ReLayout: Towards Real-World Document Understanding via Layout-enhanced Pre-training

Zhouqiang Jiang[1], **Bowen Wang**[2*], **Junhao Chen**[1], **Yuta Nakashima**[2]

Osaka University, Japan

[1]{zhouqiang, junhao}@is.ids.osaka-u.ac.jp
[2]{wang, n-yuta}@ids.osaka-u.ac.jp

Abstract

Recent approaches for visually-rich document understanding (VrDU) uses manually annotated semantic groups, where a semantic group encompasses all semantically relevant but not obviously grouped words. As OCR tools are unable to automatically identify such grouping, we argue that current VrDU approaches are unrealistic. We thus introduce a new variant of the VrDU task, real-world visually-rich document understanding (ReVrDU), that does not allow for using manually annotated semantic groups. We also propose a new method, ReLayout, compliant with the ReVrDU scenario, which learns to capture semantic grouping through arranging words and bringing the representations of words that belong to the potential same semantic group closer together. Our experimental results demonstrate the performance of existing methods is deteriorated with the ReVrDU task, while ReLayout shows superiour performance.

1 Introduction

Modern visually-rich document understanding (VrDU), which aims at automating information extraction from visually-rich documents, has become an important research direction (Liu et al., 2019a; Jaume et al., 2019; Xu et al., 2020b; Garncarek et al., 2021; Gu et al., 2022; Tu et al., 2023). Visually-rich documents, such as invoices, receipts, reports, and academic papers, not only contain a substantial volume of text data but also encode essential semantics needed for understanding them into their structures or *layout*. Figure 1(top) shows an example of such a document. People can perhaps group up "Case type" and "Plaintiff's counsel" into respective semantic groups, and also associate "Case type" and "Asbestos" as well as "Plaintiff's counsel" and the name and address on its right, even though they are spatially apart from each other, because of the knowledge on the layout (e.g., a field label and its content are typically placed next to each other). Such structural comprehension of documents provides strong prior on the semantics that respective layout groups have, enhancing the accuracy of information extraction.

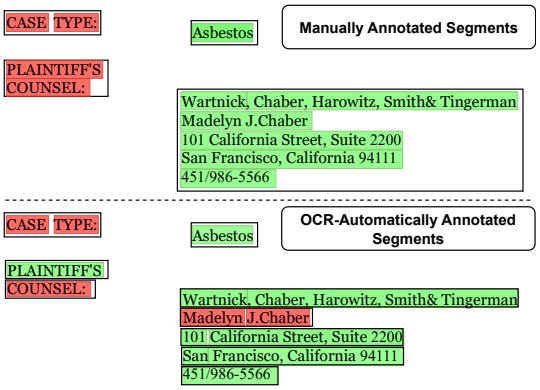


Figure 1: Result of LayoutMASK (Tu et al., 2023) for Semantic Entity Classification using different segments. :Segment, :**QUESTION**, :**ANSWER**.

Early work (Hwang et al., 2019; Denk and Reisswig, 2019) makes use of the empirical knowledge about layout that text flows from left to right and top to bottom. Such flow can be represented by global 1D positions associated with all words in OCR text. The text flow can be more precise when the document structure is obtained from auxiliary sources, such as XMLs' or PDFs' metadata (Wang et al., 2021). Recent studies have explored the interaction between OCR text and its layout in document images through pre-training model (Xu et al., 2020b; Li et al., 2021a; Hong et al., 2022; Tu et al., 2023). LayoutLM (Xu et al., 2020b) is the first to introduce word-wise 2D bounding boxes as layout embedding in the pre-training model. Similarly to BERT, LayoutLM masks words extracted by OCR but retains the corresponding layout embeddings, requiring the model to reconstruct the original words. The representations learned in this

*Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 3778–3793
January 19–24, 2025.

way exhibit excellent fine-tuning performance.

Subsequent pre-training approaches have shown that by incorporating the semantically relevant text blocks (semantic groups) and using a common segment (segment-wise) 2D bounding box as a layout embedding (Li et al., 2021a), richer semantic concepts can be provided, as illustrated in Figure 1(top), thereby significantly enhancing performance. LayoutMASK(Tu et al., 2023) further uses segment-wise 2D bounding boxes to change the global 1D position to an loacl 1D position within segments, enhancing the local information of text flow and further improving performance.

In the context of document comprehension, use of the structural information in documents faces an inherent challenge: *semantic groups can facilitate automated document understanding, while they are manually annotated, yet we aim to achieve automated document understanding.*, which forms a paradox that the costly annotation of semantic groups and automated document understanding cannot coexist. Prior works (Li et al., 2021a; Huang et al., 2022; Tu et al., 2023) avoid this problem using human-annotated semantic groups during fine-tuning (Figure 1); however, semantic groups as accurate as human annotations are not available in real-world scenarios. Off-the-shelf OCR can group some spatially consecutive words (referred to as a *text segment*) in a document as in Figure 1(bottom), but they do not necessarily align with the actual semantics that the document encompasses.

To nourish exploration toward real-world VrDU, we propose a new VrDU task, coined *ReVrDU* (**Re**al-world **VrDU**), on top of existing ones, which only allows for using information available from off-the-shelf OCR tools for both pre-training and fine-tuning, i.e., words, global 1D positions, word-wise 2D bounding boxes, and text segments, so that VrDU can be evaluated in alignment with real-world scenarios.

We also propose a new pre-training model for ReVrDU, referred to as *ReLayout* (**Re**al-world **Layout**-enhanced pre-training), use simple global 1D positions and word-wise 2D bounding boxes as layout input. In addition to the masked language modeling (MLM) strategy, ReLayout adopts 1D Local Order Prediction (1-LOP) and 2D Text Segment Clustering (2-TSC) strategies. The former reconstructs word order within each text segment. Through this task, a model learns local information about text flow as well as relationships cross text segments as it needs to predict where a text segments starts and ends in the masked global 1D positions. With the latter, a model learns to complete potential semantic groups information from text segments in a self-supervised manner. We experimentally show that pre-training a model with ReLayout demonstrates excellent downstream performance in both ideal and real-world scenarios.

