`Where Layout Meets Language`: Lightweight Spatial Enhancement to Large Language Models for Document Understanding

Nil Biescas[§1,2], Sanket Biswas[1], Josep Lladós[1], Jordy Van Landeghem[3]

[1] Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[2] Computer Vision Group, Technical University of Munich, Germany
[3] Instabase, San Francisco, USA

Abstract. Document intelligence is a multimodal challenge, requiring both textual content and visual layout cues for accurate interpretation. Traditional text-based models often struggle to capture spatial and structural information, which is essential for document understanding tasks. While decoder-only language models (e.g. GPT-4, LLaMA) excel at text-based reasoning, they often overlook layout dependencies, limiting their effectiveness in tasks like Document Visual Question Answering (DocVQA) and Key Information Extraction (KIE), where text is arranged non-linearly. To address this, we introduce `G-LLaMA`, which integrates Gaussian biases into the attention mechanism of decoder-only models to enhance spatial reasoning. Unlike prior work that applied spatial biases to encoder-based architectures, our approach conditions LLaMA's attention on spatial structure, improving layout awareness without requiring additional architectural modifications. Experimental results demonstrate that Gaussian biases significantly increase performance, yielding substantial improvements in both DocVQA and KIE tasks. These findings underscore the importance of explicit spatial conditioning in LLMs for understanding structured documents.

Keywords: Document Understanding · Layout Representation Learning · Large Language Models

1 Introduction

Visually-rich Document Understanding (VrDU) is a challenging problem stemming from the vast diversity of complex layouts, intricate visual structural elements, and heterogeneous content types that appear across different document formats. While Large Language Models (LLMs) excel in purely linguistic tasks, they often fall short in capturing the crucial spatial and structural relationships that are central to many VrDU scenarios. Documents, such as forms, invoices, recipes or infographics, are highly dependent on layout organization to convey meaning, guide the reader's attention, and allow efficient retrieval of information. Ignoring these structural cues can lead to misinterpretation, highlighting

[§] Work done during Internship at Computer Vision Center

the need for models that integrate both textual and spatial information for effective VrDU.

Existing state-of-the-art (SOTA) VrDU solutions fall under two broad categories: multimodal models and text-based models with layout features. Multimodal models (e.g. UReader [63], mPLUG-DocOwl [21,62]) integrate powerful vision encoders to process document images, achieving strong performance but incurring high computational costs and requiring extensive fine-tuning with massive datasets [30,49]. On the other hand, text-based models include "coordinate-as-token" approaches [37,43] which embed bounding box coordinates as additional tokens. Although effective in preserving spatial layout, these methods significantly increase sequence length and depend on in-context learning to generalize. Another approach, exemplified by DocLLM [55], uses disentangled spatial attention, treating spatial information as a separate modality to enhance layout-text alignment. However, they do not fully exploit the autoregressive nature of decoder-based LLMs, leading to suboptimal layout reasoning.

Against this backdrop, Structured Document Understanding (SDU) emerges as a critical research area within Document Intelligence. It aims to capture both the meaning of the text (semantics) and its presentation (structure), including the spatial arrangement of elements, their grouping, and the way they direct the reader's interaction and workflow [64]. By approaching documents through this lens, we can potentially develop models that are more robust, domain-adaptive, and ultimately better at tasks such as information extraction [25], question answering [41], and form processing [26]. Despite progress in multimodal learning [35,37,39,55], effectively combining text and layout without adding noise remains challenging, particularly in real-world industrial settings. Simple fusion methods often degrade performance by introducing irrelevant signals [39], emphasizing the need for carefully designed approaches that preserve the natural relationship between text and layout for practical and scalable SDU solutions.

Given these challenges, a key question emerges: *how can we effectively integrate spatial information into decoder-only language models [2,12,52] without compromising efficiency?* Unlike multimodal approaches [1,21,35,62] that rely on computationally expensive vision encoders, we focus on enhancing text-based models with layout awareness while preserving their autoregressive nature. Recent LLMs, particularly from the LLaMA family [16,52], have demonstrated promising results in VrDU tasks. However, these models primarily rely on attention mechanisms that lack explicit spatial awareness, limiting their ability to leverage the visual structure of documents. Previous work [66] has explored Gaussian biases over word positions to guide attention toward spatially related regions, improving layout awareness. However, these approaches have been largely restricted to encoder-based architectures [24,36,59,60] and tasks outside of DocVQA [41]. To bridge this gap, we investigate whether Gaussian biases can enhance spatial reasoning in a purely decoder-only model for DocVQA, introducing a lightweight yet effective approach for layout-aware language modeling.

The contributions of this work can be divided into three key areas. (1) We introduce `G-LLaMA`, a novel decoder-only language model that integrates Gaussian

attention to enhance layout awareness without modifying the backbone architecture, preserving efficiency while capturing spatial relationships. (2) We propose a data-efficient instruction-tuning strategy, leveraging a minimal number of samples, in contrast to existing multimodal approaches that require large-scale supervision. (3) We provide a detailed analysis of layout-aware attention within LLMs, demonstrating its impact on two core VrDU tasks (DocVQA and KIE) offering insight into how spatial conditioning enhances decoder-based architectures.

