Overcoming Data Scarcity in Domain-Specific Context-Aware MT

Anonymous ACL submission

Abstract

Neural Machine Translation (NMT) has improved translation by using Transformer-based models, but it still struggles with word ambiguity and context. Multimodal Machine Translation (MMT) offers a solution by integrating visual data to enhance translation accuracy, though it can also introduce noise. Our research explores how adding image information can improve translations in domain-specific contexts where image data might be available only for a fraction of parallel sentences. We investigate and compare different methods that are applicable to context-aware translation. We test vision-language models (VLMs) on a new e-commerce dataset and standard benchmarks like CoMMuTE and Multi30K, finding that visual context aids translation quality. Additionally, we explore the incorporation of contextual information into text-to-text models, such as the product's category path or image descriptions. The results of our study demonstrate that the incorporation of contextual information for specific scenarios leads to an improvement in the quality of machine translation.

1 Introduction

Neural Machine Translation (NMT) has significantly advanced the field of machine translation by leveraging Transformer-based models (Bahdanau et al., 2015; Vaswani et al., 2017). These models have been critical in enhancing translation quality, particularly by incorporating mechanisms like cross-attention to achieve better semantic understanding. However, despite these improvements, sentence-level translation in NMT often struggles with issues such as word ambiguity and contextual disambiguation (Rios Gonzales et al., 2017). For example, the word "bank" can refer to a financial institution or the side of a river, depending on the context. These ambiguities present a significant challenge in achieving accurate translations solely based on the level of single sentences.

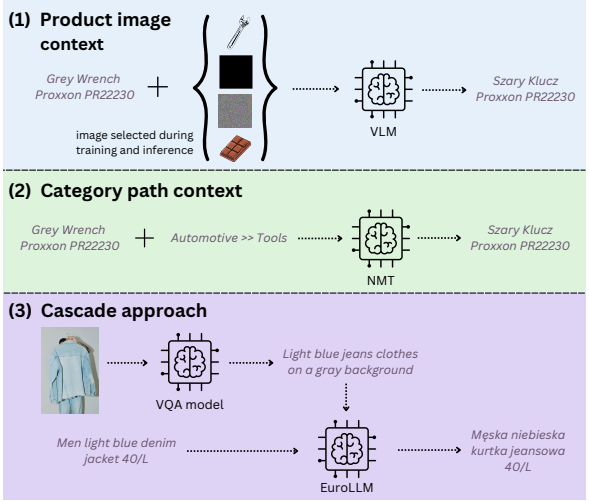

Figure 1: We evaluate different methods for tackling data scarcity in domain-specific context-aware MT: (1) adding images during both training and inference, (2) category path context, and (3) a cascade approach consisting of a vision Q&A and a text-to-text MT LLM.

In an attempt to address these limitations, Multimodal Machine Translation (MMT) has emerged as a promising paradigm (Specia et al., 2016; Shen et al., 2024). MMT integrates visual information alongside textual data to provide additional context, thereby enhancing translation quality. Studies have shown that leveraging visual cues can significantly improve the disambiguation of lexical items and contribute to more accurate translations (Yao and Wan, 2020; Wang and Xiong, 2021). For instance, visual context can help resolve ambiguities in sentences where the textual information alone is insufficient (Liu et al., 2021). However, the inclusion of visual data is not without its drawbacks. Instances of inadequate or irrelevant visual information can lead to degraded translation performance, introducing noise and potential errors (Hatami et al., 2023). Additionally, high-quality training data with aligned images is available only for a fraction of the parallel corpora.

In this work, we explore the effectiveness of integrating image information into domain-specific translation tasks, particularly in the context of e-commerce (Figure 1). We examine vision-language models (VLMs) for machine translation (MT), focusing on Czech to Polish translation, even though the models were not initially trained for this task. We evaluate performance using a newly curated e-commerce dataset, along with typical MMT benchmarks like Multi30K (Elliott et al., 2016) and CoMMuTE (Futeral et al., 2023). Our findings demonstrate that visual context can enhance translation quality in domain-specific scenarios, serving as a valuable additional feature. Furthermore, we introduce a cascade approach for encoding contextual information in text-to-text MT models, suggesting that this method can lead to further improvements in translation accuracy.

