M3T: A New Benchmark Dataset for Multi-Modal Document-Level Machine Translation

Benjamin Hsu◇, **Xiaoyu Liu**∗†, **Huayang Li**∗‡,
Yoshinari Fujinuma◇, **Maria Nadejde**◇, **Xing Niu**◇,
Yair Kittenplon◇, **Ron Litman**◇, **Raghavendra Pappagari**◇

† University of Maryland, College Park   ‡Nara Institute of Science and Technology

xliu1231@umd.edu   li.huayang.lh6@is.naist.jp

◇AWS AI Labs

{benhsu,fujinuy,mnnadejd,xingniu,yairk,litmanr,pappaga}@amazon.com

Abstract

Document translation poses a challenge for Neural Machine Translation (NMT) systems. Most document-level NMT systems rely on meticulously curated sentence-level parallel data, assuming flawless extraction of text from documents along with their precise reading order. These systems also tend to disregard additional visual cues such as the document layout, deeming it irrelevant. However, real-world documents often possess intricate text layouts that defy these assumptions. Extracting information from Optical Character Recognition (OCR) or heuristic rules can result in errors, and the layout (e.g., paragraphs, headers) may convey relationships between distant sections of text. This complexity is particularly evident in widely used PDF documents, which represent information visually. This paper addresses this gap by introducing M3T , a novel benchmark dataset tailored to evaluate NMT systems on the comprehensive task of translating semi-structured documents. This dataset aims to bridge the evaluation gap in document-level NMT systems, acknowledging the challenges posed by rich text layouts in real-world applications.

1 Introduction

Traditional machine translation (MT) systems primarily focus on textual content at the sentence level, ignoring both global context and visual layout structure of a document. Given that long-range contextual dependencies are crucial for generating high quality translations (Läubli et al., 2018; Toral et al., 2018; Hassan et al., 2018), these systems fail short of achieving human translation quality when considering entire documents (Junczys-Dowmunt, 2019). Several auhors have sought to address these gaps by including additional context, but these focused solely on text content (i.e. preceeding sentences) (Fernandes et al., 2021; Lopes et al., 2020;
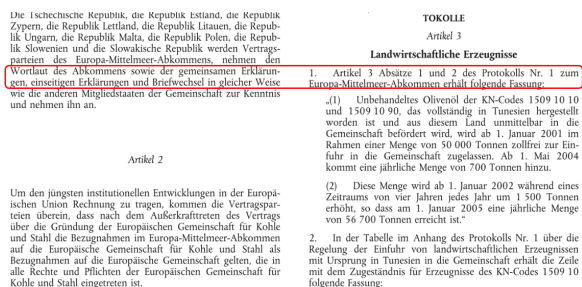

Figure 1: Example OCR errors that occur in layout-agnostic systems. Reading order is lost when extracting text and column breaks are not correctly recognized.

OCR w/o layout	
Src 1	general principles for **implemet-kit**
Src 2	**ge comma vocstonal ele poly...**
Tgt 1	allgemeiner Grundsätze für den Implementierungssatz
Tgt 2	ge Komma Gesang ele Poly...

OCR w/ layout	
Src	general principles for implementing a common vocational training policy...
Tgt	allgemeiner Grundsätze für die Umsetzung einer gemeinsamen Berufsbildungspolitik...

Table 1: Without layout information, the downstream translation system cannot recover from OCR errors resulting in garbled translations. With layout information, the reading order and contiguous blocks of text are preserved, resulting in improved translation usability and quality.

Tiedemann and Scherrer, 2017). By treating documents solely as textual content, existing methods ignore a significant portion of the information encapsulated within the visual aspects of documents.

