ConECT Dataset: Overcoming Data Scarcity in Context-Aware E-Commerce MT

Mikołaj Pokrywka[1,2] **Wojciech Kusa**[1,3] **Mieszko Rutkowski**[1] **Mikołaj Koszowski**[1]
[1]Machine Learning Research Allegro.com [2]Laniqo.com [3]NASK
Correspondence: {name}.{surname}@allegro.com

Abstract

Neural Machine Translation (NMT) has improved translation by using Transformer-based models, but it still struggles with word ambiguity and context. This problem is especially important in domain-specific applications, which often have problems with unclear sentences or poor data quality. Our research explores how adding information to models can improve translations in the context of e-commerce data. To this end we create ConECT– a new Czech-to-Polish e-commerce product translation dataset coupled with images and product metadata consisting of 11,400 sentence pairs. We then investigate and compare different methods that are applicable to context-aware translation. We test a vision-language model (VLM), 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 leads to an improvement in the quality of machine translation. We make the new dataset publicly available.[1]

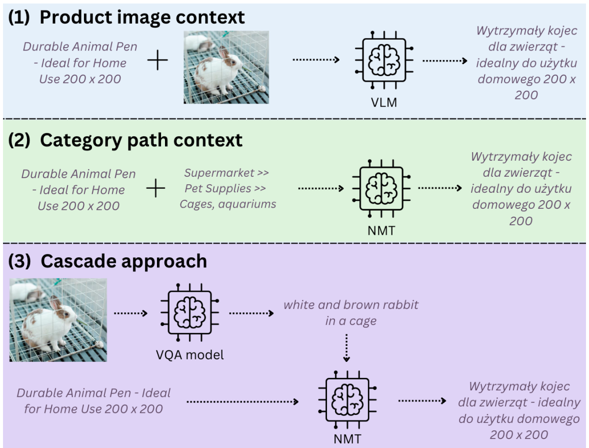

Figure 1: We evaluate three methods for contextualisation in e-commerce MT: (1) combining images with text in VLM, (2) appending category path context in NMT, and (3) a cascade approach consisting of a vision Q&A and a text-to-text NMT.

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., 2017b). These models have been critical in enhancing translation quality, particularly by incorporating mechanisms such as cross-attention to achieve better semantic understanding. However, despite these improvements, sentence-level translation in NMT often struggles with issues such as contextual disambiguation (Rios Gonzales et al., 2017). For example, the word *"pen"* can refer to a writing instrument or an enclosure for animals, depending on the context. These ambiguities present a significant challenge in achieving accurate translations based solely on the context of single sentences.

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, high-quality machine translation training data with aligned images is available only for a fraction of the parallel corpora.

In this work, we explore the effectiveness of integrating context information into a domain-specific translation task. We create a new testset for Czech-

[1]https://huggingface.co/datasets/allegro/ConECT

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 79–86
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

to-Polish translation in e-commerce, and test three approaches for context-aware MT (Figure 1). We fine-tune and evaluate a VLM, as well as traditional NMT models for which we integrate contextual information as special instruction tokens. Our findings demonstrate that visual context can enhance translation quality in domain-specific scenarios, serving as a valuable additional feature.

Our contributions are as follows:

- We release a new multimodal dataset for e-commerce MT for the under-researched cs-pl language pair;
- We fine-tune and evaluate a VLM on the domain-specific MT task;
- We propose a method for encoding category paths and image descriptions in small NMT models.

We start by describing related work (§2), then introduce the `ConECT` dataset (§3), explain our experimental setup (§4), and finally discuss the results (§5).

2 Related work

In this section we discuss multimodal and e-commerce product-oriented MT.

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 the use of more granular image contexts, such as 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 multi-modal graph linking image objects and source words, leveraging an external visual grounding model for alignment.

2.2 Benchmarks and Shared Tasks in MMT

Between 2016 and 2018, WMT organised a shared task on MMT (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.

Elliott (2018) introduced an adversarial evaluation method to assess whether multimodal translation systems effectively utilize visual context. This method evaluates the performance difference of a system when provided with either a congruent or an incongruent image as additional context.

