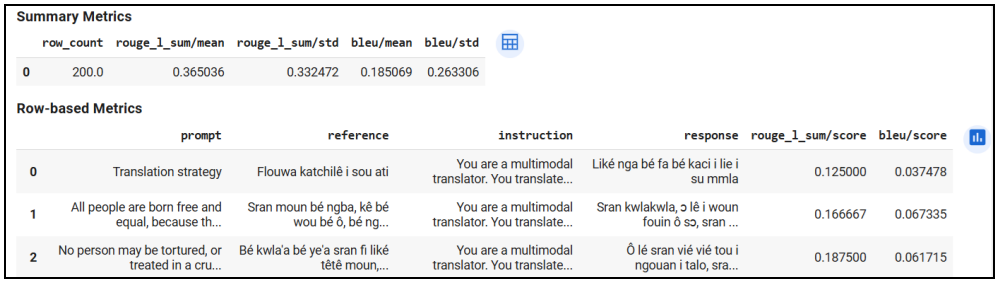


Figure 11. The Multimodal MT model's performance (accuracy)



Summary Metrics

	row_count	rouge_l_sum/mean	rouge_l_sum/std	bleu/mean	bleu/std
0	200.0	0.365036	0.332472	0.185069	0.263306

Row-based Metrics

	prompt	reference	instruction	response	rouge_l_sum/score	bleu/score
0	Translation strategy	Flouwa katchilê i sou ati	You are a multimodal translator. You translate...	Liké nga bé fa bé kaci i lie i su mmla	0.125000	0.037478
1	All people are born free and equal, because th...	Sran moun bé ngba, kê bé wou bé ô, bé ng...	You are a multimodal translator. You translate...	Sran kwlakwla, ɔ lê i woun fouin ô sɔ, sran ...	0.166667	0.067335
2	No person may be tortured, or treated in a cru...	Bé kwla'a bé ye'a sran fi liké têtê moun,...	You are a multimodal translator. You translate...	Ô lé sran vié vié tou i ngouan i talo, sra...	0.187500	0.061715

Figure 12. An illustration of the evaluation of the Multimodal Machine Translation model

We present in Table 6 below a summary of all our results (cf. the appendices for more details).

Table 6. Evaluation of baseline models on the same unseen text samples.

Models	**Multilingual Machine Translation baseline model**		**Multimodal Machine Translation baseline model**	
Metrics	**ROUGE**	**BLEU**	**ROUGE**	**BLEU**
10 samples of an unseen Evaluation set	**0,539522**	0,265949	0,50714	**0,394027**
200 samples of an unseen Evaluation set	0,349537	0,172692	**0,365036**	**0,185069**
AVG score	**0,4445295**	0,2193205	0,436088	**0,289548**

Table 7. Evaluation of the multimodal baseline model on multimodal (text and non-text) samples.

Models	**Multilingual Machine Translation baseline model**		**Multimodal Machine Translation baseline model**	
Metrics	**ROUGE**	**BLEU**	**ROUGE**	**BLEU**
190 samples of an unseen Evaluation set	N/A	N/A	**0,303208**	**0,168462**

3.3. Analysis

In this section, we provide an analysis of the results of our experiments on the multilingual and multimodal machine translation baseline models. Indeed, from the results provided above we have the following insights.

In all cases, the training curves and the model performance curves presented above (from Figure 2 to Figure 12) show that the baseline models related to our study effectively learn well, and are not enduring an overfitting nor an underfitting case. The Multilingual Machine Translation and the Multimodal Machine Translation baselines have seamlessly the same model performance (Figure 8 and 11, and Table 6), the multilingual one being only very slightly above the multimodal. However, we observe that the multimodal model produces better results. This is the reason why both have similar ROUGE values, while the multimodal model provided the highest BLEU scores in all evaluation schemes. As the number of unseen samples increases, the multimodal model shows superior performance for all provided metrics.