2 Related Work

2.1 Multimodal Pre-training

Multimodal self-supervised pre-training models (Hong et al., 2022; Wang et al., 2022; Xu et al., 2020b,a; Powalski et al., 2021; Jiang et al., 2023; Appalaraju et al., 2021; Xu et al., 2021; Li et al., 2021c; Lee et al., 2022; Huang et al., 2022; Peng et al., 2022), due to their successful application across document layout, text, and visual modalities, have propelled rapid advancements in the field of VrDU. LayoutLM (Xu et al., 2020b) first introduces each token's 2D bounding box as layout embedding to enhance MLM. On top of LayoutLM, BROS (Hong et al., 2022) proposes a more challenging MLM task that masks larger regions. StructurelLM (Li et al., 2021a) pre-trains a model by predicting positions of equally-sized regions in a document. These pre-training tasks jointly model the relationships between text and the layout in documents.

Given the richness of visual cues in image attributes such as fonts, colors, logos, and dividing lines in tables, many works incorporate the visual modality pre-training tasks (Gu et al., 2021; Li et al., 2021b; Xu et al., 2020a; Huang et al., 2022; Gu et al., 2023; Wang et al., 2023), including masked visual-language modeling (Xu et al., 2020a), masked image modeling (Huang et al., 2022), word-patch alignment (Huang et al., 2022), text-image alignment (Xu et al., 2020a), text-image match (Xu et al., 2020a), and visual contrastive learning (Gu et al., 2023). These tasks exploit the knowledge in visual components, providing additional performance boosts in the VrDU tasks.

ReLayout, however, exclusively uses text and layout modalities to evaluate the performance of models in ideal and real-world scenarios.

2.2 Layout Information

How to handle layout information is crucial for VrDU. LayoutLM first introduces spatial layout information into VrDU using word-level 2D bounding boxes. BROS proposes to encode relative

spatial relationships of word-wise 2D bounding boxes with a BERT-based model. StructureLM utilizes segment-wise 2D bounding boxes rather than word-wise to represent layout, showing promising performance improvements. LayoutMASK (Tu et al., 2023) introduces a strong prior about layout through local 1D positions and segment-wise 2D bounding boxes. It uses segment-wise 2D bounding boxes to identify semantic groups and local 1D positions to guide the model in scanning tokens in the correct order. This method achieved SOTA performance on downstream tasks with manually annotated semantic groups, but obtaining semantic groups in real-world scenarios is impractical.

ReLayout does not use complete semantic groups, which is practically unavailable, but supplies text segments that are automatically obtained with common OCR tools and do not necessarily align with actual semantic groups during pre-training. Pre-training strategy in ReLayout is designed to handle such text segments by learning to merge semantically similar text segments.

3 Task Definition: ReVrDU

Our ReVrDU task is built on top of existing VrDU tasks that supply human-annotated semantic groups as input (Jaume et al., 2019; Park et al., 2019).

Formally, traditional VrDU tasks typically provide a set $\mathcal{W} = \{w_l\}_{l=1}^{L}$ of words, a set $\mathcal{O} = \{o_l\}_{l=1}^{L}$ of the corresponding global 1D positions $o_l \in \mathbb{Z}_{\geq 0}$, where $\mathbb{Z}_{\geq 0}$ is the set of non-negative integers, semantic groups $\mathcal{S}^{true} = \{s_k^{true}\}_{k=1}^{K}$, where s_k^{true} is the k-th semantic group that contains all words in the group (i.e., $s_k^{true} \subset \mathcal{W}$), and a set $\mathcal{B} = \{b_l\}_{l=1}^{L}$ of word-wise bounding boxes. Word w_l is associated with word-wise bounding box $b_l \in \mathbb{R}^4$, represented by its top-left and bottom-right corners' positions. The semantic groups serves as a strong cue for understanding the semantics.

ReVrDU provides data in the same format, but to align with real-world scenarios, all data come from OCR results, For example, there may be missing or incorrect words and bounding boxes, and a semantic group is replaced with just a set of consecutive words in a line (i.e., a text segment), which can be obtained from OCR tools. For more details, see Appendix A.2. We denote a set of text segments by $\mathcal{S} = \{s_k\}_{k=1}^{K}$, where $s_k \subset \mathcal{W}$ is the k-th text segment.

4 Method: ReLayout

The unavailability of accurate semantic grouping in ReVrDU hinders the naive application of existing methods tailored for VrDU. Our ReLayout, a layout-enhanced multimodal pre-training model, effectively incorporates structural information in documents by pre-training a model to strengthen the understanding of local layout structures and relationships and learn semantic grouping via our proposed 1-LOP and 2-TSC strategies.

As shown in Figure 2, we use a vanilla Transformer encoder (Vaswani et al., 2017) architecture as the backbone of our model.

4.1 Tokenizers

We use byte-pair encoding (Sennrich et al., 2015) to tokenize $\mathcal{W}$ into a set $\mathcal{T} = \{t_n\}_{n=1}^{N}$ of tokens t_n. We also reassign the global 1D positions according to $\mathcal{O}$ and the tokenization result. We denote the reassigned set of the global 1D positions as $\mathcal{O}' = \{o'_n\}_{n=1}^{N}$, where $o'_n \in \mathbb{Z}_{\geq 0}$. We also remap the bounding boxes $\{b_l\}$ to $\{b'_l\}$ and the text segments so that tokens derived from the same word have the same bounding box and are included in the same text segment, as shown in Figure 2.

