

Is Multimodality still Required for Multimodal Machine Translation? A case study on English and Italian

Elio Musacchio^{1,2,*}, Lucia Siciliani¹, Pierpaolo Basile¹ and Giovanni Semeraro¹

¹Department of Computer Science, University of Bari Aldo Moro, Italy

²National PhD in Artificial Intelligence, University of Pisa, Italy

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in machine translation. A related task is multimodal machine translation, where text is paired with an image. While intuition suggests that models supporting multimodal inputs (e.g. Large Vision-Language Models or LVLMs) are essential for this task due to their image understanding, we hypothesize that, in general, text contains several clues that might be enough for effective translation. In this work, we rigorously test both LLMs and LVLMs on the multimodal machine translation task for the English and Italian languages, thoroughly analyzing the impact of text and images on translation quality.

Keywords

Large Language Models, Large Vision-Language Models, Machine Translation, Multimodal Machine Translation

1. Introduction

Large Language Models (LLMs) are being increasingly leveraged in several Natural Language Processing (NLP) tasks due to their impressive generalization capabilities [1, 2]. Several studies have demonstrated that these models, trained on massive text corpora, can perform multiple tasks without requiring further training. Among the various NLP challenges, Machine translation has always stood as a fundamental benchmark, also for its practical implications. In machine translation, given an input text, the objective is to produce an equivalent output in another target language. The translation must not only be grammatically correct but also faithful in preserving the semantics and the stylistic nuances of the original text. Numerous studies have already evaluated the ability and proficiency of LLMs in this task [3, 4].

However, despite the relevance of text-only translation, the related task of multimodal machine translation (MMT) has attracted less attention. Its formulation is similar to traditional machine translation, but the input additionally includes an image associated with the text (e.g. an image and its caption). It is thus straightforward to understand why advances in this task have proceeded more slowly: sufficiently large and high-quality image-text corpora are notoriously more scarce than their text counterparts. Another reason is that this task is more

challenging because the image often contains crucial clues and information necessary for understanding the input text and its semantics, therefore the model or algorithm must be capable of processing the additional visual input to perform the translation task. Historically, early research in MMT has typically relied on small, specialized multimodal models [5, 6].

Although traditional Large Language Models (LLMs) are limited to processing text and cannot process visual inputs, making them seemingly unsuitable for MMT, a new class of architectures known as Large Vision-Language Models (LVLMs) has emerged to bridge the gap the two modalities, extending LLMs to support both textual and visual inputs. Despite the existence of LVLMs, the rapid advancements in text-only LLMs have led to wonder whether the additional visual input is essential for effective multimodal machine translation. Intuitively, if the source text is sufficiently descriptive, a powerful LLM could already possess enough world knowledge and language understanding capabilities to generate an accurate translation without the visual input. For instance, a well-crafted image caption can often be sufficient to describe the most relevant aspects of the associated scene, making it concise and meaningful. Indeed Futeral et al. [6] leveraged a MMT model to resolve ambiguities within the input text. However, in many cases, this approach succeeds when the ambiguity is due to the low descriptiveness of the input text. For example, in the sentence "That's lots of bucks!" without further qualifiers, it is impossible to properly disambiguate the word "bucks", which could refer to deer, dollars, or be a colloquial exclamation. This highlights that, most of the time, the main challenge is represented by the vagueness or brevity of the context provided by the text, rather than the limitations of the model. Furthermore, one of the most promi-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 – 26, 2025, Cagliari, Italy

*Corresponding author.

✉ elio.musacchio@uniba.it (E. Musacchio); lucia.siciliani@uniba.it (L. Siciliani); pierpaolo.basile@uniba.it (P. Basile); giovanni.semeraro@uniba.it (G. Semeraro)

ORCID 0009-0006-9670-9998 (E. Musacchio); 0000-0002-1438-280X (L. Siciliani); 0000-0002-0545-1105 (P. Basile); 0000-0001-6883-1853 (G. Semeraro)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

nent datasets for MMT, that is Multi30k [7], only provides captions. We believe that this dataset is not enough to evaluate the capabilities of models in the MMT task.

