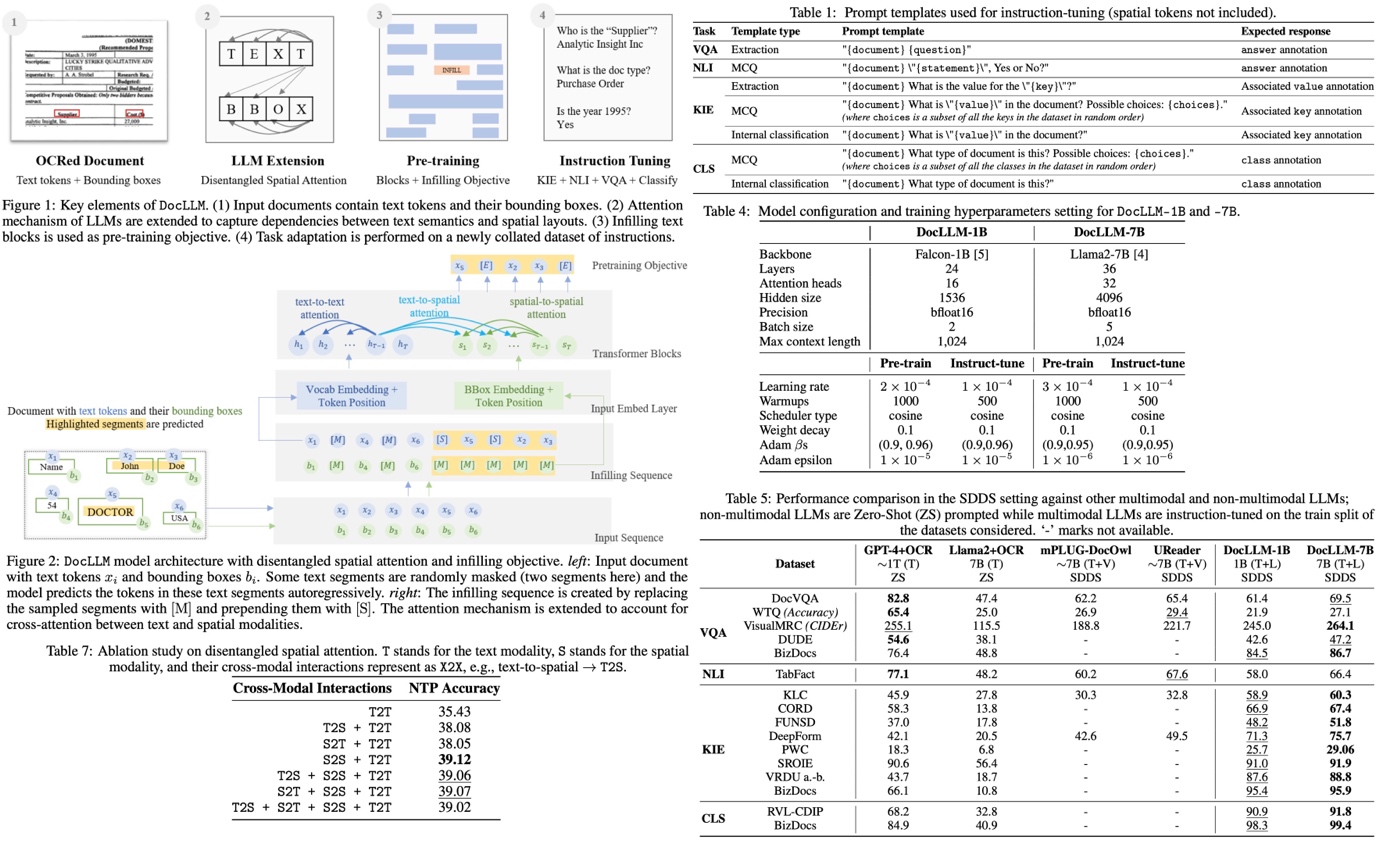
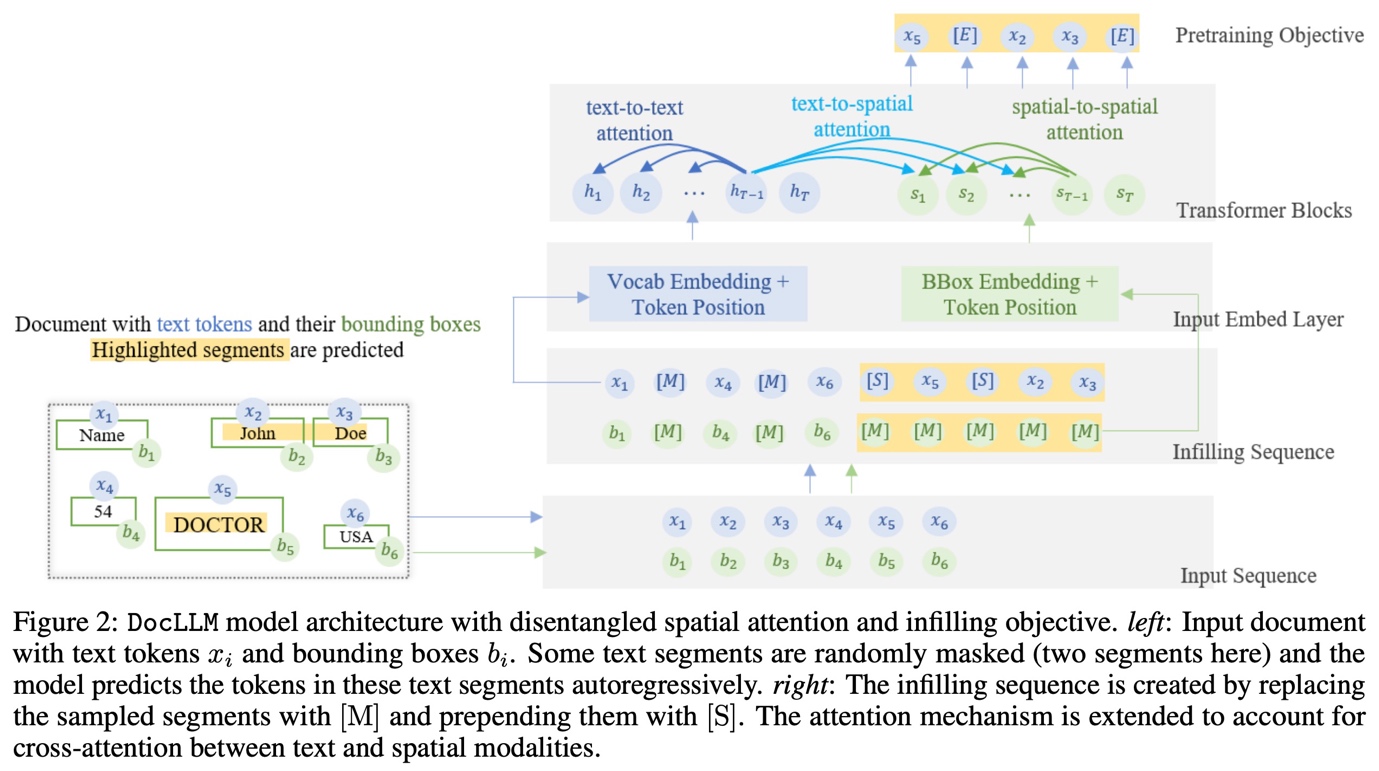
**Paper Review: DocLLM: A layout-aware generative language model for multimodal document understanding**