2 Related Work

Structured Document Understanding Transformer-based pre-trained language models such as the LayoutLM Family [24,59,60,61], DocFormer [3,4], BROS [20], LILT [56,57], TILT [45] and LAMBERT [14] have contributed to learning cross-modal interactions between textual and layout information, which has played a significant role in advancing SDU. However, these foundation models primarily rely on large-scale pretraining and integrate layout information by incorporating 2D bounding box coordinates (at word-level) extracted from OCR outputs, allowing them to capture structural relationships within text and improve layout-aware reasoning in fundamental VrDU tasks such as classification [19], key information extraction [25,26], and document visual question answering (DocVQA)[41,40]. Other approaches [17,34,53] adopt a region-level layout representation, identifying logical components such as paragraphs, figures, titles, and tables, as obtained from document layout analysis (DLA) architectures [6,8,44,65]. Region-level layout information has been found to be less effective than word-level representations for most popular VrDU tasks [25,26,41,42], which primarily involve enterprise documents (eg. administrative forms, invoices, receipts etc.) where fine-grained word positioning is crucial for accurate interpretation [15,7]. To further justify this deduction, GeoLayoutLM [38] extends LayoutLMv3 [24] by introducing geometric constraints through self-supervised pretraining, enabling better spatial reasoning between such text segments, which significantly improves entity linking performance in KIE [26]. While effective, these encoder-heavy models often rely on computationally expensive vision encoders and require extensive pretraining on multimodal datasets. Architectures like TILT [45] and Donut [28] integrate encoding with autoregressive decoding, making them effective for structured text generation and layout-sensitive tasks, but their high computational overhead and reliance on task-specific fine-tuning limit their adaptability. To address these limitations, we explore decoder-only LLMs [16,52], leveraging their scalability and autoregressive efficiency while integrating spatial awareness without additional vision encoders or architectural modifications.

Large Language Models for Visually-rich Document Understanding The rise of LLMs [2,12,16,52] has driven significant advancements in VrDU by leveraging vast textual corpora for knowledge extraction and reasoning. Before their emergence, OCR-free models like Donut [28] and Pix2Struct [32] aimed to

eliminate reliance on OCR by directly processing document images using vision-language transformers [29]. However, they often struggled with text-heavy documents, where explicit textual extraction still proved advantageous. This trade-off highlighted the ongoing challenge of balancing efficiency, layout awareness, and textual accuracy in VrDU models. LLM's such as GPT-4 [2] and LLaMA [16,52] have gained strong attraction in document understanding due to their scalability and efficiency in text generation tasks. However, these models lack explicit layout representation, as their self-attention mechanisms inherently process text in a sequential manner, failing to capture the spatial structure and positional dependencies crucial for visually-rich documents, especially for long context ones [10]. Recent efforts to enhance text-based LLMs for VrDU have explored various layout-aware encoding techniques. Recent efforts to enhance text-based Large Language Models (LLMs) for Visually-rich Document Understanding (VrDU) have explored various layout-aware encoding techniques. However, SOTA models like UReader [63], mPlug-DocOwl [21,22] and Idefics family [30,31] are multimodal LLMs (MLLMs) which are heavily dependent on expensive vision encoders that requires heavy pretraining data with visually-rich document collections [30,49,50]. In this work, we follow the direction of methods like DocLLM [55] and LayTextLLM [37] which utilize the interleaving of text and spatial layout information when evaluated in a task specific setting. Contrary to other similar instruction-tuned methods like LayoutLLM [39] and DocLayLLM [35] which utilize image patch tokens, pure text-only LLMs [37,55] rely on 2D positional tokens (cartesian coordinates) with aligned textual content. A recent work [66] highlighted the bottleneck of using cartesian system for layout modeling. They used the encoder-based LayoutLM family [24,59,60] to show some considerable improvement in model performance when using a Gaussian kernel with a handful of parameters to embed the relative spatial information with polar coordinates. Our work is the first to explore Gaussian attention biases for the decoupling of layout and text understanding for text-only LLMs to VrDU tasks.

3 Methodology

3.1 Method Overview

Using an off-the-shelf OCR system, we obtain the words and their absolute position in the document. The words are ordered in a top-to-bottom, left-to-right manner and then fed into a decoder-only transformer model. However, this ordering makes the words incoherent because lines are not necessarily grouped according to their logical structure in the document. As a result, related content is spread across the text span fed into the Transformer [54].

To provide layout cues that help contextualize the information, we incorporate layout biases during the decoding process to create representations influenced by the document's structure. Specifically, we use layout attention from Layout Attention with Gaussian Biases (LAGaBi [66]) to improve text under-