In light of this, we aim to investigate the impact of both the additional visual input and the descriptiveness of the textual input for multimodal machine translation in LLMs and LVLMs. We conducted this study in both English and Italian, using our knowledge of these languages to carry out the study carefully. Hence, the contributions of this work are the following:

- We extend an existing multimodal machine translation dataset to include the Italian language;
- We create a new multimodal machine translation dataset for English and Italian, with a focus on short texts consisting of only a few words;
- We benchmark several LLMs and LVLMs on both datasets for this task, analyzing and studying the impact of the input modalities on the output.

Furthermore, we release code and resources related to this study¹.

2. Related Works

2.1. Large Vision-Language Models

Early releases in open LLMs mainly focused on textual processing and were tailored to the English language. For example, the LLaMA 2 models [8], for which the language distribution of the train set has been officially reported, were extensively trained and tested on English text data without any mechanism to support other modalities. In light of this, several works started proposing solutions to bridge this gap. The main idea was to leverage a pre-trained LLM and extend it to an LVLM, therefore avoiding the costly procedure of multimodal pre-training from scratch. A well-known example is LLaVA [9], where visual embeddings are extracted from a pre-trained vision encoder and projected into the latent space of the LLM. This strategy has been widely adopted, and many modern LVLMs are based on this paradigm. Among these, LVLMs supporting multiple languages include: Qwen 2.5 VL [10], Gemma 3 [11] and LLaMA 4 [12]. All of them are LVLMs supporting modern strategies, for example, Qwen 2.5 VL employs dynamic resolution to decrease the number of visual tokens w.r.t. resolution of the input image, while LLaMA Scout is based on a mixture-of-experts architecture (i.e. tokens are handled by different layers according to a routing function). Finally, all of these models have been extensively trained on a multimodal and multilingual data mixture.

2.2. Multimodal Machine Translation

The most used resource for MMT is MULTI30K [7], a dataset consisting of parallel image descriptions. The dataset has been created starting from the FLICKR30K [13] dataset, which contains 31,014 images sourced from Flickr and a large number of image captions obtained through Amazon Turk. MULTI30K extended the dataset with professional manual translations from English to German. It was then further extended to French by Elliott et al. [14] and Czech by Barrault et al. [15]. The dataset has become a reliable benchmark for MMT and has been used in numerous works as their main dataset for experimentation. Researchers have proposed several solutions to tackle the challenges of the MMT task. Specifically, Yao and Wan [5] developed a multimodal transformer model, which employs a multimodal self-attention mechanism to adjust the attention score of each word w.r.t. the contents of the image. VGAMT [6] adapts a text-only encoder-decoder machine translation model to multimodality by incorporating the features of the image in the encoder-side of the model and employing guided self-attention to obtain better alignment between text and images. SOUL-Mix [16] leverages a manifold mixup method to mix the predicted translation of several text-image pairs, where the image is kept as is while the text is processed through degradation schemes. To the best of our knowledge, there are no works studying the effect of the granularity of text in MMT using modern LVLMs supporting multilingual inputs.

3. Problem Formulation

In MMT, the model is given an input comprising a text in a specified source language t_{lang_src} and an image i , semantically related to the given text. The desired output is a translated text in a target language t_{lang_tgt} . The objective is for t_{lang_tgt} to be not only syntactically correct, that is it has no grammatical errors in the target language, but also accurately aligned with t_{lang_src} both syntactically (ensuring all relevant words from the input text are present in the output) and semantically (preserving the original meaning of the input text).

As previously mentioned, research in multimodal machine translation has often focused on image captioning datasets. A caption is a short description of the image that meaningfully describes the most relevant aspects of the image. However, we argue that, despite the caption being a short text, the image does not provide additional context w.r.t. text. This is because: 1) a good caption already contains extensive information about the image; 2) the caption often contains enough words to allow for proper translation without additional context. However, if the text consists of only a few words, the task becomes much more challenging. This is because, to perform an