For the context of our study, as we explained in section 4.1 of [2], Low-Resource Languages are Languages with a digitally versatile speaker community, but very limited support in terms of language technology. This includes but is not limited to the lack of language data, and the lack of language technological support such as NLP tools for various data curation tasks like data annotation tools involving the intended language. In MT, the number of language pairs covered by low-resource languages is between 100k and 1 million. Therefore, the bci language under study is currently in a very low-resource context (< 100k pairs), compared to the high resource English language (>> 10 million, up to billions of language pairs). We also know that as mentioned in section 5.2 of [2], a baseline model that turns around the bottom of the Medium Quality would certainly work better on simpler text with about 5 tokens or words/subwords at a max. We provided in the appendices some clear illustrations and detailed information about related tests.

In conclusion, our results show that the multimodal Machine Translation (MT) model generalizes better, is effectively beneficial to the multilingual MT and even improves it because it also handles a data distribution that is not covered by the multilingual MT alone. Moreover, the multimodal MT model has the advantage of being able to also directly handle non-text modalities (image, audio, and video), as shown in Table 7. However, as the results don't clearly show the specific impact of each individual modality, further investigations and endeavors should be necessary to provide more insights and explanations about all these aspects.

4. LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

In our work, we created a Multimodal Dataset via a thorough Data Engineering strategy, and we provided a reliable evaluation on related baseline models, proving that the multimodal model consistently outperforms the performance of the traditional multilingual model as the unseen dataset increases. In this section, we show that despite the good job, some limitations still exist, and we will focus on Future Research Directions in order to overcome these challenges.

4.1. Limitations

We have provided a multimodal dataset and related models showing the superior performance of multimodal translation models over the traditional multilingual one. The results showed that from a threshold of only hundreds of samples, the multimodalities start by bringing improvement to the unimodal multilingual approach. However, the non-text modalities were just enough to detect and evaluate their impact on the trained model's performance during inference on unseen data. We don't know what modality impacted the most the multimodal model for yielding additional performance over the traditional multilingual model. This information will be beneficial for decision making about the Data Engineering strategy. For example, if the video modality has the highest impact, we could encourage people to produce short and medium-sized videos with labels

in order to support the multimodal model training. Finally, contributors will have the choice and adopt the easiest way for them to contribute.

4.2. Future Research Directions

We have created a multimodal model that showed superior performance over the traditional multilingual MT model on unseen texts. However, looking at the limitations mentioned above, we think that another thorough study will be necessary to increase the non-text modalities and individually evaluate the influence of each of them on the multimodal translation model.

Moreover, in the context of our Ph.D. program, we provided in a 2019 report first pictures of an Agent-based Multimodal Translation system for audio and video modalities (cf. [20], section IV.3- Research Objective, Fig. 4.1 Functional description of the AAVTS) for which the output keeps the original video unchanged but only the single elements including the speech and written texts are translated towards the output language. Further research should extend this approach to all available modalities in a multimodal translation context (text, image, audio, and video) and ensure that it works well. So that watching a video, the background images and texts in addition to the speeches are all automatically translated into the chosen output language, although the whole context of the video is maintained.

5. CONCLUSION

In this paper, based on our specific data engineering strategy, coding and relevant currently available multimodal LLMs, we designed a custom multimodal dataset for machine translation tasks and created a related Multimodal Machine Translation Model to evaluate it against the traditional multilingual approach. We started our experiments with the hypothesis that a multimodal Machine Translation (MT) model with four modalities (text, image, audio and video) built upon a multilingual Machine Translation baseline model covering only the text modality, should always be beneficial to the related multilingual MT (never decrease its performance) and even improve it. Our results on relevant unseen data samples confirm that, with only hundreds of additional non-text modalities, the multimodal Machine Translation (MT) model generalizes better, is effectively beneficial to the multilingual MT and even improves it.

We think that these findings are especially a hope for low resource languages for which the combination of modalities is a great solution to overcome the struggle of data scarcity in the field of Machine Translation, given that certain modalities are easier to provide by contributors to the scientific community for NLP tasks.