4.2 Embeddings

For tokens, we use a token embedding layer, denoted by $e_n^{\mathrm{t}} = \mathrm{TE}(t_n)$. The global 1D positions, represented by non-negative integers, are encoded with 1D position embedding layer, denoted by $e_n^{\mathrm{o}} = \mathrm{PE1D}(o'_n)$. The bounding box b'_n is represented by 2D position embedding $e_n^{\mathrm{b}} = \mathrm{PE2D}(b'_n)$ The n-th input embedding e_n to the backbone network is the sum of these embeddings, i.e.,

$$e_n = e_n^{\mathrm{t}} + e_n^{\mathrm{o}} + e_n^{\mathrm{b}}. \quad (1)$$

4.3 Pre-training Tasks

We employ three pre-training tasks, i.e., MLM, 1-LOP, and 2-TSC pre-training tasks.

4.3.1 Masked Language Modeling

MLM is utilized to enable the model to learn multimodal representations of text-layout interactions by combining text and layout cues. We randomly mask tokens at the word level with given probability P_{MLM}, where tokens to be masked are replaced with `[mask]` token.

The all embeddings e are then fed into the model, and the output representations pass through a non-linear MLM Head layer, obtaining logits for each

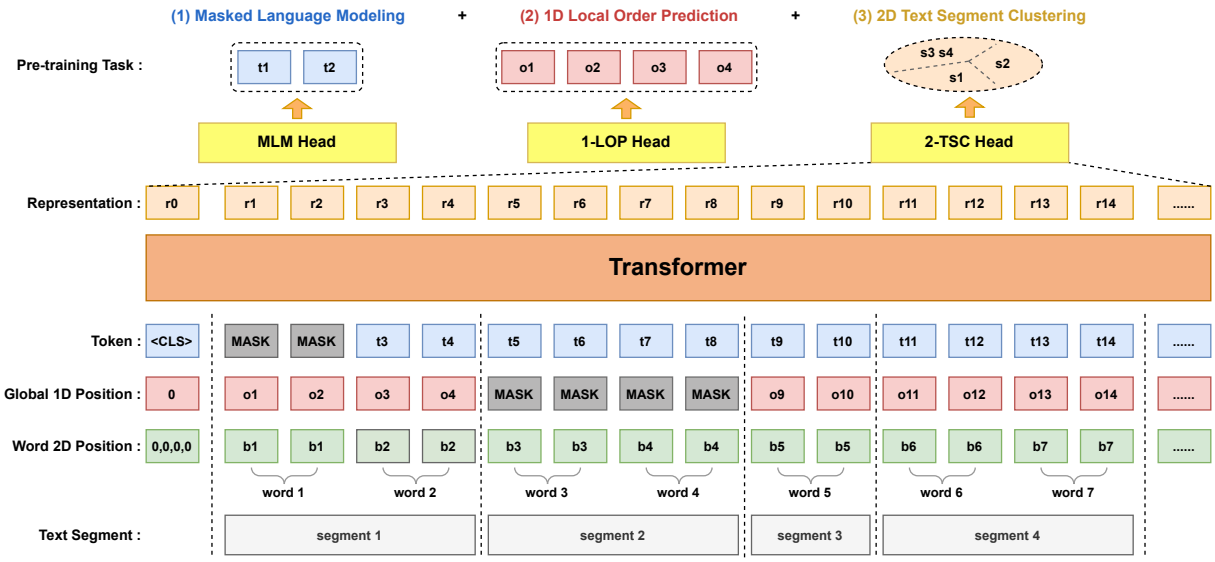


Figure 2: Architecture of ReLayout: MLM masks word-level tokens and reconstructs the original tokens. 1-LOP masks global 1D positions at the text segment and reconstructs local 1D positions. 2-TSC uses self-supervised techniques to adaptively cluster the representations of text segments that belong to the same semantic group.

masked token. With softmax, we compute the reconstructed probability p_j of the j-th masked tokens over the vocabulary ($j = 1, \ldots, J$). The loss function is thus defined as:

$$L_{\text{MLM}} = -\frac{1}{J}\sum_{j=1}^{J}\log p_j \tag{2}$$

4.3.2 1D Local Order Prediction

Local 1D positions in each semantic group give a model some ideas about how the words are arranged in the document. As semantic grouping is not available in ReVrDU, we instead predict them within each text segment through the 1-LOP pre-training task. This design choice is not optimal as text segments do not necessarily correspond to semantic groups; we still consider that they can serve as a good proxy of semantic groups for learning local structure.[1] By learning when to increment the position value and reset it to 1, our model grasps both within- and cross-segment local structures.

As shown in Figure 2, we randomly select some text segments with probability $P_{\text{1-LOP}}$ and mask all global 1D positions within all selected text segments. The output representations from the model then go through a non-linear 1-LOP head to predict the local 1D positions in the text segments. Letting M be the number of masked tokens in total and q_m the probability of the correct local position, the loss is defined as:

$$L_{\text{1-LOP}} = -\frac{1}{M}\sum_{m=1}^{M}\log q_m. \tag{3}$$

4.3.3 2D Text Segment Clustering

The 1-LOP pre-training task uses less accurate text segments provided by an OCR tool. A model pre-trained with such text segments may not be fully consistent with actual semantic groups, resulting in fragmented representations, as shown in Figure 1(top). We thus wish a model to be more aware of semantic grouping. For this, we make the mild assumption that a semantic group consists of words (or text segments) that are semantically relevant and are spatially close to each other in the document. Under this assumption, we propose the 2-TSC pre-training task to help the model learn semantic grouping.

We borrow the idea from SimSiam (Chen and He, 2021), a type of contrastive learning that do not require negative samples, to let *text segment representations* belonging to the potential same semantic group close to each other, where a text segment representation is the average pooling of the token representations in a text segment.