¹<https://github.com/swapUniba/MM-MT-ITA>

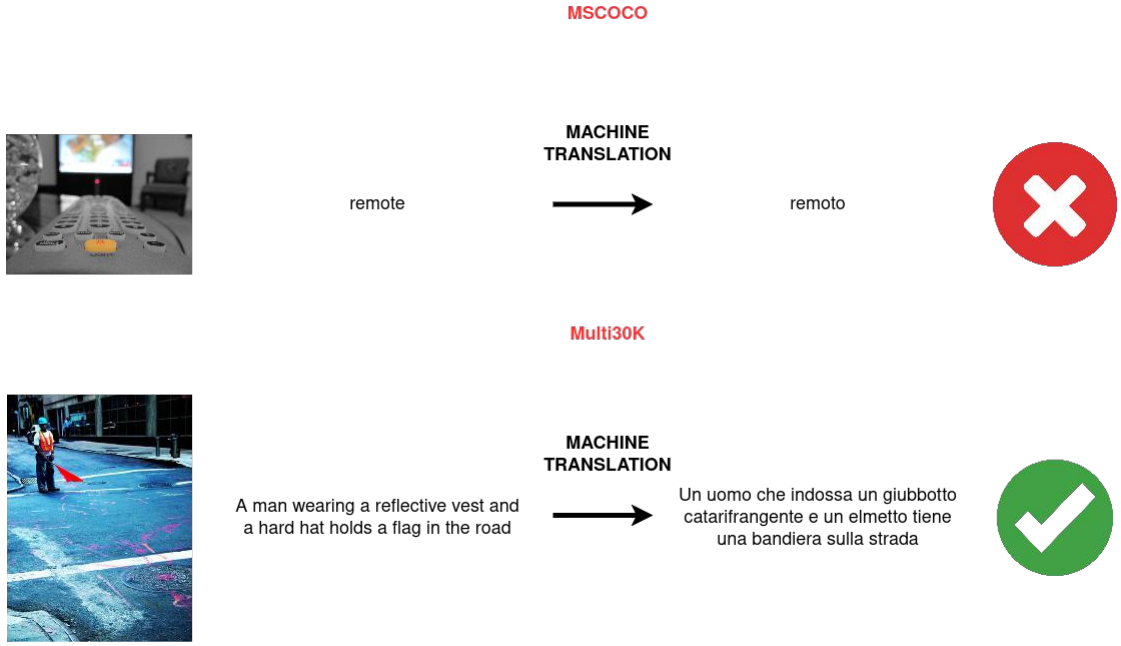


Figure 1: Example of text-only machine translation from the MSCOCO and the Multi30K datasets. We perform text-only machine translation using DeepL. In the first case, the limited textual content makes the text-only machine translation model unable to provide an optimal translation. In the second case, the additional textual content enables the model to provide an optimal translation even without providing the image as input.

optimal translation, the model is also required to understand the meaning of each word in the input sentence. Specifically, translating polysemous words requires additional context, either from the textual or visual modality. We showcase this in Figure 1, where we present an example of machine translation of two image-text pairs. In the instance from the MSCOCO [17] dataset, the word "remote" is translated as "remoto" (i.e. something that is far away) rather than its proper translation, that is "telecomando". Due to the absence of substantial textual clues, the model provides a translation that is not aligned w.r.t. the contents of the image. In the second instance from the Multi30K dataset, however, the caption is correctly translated and aligns well with the image's contents. In this case, the word "vest" is correctly translated to "giubbotto" (i.e. a jacket), thanks to the additional words present in the text. In light of this, we aim to understand the relationship between the granularity of the input text and the associated image in multimodal machine translation. To do so, we need to collect two different datasets, one made of very short texts consisting of only a few words and one made of image captions.

4. Dataset

In this section, we describe the datasets that will be used for the experimentation. Specifically, we aim to test the ability of LVMs in MMT for two different types of instances: 1) text containing a rich description of the image; 2) text containing only a few words. Going forward, we will reference the former as "long" dataset and the latter as "short" dataset.