Visual cues represent an important yet overlooked set of features which can provide contextual clues. For instance, real world documents typically include different typeface in denoting section headings to different parts of the document which indicates to human readers that parts of a document should be treated as a logically distinct. Text blocks can also indicate that sections of text are
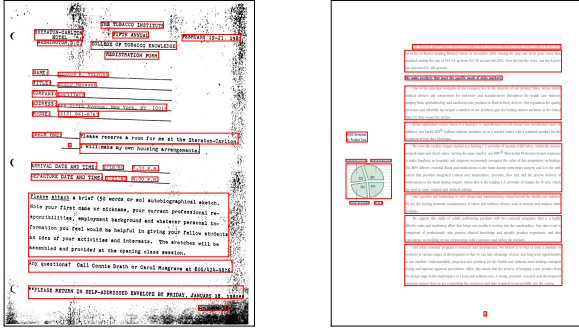
499

Figure 2: We introduce a benchmark dataset for PDF document translation. Unlike earlier works, our benchmark that focus on sentence-level text or cleanly segmented text, ours tests models on their ability to utilize visual queues for translating text. Above from left to right, are example documents sourced from the RVL-CDIP and DocLaynet datasets demonstrating the complex layout and domains included in the benchmark. Additional examples can be found in the Appendix, Figure **??**

contextually related and multi-columnar text is easily understood by human readers. Layout elements can also indicate that certain text should *not* be translated (e.g. address or name fields) or that tabular data is present. Disregarding layout and visual information can lead to catastrophic translation errors, as demonstrated in Table 1.

In recent years, there has been significant focus on Vision-Language (VL) architectures, which has led to notable advancements in multi-modal reasoning (Li et al., 2022; Alayrac et al., 2022; Ganz et al., 2023; Li et al., 2023; Ganz et al., 2024). Specifically, recent research has dedicated considerable attention to understanding structured documents, particularly in the domain of intelligent document processing and information extraction (e.g. visual question answering) tasks (Xu et al., 2020b,a; Huang et al., 2022; Tang et al., 2022). In terms of translation, earlier work has explored correcting single-word translation using visual features (Nikolov et al., 2018; Salesky et al., 2021; Wang et al., 2020; Biten et al., 2022). With recent advances in multi-modal foundation models (FMs) (Alayrac et al., 2022; OpenAI, 2023; Liu et al., 2023; Tang et al., 2022), that combine visual encoders with large language models (LLMs), development of multi-modal models to handle visually and textually challenging tasks like translation is within reach. The ability to benchmark model performance on a challenging document understanding task that takes into account long range

contextual clues is of growing importance.

Publicly available datasets for benchmarking document-level MT mainly consist of plain text documents with sentence or paragraph level alignment. The widely used JRC-Acquis consists of 8,000 documents in 20 official EU languages aligned at sentence and paragraph level (Steinberger et al., 2006). Similarly, the Web Fiction benchmark (Wang et al., 2023) consists of manually aligned sentence pairs that capture book and chapter-level contexts. These benchmarks address the text domain, assuming perfect segmentation of document structure (e.g. document and paragraph boundaries).

In this work, we present a multi-modal benchmark dataset for evaluating MT models on translating visually rich PDF documents. Our dataset is unique in that it focuses on machine translation at the document level and test models on both their ability to translate and their ability to use visual features as contextual clues. Our contributions are: (1) We introduce M3T , a new challenging benchmark testset for evaluating end-to-end **Multi-Modal MT** of structured documents (Figure 2). To further aid in research in document understanding, we have also annotated layout and reading order of the extracted text; (2) We conduct initial experiments using a multi-modal foundation model, LLaVa-v1.5 (Liu et al., 2023), and find that multi-modal features improve translation quality, though there is significant room for improvement; (3) Finally, we release synthetically generated parallel data to aid future model development by the community.[1]

2 Dataset Description

M3T focuses on PDF documents which are a commonly utilized format that pose several challenges for modern language models. Even in digitally generated PDFs, certain artifacts, such as white characters strategically placed for spacing adjustments or the duplication of characters in the absence of bold-face options, can complicate straightforward text extraction.