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 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.

2.3 E-commerce product-oriented MT

An e-commerce product-oriented machine translation task uses product images and product metadata as inputs. Calixto et al. (2017c) pioneered this task with a bilingual product description dataset, evaluating models such as phrase-based statistical MT (PBSMT), text-only MT, and MMT. Their results highlighted PBSMT's superior performance, with MMT models enhancing translation when reranking PBSMT outputs. Calixto et al. (2017b) highlight the potential impact of multi-modal NMT in the context of e-commerce product listings. With only a limited amount of multimodal and multilingual training data available, both text-only and multi-modal NMT models failed to outperform a productive SMT system.

Song et al. (2021) introduced a large-scale dataset along with 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 of source words conveying product attributes. This framework contributed to more robust translation models.

2.4 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 `ConECT` dataset

The `ConECT` dataset (**Con**textual **E**-**C**ommerce **T**ranslation) is designed to support research on context-aware MT in the e-commerce domain. To create this dataset, we extracted 11,000 sentences in Polish from the `allegro.pl` e-commerce platform. Next, we aligned a primary product image and category path. The sentences were then manually translated into Czech by professional translators. Each translation was reviewed to ensure accuracy and contextual relevance. A detailed breakdown of the dataset statistics is provided in Table 1.

The `ConECT` dataset is divided into various content types to provide comprehensive coverage of e-commerce translation contexts, as summarized below.

Product names Product names are typically short phrases that precisely identify a product. They often contain specific terminology, brand names, and product specifications.

Product descriptions Product descriptions are longer texts that provide detailed information about a product, including its features, specifications, usage instructions, and benefits. These texts can be more descriptive and less structured than product names.

Split	Content type	#Sent.	#Img.	#category paths	#Tokens	CS Len.	Vocab
Test	Offer titles	1,924	1,920	840	13,898,	7.2	6,206
	Prod. desc.	3,680	2,905	1,090	38,885,	10.6	13,286
	Prod. names	4,691	4,659	1,361	34,886,	7.4	10,422
	ALL	10,295	6,146	1,542	87,669,	8.5	22,121
Valid	Offer titles	203	203	165	1,449	7.1	1,080
	Prod. desc.	403	389	285	4,118	10.2	2,472
	Prod names	505	505	360	3,757	7.4	2,116
	ALL	1,111	1,042	596	9,324	8.4	4,822

Table 1: `ConECT` dataset statistics. *Len.* denotes average length in words. Polish sentences have similar statistics.

Offer titles Offer titles are concise and attractive phrases crafted by marketers to engage potential buyers, often including promotional language, discounts, or special offers. This category can be challenging due to the need to maintain both the persuasive tone and the specific promotional content during the translation process.

4 Experimental setup

In this section we describe the training and evaluation data, the models used, and the evaluation procedure.

4.1 Training data

To create a parallel e-commerce dataset, we aligned Polish and Czech product names and descriptions for corresponding products that were listed on both e-commerce platforms `allegro.pl` and `mall.cz`. Merchants manually translated original Czech product names and descriptions into Polish. To create sentence-level pairs of multi-sentence descriptions, we split them into sentences and aligned them using a language-agnostic BERT sentence embedding model (Feng et al., 2022). We compared every sentence in one description with every sentence in the corresponding description in the other language to align the sentence pairs. That procedure resulted in a dataset containing 230,000 parallel sentences, each paired with product category paths and one of 38,000 unique images.

Data with image context For experiments involving translation with image context, we additionally collected 440,000 Polish product names paired with 430,000 unique images from the `allegro.pl` e-commerce platform. These product 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.

Text-to-text models For the baseline text-to-text models, we used 53 million sentence pairs, primarily drawn from the OPUS corpora (Tiedemann and Nygaard, 2004) and internal e-commerce domain data.