4.1. Dataset Collection

For the "long" dataset, we collect the English 2016 Flickr test set from the Multi30K dataset. Specifically, we leverage a version uploaded on HuggingFace. For the "short" dataset, we collect lemmas from BabelNet [18]. BabelNet is a semantic network organized according to a synset hierarchy. A synset is a synonym set, containing all possible words that can be associated with that concept. Additionally, in BabelNet, each synset is linked with one or more images, providing useful resources for multimodality. It also provides lemmas in multiple languages, allowing access to the lemmas for all required languages. In our case, we collect both the first lemma in English and Italian, as well as the best image for each synset. However, these datasets cannot be used directly after

collecting them as they are. In fact, Multi30K does not provide labels in Italian, and BabelNet lemmas are not precise translations from English to Italian and vice versa. For example, the English lemma "economy of resources" is paired with the Italian lemma "efficienza", which is not a literary translation of the original text. In light of this, we perform manual annotation for "long" dataset and manual verification for the "short" dataset.

4.2. Dataset Annotation

For the "long" dataset, we begin by performing a preliminary Italian translation of the data with LLaMA 3.3 70B Instruct, which helps reduce the editing overload. After that, we manually check each translated instance and correct any machine translation errors that are present in the dataset. Specifically, we follow these guidelines when correcting the translated text: 1) we use Italian figures of speech whenever possible (e.g. we translate "shirtless man" as "uomo a torso nudo" instead of "uomo senza maglietta"); 2) we only keep English words when they represent commonly used terms across languages (e.g. we keep the word "cowboy" as is). For the "short" dataset, we manually filter each pair of lemmas in Italian and English to include only those that are proper translations of one another. After performing the previously described steps, we obtain the final versions of the "long" and "short" datasets. The "long" dataset consists of 1,000 instances, the same cardinality as the original Multi30k dataset, while the "short" dataset consists of 400 instances.

5. Evaluation

In this section, we describe the evaluation setting that has been considered for all models (e.g. generation strategy), we discuss the obtained results and present some interesting additional experiments.

Additionally, we aim to answer the following research questions: 1) Are LVLMs capable of performing MMT for both the "short" and "long" dataset? 2) Is performance affected by the presence of the image in the input? 3) Are LLMs as capable as LVLMs in MMT? 4) Does the generation strategy impact the quality of MMT?

5.1. Evaluation Setting

We use the same metrics as the original Multi30K dataset for the "long" dataset, namely BLEU and METEOR. Additionally, we also include COMET, since it has been widely used in machine translation. For our short dataset, since it consists of only a few words, we perform an exact match, that is, we verify that the generated output is identical to the ground truth label. However, to have a

more precise evaluation, we perform an exact match for each possible lemma associated with the synset of the instance. If at least one of the labels exactly matches the generated output, the translation is considered correct. For example, for the synset "bn:00109359a" with English lemma "quiet" and Italian lemmas "tranquillo", "calmo", "silenzioso", "quieto", the translation from English to Italian is correct as long as the generated output is one of the Italian lemmas of the synset (and viceversa for translation from Italian to English). Thanks to the multiple labels, we cover cases where the model may translate the input lemma with a word that has the same meaning. All models are evaluated using greedy decoding, which makes the inference process reproducible and removes any randomness from the outputs. In all cases, the chat template associated with each model is used during inference. We consider the following models for evaluation: Qwen 2.5 VL and LLaMA Scout. Both models support multimodal and multilingual inputs. For Qwen 2.5 VL, we consider the 3B, 7B and 72B checkpoints, while for LLaMA Scout, we consider the only available checkpoint (17B with 16 experts). Inference is performed locally for Qwen 2.5 VL 3B and 7B, while we rely on a cloud service² for Qwen 72B and LLaMA Scout. All models are prompted using the following input strings if the image associated to the text is provided: "Translate the following text from [*src_lang*] to [*tgt_lang*]: "[TEXT]". Use the image as additional context for the translation. Provide only the translated text.", otherwise the input string is "Translate the following text from *src_lang* to [*tgt_lang*]: "[TEXT]". Provide only the translated text.". [*src_lang*] and [*tgt_lang*] are placeholders for representative strings of the source and target languages; in this case, they are either "English" or "Italian", while [TEXT] is a placeholder for the text of the instance.