Let $\mathcal{R}_k = \{r_{ki}\}_{i=1}^{I}$ denote the set of representation vectors for the i-th token in the k-th text segment s_k. We represent the semantics of s_k by

[1] We experimentally validate this assumption.

average-pooling its token representations, i.e.:

$$v_k = \frac{1}{|\mathcal{R}_k|} \sum_{i=1}^{I} r_{ki}. \quad (4)$$

We can find a set $K = \{(k, k')\}$ of semantically close text segment indices k and k', which satisfies the following two conditions with predefined thresholds θ_{dis} and θ_{sim}:

$$\text{Dist}(s_k, s_{k'}) < \theta_{\text{dis}}, \ \text{Sim}(v_k, v_{k'}) > \theta_{\text{sim}}. \quad (5)$$

$\text{Dist}(s_k, s_{k'})$ gives the Euclid distance between the centers of bounding boxes that encompasses all tokens in s_k and $s_{k'}$, which can be computed based on the merged bounding box in the k-th text segment. Sim gives the cosine similarity. Given $(k, k') \in \mathcal{K}$, v_k and $v_{k'}$ should be close to each other. We thus introduce a predictor (Chen and He, 2021) to map v_k to the same dimensional space as

$$z_k = \frac{1}{|\mathcal{R}_k|} \sum_{i=1}^{I} f(r_{ki}), \quad (6)$$

and bring them closer by

$$L_{\text{2-TSC}} = -\text{Sim}(z_k, \text{stopgrad}(v_{k'})), \quad (7)$$

With the above three pre-training objectives, the model is pre-trained with the following loss:

$$L_{\text{total}} = L_{\text{MLM}} + \alpha L_{\text{1-LOP}} + \gamma L_{\text{2-TSC}} \quad (8)$$

where α and γ are hyper-parameters. $L_{\text{2-TSC}}$ is used only in the final epoch of pre-training.

5 Experiments

5.1 Pre-training Settings

We pre-train ReLayout on the IIT-CDIP Test Collection (Lewis et al., 2006), which contains over 11 million scanned document pages. We extract words and global 1D positions, word-wise 2D bounding boxes, and text segments from document page images with an open-source OCR tool, PaddleOCR.

ReLayout's model architecture is almost the same as RoBERTa (Liu et al., 2019b) but with an additional 2D embedding layer. All parameters, except for the 2D embedding layer, are initialized with RoBERTa's parameters. We use AdamW optimizer (Loshchilov, 2017) with a batch size of 32 for 5 epochs. The base learning rate is set to 5e-5, with weight decay of 1e-2 and $(\beta_1, \beta_2) = (0.9, 0.999)$. The learning rate changes with a linear decay strategy. We evaluated two variants based on RoBERTa variants, i.e., $\text{ReLayout}_{\text{Base}}$ and $\text{ReLayout}_{\text{Large}}$. The former has 12 layers with 16 heads; the latent dimensionality is 768. The latter has 24 layers with 16 heads where the latent dimensionality is 1024.

As for the hyper-parameters, $P_{\text{MLM}} = 20\%$ and $P_{\text{1-LOP}} = 30\%$. For $L_{\text{2-TSC}}$, the thresholds are $\theta_{\text{dis}} = 120$ and $\theta_{\text{sim}} = 0.9$. The coefficients for balancing the objectives are $\alpha = 0.5$ and $\gamma = 0.5$.

5.2 Fine-tuning Settings

FUNSD and CORD. FUNSD (Jaume et al., 2019) and CORD (Park et al., 2019) are used for semantic entity classification tasks in complex forms and receipt documents, aiming to classify words into a set of predefined semantic entities. The FUNSD dataset contains 199 documents with annotations for 9,707 semantic entities, which are among "question," "answer," "header," and "other". The training and test splits contain 149 and 50 samples, respectively. CORD is a dataset for information extraction in receipts with 30 semantic labels in 4 categories. It contains 1,000 receipts, 800 for training, 100 for validation, and 100 for test. For these two datasets, we use BIO tags (Xu et al., 2020b) and formalize semantic entity classification as a sequential labeling task.

We fine-tune ReLayout for 1,000 steps with learning rate of 4.5e-5 and batch size of 64 for FUNSD, while learning rate of 7e-5 and batch size of 32 for CORD. Similarly to the existing methods, we use officially-provided OCR annotations (including words, word-wise bounding boxes, and global 1D positions) on the training set and report word-level F1 scores on the test set. For models that use manually annotated semantic groups, which are used to set segment-wise bounding boxes as the model's 2D position input (shaded rows in Table 1), we also report their performance when semantic groups are replaced by text segments by Microsoft Read API (MSR)[2].

DocVQA. Visual question answering on document images requires a model to take a document image (if need), OCR annotations, and a question as input and output an answer. The DocVQA dataset (Mathew et al., 2021) offers 10,194/1,286/1,287 images and 39,463/5,349/5,188 questions in training/validation/test splits, respectively. We formal-