We study the following research questions:

RQ1 Can image information improve translation quality in domain-specific data?

RQ2 How do visual inputs affect training and inference in multimodal models?

RQ3 Can a cascade approach improve context encoding for text-to-text MT models?

We start by looking at related work (§2), then explain our experimental setup (§3), and finally discuss the results (§4).

2 Related Work

In this section we discuss multimodal MT and vision-enhanced language models. Literature review of multimodal MT datasets and e-commerce MT is provided in Appendix A.

2.1 Multimodal machine translation

Multimodal Machine Translation (MMT), which integrates visual context into machine translation, has garnered significant attention in recent years (Elliott et al., 2016; Shen et al., 2024). Research has shown that visuals can effectively bridge linguistic gaps between languages (Chen et al., 2019; Sigurdsson et al., 2020; Su et al., 2019). Early efforts by Calixto and Liu (2017) incorporated global image features into the encoder or decoder of NMT models. Subsequent studies have explored using more granular image contexts, like spatial image regions with attention mechanisms (Calixto et al., 2017a; Caglayan et al., 2016). Yang et al. (2020) introduced a method for joint training of source-to-target and target-to-source models to promote visual agreement. Yin et al. (2020) built a multimodal graph linking image objects and source words, leveraging an external visual grounding model for alignment.

2.2 Visual information integration in LLMs

The field of NLP has evolved with the introduction of large language models (LLMs) (Ouyang et al., 2022; Achiam et al., 2023; Touvron et al., 2023). Since traditional text-only MT has advanced to LLM-based methods (Xu et al., 2023), leveraging LLMs for MMT is a promising direction. Current multi-modal LLMs, whether using linear models (Liu et al., 2024) or Query Transformers (Li et al., 2023), often suffer from visual information loss. Enhancing alignment between textual and visual modalities and adaptively extracting relevant visual information are critical for optimizing LLM performance in MMT tasks.

3 Experiment Setup

In this work, we examine two types of experiments. First, we focus on the *Czech to Polish e-commerce domain*, evaluating the translation of product-related texts, including product names and descriptions. Our approach uses fine-tuned VLMs with images as context, alongside text-to-text NMT models that incorporate product category paths. Secondly, we assess a *cascade approach* that integrates descriptions of images in general domain translations using VLM and LLM.

3.1 Czech to Polish E-commerce domain

3.1.1 Training data

By aligning Polish and Czech product names and descriptions from e-commerce platform, we collected 230,000 parallel sentences paired with product category paths and 38,000 unique images creating a parallel e-commerce dataset. This alignment was performed using a language-agnostic BERT sentence embedding model (Feng et al., 2022). For experiments involving translation with image context, we also gathered 440,000 product names in Polish with 430,000 unique images. These names were back-translated into Czech, creating a synthetic dataset of product names with image context. The images were in JPEG format with varying sizes and were resized to 224x224 pixels.

For experiments incorporating category paths

as context, in addition to the 230,000 parallel sentences from the aforementioned corpus, we prepared 7 million back-translated product names and supplemented them with 7 million sentences from product descriptions. All back-translations were paired with category paths. Additionally, we expanded the training set by adding 15 million parallel sentences without category paths from Czech-Polish dataset (comprising primarily the OPUS corpora (Tiedemann and Nygaard, 2004) and e-commerce domain data). The path of the categories was provided as text, listing the subcategories of each product ("*Sports* » *Bicycles* » *Tyres*").

3.1.2 Evaluation data

From the aforementioned e-commerce parallel training dataset, we randomly extracted 640 sentences for the validation set and 3,405 sentences for the test set. We ensured that neither the sentences nor the images overlap with the training set.

Furthermore, we evaluated our model on a test set that contains 5,212 human-translated product names. We believe that this type of test set is better suited for image or category context translation due to the lack of rich textual content in these segments.