In our work that presents an approach to handle language translations implying four (04) modalities, we found valuable insights for researchers and practitioners interested in the field of Multimodal Machine Translation, that is still at its early stage. However, the specific impact of each non-text modality remains an aspect to be fully covered. Indeed, our study didn't explicitly show the impact of each modality, separately taken, on the overall results provided by the multimodal MT model. Further research works should focus on showing which modality among the image, audio and video modalities provides the best additional performance over the traditional multilingual model. Moreover, we should extend our work to a multimodal translation context with text, image, audio, and video inputs, where all these input elements are automatically and contextually translated towards the output language, keeping the original input context unchanged but only translated. Finally, let's note that, regardless of the specific case study that supported our work (English and Baoulé), the proposed methods are adaptable and reproducible to all other languages worldwide.

ACKNOWLEDGEMENT

We would like to thank Professor William Martin our Supervisor, Professor ABDO Miled Abou Jaoude and the BIU staff for the reviews of the reports of our Ph.D. program, and for the access to a partial scholarship leading to this highest level of specialization in Artificial Intelligence. Invaluable things we learned there are the supports for writing this paper.

We are also thankful to the Ivoirian Civil Service, as we got access to this Ph.D. program in 2018 as a Computer Scientist Civilian (Civil Servant or Official) from Côte d'Ivoire. Local initiatives encourage commitment to training and to the field of research for civilians.

DECLARATIONS

Funding	No funding was received to assist with the preparation of this manuscript.
Conflicts of interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	The French Latin Script Converter model is available for scientific experimentations (cf. APPENDICES).
Code availability	Yes, for the converter only (cf. APPENDICES).
Authors Contributions	One author article.

REFERENCES

[1] Russell, Stuart J. & Norvig, Peter (2018) *Artificial Intelligence: A Modern Approach*, Pearson India Education Services Pvt. Ltd., India, Third Edition, twelfth Impression 2018.

[2] Konan Jean-Claude, Kouassi, (2023) "Understanding the Worldwide Paths Towards the Creation of True Intelligence for Machines", *International Journal of Computer Science & Information Technology (IJCSIT)*, Vol. 15, No. 1, pp. 43-68, Academy and Industry Research Collaboration Center (AIRCC), doi:10.5121/ijcsit.2023.15104.

[3] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, Sukrut Oak. "*The AI Index 2025 Annual Report*," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2025.

[4] Parthasarathy, Venkatesh et al., (2024) "The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities (Version 1.0)", *arXiv:2408.13296v2*.

[5] Hatami, Ali et al., (2024) "*English-to-Low-Resource Translation: A Multimodal Approach for Hindi, Malayalam, Bengali, and Hausa*." Conference on Machine Translation.

[6] Hatami, Ali et al., (2024) "Enhancing Translation Quality by Leveraging Semantic Diversity in Multimodal Machine Translation", *In Proceedings of the 16th Conference of the Association for Machine Translation in the Americas*, Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers.

[7] Jeong, Cheonsu, (2025) "Beyond Text: Implementing Multimodal Large Language Model-Powered Multi-Agent Systems Using a No-Code Platform", *arXiv:2501.00750v2*.

[8] Omniscien Technologies. (2025, January 19). AI and Language Processing Predictions for 2025 [Video]. YouTube. https://youtu.be/tDa2jH3lboY.

[9] Choi, Jeongsoo et al., (2024) "AV2AV: Direct Audio-Visual Speech to Audio-Visual Speech Translation with Unified Audio-Visual Speech Representation", *arXiv:2312.02512v2*.

[10] Dougban Monsia. (2022, June 23) Baule Speech Dataset. 1.0, Zenodo, doi:10.5281/zenodo.6705861.