5.2. Results

We report results on the Multi30k test set in Table 1 while results for the BabelNet test set can be found in Table 2. Overall, both the "long" and "short" datasets are sensitive to the scale of the model, with larger models achieving better results on every metric. Furthermore, the translation from English to Italian makes the task more challenging for smaller models. As a matter of fact, Qwen 2.5 VL 7B Instruct achieves a score of .4800 in BLEU for the "long" dataset in translation from English to Italian, while it achieves a score of .5839 in translation from Italian to English. The same pattern is also present for the "short" dataset, where the model achieves a score of .4700 in exact match in translation from English to Italian, while it achieves a score of .5900 in translation from Italian to English. This pattern is less prevalent for

²<https://api.together.ai/>

Model	With Image	EN → IT			IT → EN		
		BLEU	METEOR	COMET	BLEU	METEOR	COMET
Qwen2.5-VL-3B-Instruct	X	<u>.4332</u>	<u>.6871</u>	.8627	.5551	.8233	.8987
	✓	.4316	.6813	<u>.8637</u>	<u>.5644</u>	<u>.8266</u>	<u>.9026</u>
Qwen2.5-VL-7B-Instruct	X	.4793	<u>.7213</u>	<u>.8751</u>	.5680	.8308	.9037
	✓	<u>.4800</u>	.7211	.8745	<u>.5839</u>	<u>.8398</u>	<u>.9069</u>
Qwen2.5-VL-72B-Instruct	X	.6064	.8021	.8994	.5803	.8473	.9048
	✓	<u>.6186</u>	<u>.8130</u>	<u>.9056</u>	<u>.6027</u>	<u>.8589</u>	<u>.9103</u>
Llama-4-Scout-17B-16E-Instruct	X	<u>.5441</u>	<u>.8084</u>	<u>.8815</u>	<u>.5346</u>	.8364	.8895
	✓	.5413	.8043	.8797	.5311	<u>.8396</u>	.8839

Table 1

Results for the BLEU and METEOR metrics on the "long" dataset for translation from English to Italian and viceversa. The "With Image" column indicates whether the input text is provided to the model along with the associated image for each instance. For each model, the best score for each metric is underlined. The best result for each metric across all models is in bold.

Model	With Image	EN → IT	IT → EN
		EM	EM
Qwen2.5-VL-3B-Instruct	X	<u>.3825</u>	.4550
	✓	.3625	<u>.4800</u>
Qwen2.5-VL-7B-Instruct	X	<u>.4700</u>	.5150
	✓	<u>.4700</u>	<u>.5900</u>
Qwen2.5-VL-72B-Instruct	X	.6150	.6175
	✓	<u>.6750</u>	<u>.6775</u>
Llama-4-Scout-17B-16E-Instruct	X	<u>.5950</u>	<u>.5275</u>
	✓	.5375	.3675

Table 2

Results for the exact match metric on the "short" dataset for translation from English to Italian and viceversa. The "With Image" column indicates whether the input text is provided to the model along with the associated image for each instance. For each model, the best score for each metric is underlined. The best result for each metric across all models is in bold.

bigger models, for example, Qwen 2.5 VL 72B Instruct achieves a score of .6186 in BLEU for the "long" dataset in translation from English to Italian and a score of .6027 in translation from Italian to English. This showcases that natural language generation capabilities of smaller models are limited in a multilingual use case w.r.t. bigger models, since they achieve better performance when generating English text. Finally, results also showcase that, in general, the presence of the image in the input is better for translation. For example, Qwen 2.5 VL 7B Instruct achieves an exact match score of .5900 on the "short" dataset for translation from Italian to English when the image is provided in the input, while it achieves a score of .5150 when it is not provided. However, there are some exceptions, for example, LLaMA Scout performs better when the image is not provided as part of the input, which highlights the importance of testing the behaviour

of different models for this task.

5.3. Evaluation of LLMs against LVLMs

All models considered so far are LVLMs, that is, they have been extensively trained on a multimodal data mixture. However, since we have also studied these models for MMT without providing the input image, the underlying vision encoder used by LVLMs becomes useless, as no visual input is provided. In light of this, we compare the performance of two models of the same size and architecture, where one is an LLM and the other is an LVLM. This allows us to determine whether multimodal training can still be beneficial for MMT even when an image is not provided as additional input. To perform this experiment, we rely on Qwen 2.5 VL 7B and Qwen 2.5 7B, which guarantees fairness of the experiment between the two