3.1.3 Models

Models with image context Experiments involving translation with image context were conducted using two VLMs: LLaVA-v1.6-Mistral-7b (Liu et al., 2024) and PaliGemma-3b-pt-224-pt-224 (Beyer et al., 2024). Our primary goal was to evaluate the influence of images on translation. To achieve this, we fine-tuned the models for translating product names and descriptions with two types of image data: (1) original corresponding product images, and (2) images unrelated to the text input. For the LLaVA model, we tested three types of unrelated image inputs: white, RGB noise, and random images selected from the training set. For PaliGemma, we experimented only with correct images and black images. Details on experiment and fine-tuning are presented in Appendix B.2.

Models with category context We conducted experiments by fine-tuning e-commerce text-to-text model, developed using the Marian framework (Junczys-Dowmunt et al., 2018). We implemented two fine-tuning approaches: one that incorporates category context and one that does not.

The category context was integrated by adding a prefix containing the product's category path to the source sentences. The model without category context was trained using the same configuration and data setup, but without the category path prefixes. Detailed data and training configuration are available in the Appendix B.3

3.2 Image context – cascade approach

To address the challenge of translating with image context, given the limited availability of data and models, we adopted an approach using a cascade methodology. Instead of applying a VLM directly for translation, we first generate an image description, and then employ a LLM to translate the text with image description incorporated into the translation prompt as a text. In this experiments we did not perform any fine-tuning.

3.2.1 Evaluation data

Evaluations were performed using the CoMMuTE test set (Futeral et al., 2023), specifically designed to assess the impact of images on translation quality. We tested the cascade approach on the following language pairs: English to German, English to French, and English to Czech. Furthermore, for the English-to-German language pair, we assessed inference by providing prompts containing image descriptions that were unrelated to the source segment. This approach aimed to evaluate the robustness of LLM. In addition to testing on the CoMMuTE dataset, we also conducted experiments using the WMT2018 test set from Multi30k benchmark (Elliott et al., 2016).

3.2.2 Models

The image descriptions were generated in the target language using the paligemma-3b-mix-224 model. Details of the inference process are provided in Appendix C.1. For the translation process, we used EuroLLM (Martins et al., 2024) model. We compared translations generated with and without the inclusion of image descriptions in the prompt. The inference process and prompt format used in the comparison is shown in the Appendix C.2.

3.3 Evaluation metrics

We use sacreBLEU (Post, 2018) to calculate chrF[1] (Popović, 2015) score and Unbabel/wmt22-comet-da[2] (Rei et al., 2022) model to calculate COMET metric. We used sacreCOMET (Zouhar et al., 2024) to create the COMET setup signature.

[1] chrF signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1

[2] Python3.9.19|Comet2.2.2|fp32|Unbabel/wmt22-comet-da

Model name	Train	Inference	chrF	COMET
Tatoeba ces-pol	OPUS	—	72.80	0.8870
E-commerce	—	—	83.61	0.9289
E-commerce-ft	no category	no category	84.10	0.9342
	category	category	**84.20**	**0.9347**
llava-v1.6-Mistral-7b	real_img	real_img	**82.13**	**0.9254**
		white_img	81.06	0.9198
		noise_img	80.95	0.9204
		random_img	80.04	0.9149
	white_img	real_img	81.75	0.9245
		white_img	81.80	0.9242
		noise_img	81.72	0.9241
		random_img	81.76	0.9245
PaliGemma-3b-pt-224	real_img	real_img	82.46	**0.9292**
		black_img	81.14	0.9206
	black_img	real_img	**82.53**	0.9271
		black_img	82.10	0.9255

Table 1: Results on the product names test corpora. Model with context achieved better results on automatic metrics.

	En-De		En-Fr		En-Cs	
Metric	chrF	COMET	chrF	COMET	chrF	COMET
EuroLLM	57.1	80.94	56.66	80.07	43.13	82.64
EuroLLM+desc	58.6	82.37	58.67	80.96	44.72	83.23
Δ	*+1.5*	*+1.43*	*+2.01*	*+0.89*	*+1.59*	*+0.59*

Table 2: Results of the cascade approach on the CoMMuTE dataset.