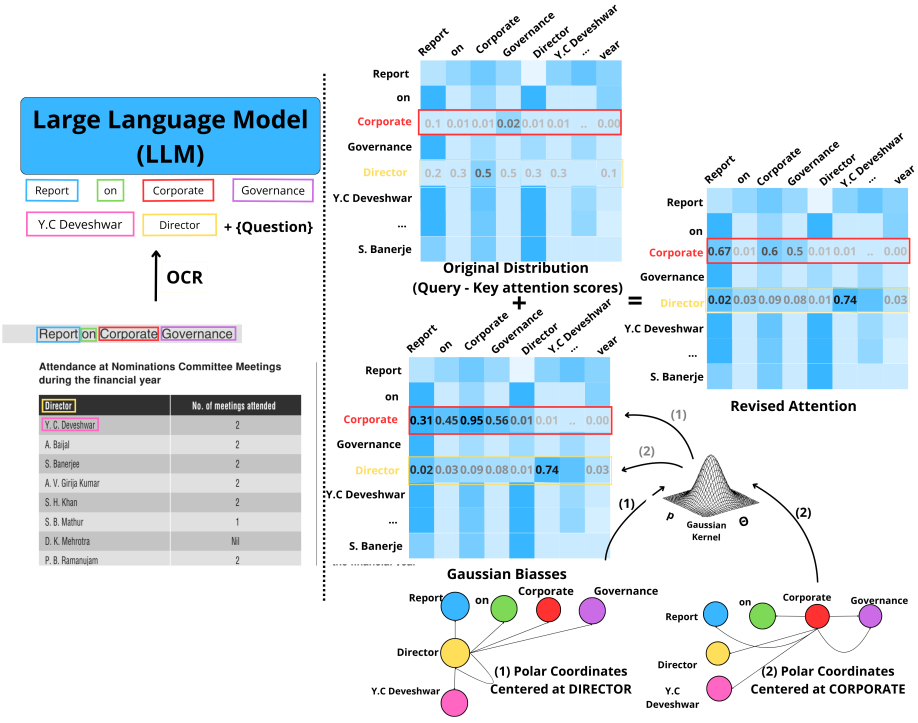

Fig. 1: **Overview of G-LLaMA's architecture and attention refinement.** The left side depicts the document processing pipeline, where an OCR system extracts text and bounding box data, which is then tokenized and input into a decoder-only LLM with a query. The right side shows how Gaussian biases refine attention, adjusting the original distribution (middle) using spatially-aware conditioning (right). The Gaussian kernel applies polar coordinate-based biases, enhancing layout-aware reasoning for DocVQA and KIE.

standing and enable more accurate question answering. The process is shown in Figure 1 and is as follows:

1. **Euclidean to Polar Conversion** – The coordinates for each word are transformed from Euclidean to polar coordinates. This efficiently represents the spatial relationships between words.
2. **Layout Attention with Gaussian Biases + Standard Attention** – Layout attention incorporates Gaussian biases along with standard attention to influence the decoding process. This helps predict the next token based on both textual information and layout cues. Specifically, the polar coordinates are converted into attention biases, which adjust the original semantic query-key attentions to better reflect the document's layout structure.

3.2 Spatial Relationships with Polar Coordinates

Layout follows a hierarchical structure, with words as one of the first semantic units. Modeling the relationships between words is key to merging individual layout details into higher-level abstractions. For example, in a table —even without dividing lines— words are arranged in a way that communicates organized information. Words that are close together tend to share stronger semantic links, while words that are further apart or on different lines have weaker associations. Previous works [33,45,59] have modeled relative distances using learnable attention biases. However, these methods still depend on learnable positional embeddings and only consider horizontal and vertical distances in a Cartesian system.

Polar coordinates can better capture inter-word spatial relationships by preserving both orientation and distance. For each query token, we set up a polar coordinate system centered at its position. The token's position acts as the reference point (pole), and the horizontal direction in the Cartesian system is used as the reference direction (polar axis), following the typical reading order of left-to-right and top-to-bottom. Formally, given a query token with its 2-D coordinates c_i as the pole, the polar coordinates $u_{ij} = (\rho_{ij}, \theta_{ij})$ of the jth key on the document page can be calculated as follows:

$$\rho_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \tag{1}$$

$$\theta_{ij} = \tan^{-1}\left(\frac{(y_j - y_i)}{(x_j - x_i)}\right) \tag{2}$$

where $\rho_{ij} \in [0, 1]$ and $\theta_{ij} \in [-\pi/2, \pi/2]$ denote the distance and angle (orientation) from the ith token to the jth token, respectively. The coordinates (x_i, y_i) represent the normalized position of the top-left point of the ith bounding box.

For instance, in Figure 1, the spatial relationship between the words "CORPORATE" and "REPORT" can be represented as a polar coordinate $(0.064, 0)$, indicating a distance of 0.064 and an angle of 0 degrees. Similarly, the relationship between the words "DIRECTOR" and "REPORT" is given by $(0.297, 1.432)$. When taking "REPORT" as the reference point, the spatial relationship between the words "CORPORATE" and "REPORT" remains $(0.064, 0)$, while the relationship between "DIRECTOR" and "REPORT" is $(0.297, 1.432)$.

3.3 Layout aware Decoding with Gaussian Biases

It has been shown that LLMs learn complex, non-linear representations of textual data by autoregressively predicting the next word over long sequences of tokens [47,12]. This process enables the models to capture detailed semantic information embedded in text. Effectively leveraging these semantic capabilities is essential for improving Document Intelligence tasks.

However, incorporating additional information such as layout cues can introduce risks: if these new representations are noisy or poorly aligned, they might

degrade the model's existing high-quality semantic knowledge. To avoid this issue, it's crucial to introduce appropriate inductive biases. These biases constrain the learning process, ensuring that layout information complements the existing learned representations rather than interfering with them.

Several approaches have explored how to fuse information from different modalities for document Intelligence, such as DocLLM [55] and LayoutLM [24]. In particular, we focus on the role of ALiBi [46] and T5 [48] due to their use of biases for enhancing model performance.

Inductive biases within attention mechanisms have previously been explored in approaches such as ALiBi and T5, primarily to influence attention based on positional or structural cues. Specifically, Layout Attention with Gaussian Biases (LAGaBi [66]) conditions attention scores on spatial relationships derived from the document's two-dimensional layout. `G-LLaMA` builds upon LaGABI by explicitly incorporating these spatial biases, effectively guiding attention during the decoding process to capture hierarchical semantic structures based on the document layout.

Specifically, the single-head self-attention mechanism can incorporate layout information by modifying the attention score as follows:

$$a_{ij} = \frac{\exp\left(q_i k_j^T / \sqrt{d_k} + \alpha\left(g(u_{ij}) - 1\right)\right)}{\sum_{j=1}^{N} \exp\left(q_i k_j^T / \sqrt{d_k} + \alpha\left(g(u_{ij}) - 1\right)\right)} \tag{3}$$

where q_i and k_j represent the query and key vectors at positions i and j, respectively, and d_k is the dimensionality of the attention head. The term $g(u_{ij})$ represents layout-based biases computed from a two-dimensional Gaussian function with learnable parameters. The function is applied to the polar coordinates u_{ij} that capture spatial relationships between words. This Gaussian function decreases with greater spatial distance, reducing the attention given to word pairs that are farther apart. Thus, by using $(g(u_{ij}) - 1)$, we penalize distant query-key pairs more strongly while minimally altering scores of closely positioned pairs.