Data Sourcing We sourced documents from several public datasets to cover a wide range of documents. First, we used EUR-Lex[2] to source documents which are translated by professional trans-

---

[1]Dataset and scripts can be found here: http://github.com/amazon-science/m3t-multi-modal-translation-bench

[2]https://eur-lex.europa.eu/homepage.html

	RVL-CDIP	DocLayNet	EUR-Lex
Layout Elements			
Text	2780	6679	1325
Section-header	144	1070	999
Page-footer	92	685	0
Vertical	68	133	1088
Table-cell	1332	16292	16451
Table	44	256	445
Page-header	220	525	1532
Picture	28	444	0
List-item	204	2406	1317
Formula	0	23	2
Footnote	0	93	391
Code	0	13	0
Title	0	74	30
Caption	0	76	24
Language Pairs			
en→de	20	126	0
de→en	0	34	54
en→es	20	126	0
es→en	0	0	54
en→fr	20	126	0
fr→en	0	11	54
en→zh	20	126	113
zh→en	0	19	0
Document Domains			
Scanned	80	-	-
Govt. tenders	-	102	-
Patents	-	36	-
Legal	-	71	383
Fin. reports	-	344	-

Table 2: Distribution of the layout elements, language pairs, and document domains contained in the benchmark dataset. Language pairs indicate the original language of the source document.

lators into all the European languages while also preserving the document layout. We sample a subset of documents which we annotated with layout information described in the next section.

Second, we sourced documents from DocLayNet (Pfitzmann et al., 2022) and RVL-CDIP (Harley et al., 2015), which consist of documents annotated with layout information. We sourced annotations on the documents from their respective test sets to preserve the train/test split that might be used in training layout/vision models (e.g. LayoutLM).

For the sampling strategy, we focused on covering a wide range of layout elements. For RVL-CDIP, were randomly selected, excluding fax cover sheets. For DocLayNet we sampled documents containing flat layouts, multi-columnar data, images and tabular information. For EUR-Lex, we used an initial set of automatically labeled layout el-

ements for sampling documents, and selected documents containing more than just text elements (e.g. tables, captions, footnotes) that were not previously translated and represent more complex layout features. Initial automatic labelling was done using a Faster R-CNN classifier on the DocLayNet dataset using the standard `dectrctron2`[3] following Pfitzmann et al. (2022). Our classifier had an F1 of 73.4 averaged across all layout elements.

Document Annotation We annotated the sampled documents in two stages: (1) Annotators first extracted the text and labeled the layout elements; (2) Annotators translated the text extracted in each layout element. Table 2 reports the final distribution of layouts and domain coverage.

For the layout information, annotators were asked to provide bounding box coordinates and labels for each element. Bounding box labeling means grouping the elements on the page into reasonable paragraphs, headings, tables, pictures, etc. areas on the page, in non-overlapping rectangular areas (see Figure 2). For the set of labels, we extended the DocLayNet ontology to include reading order and table cells. Specifically, annotators labelled the layout elements described in Table 4.

For translating the extracted text, we provided professional translators with the original document and asked to take into account the context of the document when translating. Extracted text was then translated using a selection of commercial segment level MT systems. Translators were then instructed to post-edit these translations with the additional requirement that the length of the translation within $\pm 10\%$ of the source text length, which is useful for evaluating how well systems may preserve the original layout of the source document. In order to facilitate length compliance, we sourced documents in de, es, and fr and asked to translate them into en. We also asked to translate en documents to into zh. Additional annotation details can be found in Appendix A.

Supplementary Data We also provide automatically annotated EUR-Lex documents for future research (e.g., model training). First, we downloaded the plain text parallel corpus based on EUR-Lex (Steinberger et al., 2006, JRC-Acquis) and extracted text from EUR-Lex documents using Tesseract OCR.[4] Next, we aligned text in JRC-

---

[3]https://github.com/facebookresearch/detectron2
[4]https://github.com/tesseract-ocr/tesseract

Text	Visual Features			doc-COMET	
	Rand.	Block	Doc	EUR-Lex	DocLayNet
OCR	✓			0.0173	0.0120
			✓	**0.0174**	**0.0126**
		✓		0.0174	0.0122
Gold	✓			0.0214	0.0193
			✓	**0.0218**	**0.0194**
		✓		0.0215	0.0192

Table 3: Results from our experiments using LLaVa-v1.5 using different visual features such as random images (Rand.), images of text blocks (Block), and images of the whole document (Doc). We tested how these features are used to improve OCR'd text (OCR) and if they provide additional context for clean text (Gold). We found that visual features improved document level translation quality in terms of doc-COMET scores. Results are averaged over de, fr→en language pairs. We found LLaVa is capable of using visual features to address OCR errors. It also appears to generate better translations when given the document image as context.