DocLLM is a new extension to LLMs designed for processing visual documents such as forms and invoices. It uniquely uses bounding box information instead of image encoders to understand document layouts. The model modifies the attention mechanism in transformers to separately handle text and spatial information. It is pre-trained to fill in text segments, aiding in managing diverse layouts and content in visual documents. After pre-training, DocLLM is fine-tuned on a large dataset for four core document intelligence tasks. It outperforms state-of-the-art LLMs in most tested datasets and shows strong generalization capabilities to new datasets.

**DocLLM Framework**



**Disentangled Spatial Attention**

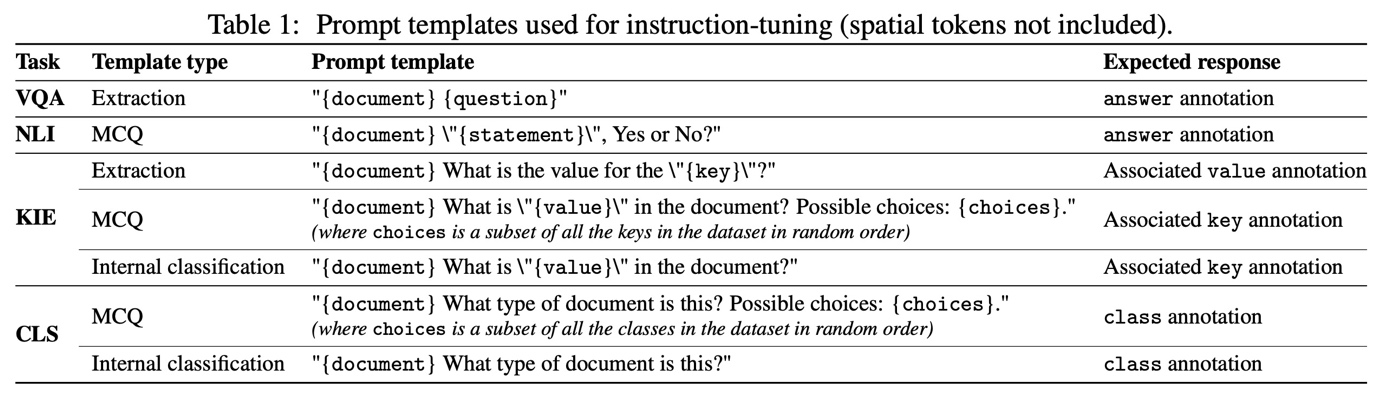
DocLLM represents inputs as pairs of text tokens and their corresponding bounding boxes. It encodes bounding boxes into separate hidden vectors and decomposes the attention mechanism into four scores: text-to-text, text-to-spatial, spatial-to-text, and spatial-to-spatial, using projection matrices and hyperparameters to balance the importance of each score. The hidden vectors for spatial information are reused across layers, reducing the number of extra parameters compared to image-based models. By not simply adding spatial information to text (which would couple layout with semantics), DocLLM maintains a disentangled representation, enabling selective focus on either modality when necessary.

**Pretraining**

DocLLM is pre-trained in a self-supervised manner on a large dataset of unlabeled documents. Instead of focusing on individual tokens, DocLLM considers larger segments or blocks of related tokens, such as text blocks or linear sequences. Additionally, DocLLM employs a text infilling objective, where predictions are based on both the preceding and following tokens rather than just the preceding ones. This approach is particularly effective for handling OCR noise, misalignments, and relationships between different document fields.

During pre-training, text blocks identified from OCR information are randomly masked and shuffled for reconstruction. This block infilling is done autoregressively, with special tokens indicating the start and end of blocks. This method is used only in pre-training and not in subsequent fine-tuning or downstream tasks. The model minimizes a cross-entropy loss.