The hyperparameter α balances semantic relevance and spatial layout influence, determining how strongly spatial positions affect the attention scores. Specifically, $g(u)$ can be formulated as:

$$g(u) = \exp\left(-\frac{1}{2}(u-\mu)^T \Sigma^{-1}(u-\mu)\right) \tag{4}$$

where μ and Σ are learnable parameters, representing the mean vector (2×1) and covariance matrix (2×2) of the Gaussian distribution, respectively. For computational efficiency and simplicity, the covariance matrix Σ is constrained to be diagonal, reducing the number of parameters per Gaussian kernel to four. The transformer architecture comprises multiple self-attention layers, each containing N_{heads} attention heads. For every head $i \in [1, N_{\text{heads}}]$, there is a dedicated learnable mean vector $\mu_i \in \mathbb{R}^2$ and diagonal covariance matrix $\Sigma_i \in \mathbb{R}^{2\times2}$. Importantly, these parameters are shared across all layers; thus, the i-th head across every layer uses the same μ_i and Σ_i. Consequently, the total number of Gaussian parameters in the model is $2 \times 2 \times N_{\text{heads}}$.

4 Experimental Validation

4.1 Datasets

We evaluate our method on two widely-used datasets: DocVQA [41] and FUNSD. DocVQA is a large-scale dataset specifically designed for Visual Question Answering on document images. It contains approximately 50,000 questions generated from over 12,000 diverse, real-world document images. Questions within DocVQA require varying levels of reasoning, including layout understanding, text comprehension, and inference based on document content.

The FUNSD [27] dataset is targeted at form understanding tasks. It comprises 199 scanned documents, fully annotated to include semantic labels such as questions, answers, headers, and miscellaneous information. These documents are carefully selected from the larger RVL-CDIP collection, which encompasses 400,000 grayscale images of various document types. FUNSD is divided into a training set of 149 documents and a test set of 50 documents.

In our evaluation, we follow existing methodologies by transforming the task of key information extraction into a natural language processing task, consistent with frameworks such as T5 [48].

4.2 Model Setup and Training Details

We have explored LAGaBi for decoder-only models using Llama 3.1-8B [16], which we named `G-LLaMA`. While conducting our experiments, we specifically refer to its base and instruct versions as $\texttt{G-LLaMA}_{\texttt{Base}}$ and $\texttt{G-LLaMA}_{\texttt{Instruct}}$, respectively. Llama 3.1-8B comprises 32 layers, each with 32 attention heads, and a hidden size of 4096. We use pre-trained weights as the backbone for the text modality. As described in 3.3, we can calculate the number of parameters added when using LAGaBi, which extends the Llama 3.1-8B models by adding a total of 128 parameters for learning the Gaussian distributions over the document layout, trained jointly with the LLM. The α termed used in equation (3) is tuned on the CORD validation set and fixed it to 4 in all experiments, as this gave the best balance between text and layout signals.

For training `G-LLaMA`, we use LoRA [23] with a low rank of 2 and an alpha of 4. We experimented with higher ranks and alphas, but the increased number of parameters made it challenging to prevent overfitting. Lower ranks worked best, resulting in a total of 5,243,008 trainable parameters, of which 128 parameters are specifically introduced by the Layout Attention with Gaussian Biases (LAGaBi [66]), with the remainder from the underlying large language model. We used a learning rate of 1×10^{-4} with a weight decay of 0.01, and the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and an Adam epsilon of 1×10^{-6}. Notably, the Gaussian layout parameters (LAGaBi) are trained separately with a higher learning rate compared to the LoRA parameters. This difference arises because the pretrained LLM already provides robust semantic representations, thus requiring only minimal adjustments. In contrast, Gaussian parameters necessitate

more substantial updates to accurately model spatial relationships specific to document layouts.

The model is trained on a single A40 GPU. For DocVQA, we ran for 8K iterations, while for FUNSD we ran for 1K iterations.

Metrics Following previous work [11,32,62,63], we evaluate DocVQA using Average Normalized Levenshtein Similarity (ANLS) [9]. We evaluate key information extraction (KIE) with the F1 score.

Table 1: Prompt templates used for instruction-tuning (spatial tokens not included).

Task	Template type	Prompt template	Expected response
VQA	Extraction	"{document} {question}"	answer annotation
KIE	Extraction	"What is the value for the "{key}"?"	answer annotation

4.3 Instruction Tuning

In line with recent advances in Vision-rich Document Understanding (VrDU) [51,62] and prior research in NLP [13,58], we perform instruction tuning on the `G-LLaMA` 8B model. We utilize task-specific instructions constructed from established Document AI datasets and corresponding templates.

We employ distinct prompts tailored to Visual Question Answering (VQA) and Key Information Extraction (KIE), both adhering to a unified prompt structure detailed in Table 1.

Visual Question Answering An illustrative prompt derived from the DocVQA dataset is as follows:

```
{document} What is the deadline for scientific abstract submission
for ACOG - 51st annual clinical meeting?
```

Key Information Extraction An example prompt taken from the FUNSD dataset is provided below:

```
{document} What is the value for the "Advertiser"?
```

For the KIE tasks in FUNSD, cases involving multiple values associated with a single key are formatted such that all corresponding values are concatenated using a special delimiter. An example is shown below:

```
{document} What is the value for the "inclusive date"? '01/01/94'
|| '01/01/95' || '01/01/96'
```

Table 2: Comparison with other OCR-based methods for DocVQA.