[1] https://github.com/PaddlePaddle/PaddleOCR

[2] https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/concept-ocr

Method	#Params	Mod.	FUNSD(F1↑)	CORD(F1↑)	DocVQA(ANLS↑)
$\text{BERT}_{\text{Base}}$ (Devlin et al., 2018)	110M	T	60.26	89.68	63.72
$\text{RoBERTa}_{\text{Base}}$ (Liu et al., 2019b)	125M	T	66.48	93.54	66.42
$\text{UniLMv2}_{\text{Base}}$ (Bao et al., 2020)	125M	T	68.90	90.92	71.34
$\text{LayoutLM}_{\text{Base}}$ (Xu et al., 2020b)	160M	T+L+I	79.27	-	69.79
$\text{LayoutLMv2}_{\text{Base}}$ (Xu et al., 2020a)	200M	T+L+I	82.76	94.95	**78.08**
$\text{DocFormer}_{\text{Base}}$ (Appalaraju et al., 2021)	183M	T+L+I	83.34	96.33	-
$\text{BROS}_{\text{Base}}$ (Hong et al., 2022)	110M	T+L	83.05	95.73	71.92
$\text{LiLT}_{\text{Base}}$ (Wang et al., 2022)	-	T+L	88.41	96.07	-
$\text{LayoutLMv3}_{\text{Base}}$ (Huang et al., 2022)	133M	T+L+I	81.61†	94.64†	74.56†
$\text{LayoutLMv3}_{\text{Base}}$ (Huang et al., 2022)	133M	T+L+I	90.29	96.56	78.76
$\text{LayoutMask}_{\text{Base}}$ (Tu et al., 2023)	182M	T+L	73.97†	82.37†	70.79†
$\text{LayoutMask}_{\text{Base}}$ (Tu et al., 2023)	182M	T+L	92.91	96.99	-
$\text{ReLayout}_{\text{Base}}$ (Ours)	125M	T+L	**84.64**	**96.82**	76.02
$\text{BERT}_{\text{Large}}$ (Devlin et al., 2018)	340M	T	65.63	90.25	67.45
$\text{RoBERTa}_{\text{Large}}$ (Liu et al., 2019b)	355M	T	70.72	93.80	69.52
$\text{UniLMv2}_{\text{Large}}$ (Bao et al., 2020)	355M	T	72.57	92.05	77.09
$\text{LayoutLM}_{\text{Large}}$ (Xu et al., 2020b)	343M	T+L	77.89	-	72.59
$\text{LayoutLMv2}_{\text{Large}}$ (Xu et al., 2020a)	426M	T+L+I	84.20	96.01	**83.48**
$\text{DocFormer}_{\text{Large}}$ (Appalaraju et al., 2021)	536M	T+L+I	84.55	96.99	-
$\text{BROS}_{\text{Large}}$ (Hong et al., 2022)	340M	T+L	84.52	97.40	74.70
$\text{StructuralLM}_{\text{Large}}$ (Li et al., 2021a)	355M	T+L	85.14	-	83.94‡
$\text{LayoutLMv3}_{\text{Large}}$ (Huang et al., 2022)	368M	T+L+I	84.13†	96.88†	78.26†
$\text{LayoutLMv3}_{\text{Large}}$ (Huang et al., 2022)	368M	T+L+I	92.08	97.46	83.37
$\text{LayoutMask}_{\text{Large}}$ (Tu et al., 2023)	404M	T+L	78.12†	84.67†	74.06†
$\text{LayoutMask}_{\text{Large}}$ (Tu et al., 2023)	404M	T+L	93.20	97.19	-
$\text{ReLayout}_{\text{Large}}$ (Ours)	355M	T+L	**86.11**	**97.42**	80.14

Table 1: Comparison of existing models on the FUNSD, CORD, and DocVQA datasets. T/L/I denotes the "text/layout/image" modality. Grids in ▒ indicate that the model uses manually-annotated semantic groups. The superscript † indicates that the model uses text segments provided by Microsoft Read API. The superscript ‡ indicates that the model was trained with additional QA data to achieve higher scores, it isn't directly comparable.

ize this task as an extractive QA problem, wherein a model predicts the start and end positions with binary classifiers. We fine-tune models on the training set and report ANLS (average normalized Levenshtein similarity), a commonly-used edit distance-based metric, on the test set. Unfortunately, the OCR annotations provided in this dataset are of low quality. We thus use the MSR to extract words, word-wise bounding boxes, and global 1D positions. For models that use segment-wise bounding boxes, we employ text segments' bounding boxes by MSR (marked with † in Table 1). We fine-tune all models for 40 epochs with learning rate of 2e-5 and batch size of 32.

Besides the experiments above, the scores of all other models come from previous papers (Huang et al., 2022; Tu et al., 2023).

5.3 Results

As shown in Table 1, when officially-providedOCR annotations are used (and so the task is VrDU), ReLayout surpasses all models that do not use manually annotated semantic groups. The models that use manually annotated semantic groups (referred to as segment-dependent models, shaded rows in Table 1) yielded higher scores. Their performance significantly drop if semantic groups are replaced with text segments provided by a commercial OCR tool (e.g., the performance of $\text{LayoutLMv3}_{\text{Base}}$ and $\text{LayoutMASK}_{\text{Base}}$ drop by -8.68 and -18.94 respectively on FUNSD.). This implies that segment-dependent models heavily rely on manually annotated semantic groups to capture semantic structures in documents, which may contradict the initial purpose of automated document understanding. Also, the results reinforce the necessity to reexamine the choice of semantic grouping in the VrDU tasks, which is previously proven to be a shortcut (Li et al., 2021a).

For the DocVQA dataset, it is fair to compare ReLayout with LayoutLMv3 and LayoutMASK in the ReVrDU setting (i.e., without manually annotated semantic grouping, marked with † in the DocVQA column of Table 1), and ReLayout outperforms them. Yet, it consistently falls behind LayoutLMv2. We believe this gap primarily stems from the absence of the visual modality, as LayoutLMv2 additionally leverages a visual encoder.

	Our	LMv2	LMv3	LayoutMASK
FUNSD	84.64	82.76	90.29	**92.91**
-M	**83.13**	80.13	81.46	73.61
-P	**82.87**	78.69	80.27	71.15
CORD	96.82	94.95	96.56	**96.99**
-M	**96.23**	93.78	94.37	82.33
-P	**95.92**	92.24	93.87	81.47
DocVQA	-	78.08	**78.76**	-
-M	76.02	**76.33**	74.56	70.79
-P	**64.25**	63.73	63.26	60.81
-O	**74.19**	72.56	71.17	69.96

Table 2: Comparison in the ReVrDU setting with different OCR parsing results. M and P respectively indicate that the OCR parsing results are from MSR and PaddleOCR. O is the officially provided OCR parsing results by an OCR tool (not manually annotated).