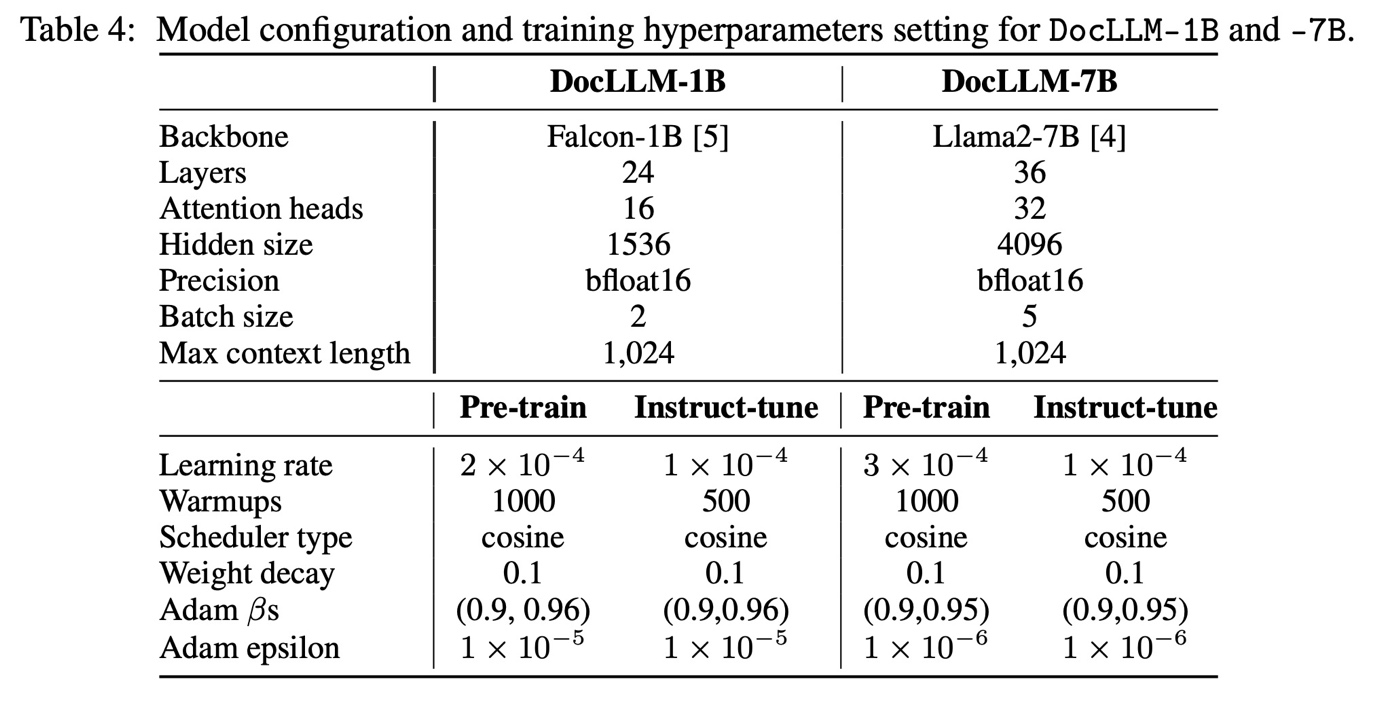
**Instruction Tuning**



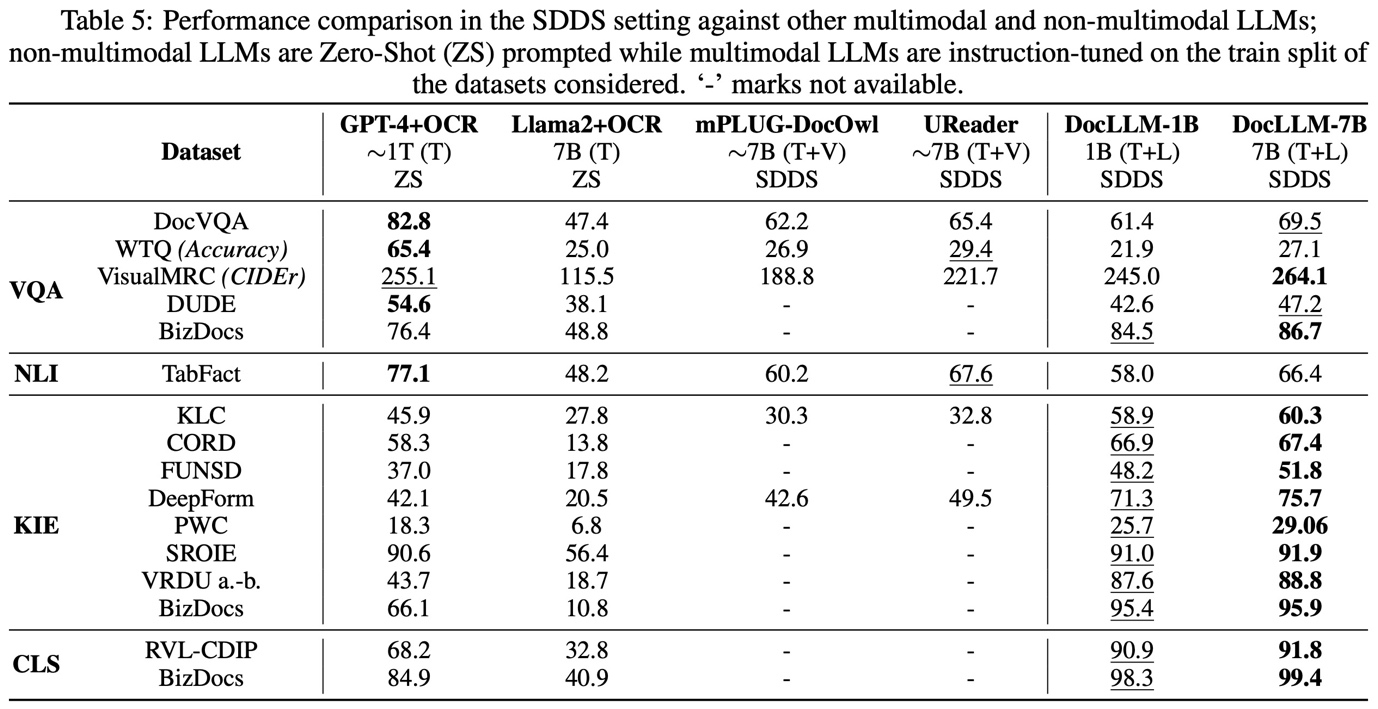
DocLLM is fine-tuned using instruction-based methods on 16 datasets covering four DocAI tasks: Visual Question Answering (VQA), Natural Language Inference (NLI), Key Information Extraction (KIE), and Document Classification (CLS). Different templates are used for each task, the examples include the following ones:

* VQA: "{document} What is the deadline for scientific abstract submission for ACOG - 51st annual clinical meeting?"
* NLI: "{document} \"The UN commission on Korea include 2 Australians.\", Yes or No?"
* KIE: "{document} What is the value for the \"charity number\"?"
* CLS: "{document} What type of document is this? Possible answers: [budget, form, file folder, questionnaire]."

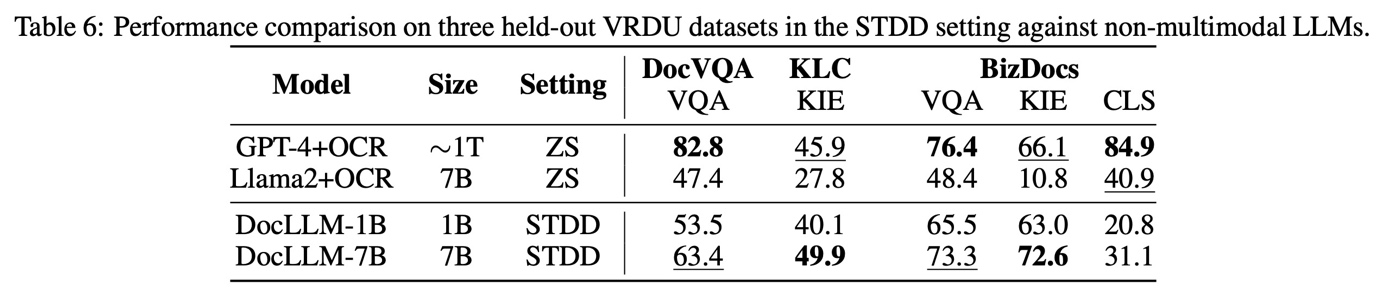
**Experiments**



The models are trained on 24GB A10g GPUs.