Model	Multi30K						BabelNet	
	EN $\rightarrow$ IT			IT $\rightarrow$ EN			EN $\rightarrow$ IT	IT $\rightarrow$ EN
	BLEU	METEOR	COMET	BLEU	METEOR	COMET	EM	EM
Qwen2.5-7B-Instruct	.4132	.6887	.8867	.5153	.8211	.8530	.3875	.4925
Qwen2.5-VL-7B-Instruct (no image)	.4793	.7213	.8751	.5680	.8308	.9037	.4700	.5150

Table 3

Results for the Qwen 2.5 models (with and without multimodal input support) for the "long" and "short" dataset using their related metrics. Best result between the two models for each metric is in bold.

models (since they share the same number of parameters and underlying architecture). Results are reported in Table 3. Interestingly, the LVLM performs better than the LLM on both the "short" and "long" datasets. This highlights that multimodal training still helps in MMT when the image input is not provided. This is probably due to the style of the text that LVLMs are trained on. For example, LVLM training includes data containing image captions, which still affects the model even when no image is provided in the input during inference.

5.4. Evaluation of generation strategy

All results considered so far used greedy decoding as the generation strategy. In greedy decoding, each new token that is generated is selected according to the highest probability out of all the ones available in the model’s vocabulary. However, beam search has been widely considered as the standard generation strategy for the machine translation task [19]. In beam search, the model considers the n possible paths with the highest probability at each generation step, instead of only considering the path of the highest probability token for each generation step. This strategy enables the model to avoid greedy predictions, where the overall probability of a greedy-generated path is lower than the overall probability of another path that wasn’t considered due to greedy generation. However, in modern LLMs, this strategy has been widely disregarded. Even popular frameworks used for inference and deployment of LLMs are considering dropping support for this generation strategy³, since most models leverage sampling-based strategies, where the next token is sampled from the probability distribution learned from the model. This is due to computational efficiency, since beam search considers multiple possible generation paths it takes more time than greedy decoding. Therefore, we are interested in understanding how relevant is beam search in modern LVLMs for the MMT task. In this case, we only consider the Qwen 2.5 VL 7B model and all previously considered settings on this model. We perform beam search decoding with a number of beams equal to 3. Note that there is still no sampling when using this approach, since the strategy still relies on navigating

the paths with the highest probability. Therefore, the results are still reproducible, and randomness is not present. Results for the "long" and "short" datasets are reported in Table 4. Results indicate that performance improves when using beam search, both for inference with and without the image associated with the text. Remarkably, performance is also better for the "short" dataset, indicating that even for the generation of a short sequence of tokens, beam search still proves more effective than greedy decoding.

5.5. Error Analysis

We perform manual verification of a subset of instances for both the "long" and "short" datasets. We aim to find types of errors in instances where the generated lemma is not correct (for the "short" dataset) and where the generated translated sentence is not correct (for the "long" dataset). For LLaMA Scout, most error cases for the "short" dataset are related to the model generating longer outputs to describe the reasoning process or alternative options. For example, the model may provide a list of possible alternatives, separated by a newline character, instead of a single string. This highlights that the model is not as capable of following instructions embedded within the prompt (that is, the string "Provide only the translated text") when the text to translate only contains a few words. This behavior is not as prevalent for the "long" dataset where the model only provides the translated sentence directly. Additionally, this pattern is more present for outputs obtained when performing inference using the image, rather than text alone. This explains the lower result for exact match on the "short" dataset in translation from Italian to English for LLaMA Scout as shown in Table 2. However, this does not seem to affect Qwen 2.5 VL 72B as much, since there is no instance of generated text showcasing the previously described problem. Finally, we also showcase a relevant problem in MMT for the "long" dataset. That is, properly evaluating domain-specific knowledge is complex in the MMT task. For example, several instances within the original dataset refer to the "football" sport (e.g. "A young man about to throw a football."). When translating these instances from Italian to English with the image paired to it, even when

³<https://github.com/vllm-project/vllm/issues/6226>

Model	With Image	Multi30K						BabelNet	
		EN → IT			IT → EN			EN → IT	IT → EN
		BLEU	METEOR	COMET	BLEU	METEOR	COMET	EM	EM
Qwen2.5-VL-7B-Instruct GD	X ✓	.4793 .4800	.7213 .7211	.8751 .8745	.5680 .5839	.8308 .8398	.9037 .9069	.4700 .4700	.5150 .5900
Qwen2.5-VL-7B-Instruct BS	X ✓	.5169 .5103	.7462 .7408	.8856 .8842	.5745 .5961	.8380 .8467	.9049 .9086	.4800 .4925	.5350 .5950

Table 4

Results for the Qwen 2.5 VL 7B model on the greedy decoding (GD) and beam search (BS) generation strategies for the "long" and "short" dataset using their related metrics. Best result between the two models for each metric is in bold.