For fine-tuning with category paths as context, we extended the 230,000 parallel sentences from the original dataset with 7 million back-translated product names and 7 million back-translated product description sentences, paired with their respective category paths. Additionally, we extended the fine-tuning dataset by incorporating 7 million parallel sentences without category paths from the baseline model's training set. The category paths were represented as text, listing the hierarchical subcategories for each product (e.g., "*Sports* » *Bicycles* » *Tires*").

For experiments involving image descriptions, we generated image descriptions in Czech using the `paligemma-3b-mix-224`[2] (Beyer et al., 2024) model on all of the previously mentioned data with image context. Additionally, for fine-tuning, we included 700,000 sentences without image descriptions, extracted from the baseline model's training set.

4.2 Models

Models with image context Experiments involving translation with image context were conducted on the `paligemma-3b-pt-224`[3] model. Our primary goal was to evaluate the influence of images on translation. To achieve this, we fine-tuned the models for the translation task with two types of image data: (1) original corresponding product images, and (2) a black image unrelated to the text input. Further details of the experimental setup are given in Appendix A.1.

Text-to-text baseline model We developed a sentence-level text-to-text baseline model using the Transformer (big) architecture (Vaswani et al., 2017a), trained with the Marian framework (Junczys-Dowmunt et al., 2018). Details of the experimental setup are given in Appendix A.2.

Models with category context To ensure a fair comparison, we implemented two fine-tuning approaches on the baseline model: 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 in Polish to the source sentences. The category path was enclosed with special tokens <SC> and <EC> to separate it clearly from the source sentence. The model without category context was trained using the same configuration and data setup, but without the category path prefixes. Details of the experimental setup are given in Appendix A.3.

Models with image descriptions Following a similar approach as in the category context experiments, we fine-tuned the baseline model with and without prefixes containing image descriptions. The image descriptions were added as prefixes wrapped with <SD> and <ED> tokens. This model was trained exclusively on data with image context converted into image descriptions and did not include any data from category context experiments. Details of the experimental setup are given in Appendix A.4.

NLLB-baseline For comparison, we report our results alongside the NLLB-200-Distilled-600M model.[4]

Evaluation metrics We use sacreBLEU (Post, 2018) to calculate the chrF[5] (Popović, 2015) score, and the Unbabel/wmt22-comet-da[6] (Rei et al., 2022) model to calculate the COMET metric. We used sacreCOMET (Zouhar et al., 2024) to create the COMET setup signature.

5 Results and discussion

A performance comparison of models on `ConECT` test sets is presented in Table 2.

Models with image context For the PaliGemma models, those fine-tuned and evaluated with appropriate images outperformed those with unrelated images. Notable improvements were observed in the product names and offer titles sets. However, while the COMET metric showed a decrease for the product descriptions, the chrF metric showed a slight increase.

Models with category context Both fine-tuned models showed improved performance, with the

[2] `https://huggingface.co/google/paligemma-3b-mix-224`

[3] `https://huggingface.co/google/paligemma-3b-pt-224`

[4] `https://huggingface.co/facebook/nllb-200-distilled-600M`

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

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

Model	Train	Inference	Product names		Offer titles		Product desc		All sets	
			chrF	COMET	chrF	COMET	chrF	COMET	chrF	COMET
NLLB-600M	–	–	48.46	0.7214	38.01	0.6537	48.50	0.7774	46.85	0.7288
PaliGemma-3b	real img	real img	**83.48**	**0.9310**	**79.41**	**0.9083**	**61.92**	0.8987	**72.31**	**0.9152**
	real img	black img	81.36	0.9224	77.10	0.8972	61.75	**0.8994**	71.12	0.9095
	black img	real img	82.69	0.9275	77.86	0.9009	60.57	0.8891	71.15	0.9088
	black img	black img	82.49	0.9268	77.97	0.9009	60.87	0.8908	71.24	0.9091
Baseline	–	–	84.83	0.9326	83.73	0.9227	70.76	0.9335	77.74	0.9311
Category paths experiements	no category context		85.27	0.9372	83.66	0.9242	**72.78**	0.9389	**78.87**	0.9354
	category context		**85.51**	**0.9385**	**83.73**	**0.9248**	71.95	**0.9393**	78.56	**0.9362**
Image desc. experiments	no description context		**85.10**	**0.9367**	**83.99**	**0.9246**	**70.81**	**0.9358**	**77.90**	**0.9341**
	description context		83.25	0.8673	82.63	0.8974	48.26	0.7243	65.97	0.8219