Acquis with recognized (OCR'd) bounding boxes in EUR-Lex documents. Finally, we erased unaligned bounding boxes and produced tuples of `<source-text, OCR'd-text, bounding-box, target-text>`.

3 Multi-Modal Document Translation

Multi-modal document translation is under explored because of a lack of annotated benchmark datasets. Given the creation of a novel benchmark dataset in this work, we conducted experiments testing how recent multi-modal FMs perform on multi-modal document translations. To motivate future research, we experiment with a recent LLaVa-v1.5 model (Liu et al., 2023), a large multi-modal model that combines a CLIP vision encoder ViT-L (Radford et al., 2021) and Vicuna-13b (Zheng et al., 2023) for general-purpose visual and language understanding.

We conduct experiments adding various visual and layout features to the model. For visually rich documents, evaluating OCR text translation directly is difficult since an alignment between extracted source segments with clean source segments needs to be found. We leave the development of such an alignment model to future work. Instead, our experiments focus on whether the visual features are utilized by a multi-modal FM to improve translations by (1) recovering from OCR errors or (2) understanding the context of the document. We focused our experiments on X→en documents for reasons discussed below.

To test these hypothesizes, we controlled for the effects of incorrect layout prediction by conducting experiments using the gold annotated layout elements (e.g., annotated bounding boxes). For hypothesis (1), we experimented with different granularity of visual features (at the text block level or at the document level). We used Tesseract for OCR in our experiments to extract text blocks and passed a single text block at a time to the LLaVA model. We also compared with using translations gold source text translation to understand the upper bound on translation quality. Finally to test hypothesis (2), we experimented with translations using the gold text to see if less granular visual features enable a multi-modal FM to understand contextual clues better. As a baseline, we tested random images from the MS COCO dataset (Lin et al., 2014) into the LLaVa model which allows us to measure improvements from including relevant visual features. We conducted ablations using images of identified text blocks or images of the entire document. Lastly for evaluation, since the reading order of the text was annotated, we used document level COMET scores[5] (Vernikos et al., 2022) with a context size of two (i.e., two previous segments as context).

Results Table 3 summarizes the results from our experiments. We found that visual features helped the model improve translations of OCR'd text. We found evidence for this comparing performance on different languages. Tesseract OCR performance is known to vary significantly for none Latin script languages (Gupte et al., 2021). In particular, on Chinese text, we found that including images in the model improved performance by +0.02 when including images of the entire document.

We found that including relevant visual features improved upon random visual features from the MS COCO dataset. Moreover, features focused on the relevant text blocks improved translation quality (as measured by doc-COMET score) on both EUR-Lex and DocLayNet subsets of data. When looking at document level visual features, we found that these universally improved translation quality beyond the baseline with random visual features.

While there were improvements, visual features did not help translation appreciably. This presents an opportunity for future researchers to find methods to leverage contextual clues in the visual domain more fully. One issue was that the Vicuna-13b model suffered from hallucinations especially

---

[5]https://github.com/amazon-science/doc-mt-metrics

when the inputs are short, single word, or numerical segments which is the case for many of the tabular data from both EUR-Lex and DocLayNet documents. In the case of en→X translation, we found that the model misinterpreted the prompt and did not produce translations.

4 Conclusions

In conclusion, we release M3T , a new benchmark for assessing multi-modal machine translation of visually-rich documents. We conducted experiments using a multi-modal FM built ontop of Vicuna-13b and found that while these models attain impressive results on visual question-answering and captioning tasks, multi-modal document translation is still an area for future research.