5.4 Comparison with Different OCR Tools

To evaluate how differences in real-world OCR tools affect the ReVrDU performance, we used both open-source and commercial OCR tools to extract OCR parsing results to create revised datasets. We assess ReLayout, LayoutLMv2 (LMv2), LayoutLMv3 (LMv3), and LayoutMASK (all of them are the base variant) over multiple OCR parsing results on the FUNSD, CORD, and DocVQA dataset, where the models are fine-tuned on the revised training sets and evaluated on the revised test sets. Table 2 summarizes the scores.

On the FUNSD-M and FUNSD-P datasets, segment-dependent models, i.e., LayoutLMv3 and LayoutMASK, show a significant performance drop, especially LayoutMASK. As we discussed in the previous section, manually annotated semantic groups offer a strong cue to capture the semantics in the document as the text within each group is complete (as shown in Figure 1(top)), casting the word-level semantic entity classification into segment-level classification. When text segments provided by OCR tools are used, the models struggle to understand the semantic structure spanned over multiple text segments, leading to classification failures. The models that do not rely on manually annotated semantic groups, like ReLayout and LayoutLMv2, show only a slight performance drop. The smaller performance decline of ReLayout demonstrates its robustness against imperfect layout information provided by various OCR tools.

On the CORD-M and CORD-P datasets, the four

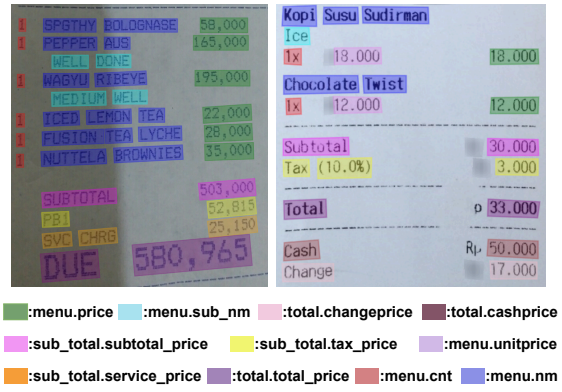
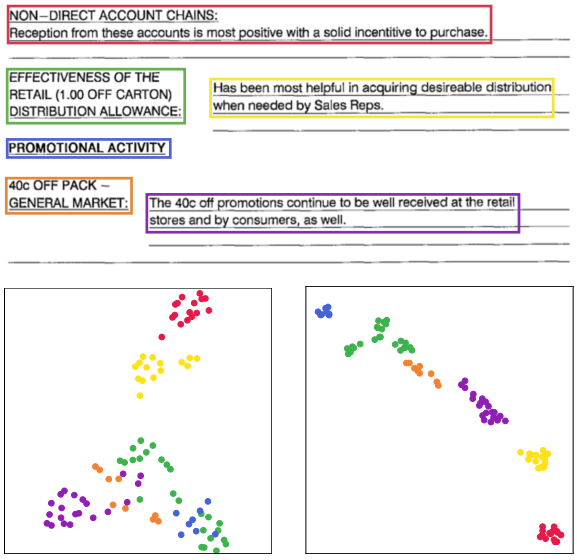


Figure 3: Two examples document images from CORD.



(a) MLM and 1-LOP (b) MLM, 1-LOP and 2-TSC

Figure 4: Visualization of pre-trained representations.

models respectively show performance declines. ReLayout still maintains the best robustness with real-world layout input. LayoutLMv3 shows an acceptable performance drop, while LayoutMASK still experiences a significant decline. The visual modality may serve as a beneficial complement when semantic grouping is inaccurate. Overall, the performance declines on CORD are smaller compared to FUNSD, possibly because the semantic ststructures receipts are basically complete in a single line as shown in Figure 3, which is much simpler than those in the FUNSD.

For DocVQA, the scores in the first row of the table originate from the original reports in the LayoutLMv2 and LayoutLMv3 papers (Xu et al., 2020a; Huang et al., 2022), while LayoutMASK was not evaluated on this dataset. We evaluate the models on DocVQA in the ReVrDU setting with three types of automatically acquired OCR parsing

#	Pre-training Setting			Datasets					
	MLM	1-LOP	2-TSC	FUNSD	CORD	DocVQA	FUNSD-P	CORD-P	DocVQA-P
1	√			81.55±0.06	95.45±0.03	73.85±0.12	79.96±0.06	94.65±0.02	62.97±0.10
2	√	√		83.84±0.18	96.43±0.15	74.30±0.07	82.94±0.08	95.49±0.13	63.17±0.04
3	√		√	81.73±0.05	95.82±0.07	74.07±0.05	81.52±0.03	95.68±0.10	62.44±0.17
4	√	√	√	**84.37±0.12**	**96.64±0.12**	**74.82±0.05**	**83.15±0.05**	**95.91±0.06**	**63.58±0.08**

Table 3: Ablation experiments of different pre-training methods.

results. ReLayout shows the best performance on the revised datasets (the -P and -O variants) with poorer OCR quality, but when using the commercial MSR, ReLayout performs slightly lower than LayoutLMv2. This still demonstrates ReLayout's robustness in extracting semantically meaningful structural information even in complex documents and potentially erroneous OCR. We acknowledge that integrating the visual modality is effective in enhancing performance on the ReVrDU task.

6 Ablation Study

We ablate newly added pre-training losses which learn comprehensively layout information in the VrDU and ReVrDU setting.