Method	#Params	Modalities	DocVQA
			ANLS% / CIDEr
$\text{Llama3.1}_{\text{Instruct}}$ [16]	8B	Text	71.5
$\text{Llama3.1}_{\text{base}}$ [16]	8B	Text	73.8
DocLLM [55]	7B	Text + Layout	69.5
$\text{LayTextLLM}_{\text{vqa}}$ [37]	7B	Text + Layout	75.6
Qwen-VL-7B [5]	7B	Text + Visual	65.1
TILT [45]	780M	Text + Visual + Layout	87.0
ScreenAI	5B	Text + Visual + Layout	89.9
$\texttt{G-LLaMA}_{\texttt{Instruct}}$ (Ours)	8B + 128	Text + Layout	74.6
$\texttt{G-LLaMA}_{\texttt{base}}$ (Ours)	8B + 128	Text + Layout	76.5

4.4 Results

Table 2 provides a comparison of various methods for integrating layout information into Transformer based models for DocVQA. $\texttt{G-LLaMA}_{\texttt{base}}$ demonstrates strong performance, achieving an average ANLS of 76.46 on the DocVQA test set. In contrast, DocLLM, which treats layout as a separate dimension and applies late fusion with textual features, reaches an ANLS of 69.5 after instruction tuning across multiple DocAI datasets including DocVQA. $\text{LayTextLLM}_{\text{vqa}}$, which embeds bounding box coordinates as additional tokens to integrate layout and textual information within an autoregressive framework, achieves 75.6 ANLS. Our approach is less intrusive, modifying only the attention mechanism of the LLM without introducing additional tokens.

Models incorporating both layout, text and vision achieve stronger results, yet invalidate the comparison, since we only rely on text and layout. While this methods rely on large vision encoders and extensive text–vision data, our work instead explores how to add layout bias directly to the text embeddings of a decoder LLM. We operate in the pretrained LLM embedding space and learn that bias with just 128 extra parameters—no heavy vision encoders needed. This is not a strict one-to-one comparison, since this models use visual information, but it highlights a lighter, more data-efficient way to handle document layouts.

Notably, incorporating LAGaBi, both in $\texttt{G-LLaMA}_{\texttt{Instruct}}$ and $\texttt{G-LLaMA}_{\texttt{base}}$ results in ANLS improvements of 3.1 and 2.7 points, respectively, compared to their counterparts without LAGaBi. These results highlight the effectiveness of modeling inductive biases for document understanding without relying on additional tokens.

Table 3 provides a detailed comparison of methods for key information extraction on the FUNSD dataset, highlighting the importance of integrating layout information in large language models. Purely textual methods, such as $\texttt{Llama3.1}_{\texttt{base}}$, yield an F-score of 29.15%. Instruction tuning provides a moderate boost, as seen with $\texttt{Llama3.1}_{\texttt{Instruct}}$ (33.87%), yet it alone does not sufficiently enhance structured extraction without explicit spatial context.

Table 3: Comparison with other OCR-based methods for FUNSD

Method	**#Params**	**Modalities**	**FUNSD**
			F-score%
$\text{Llama3.1}_{\text{base}}$	8B	Text	29.15
$\text{Llama3.1}_{\text{Instruct}}$	8B	Text	33.87
DocLLM [55]	7B	Text + Layout	51.8
$\text{LayTextLLM}_{\text{vqa}}$ [37]	7B	Text + Layout	52.6
$\texttt{G-LLaMA}_{\texttt{base}}$ (Ours)	8B + 128	Text + Layout	31.20
$\texttt{G-LLaMA}_{\texttt{Instruct}}$ (Ours)	8B + 128	Text + Layout	48.79

Our $\texttt{G-LLaMA}_{\texttt{base}}$ incorporates Layout-Aware Gaussian Biases (LAGaBi) and achieves an F-score of 31.20%, slightly outperforming the text-only baseline. The instruct variant without layout enhancements, $\texttt{LLaMA3.1}_{\texttt{Instruct}}$, offers a minor improvement over the base model. However, when combining LAGaBi with instruction tuning in $\texttt{G-LLaMA}_{\texttt{Instruct}}$, the F-score significantly increases to 48.79%. We attribute this substantial improvement to instruction-tuned models ability to learn more efficiently [18], granting an advantage on the smaller FUNSD dataset. Additionally, FUNSD's compact, instruction-oriented format aligns well with the tuning data, further benefiting the instruct model.

This substantial improvement does not occur in DocVQA (Table 2). DocVQA larger size and abundant examples enable the base model to catch up, reducing the advantage of the instruct variant.

A direct comparison with established methods further emphasizes our approach. DocLLM [55] obtains an F-score of 51.8% and $\text{LayTextLLM}_{\text{vqa}}$ [37] reaches 52.6%.

DocLLM benefits from direct fine-tuning on FUNSD, and $\text{LayTextLLM}_{\text{vqa}}$, although not trained on FUNSD, leverages spatial pretraining. In contrast, $\texttt{G-LLaMA}_{\texttt{Instruct}}$ achieves comparable performance by incorporating Layout Attention with Gaussian Biases.

These findings confirm that explicitly encoding spatial layout is crucial for tasks involving visually structured documents. The integration of Gaussian layout biases improves the model's spatial awareness without major architectural changes, demonstrating the practical benefits of the proposed approach.

Table 4: Effect of LAGaBi on model performance compared to the baseline.

	Figure / Diagram	Form	Table/ List	Layout	Free Text	Image/ Photo	Hand-written	Yes/ No	Others	Score
$\text{Llama3.1-8B}_{\text{base}}$	0.45	0.81	0.70	0.78	0.78	0.43	0.59	0.55	0.66	0.74
$\text{G-LLaMA}_{\text{base}}$	0.47	0.84	0.75	0.79	0.78	0.55	0.63	0.66	0.67	0.76

Table 5: Few-shot comparison of OCR-based methods on FUNSD using checkpoints obtained after fine-tuning on DocVQA.