5 Limitations

In this work, we used available open source models which were not specifically tuned for this benchmark task. Finally, while we sourced documents from large set of domains, the benchmark is ultimately limited to a few cases and heavily focused on legal and financial text in a few high resource languages (de, es, fr, and zh) which can introduce bias in future model evaluation and development. Furthermore, we did not experiment extensively with prompts, and an interesting future direction could be an investigation into whether their are better prompts to use for multi-modal models.

Acknowledgements

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R Manmatha. 2022. Latr: Layout-

aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16548–16558.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. 2024. Question aware vision transformer for multi-modal reasoning. *arXiv preprint arXiv:2402.05472*.

Roy Ganz, Oren Nuriel, Aviad Aberdam, Yair Kittenplon, Shai Mazor, and Ron Litman. 2023. Towards models that can see and read. *arXiv preprint arXiv:2301.07389*.

Amit Gupte, Alexey Romanov, Sahitya Mantravadi, Dalitso Banda, Jianjie Liu, Raza Khan, Lakshmanan Ramu Meenal, Benjamin Han, and Soundar Srinivasan. 2021. Lights, camera, action! a framework to improve nlp accuracy over ocr documents. *Document Intelligence Workshop at KDD 2021*.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 991–995.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings*

of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*. Oral Presentation Project Page: https://llava-vl.github.io/.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Nikola I. Nikolov, Yuhuang Hu, Mi Xue Tan, and Richard H.R. Hahnloser. 2018. Character-level Chinese-English translation through ASCII encoding. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 10–16, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3743–3751, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Elizabeth Salesky, David Etter, and Matt Post. 2021. Robust open-vocabulary translation from visual text

representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7235–7252, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Chao-Yue Zhang, and Mohit Bansal. 2022. Unifying vision, text, and layout for universal document processing. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19254–19264.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Haohan Wang, Peiyan Zhang, and Eric P. Xing. 2020. Word shape matters: Robust machine translation with visual embedding.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. Findings of the wmt 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of llms. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. Layoutlm: Pre-training of text and layout for document image understanding.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Layout	Description
Captions	Special text outside a picture or table that describes the picture or table.
Footnotes	Typically small text at the bottom of a page, with a number or symbol that is referred to in the text above.
Formula	Mathematical equation or chemical formula on its own line.
List-item	One element of a list, in a hanging shape, i.e., from the second line onwards the paragraph is indented more than the first line.
Page-footer	Repeating elements like page number at the bottom, outside of the normal text flow.
Page-header	Repeating elements like page number at the top, outside of the normal text flow.
Picture	A graphic or photograph.
Section-header	Any kind of heading in the text, except overall document title.
Table	Material arranged in a grid alignment with rows and columns, often with separator lines
Table-cell	Text contained in a cell of a table.
Text-block	Regular text paragraphs or blocks of text that should be logically grouped together.
Title	Overall title of a document typically appearing in large font or denoted with special typeface to distinguish it from the rest of the document.
Order	The order number that the bounding box should be read on the page (i.e. the reading order). For European languages, this is generally left to right, top to bottom. For tables, we adopt a left to right, top to bottom ordering of the cells.
Vertical	Text is oriented vertically.

Table 4: Layout elements and their labels annotated in our benchmark following the DocLayNet typology.

A   Annotation Details

For layout annotations, professional annotators were instructed to annotate bounding box coordinates and assign a label to the type of text given the definitions outlined in Table 4. A random selection 10% of annotation per annotators was selected and manually reviewed. If more than 5% of those labels did not follow guidelines, the set was then re-annotated.

For translations, we followed a similar protocol. We used multiple systems to generate initial translations for each sentence per document at the segment level. Professional translators were then asked to post-edit translations based on the context of the document. A 10% of translations were sampled and checked for errors. If there were errors, these samples were re-translated.