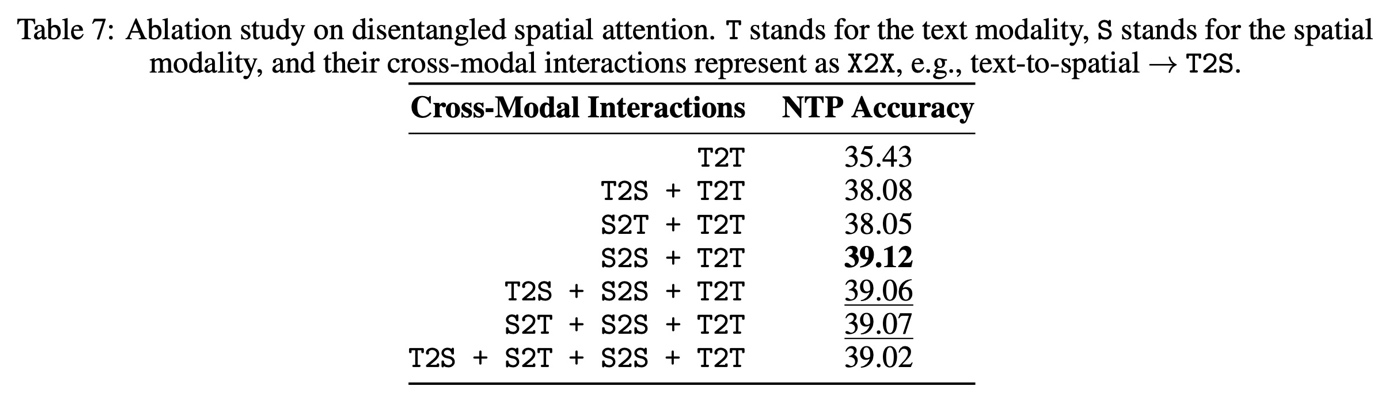


In the SDDS (Same Datasets, Different Splits), DocLLM-7B outperforms other models in 12 out of 16 datasets, including GPT4 and Llama2 in zero-shot settings. Particularly, it excels in layout-intensive tasks like KIE and CLS. However, in the other two tasks, it is outperformed by GPT-4, likely due to GPT-4’s better reasoning and abstraction capabilities.