Dataset: FUNSD	**Modalities**	**0-shot**	**2-shot**
		F-score (%)	
$\text{Llama3.1}_{\text{base-vqa}}$	Text	1.55	0.24
$\text{Llama3.1}_{\text{Instruct-vqa}}$	Text	9.77	3.45
$\texttt{G-LLaMA}_{\texttt{base-vqa}}$	Text + Layout	2.50	0.12
$\texttt{G-LLaMA}_{\texttt{Instruct-vqa}}$	Text + Layout	7.15	2.86

Improved Layout intensive questions: Table 4 illustrates the effect of LAGaBi based on the question-answer types in DocVQA [41]. The integration of LAGaBi, which employs Gaussian biases to encode spatial layout information into the attention mechanism of the LLM, consistently improves results across almost all categories. Particularly notable improvements occur in categories strongly tied to spatial understanding, such as Image/Photo, where the score significantly increases from 0.43 to 0.55, and Table/List, with a marked improvement from 0.70 to 0.75. These question types inherently depend on recognizing and interpreting document layout, highlighting the effectiveness of LAGaBi's approach to incorporating layout cues directly into the language model's attention computation.

Furthermore, LAGaBi provides moderate yet consistent improvements in other challenging question categories, including Handwritten (from 0.59 to 0.63) and Yes/No questions (from 0.55 to 0.66), demonstrating the broader applicability of spatial biases even beyond purely layout-dependent tasks. Overall, the results indicate a clear correlation between the inclusion of Gaussian layout biases and improved performance, underscoring the importance of explicitly modeling spatial relationships within document understanding tasks.

Table 5 shows few-shot evaluation results on the FUNSD dataset using checkpoints fine-tuned on DocVQA. The instruction-tuned model consistently achieves better results in both 0-shot and 2-shot scenarios compared to the base model, indicating stronger generalization ability. Specifically, the instruction-tuned model without LAGaBi ($\text{Llama3.1}_{\text{Instruct-vqa}}$) reaches an F-score of 9.77% in 0-shot and 3.45% in 2-shot scenarios, clearly surpassing the base model (1.55% and 0.24%, respectively). However, when layout-aware Gaussian biases (LAGaBi) are introduced, the instruction-tuned variant ($\texttt{G-LLaMA}_{\texttt{Instruct-vqa}}$) experiences a slight performance drop (7.15%) compared to $\text{Llama3.1}_{\text{Instruct-vqa}}$ (9.77%) in the 0-shot case. Despite this reduction, the instruct variant significantly outperforms the base model with LAGaBi in both few-shot settings.

These findings suggest that instruction tuning facilitates rapid adaptation to new data distributions. Although the integration of LAGaBi slightly penalizes zero-shot transfer for the instruct variant, the combined benefits of instruction tuning and layout attention still substantially outperform the base model under

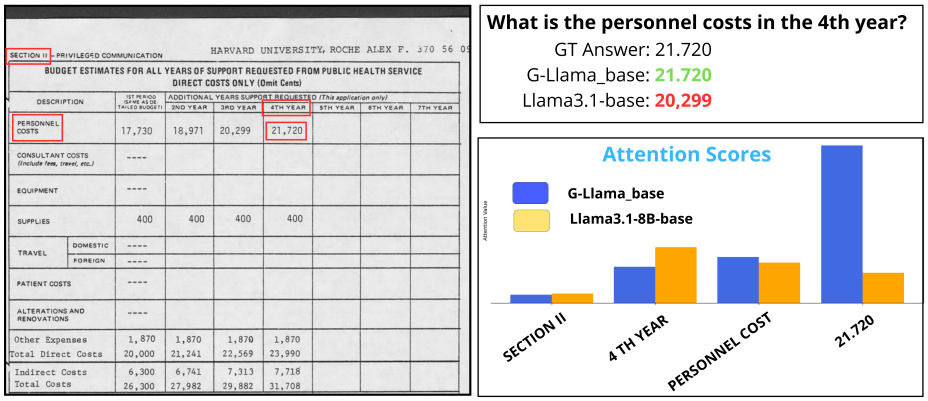

Fig. 2: Comparison of attention scores with and without LAGaBi for spatially relevant tokens when answering the question: "What are the personnel costs in the 4th year?"

minimal supervision. This emphasizes the advantage of instruction-tuned models enhanced with spatial biases for effective generalization in visually structured tasks.

Qualitative Evaluation: Improved Layout Understanding in Attention Scores Figure 2 shows the attention weights for the final token before the model predicts its answer. In a decoder-only setup, these attention weights guide how the model selects the next tokens. With `LAGaBi`, the model identifies the document's structure more precisely, focusing attention on the relevant token that leads to the correct answer.

For example, when asked "What is the personnel costs in the 4th year?", the model must locate the "personnel costs" row and align it with the column for the "4th year". While both models highlight these key parts, `LAGaBi` combines them more effectively, improving accuracy. Without LAGaBi, the model fails to use layout-specific attention and mistakenly spreads attention to unrelated text, missing the right answer.

5 Conclusion and Future Work

In conclusion, we demonstrate the effective application of Layout Attention with Gaussian Biases (LAGaBi) to decoder-only language models. By incorporating LAGaBi into the attention mechanism of existing models, our approach enhances layout understanding for document understanding without requiring major architectural modifications. Experiments on DocVQA and FUNSD show that applying layout-aware inductive biases significantly improves document understanding for autoregressive large language models, particularly in tasks that

rely on document structure. Furthermore, when combined with instruction tuning, our method substantially improves the extraction of key information even when training data is limited. This work highlights the use of gaussian biases for layout understanding to boost the performance of decoder-based LLMs in real-world applications where computational constraints limit the feasibility of fully multimodal models.