B   Detailed Results

|	Visual Features			doc-COMET					
				EUR-Lex			DocLayNet		
Text	Rand	Block	Doc	de→en	fr→en	es→en	de→en	fr→en	zh→en
OCR	✓			0.0159	0.0187	0.0226	0.0123	0.0118	-0.011
			✓	0.0160	0.0187	0.0225	0.0125	0.0128	0.013
		✓		0.0159	0.0189	0.0226	0.0123	0.0120	-0.010
Gold	✓			0.0190	0.0238	0.0281	0.0219	0.0166	0.013
			✓	0.0195	0.0241	0.0282	0.0217	0.0170	0.014
		✓		0.0191	0.0238	0.0281	0.0219	0.0166	0.014

Table 5: Results from our experiments using LLaVa-v1.5 using different visual features such as random images (Rand.), images of text blocks (Block), and images of the whole document (Doc). We tested how these features are used to improve OCR'd text (OCR) and if they provide additional context for clean text (Gold). We found that visual features improved document level translation quality in terms of doc-COMET scores. Results are averaged over de, fr→en language pairs. We found LLaVa is capable of using visual features to address OCR errors. It also appears to generate better translations when given the document image as context.
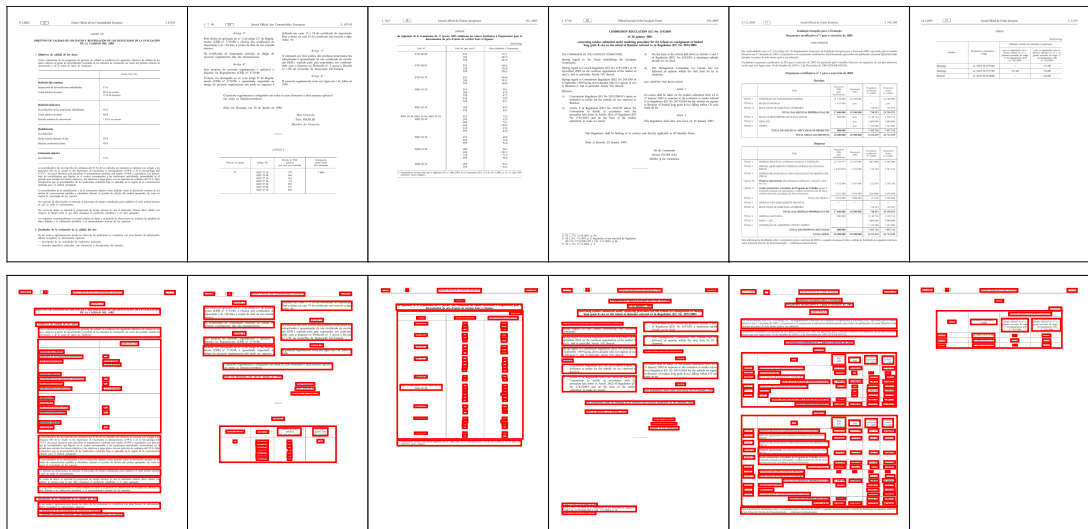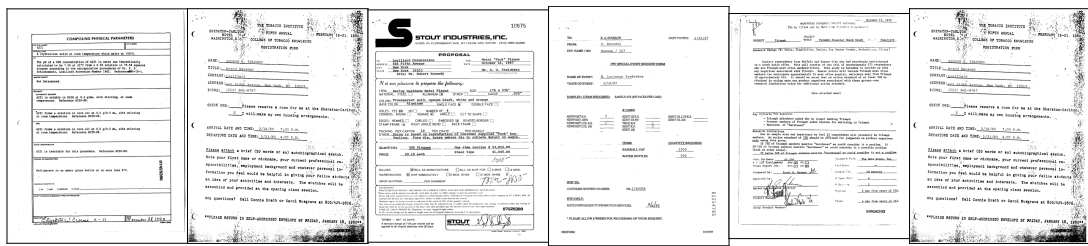
Figure 3: In this work, we introduce a benchmark dataset for PDF document translation. Example documents sourced from (a) EUR-Lex (b) DocLayNet and (c) RVL-CDIP datasets demonstrating the complex layout and domains included in the benchmark. Unlike earlier works, our benchmark focuses on machine translation which depends more heavily on contextual information than earlier works centered around entity extraction.