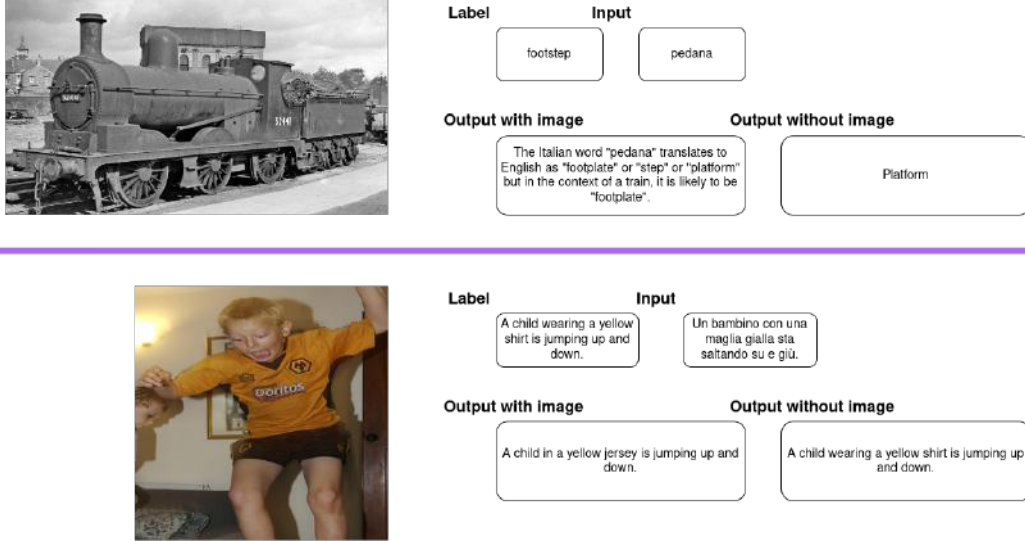


Figure 2: Example of the two types of errors that have been manually verified. The first example refers to an instance of the "short" dataset generated by LLaMA Scout, while the second refers to an instance of the "long" dataset generated by Qwen 2.5 VL 72B. In the first example, the translation with the image input is correct, but due to the reasoning generated by the model, is flagged as incorrect by the exact match metric. In the second case, the translation that is obtained using the additional input image is more faithful to the contents of the image w.r.t. the contents of the input sentence.

the word "football" was kept in the translated text (e.g. "Un ragazzo pronto a lanciare un pallone da football."), the model translated it with "rugby" (e.g. "A boy ready to throw a rugby ball."). Interestingly, this pattern is not as prevalent when the image is not provided to the model, which tends to follow the terminology used in the input sentence (e.g. "A boy ready to throw a football."). This pattern was also evident for the Qwen 2.5 VL 72B model, which is the best-performing model on the benchmark. This highlights that the models tend to prefer specific terminology and are overall deeply affected by the image that is paired with the input text. In Figure 2 we provide visual examples of these two types of errors we found during manual verification.

6. Conclusions

In this work, we have extended the current state-of-the-art in MMT by providing a study on the English and Italian languages for the task. Specifically, we extended the most relevant dataset in the state-of-the-art for MMT, that is Multi30K and introduced a new benchmark based on BabelNet, which allows to study the effectiveness of MMT when the text only consists of few words. Moreover, we have conducted extensive experimentation with several modern LVLMs, evaluating their performance in MMT across two different use cases ("long" and "short" input text). Finally, we have studied and discussed the impact of several factors on the performance of the models for MMT, namely the presence of an image along with the input text, the scale of the model, the use of LLMs instead of LVLMs, and the generation strategy. In the fu-

ture, we plan to further extend this study to more models and to consider additional languages, like German and French that are present in the original Multi30K dataset.