	CoMMuTE		WMT2018 test	
Metric	chrF	COMET	chrF	COMET
EuroLLM	57.1	0.8094	59.9	0.8374
EuroLLM+ noise description	56.6	0.8035	59.9	0.8375
EuroLLM+ description	**58.6**	**0.8237**	60.2	**0.8395**
NLLB-1.3b	54.9	0.8057	**60.4**	0.8235

Table 3: En–De evaluations on the CoMMuTE and the WMT2018 test sets.

4 Results and discussion

The baseline models – Tatoeba ces-pol, E-Commerce, and NLLB-1.3b are described in detail in Appendix B.1.

4.1 Czech to Polish e-commerce domain

The results on the product names test set, shown in Table 1, demonstrate that both fine-tuned versions of the e-commerce text-to-text model improved performance, and the version incorporating the category path context achieved a slight advantage.

In the case of VLMs, those trained and evaluated with correctly paired images outperformed those using unrelated images. The results confirm our hypothesis that the worst-performing models are those using randomly selected images, compared to random noise and white background. Even models trained with unrelated images showed a slight quality improvement when inference was performed on the correct images. Additional experimental results can be found in Appendix C.

4.2 Cascade approach on general-domain data

The results of the cascade approach on the CoMMuTE test set showed that translations incorporating image descriptions across all language pairs achieved higher quality, based on automatic metrics, as shown in Table 2. In this type of test set, image context is essential for effective machine translation, and image descriptions provided the necessary information for EuroLLM to enhance the translation.

The evaluation of the cascade approach, compared with adding noisy image descriptions as shown in Table 3, demonstrates that while unrelated descriptions led to a slight reduction in MT quality, the differences were smaller than those observed between translations with and without proper image descriptions. Nevertheless, prompts with image descriptions continue to deliver better performance.

5 Conclusion

This study investigates methods for incorporating context into MT, with a focus on Czech-to-Polish translations in the e-commerce domain. First, we explored fine-tuning VLMs for machine translation to leverage image-based context for improved translation quality. Secondly, we analysed the effect of product category paths on translation performance in text-to-text models for e-commerce data. Both experiments showed that the models benefited from contextual information when tested on a human-created product names dataset.

Additionally, we evaluated a cascade approach for general MT that incorporates image descriptions as contextual information. This method showed considerable translation quality improvement, even without fine-tuning, making it a viable alternative for machine translation tasks that directly leverage image-based context.

Limitations

Our approaches rely heavily on the quality of training data and the suitability of the test set for context-aware translation. In many cases, the text alone is sufficient without additional context. Moreover, incorporating extra context can sometimes reduce translation quality, especially with LLMs, where hallucinations may introduce critical errors for users. The experiments with VLMs and LLMs discussed in this paper require significantly more computational resources compared to text-to-text NMT models, due to the larger model and data sizes. We used AI assistance exclusively to enhance the text style and identify grammatical errors in this manuscript.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *arXiv preprint arXiv:1609.03976*.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

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

Iacer Calixto, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. 2017b. Using images to improve machine-translating E-commerce product listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 637–643, Valencia, Spain. Association for Computational Linguistics.

Shizhe Chen, Qin Jin, and Jianlong Fu. 2019. From words to sentences: A progressive learning approach for zero-resource machine translation with visual pivots. *arXiv preprint arXiv:1906.00872*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. 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)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Benoît Sagot, and Rachel Bawden. 2024. Towards zero-shot multimodal machine translation. *arXiv preprint arXiv:2407.13579*.

Ali Hatami, Paul Buitelaar, and Mihael Arcan. 2023. A filtering approach to object region detection in multimodal machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 393–405, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Pengbo Liu, Hailong Cao, and Tiejun Zhao. 2021. Gumbel-attention for multi-modal machine translation. *arXiv preprint arXiv:2103.08862*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Huangjun Shen, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. A survey on multi-modal machine translation: Tasks, methods and challenges. *arXiv preprint arXiv:2405.12669*.