[11] RTI Officiel. (2019). *Les Trésors Du Monde : Apprenons nos langues "La langue Agba (Baoulé de Dimbokro)* [Video]. YouTube. https://www.youtube.com/watch?v=DWAtiTCCH60.
[12] RTI Officiel. (2021, March 4). *Les Trésors du monde | Apprenons nos langues* [Video]. YouTube. https://www.youtube.com/watch?v=iYdMacPy1qg.
[13] Djeh Pyco. (2021, March 4). *Salutations chez le peuple Baoulé* [Video]. YouTube. https://www.youtube.com/watch?v=hJWThBDXX0U.
[14] BAOULE FACILE. (2017, Dec 15). *BAOULE FACILE APPRENDRE A SALUER EN BAOULE* [Video]. YouTube. https://www.youtube.com/watch?v=j-XNIOq0CN0.
[15] BAOULE FACILE. (2023, Apr 17). *COMMENT DEMANDER LES NOUVELLES EN BAOULÉ FACILE* [Video]. YouTube. https://www.youtube.com/watch?v=8m1mdy9H8o0.
[16] BAOULE FACILE. (2023, Apr 20). *FAIRE LA PAIX EN BAOULE FACILE* [Video]. YouTube. https://www.youtube.com/watch?v=Aq8SgGp9lQc.
[17] BAOULE FACILE. (2024, Nov 18). *Parlons baoulé: Sapka, dan, sran dan, adinanwlè* [Video]. YouTube. https://www.youtube.com/watch?v=Q0_lbyGFFfE.
[18] BAOULE FACILE. (2024, Jun 2). *LES DIFFÉRENTES PARTIE AU NIVEAU DE LA TÊTE EN BAOULE FACILE* [Video]. YouTube. https://www.youtube.com/watch?v=xYuBYtzAgbM.
[19] BAOULE FACILE. (2024, Nov 21). *Expression simple en baoulé: niman* [Video]. YouTube. https://www.youtube.com/watch?v=cskYh0S_qgs.
[20] Kouassi Konan Jean-Claude, Ph.D. Student in Artificial Intelligence at Bircham International University (BIU) – Madrid since 2018. (2019, April 1). *Case Study of the 3rd Book Report, 30 pp*., section IV.3- Research Objective, Fig. 4.1 Functional description of the AAVTS [Figure], p.29.

APPENDICES

1. French Latin Script Converter Model and Bci-Related Data Creation

French Latin Script Converter model on Kaggle: https://www.kaggle.com/code/kjeanclaude/inference-french-latin-script-converter-gpu.
BCI Wikimedia Incubator: https://incubator.wikimedia.org/wiki/Wp/bci.
BCI Translatewiki: https://translatewiki.net/, the bci language is mentioned here as wawle (Thank you very much to Amir Aharoni for the translation rights.)
Open-ended Data Crowdsourcing website: https://pairing.excellence-integration.org.
Github: https://github.com/Kjeanclaude/mtBCI-1.0-Corpus.

2. Additional Information about the Models' Evaluation

Table 8. Top 3 prediction samples from the multilingual and multimodal baseline models, on 2 longer texts (above 5 tokens or words/subwords) and 14 simpler texts (about 5 tokens or words/subwords at a max), taken from section 5.3 of [2], our previous paper.

N°	Unseen prompts in English or Baoulé	Best predictions from the **Multilingual baseline model** (Top 3)	Best predictions from the **Multimodal baseline model** (Top 3)	Expected results from a human translator
1	My name is Jean-Claude, and you, what is your name?	Be flê mi Jean-Claude, amoun bé flê amoun sê nan?	Be flê mi Jean-Claude, amoun liè o?	Bé flê mi Jean-Claude, yê ô li, bé flê wô sê ?
2	Bé flê mi Jean-Claude, yê ô li, bé flê wô sê ?	My name is John-Claude, and what's your name ?	My name is Jean-Claude, and what is your name?	My name is Jean-Claude, and you, what is your name?
3	Bé flê mi Jean-Claude.	My name is Jean-Claude.	My name is Jean-Claude.	My name is Jean-Claude.