Acknowledgments

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007) under the NRRP MUR program funded by the NextGenerationEU.

References

- [1] P. Kumar, Large language models (llms): survey, technical frameworks, and future challenges, *Artificial Intelligence Review* 57 (2024) 260.
- [2] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, X. Hu, Harnessing the power of llms in practice: A survey on chatgpt and beyond, *ACM Transactions on Knowledge Discovery from Data* 18 (2024) 1–32.
- [3] W. Zhu, H. Liu, Q. Dong, J. Xu, S. Huang, L. Kong, J. Chen, L. Li, Multilingual machine translation with large language models: Empirical results and analysis, in: K. Duh, H. Gomez, S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2765–2781. URL: <https://aclanthology.org/2024.findings-naacl.176/>. doi:10.18653/v1/2024.findings-naacl.176.
- [4] M. Cui, P. Gao, W. Liu, J. Luan, B. Wang, Multilingual machine translation with open large language models at practical scale: An empirical study, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 5420–5443. URL: <https://aclanthology.org/2025.naacl-long.280/>.
- [5] S. Yao, X. Wan, Multimodal transformer for multimodal machine translation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 4346–4350. URL: <https://aclanthology.org/2020.acl-main.400/>. doi:10.18653/v1/2020.acl-main.400.
- [6] M. Futeral, C. Schmid, I. Laptev, B. Sagot, R. Bawden, Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 5394–5413.
- [7] D. Elliott, S. Frank, K. Sima'an, L. Specia, Multi30K: Multilingual English-German image descriptions, in: A. Belz, E. Erdem, K. Mikolajczyk, K. Passtra (Eds.), *Proceedings of the 5th Workshop on Vision and Language*, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 70–74. URL: <https://aclanthology.org/W16-3210/>. doi:10.18653/v1/W16-3210.
- [8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, *Llama 2: Open foundation and fine-tuned chat models*, 2023. URL: <https://arxiv.org/abs/2307.09288>. arXiv:2307.09288.
- [9] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, *Advances in neural information processing systems* 36 (2023) 34892–34916.
- [10] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, *Qwen2.5-vl technical report*, 2025. URL: <https://arxiv.org/abs/2502.13923>. arXiv:2502.13923.
- [11] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Keanealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebrury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A.

- Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Bal-lantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Pöder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, Gemma 3 technical report, 2025. URL: <https://arxiv.org/abs/2503.19786>. arXiv:2503.19786.
- [12] M. AI, The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025. URL: <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.
- [13] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Transactions of the Association for Computational Linguistics* 2 (2014) 67–78. URL: <https://aclanthology.org/Q14-1006/>. doi:10.1162/tac1_a_00166.
- [14] D. Elliott, S. Frank, L. Barrault, F. Bougares, L. Specia, Findings of the second shared task on multimodal machine translation and multilingual image description, in: *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 215–233. URL: <http://www.aclweb.org/anthology/W17-4718>.
- [15] L. Barrault, F. Bougares, L. Specia, C. Lala, D. Elliott, S. Frank, Findings of the third shared task on multimodal machine translation, in: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 2018, pp. 304–323.
- [16] X. Cheng, Z. Yao, Y. Xin, H. An, H. Li, Y. Li, Y. Zou, Soul-mix: Enhancing multimodal machine translation with manifold mixup, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 11283–11294.
- [17] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft coco: Common objects in context, 2015. URL: <https://arxiv.org/abs/1405.0312>. arXiv:1405.0312.
- [18] R. Navigli, S. P. Ponzetto, Babelnet: Building a very large multilingual semantic network, in: *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 216–225.
- [19] M. Freitag, Y. Al-Onaizan, Beam search strategies for neural machine translation, in: T. Luong, A. Birch, G. Neubig, A. Finch (Eds.), *Proceedings of the First Workshop on Neural Machine Translation*, Association for Computational Linguistics, Vancouver, 2017, pp. 56–60. URL: <https://aclanthology.org/W17-3207/>. doi:10.18653/v1/W17-3207.