Quantitative analysis. Table 3 shows the performance scores for all possible combinations of losses (the masked language model loss cannot be removed as it is the basis for pre-training). The use of 1-LOP significantly enhances model performance, particularly on the FUNSD dataset. This is because forms contain densely packed local text structures, and using 1-LOP not only helps the model enhance the comprehension of the text flow but also aids in capturing cross-segment relationships. Comparison between the first and third rows shows that 2-TSC can bring a certain, though limited, performance improvement. The combination of 2-TSC and 1-LOP improves the performance by a larger margin. We can guess that 2-TSC hardly stand by itself as it mainly relies on the local layout information of tokens when determining relevant local text segments to bring closer. The 1-LOP loss may give ideas about the local layout information, ending up with better representations that helps 2-TSC. This is also why we only add the 2-TSC loss in the final epoch.

Qualitative analysis. To evaluate whether the 2-TSC pre-training task can effectively learn semantic groups, we input words, global 1D positions, and word-wise 2D bounding boxes from official annotations of a FUNSD form into the MLM and 1-LOP pre-training models with/without the 2-TSC loss. For the document shown in Figure 4(top), we visualize the respective models' token representations using UMAP (McInnes et al., 2018). The document has six (manually annotated) semantic groups, identified by boxes in different colors. Figure 4a displays the representations without 2-TSC, while Figure 4b shows those with 2-TSC (Representations learned solely from MLM pre-training can be seen in the Appendix A.). The representations with MLM and 1-LOP form reasonable clusters, though the green, blue, orange, and purple semantic groups overlap to some extent. This may be because the 1-LOP loss introduces strong local layout information into the representations (as we also visualize the representations using only the MLM loss under the same input, which showed a very chaotic distribution, but due to space limitations, we did not display it). The 2-TSC loss leads to more clear-cut clusters, without using manually annotated semantic groups. This difference does not directly explain the better performance of the model with 2-TSC, but it still demonstrates that 2-TSC can be a reasonable proxy of manually annotated semantic grouping.

7 Conclusion

This paper introduces the ReVrDU task that align more with real-world scenarios compared to the original VrDU tasks, shedding light on the problem of using manually annotated semantic grouping for document understanding. Our experimental results showed that the existing models worsen performance scores when accurate semantic groups are unavailable. We also propose pre-training losses, 1-LOP and 2-TCS, to aid the lack of semantic grouping, showing superior performance compared to the existing models but the reliance on semantic grouping removed. We believe the ReVrDU task brings a new dimension of challenge into document understanding and contributes its progress.

8 Limitation and Future Work

Introducing the visual modality: Incorporating information from the visual modality through a visual encoder is a common approach to enhance document understanding models. This requires considering the interactions between the three modalities: text, layout, and image. Future work will explore compatible ways to introduce visual modality information, further improving the performance of document understanding models.

Dependency on OCR tools: Most state-of-the-art pre-trained document understanding models rely on OCR annotations, and our model is no exception. However, this two-stage data processing approach means that the performance of OCR can significantly affect the subsequent model's results. Therefore, exploring effective OCR-free models is an important direction to reduce accumulated errors, speed up processing, and lower computational costs.

9 Ethics Statement

After careful consideration, we believe that our paper does not introduce additional ethical concerns. We declare that our work complies with the ACL Ethics Policy.

Acknowledgments

This work was supported by World Premier International Research Center Initiative (WPI), MEXT, Japan. This work was also supported by JST ACT-X Grant Number JPMJAX24C8 and JSPS KAKENHI No. 24K20795. This work was partly supported by DAIKIN INDUSTRIES, LTD.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International conference on machine learning*, pages 642–652. PMLR.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Timo I Denk and Christian Reisswig. 2019. Bertgrid: Contextualized embedding for 2d document representation and understanding. *arXiv preprint arXiv:1909.04948*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Łukasz Garncarek, Rafał Powalski, Tomasz Stanisławek, Bartosz Topolski, Piotr Halama, Michał Turski, and Filip Graliński. 2021. Lambert: layout-aware language modeling for information extraction. In *International conference on document analysis and recognition*, pages 532–547. Springer.

Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. 2021. Unidoc: Unified pretraining framework for document understanding. *Advances in Neural Information Processing Systems*, 34:39–50.

Jiuxiang Gu, Ani Nenkova Nenkova, Nikolaos Barmpalios, Vlad Ion Morariu, Tong Sun, Rajiv Bhawanji Jain, Jason Wen Yong Kuen, and Handong Zhao. 2023. Unified pretraining framework for document understanding. US Patent App. 17/528,061.

Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4583–4592.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. 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*, volume 36, pages 10767–10775.

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.

Wonseok Hwang, Seonghyeon Kim, Minjoon Seo, Jinyeong Yim, Seunghyun Park, Sungrae Park, Junyeop Lee, Bado Lee, and Hwalsuk Lee. 2019. Post-ocr parsing: building simple and robust parser via bio tagging. In *Workshop on Document Intelligence at NeurIPS 2019*.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Zhouqiang Jiang, Bowen Wang, Tong Xiang, Zhaofeng Niu, Hong Tang, Guangshun Li, and Liangzhi Li. 2023. Concatenated masked autoencoders as spatial-temporal learner. *arXiv preprint arXiv:2311.00961*.

Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, and Tomas Pfister. 2022. Formnet: Structural encoding beyond sequential modeling in form document information extraction. *arXiv preprint arXiv:2203.08411*.

David Lewis, Gady Agam, Shlomo Argamon, Ophir Frieder, David Grossman, and Jefferson Heard. 2006. Building a test collection for complex document information processing. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. 2021a. Structurallm: Structural pre-training for form understanding. *arXiv preprint arXiv:2105.11210*.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021b. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5652–5660.

Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021c. Structext: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1912–1920.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019a. Graph convolution for multimodal information extraction from visually rich documents. *arXiv preprint arXiv:1903.11279*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, et al. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *arXiv preprint arXiv:2210.06155*.