N°	Unseen prompts in English or Baoulé	Best predictions from the **Multilingual baseline model** (Top 3)	Best predictions from the **Multimodal baseline model** (Top 3)	Expected results from a human translator
4	N tran Côte d'Ivoire lô.	I live in Côte d'Ivoire.	I live in Côte d'Ivoire.	I live in Côte d'Ivoire.
5	Aublé ti awoundjôê i nzôliê.	Olive branch is a symbol of peace.	White represents peace.	The dove is a symbol of peace.
6	A fin ni ?	Where do you come from?	Where do you come from?	Where do you come from?
7	Man mi kpaoun.	Give me some bread.	Give me some bread.	Give me some bread.
8	A wla tralê klanman kpa.	You are perfectly dressed.	You are wearing a beautiful dress.	You wear beautiful clothes. ("Tu portes de jolies habits")
9	N sou di aliê wié.	I'm eating.	I am eating some food.	I am eating some food.
10	My name is Jean-Claude.	Be flê mi Jean-Claude.	Be flè mi Jean-Claude.	Bé flê mi Jean-Claude.
11	I live in Côte d'Ivoire.	N o Côte d'Ivoire.	N tran Côte d'Ivoire.	N tran Côte d'Ivoire lô.
12	The dove is a symbol of peace.	Afian sissalê i nzôliê ko kpakpa sou .	Atin mlanlê yɛ i nzôliê ni.	Aublé ti awoundjôê i nzôliê.
13	Where do you come from?	A fi nin ?	Amoun fi nin?	A fin ni ? / Amoun fin nin?
14	Give me some bread.	Man mi kpa nnian.	Man mi kpa blé yé.	Man mi kpaoun.
15	You wear beautiful clothes.	A fataa tralê kpakpa nan blêbla blêblê.	A wounnfouê sié a tralê kpa.	A wla tralê klanman kpa (moun).
16	I am eating some food.	N di liké wié.	N di aliê wié.	N sou di aliê wié.

Our results on Table 8 samples show and confirm that the multimodal Machine Translation (MT) model generalizes better, is effectively beneficial to the multilingual MT (never decrease its performance) and even improves it because it also handles a data distribution that is not covered by the multilingual MT only. On simpler texts (on which they can output best human-like predictions looking at the model performance that turns around 70% of accuracy and also at ROUGE scores) both models seem to provide similar translations. But the multimodal model presents a better syntactic and semantic understanding, hence its overall highest BLEU scores (cf. Table 6).

3. Artefacts

3.1. Multilingual Machine Translation Interface

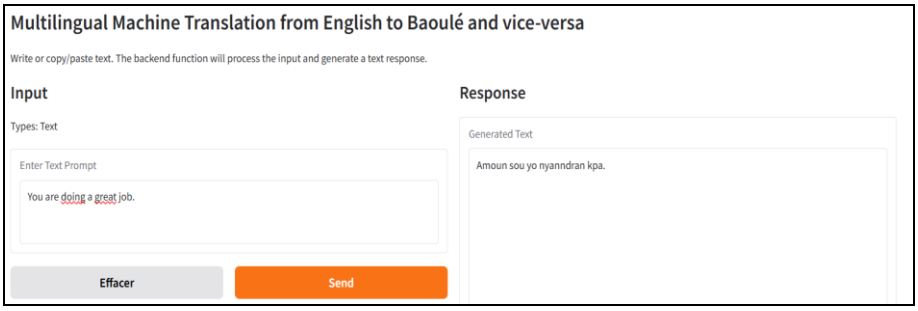


Figure 13. An illustration of the Multilingual Machine Translation Interface for a translation from English to Baoulé

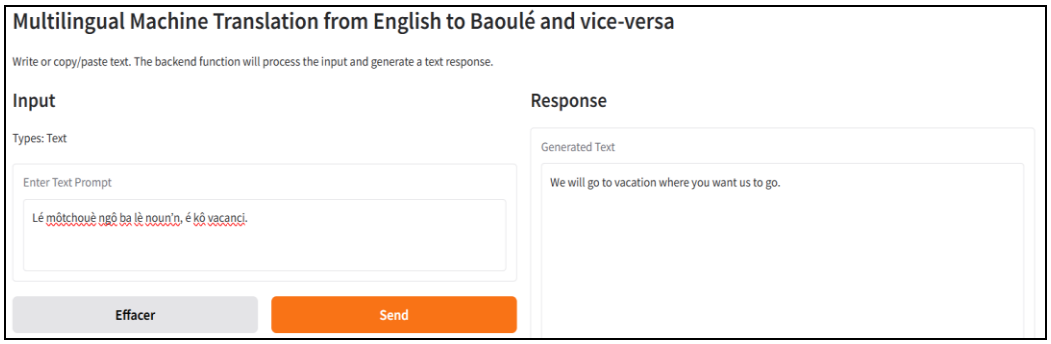


Figure 14. An illustration of the Multilingual Machine Translation Interface for a translation from Baoulé to English

The Multilingual Machine Translation model behind this Interface was trained only on the text modality of our multimodal dataset (cf. Table 2). This English/Baoulé Multimodal Dataset has been thoroughly curated from the mtBCI-1.0-Corpus provided in our previous paper [2]. *The language pairs have also been somehow reviewed for reaching high quality translations*. This text modality has been combined with non-text modalities for training the multimodal model.