Gunnar A Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10850–10859.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2021. Product-oriented machine translation with cross-modal cross-lingual pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2843–2852.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. 2019. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International*

Conference on Neural Information Processing Systems, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2720–2728.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. 2020. Visual agreement regularized training for multi-modal machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9418–9425.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, Online. Association for Computational Linguistics.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using comet. *Preprint*, arXiv:2408.15366.

A Expanded literature review

A.1 Multimodal MT shared tasks and benchmarks

Between 2016 and 2018, WMT organised a shared task on multimodal MT (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). The organisers primarily used the Multi30K dataset (Elliott et al., 2016) and metrics like BLEU, METEOR, and TER. In the first edition of the shared task, Specia et al. (2016) found that neural MMT models initially underperformed compared to text-only SMT models. However, Elliott et al. (2017) and Barrault et al. (2018) expanded languages and test datasets, noting performance improvements in MMT models, especially when incorporating external resources.

Futeral et al. (2023) introduced the *CoMMuTE* (Contrastive Multilingual Multimodal Translation Evaluation) dataset to evaluate multimodal machine translation systems with a focus on resolving ambiguity using images. Their new approach, which includes neural adapters and guided self-attention, showed significant improvement over text-only models, particularly in English-French, English-German, and English-Czech translations. Futeral et al. (2024) extended the setup with *ZeroMMT*, a technique for zero-shot multimodal machine translation that does not rely on fully supervised data. This method, which uses visually conditioned masked language modelling and Kullback-Leibler divergence training, demonstrated near state-of-the-art performance and was extended to Arabic, Russian, and Chinese to show broader applicability.

A.2 E-commerce product-oriented MT

E-commerce Product-oriented Machine Translation task uses both product images and their descriptions as inputs, outputting translated descriptions essential for cross-border e-commerce. It faces unique challenges such as data scarcity and translating product-specific jargon without relevant visual context. Calixto et al. (2017b) pioneered this task with a bilingual product description dataset, evaluating models like phrase-based statistical MT (PBSMT), text-only MT, and MMT. Their results highlighted PBSMT's superior performance, with MMT models enhancing translation when re-ranking PBSMT outputs.

Song et al. (2021) expanded the dataset and introduced a unified pre-training and fine-tuning framework, proposing pre-training tasks for aligning bilingual texts and product images. These tasks include masked word reconstruction with bilingual and image context, semantic matching between text and image, and masking source words conveying product attributes. This framework contributed to more robust translation models.

B Details on Experiments

We conducted all our experiments on a single server equipped with four Nvidia A100 GPUs, each with 80GB of RAM.

B.1 Baselines

We compare our results to the following baseline models:

1. Tatoeba ces-pol – an open-source model from the Tatoeba-Challenge, built using the Marian

Model name	Train	Test	chrF	COMET
Tatoeba ces-pol	OPUS	—	63.10	0.8893
E-commerce	—	—	70.6	0.9176
E-commerce-ft	no category	no category	**71.1**	**0.9203**
	category	category	70.1	0.9178
llava-v1.6-Mistral-7b	real_img	real_img	69.99	0.9112
		white_img	68.97	0.9063
		noise_img	68.32	0.9006
		random_img	67.84	0.8994
	white_img	real_img	69.91	**0.9141**
		white_img	69.97	0.9139
		noise_img	69.73	0.9140
		random_img	**70.00**	0.9140
PaliGemma-3b-pt-224	real_img	real_img	**70.62**	**0.9209**
		black_img	68.52	0.9119
	black_img	real_img	70.10	0.9178
		black_img	70.01	0.9177

Table 4: Results on the e-commerce parallel test corpus showed that only the PaliGemma model achieved better performance with context-aware translation. The LLava experiments yielded inconsistent outcomes on this test set compared to the one with product names, while the E-commerce-ft model without category context outperformed the other text-to-text models.

framework and primarily trained on OPUS corpora.