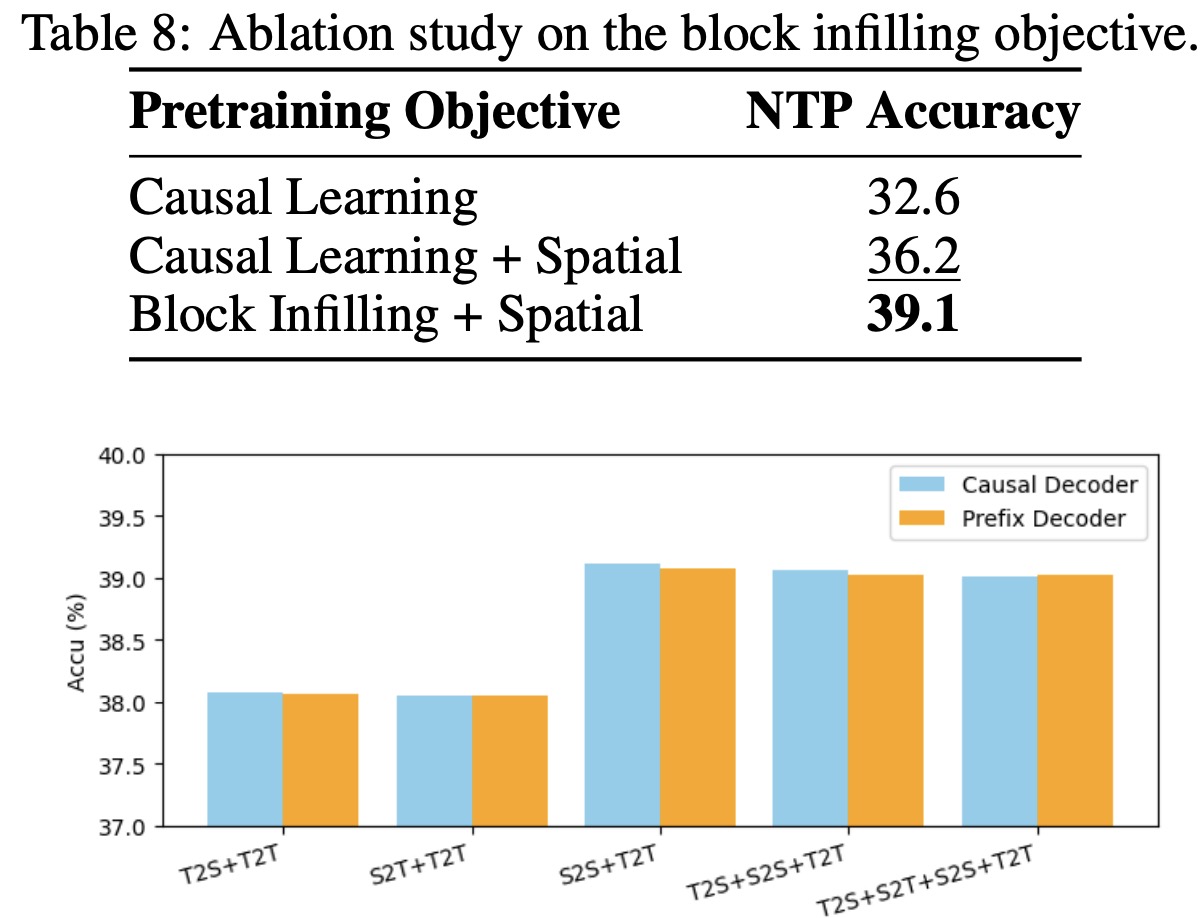


In the STDD (Same Tasks, Different Datasets), DocLLM surpasses Llama2 in four out of five datasets and achieves the best scores in two, especially in the KIE task. However, DocLLM’s classification accuracy is lower, possibly due to its training on only one classification dataset, which might limit its generalization to new datasets.

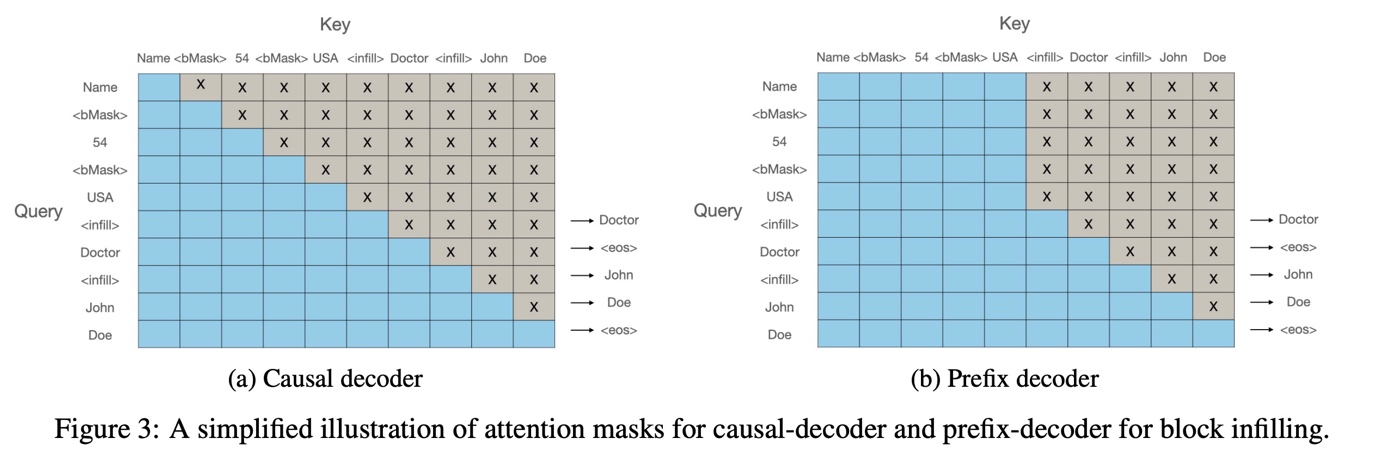
**Ablation Studies**



Disentangled Spatial Attention: Focusing solely on spatial-to-spatial interaction led to the highest accuracy in understanding documents with rich layouts. This finding emphasizes the importance of incorporating spatial features in document analysis.



Autoregressive Block Infilling: The block infilling approach with spatial modality demonstrated the best performance, underscoring the value of spatial information in the learning process.



Prefix Decoder vs. Causal Decoder: The authors compared a prefix decoder, which allows bidirectional visibility of the whole document, with a conventional causal decoder. The experiments showed only marginal differences between the two decoders across various configurations, with the causal decoder slightly outperforming the prefix decoder.