Future Scopes and Challenges While our application of LAGaBi to decoder-only language models demonstrates promising improvements, several challenges and limitations remain. First, the approach relies on accurate OCR outputs, and any errors in text extraction may propagate to downstream tasks. Additionally, although Gaussian biases effectively enhance spatial reasoning, they may not fully capture non-textual visual elements such as table borders, shading, or font variations, which often provide additional cues in document layouts. Adapting this approach to domain-specific layouts—where structural variations across different document types are significant—may require further fine-tuning. Moreover, although our method is computationally efficient, it might still fall short of state-of-the-art multimodal approaches that combine both text and vision for richer document representations.

Future research could explore more advanced layout representations that integrate additional structural features beyond simple bounding boxes, such as region-based spatial embeddings or hierarchical document segmentation. Another promising direction is extending this approach to multilingual settings, where variations in text flow and reading orders across languages demand more adaptable spatial reasoning mechanisms. In addition, investigating hybrid approaches that integrate lightweight vision modules—such as line detection or structural tokenization—could further enhance the model's ability to process complex layouts without a significant increase in computational cost.

Acknowledgements

The authors acknowledge the financial support of the Department of Research and Universities of the Generalitat of Catalonia to the DocAI Research Group: Group on Document Intelligence (2021 SGR 01559), Grant PID2021-126808OB-I00 (GRAIL) and Grant CNS2022-135947 (DOLORES) funded by MCIN/AEI/10.13039/501100011033, and by ERDF/EU and Ph.D. Scholarship from AGAUR (2023 FI-3-00223). The Computer Vision Center is part of the CERCA Program/Generalitat de Catalunya.

References

1. Abramovich, O., Nayman, N., Fogel, S., Lavi, I., Litman, R., Tsiper, S., Tichauer, R., Appalaraju, S., Mazor, S., Manmatha, R.: Visfocus: Prompt-guided vision encoders for ocr-free dense document understanding. In: European Conference on Computer Vision. pp. 241–259. Springer (2024) 2

2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 2, 3, 4
3. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: End-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 993–1003 (2021) 3
4. Appalaraju, S., Tang, P., Dong, Q., Sankaran, N., Zhou, Y., Manmatha, R.: Docformerv2: Local features for document understanding. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 709–718 (2024) 3
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 **1**(2), 3 (2023) 10
6. Banerjee, A., Biswas, S., Lladós, J., Pal, U.: Swindocsegmenter: an end-to-end unified domain adaptive transformer for document instance segmentation. In: International Conference on Document Analysis and Recognition. pp. 307–325. Springer (2023) 3
7. Biescas, N., Boned, C., Lladós, J., Biswas, S.: Geocontrastnet: Contrastive key-value edge learning for language-agnostic document understanding. In: International Conference on Document Analysis and Recognition. pp. 294–310. Springer (2024) 3
8. Biswas, S., Riba, P., Lladós, J., Pal, U.: Beyond document object detection: instance-level segmentation of complex layouts. International Journal on Document Analysis and Recognition (IJDAR) **24**(3), 269–281 (2021) 3
9. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4291–4301 (2019) 9
10. Borchmann, Ł.: Notes on applicability of gpt-4 to document understanding. arXiv preprint arXiv:2405.18433 (2024) 4
11. Borchmann, Ł., Pietruszka, M., Stanislawek, T., Jurkiewicz, D., Turski, M., Szyndler, K., Graliński, F.: Due: End-to-end document understanding benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021) 9
12. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 2, 3, 6
13. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research **25**(70), 1–53 (2024) 9
14. Garncarek, Ł., Powalski, R., Stanisławek, T., Topolski, B., Halama, P., Turski, M., Graliński, F.: Lambert: layout-aware language modeling for information extraction. In: International Conference on Document Analysis and Recognition. pp. 532–547. Springer (2021) 3
15. Gemelli, A., Biswas, S., Civitelli, E., Lladós, J., Marinai, S.: Doc2graph: a task agnostic document understanding framework based on graph neural networks. In: European Conference on Computer Vision. pp. 329–344. Springer (2022) 3
16. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 2, 3, 4, 8, 10

17. Gu, J., Kuen, J., Morariu, V.I., Zhao, H., Jain, R., Barmpalios, N., Nenkova, A., Sun, T.: Unidoc: Unified pretraining framework for document understanding. Advances in Neural Information Processing Systems **34**, 39–50 (2021) 3
18. Gupta, H., Sawant, S.A., Mishra, S., Nakamura, M., Mitra, A., Mashetty, S., Baral, C.: Instruction tuned models are quick learners. arXiv preprint arXiv:2306.05539 (2023) 11
19. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: International Conference on Document Analysis and Recognition (ICDAR) (2015) 3
20. Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., Park, S.: Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 10767–10775 (2022) 3
21. Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Li, C., Zhang, J., Jin, Q., Huang, F., et al.: mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. arXiv preprint arXiv:2403.12895 (2024) 2, 4
22. Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. arXiv preprint arXiv:2409.03420 (2024) 4
23. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022) 8
24. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4083–4091 (2022) 2, 3, 4, 7
25. Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.: Icdar2019 competition on scanned receipt ocr and information extraction. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1516–1520. IEEE (2019) 2, 3
26. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 1–6. IEEE (2019) 2, 3
27. Jaume, G., Ekenel, H.K., Thiran, J.P.: Funsd: A dataset for form understanding in noisy scanned documents. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 1–6. IEEE (2019) 8
28. Kim, G., Hong, T., Yim, M., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Donut: Document understanding transformer without ocr. arXiv preprint arXiv:2111.15664 **7**(15), 2 (2021) 3
29. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International conference on machine learning. pp. 5583–5594. PMLR (2021) 4
30. Laurençon, H., Marafioti, A., Sanh, V., Tronchon, L.: Building and better understanding vision-language models: insights and future directions. In: Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models (2024) 2, 4
31. Laurençon, H., Tronchon, L., Cord, M., Sanh, V.: What matters when building vision-language models? Advances in Neural Information Processing Systems **37**, 87874–87907 (2024) 4
32. Lee, K., Joshi, M., Turc, I.R., Hu, H., Liu, F., Eisenschlos, J.M., Khandelwal, U., Shaw, P., Chang, M.W., Toutanova, K.: Pix2struct: Screenshot parsing as pretrain-