3.2. Multimodal Machine Translation Interface

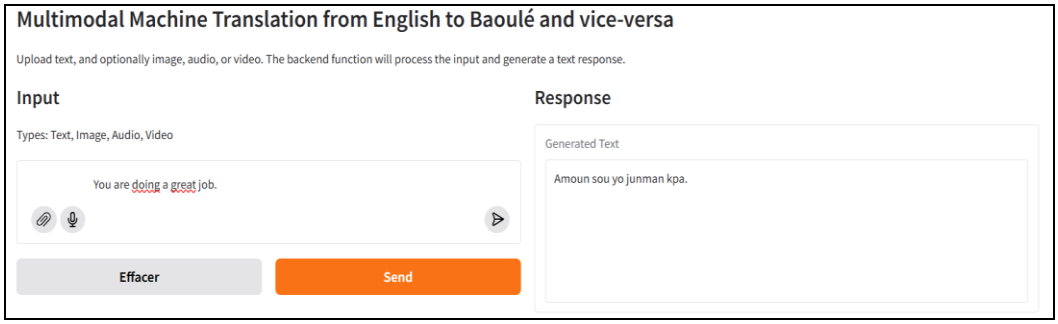


Figure 15. An illustration of the Multimodal Machine Translation Interface for a translation from English to Baoulé (better compared to the translation of Figure 13)

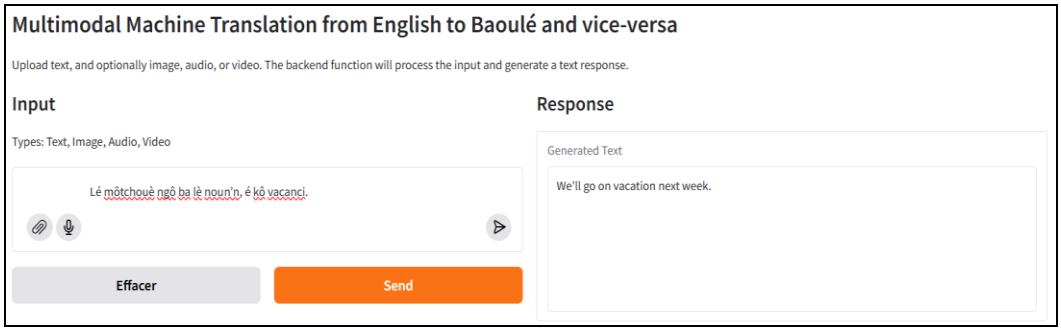


Figure 16. An illustration of the Multimodal Machine Translation Interface for a translation from Baoulé to English (better compared to the translation of Figure 14)