2. NLLB-1.3b – the nllb-200-distilled-1.3B[3] is an MT model designed primarily for research in low-resource languages. It supports single-sentence translation across 200 languages.

3. E-Commerce – MT model based on the Transformer (big) architecture (Vaswani et al., 2017), trained specifically for the e-commerce domain using the Marian framework. It employs a shared vocabulary created by the SentencePiece toolkit (Kudo and Richardson, 2018), consisting of 32,000 subword tokens, with all embeddings tied during training.

B.2 VLM fine-tuning

Both VLM's were fine-tuned using LoRA (Hu et al., 2021). For LLaVA we use rank $r = 64$ and alpha $\alpha = 128$ and select `llava-v1.6-mistral-7b`[4] as a starting model checkpoint. We fine-tune the model for one epoch with a learning rate of $1e^{-4}$ and a batch size of 32 without early stopping using the original fine-tuning scripts[5]. PaliGemma was fine-tuned with a learning rate of $1e^{-4}$ and a patience parameter of 5 epochs, with early stopping based on the chrF metric. The validation set consisted of the previously mentioned 640 sentences from the e-commerce parallel dataset. Models were compressed using 4-bit quantization, and LoRA fine-tuning was conducted in float16 precision.

B.3 Models with category context

For fine-tuning with prefixes we employed the following special tokens to mark category context in the SentencePiece vocabulary:

- [SC] — start of the category path
- [SEP] — separator of subcategories
- [EC] — end of the category path

The subcategories were provided in the target language and were not included as special tokens in the vocabulary. An example of a source sentence appears as follows: *[SC] Elektronika [SEP] Telefony i Akcesoria [SEP] Akcesoria GSM [SEP] Etui i pokrowce [EC] Pouzdro TelForceOne do Redmi Note 11 / 11 S Smart Magnet černý*. The target sentences remained unchanged.

The model without category context was trained using the same configuration and data setup, but without the category path prefixes. Fine-tunings was performed with a learning rate of 5e-6 and early stopping was applied based on the chrF metric. The validation set which contained 500 product names and 500 product descriptions, was extracted from the training data.

C Additional E-commerce Data Results

The results of the test set from the alignment of e-commerce platforms shown in Table 4 were more inconsistent compared to the evaluation of the product names. First, fine-tuning without the category context prefix yielded better results than with the context. Second, training with white images and performing inference with real images on the LLava model produced the highest score on the COMET metric. However, the PaliGemma model, when trained and evaluated with related images, achieved the best automatic metric scores among the VLMs. It should be acknowledged that this test set was automatically extracted from the training set and was not designed to address issues with text lacking context, which underscores the importance of using an appropriate test set for context-aware translation.

[3] https://huggingface.co/facebook/nllb-200-distilled-1.3B

[4] https://huggingface.co/liuhaotian/llava-v1.6-mistral-7b

[5] https://github.com/haotian-liu/LLaVA

Target language	Prompt used for image description
Czech	Popište obrázek na českém
German	Bild auf Deutsch beschreiben
French	Décrire l'image en français

Table 5: The prompt for image description generation used with paligemma-3b-mix-224 model.

> You are a professional {src_lang} to {tgt_lang} translator. Your goal is to accurately convey the meaning and nuances of the original {src_lang} text while adhering to {tgt_lang} grammar, vocabulary, and cultural sensitivities. **Additionally, consider the following image description in {tgt_lang}: {img_desc}, and ensure that your translation reflects the image's details and overall context**.

Figure 2: The prompt used to EuroLLM inference. For experiments without image context we did not include bolded text.

C.1 Generating description via VLM

The image descriptions were generated in the target language using the paligemma-3b-mix-224 model with prompts structured as simple tasks, for instance: "*Describe the image in English*". The exact prompts, depending on target language, are shown in Table 5. Images for inference were resized to 224x224 pixels.

C.2 EuroLLM setup

Prompt used for the EuroLLM inference is presented in Figure 2. Inference was performed with beam-search decoding algorithm with a beam-size set to 8.