ing for visual language understanding. In: International Conference on Machine Learning. pp. 18893–18912. PMLR (2023) 3, 9
33. Li, C., Bi, B., Yan, M., Wang, W., Huang, S., Huang, F., Si, L.: Structurallm: Structural pre-training for form understanding. arXiv preprint arXiv:2105.11210 (2021) 6
34. Li, P., Gu, J., Kuen, J., Morariu, V.I., Zhao, H., Jain, R., Manjunatha, V., Liu, H.: Selfdoc: Self-supervised document representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5652–5660 (2021) 3
35. Liao, W., Wang, J., Li, H., Wang, C., Huang, J., Jin, L.: Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. arXiv preprint arXiv:2408.15045 (2024) 2, 4
36. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019) 2
37. Lu, J., Yu, H., Wang, Y., Ye, Y., Tang, J., Yang, Z., Wu, B., Liu, Q., Feng, H., Wang, H., et al.: A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. arXiv preprint arXiv:2407.01976 (2024) 2, 4, 10, 11
38. Luo, C., Cheng, C., Zheng, Q., Yao, C.: Geolayoutlm: Geometric pre-training for visual information extraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7092–7101 (2023) 3
39. Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., Yao, C.: Layoutllm: Layout instruction tuning with large language models for document understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15630–15640 (2024) 2, 4
40. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022) 3
41. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021) 2, 3, 8, 12
42. Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., Lee, H.: Cord: a consolidated receipt dataset for post-ocr parsing. In: Workshop on Document Intelligence at NeurIPS 2019 (2019) 3
43. Perot, V., Kang, K., Luisier, F., Su, G., Sun, X., Boppana, R.S., Wang, Z., Wang, Z., Mu, J., Zhang, H., et al.: Lmdx: Language model-based document information extraction and localization. arXiv preprint arXiv:2309.10952 (2023) 2
44. Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.: Doclaynet: A large human-annotated dataset for document-layout segmentation. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. pp. 3743–3751 (2022) 3
45. Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Pałka, G.: Going full-tilt boogie on document understanding with text-image-layout transformer. In: International Conference on Document Analysis and Recognition. pp. 732–747. Springer (2021) 3, 6, 10
46. Press, O., Smith, N.A., Lewis, M.: Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409 (2021) 7
47. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) 6

48. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020) 7, 8
49. Rodriguez, J., Jian, X., Panigrahi, S.S., Zhang, T., Feizi, A., Puri, A., Kalkunte, A., Savard, F., Masry, A., Nayak, S., et al.: Bigdocs: An open and permissively-licensed dataset for training multimodal models on document and code tasks. arXiv preprint arXiv:2412.04626 (2024) 2, 4
50. Tanaka, R., Iki, T., Nishida, K., Saito, K., Suzuki, J.: Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 19071–19079 (2024) 4
51. Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., Bansal, M.: Unifying vision, text, and layout for universal document processing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19254–19264 (2023) 9
52. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) 2, 3, 4
53. Van Landeghem, J., Maity, S., Banerjee, A., Blaschko, M., Moens, M.F., Lladós, J., Biswas, S.: Distildoc: Knowledge distillation for visually-rich document applications. In: International Conference on Document Analysis and Recognition. pp. 195–217. Springer (2024) 3
54. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 4
55. Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., Pei, Y., Nourbakhsh, A., Liu, X.: Docllm: A layout-aware generative language model for multimodal document understanding. arXiv preprint arXiv:2401.00908 (2023) 2, 4, 7, 10, 11
56. Wang, J., Jin, L., Ding, K.: Lilt: A simple yet effective language-independent layout transformer for structured document understanding. arXiv preprint arXiv:2202.13669 (2022) 3
57. Wang, J., Lin, Z., Huang, D., Xiong, L., Jin, L.: Liltv2: Language-substitutable layout-image transformer for visual information extraction. ACM Transactions on Multimedia Computing, Communications and Applications (2024) 3
58. Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021) 9
59. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740 (2020) 2, 3, 4, 6
60. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192–1200 (2020) 2, 3, 4
61. Xu, Y., Lv, T., Cui, L., Wang, G., Lu, Y., Florencio, D., Zhang, C., Wei, F.: Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. arXiv preprint arXiv:2104.08836 (2021) 3
62. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., et al.: mplug-docowl: Modularized multimodal large language model for document understanding. arXiv preprint arXiv:2307.02499 (2023) 2, 9

63. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., Li, C., Tian, J., Qian, Q., Zhang, J., et al.: Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. arXiv preprint arXiv:2310.05126 (2023) 2, 4, 9
64. Zhang, C., Tu, Y., Zhao, Y., Yuan, C., Chen, H., Zhang, Y., Chai, M., Guo, Y., Zhu, H., Zhang, Q., et al.: Modeling layout reading order as ordering relations for visually-rich document understanding. arXiv preprint arXiv:2409.19672 (2024) 2
65. Zhong, X., Tang, J., Yepes, A.J.: Publaynet: largest dataset ever for document layout analysis. In: 2019 International conference on document analysis and recognition (ICDAR). pp. 1015–1022. IEEE (2019) 3
66. Zhu, X., Han, X., Peng, S., Lei, S., Deng, C., Feng, J.: Beyond layout embedding: layout attention with gaussian biases for structured document understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 7773–7784 (2023) 2, 4, 7, 8