Table 2: Comparison of the results on the ConECT test set shows that the VLM model with image context and the NMT model with category paths achieved improved performance due to the added context. However, experiments with synthetic image descriptions led to a decrease in metrics.

version incorporating category path context achieving a notable advantage in the COMET metric across all datasets. The most significant improvement was observed with product names, while the smallest gain in the COMET metric occurred with product descriptions, where the chrF metric actually decreased.

Models with image descriptions The model using image description prefixes showed a significant decrease in quality, especially on the product description dataset. It is important to note that the fine-tuning was performed on synthetic image descriptions and used a smaller dataset than the models incorporating category context. In this case, the added context had a negative impact on performance, highlighting that fine-tuning an NMT model with prefixes can degrade its quality in some scenarios.

6 Conclusion

This study explores methods for incorporating context into MT using the ConECT dataset. We are making this dataset publicly available to support research into context-aware translation tasks. We investigated the fine-tuning of VLM for machine translation to exploit 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. We also report negative results from fine-tuning with image description prefixes, highlighting that added context can sometimes impair model quality and that this straightforward approach requires further refinement.

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 VLM discussed in this paper require significantly more computational resources than 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, Sheila Castilho, and Andy Way. 2017b. Human evaluation of multi-modal neural machine translation: A case-study on E-commerce listing titles. In *Proceedings of the Sixth Workshop on Vision and Language*, pages 31–37, Valencia, Spain. Association for Computational Linguistics.

Iacer Calixto, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. 2017c. 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. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

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*.

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*.

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, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. 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 Details on experiments

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

A.1 VLM setup

PaliGemma was fine-tuned using LoRA (Hu et al., 2021) with rank $r = 8$ and alpha $\alpha = 8$. The fine-tuning was performed on a single A100 GPU for 4 epochs with a learning rate of $1e^{-4}$ and batch size set to 16.

A.2 Text-to-text baseline model

The model was trained on four NVIDIA A100 GPUs. It employs a shared vocabulary of 32,000 subword tokens, generated using the SentencePiece toolkit (Kudo and Richardson, 2018), with all embeddings tied during training. Early stopping was set to 10, with the validation frequency set to 3000 steps and based on the chrF metric on the `ConECT` validation set.

A.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 is as follows: *<SC> Moda <SEP> Odzież, Obuwie, Dodatki <SEP> Obuwie <SEP> Męskie <SEP> Sportowe <EC> Big Star pánské sportovní boty JJ174278 černé 44*. 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. Both fine-tunings were performed on two NVIDIA A100 GPUs with a learning rate of $5e^{-6}$, and early stopping was applied based on the chrF metric on the concatenated `ConECT` validation set.

A.4 Models with image descriptions

The image descriptions were generated in the Czech and Polish languages using the `google/paligemma-3b-mix-224` model and the `transformers` library. However, only image descriptions in Czech were used in the experiments. The prompts were structured as simple tasks as found in the original PaliGemma paper (Beyer et al., 2024). The exact prompts are shown in Table 3. Images for inference were resized to 224x224 pixels. We include generated Czech and Polish captions in the dataset. For fine-tuning with image description we employed the following special tokens in the SentencePiece vocabulary:

- <SD> – start of the image description
- <ED> – end of the image description

An example of a source sentence is as follows: *<SD> Černé a bílé boty s nápisem " big star " na boku. <ED> Big Star pánské sportovní boty JJ174278 černé 44*

The training configuration was identical to the experiments with category context, except that the validation frequency was reduced to every 300 steps.

Target language	Prompt used for image description
Czech	`popsat obrázek v češtině`
Polish	`opisz obrazek po polsku`

Table 3: Prompts for image description generation used with the paligemma-3b-mix-224 model.