3.3. An illustration of non-text Machine Translation cases

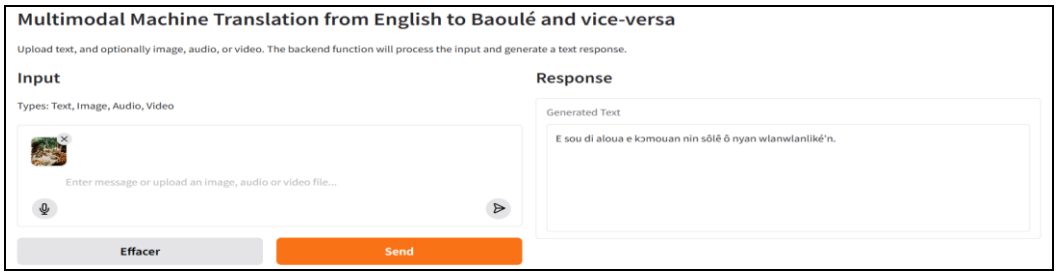


Figure 17. An illustration of the Multimodal Machine Translation with images

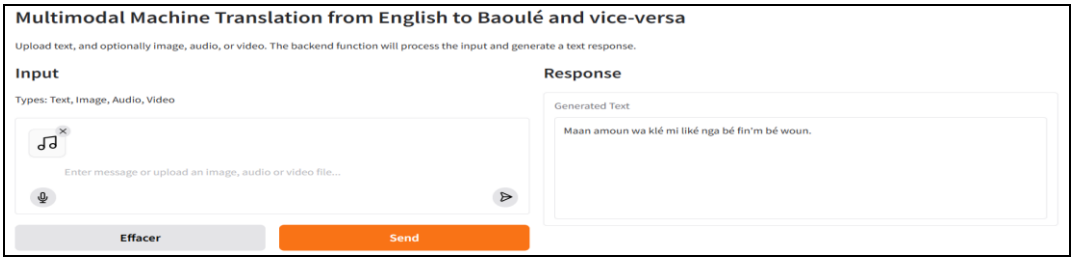


Figure 18. An illustration of the Multimodal Machine Translation with audio data

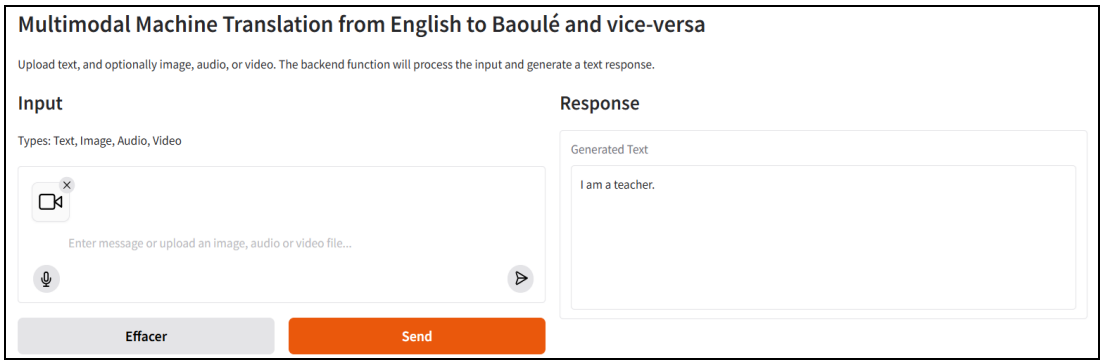


Figure 19. An illustration of the Multimodal Machine Translation with video data

For the translation with non-text modalities, although we evaluated their global impact on the multilingual model leading to better generalization, we didn't study the specific impact of each non-text modality (image or audio or video) on the multimodal model performance. Moreover, despite the multimodal model handles non-text modalities (cf. Figure 6 and Table 7), **we still need to know the threshold from which each modality starts by providing at least medium quality translations on its own channel** (we only know the threshold for globally and positively influencing the multilingual model; that is of hundreds of non-text samples per modality, with respect to the ratio).

AUTHOR

Kouassi Konan Jean-Claude

- Senior Computer Scientist, Civilian (Public Sector) in Côte d'Ivoire since February 2013.
- Currently (May 2025) in service at the Ministry of Environment, Sustainable Development and Ecological Transition. Head of the Study, Development and Environmental Information System Service (inside the IT Department).
- Totalizing 12+ years of cumulated experiences in the field of Computer Science, in private enterprises as well as in the Ivorian Public Sector as a Civilian (Civil Servant).
- From Junior AI Expert (0-2 years of experience) to Middle-Level AI Expert (2-5 years of experience), I am now a Confirmed or Advanced AI Expert (5-10 years of experience) with Deep expertise in specific AI domains, strong research background, and ability to innovate new algorithms. But I am not yet fully what I am called to be; a Senior or Very Advanced AI Expert (10+ years of experience in AI) with Extensive knowledge across multiple AI domains, strategic thinking, and leadership abilities.

.