Rafał Powalski, Łukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michał Pietruszka, and Gabriela Pałka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 732–747. Springer.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Yi Tu, Ya Guo, Huan Chen, and Jinyang Tang. 2023. Layoutmask: Enhance text-layout interaction in multi-modal pre-training for document understanding. *arXiv preprint arXiv:2305.18721*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. 2023. Learning bottleneck concepts in image classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 10962–10971.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. *arXiv preprint arXiv:2202.13669*.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. Layoutreader: Pre-training of text and layout for reading order detection. *arXiv preprint arXiv:2108.11591*.

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. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2021. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*.

A Appendix

A.1 Visualization of Representations

In Table 5, we visualize the representations learned from different pre-training task combinations.

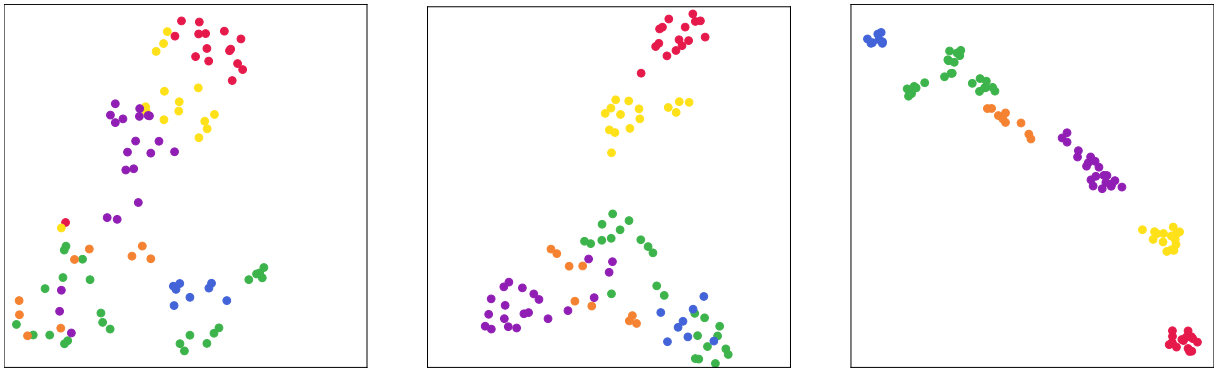
A.2 Visualization of VrDU and ReVrDU

Figures 6-8 display OCR annotation visualizations on the FUNSD sample, categorizing bounding boxes into word-wise and segment-wise types, and annotations also are classified into three types: manual, MSR, and PaddleOCR annotations. In Figure 9, the images, questions, and answers from the DocVQA sample are visualized.

A.3 BIO Tags

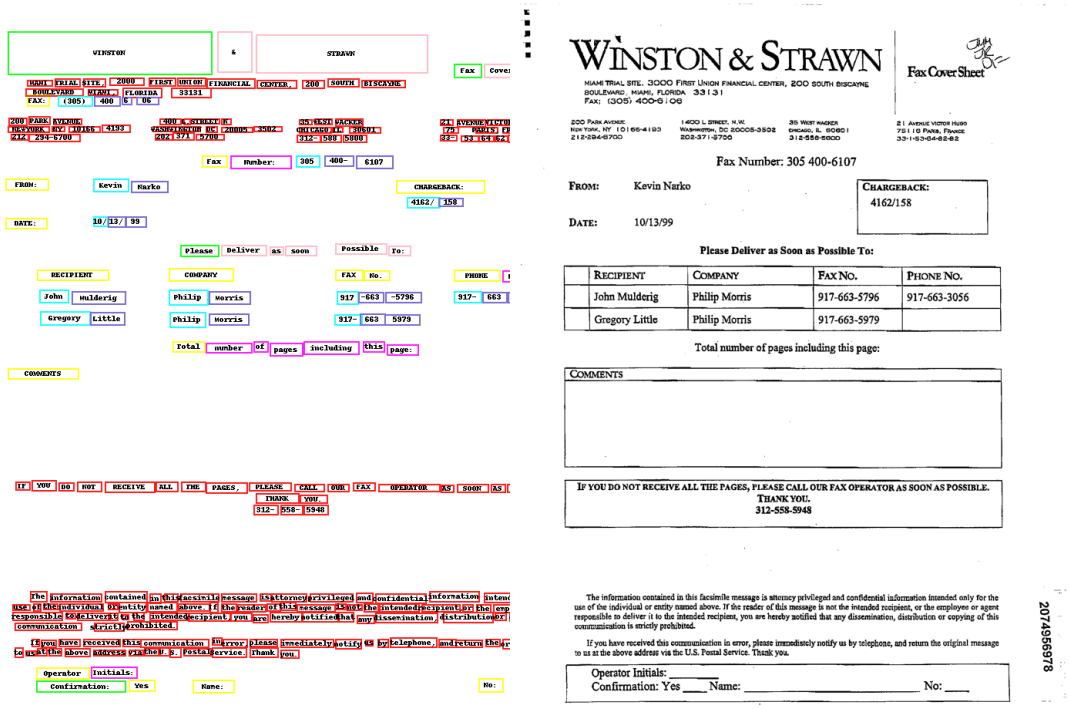
The BIO tagging scheme is a method used for marking up text in sequence labeling tasks, commonly applied in Named Entity Recognition (NER) and other forms of linguistic annotation. In this scheme, each token of the text is tagged with one of three prefixes: "B-" (Beginning), "I-" (Inside), and "O" (Outside). The "B-" prefix indicates the beginning of an entity, "I-" marks the continuation of an entity, and "O" denotes a token that does not belong to any entity. This method helps in clearly differentiating the boundaries of entities within the text, making it easier for models to recognize and categorize text segments accurately. For example, in the entity "New York," "New" would be tagged as "B-Location" and "York" as "I-Location," clearly identifying the entire phrase as a geographical entity.

Therefore, learning to use 1-LOS to learn the local reading order is beneficial for the model to solve entity classification tasks based on BIO tagging.

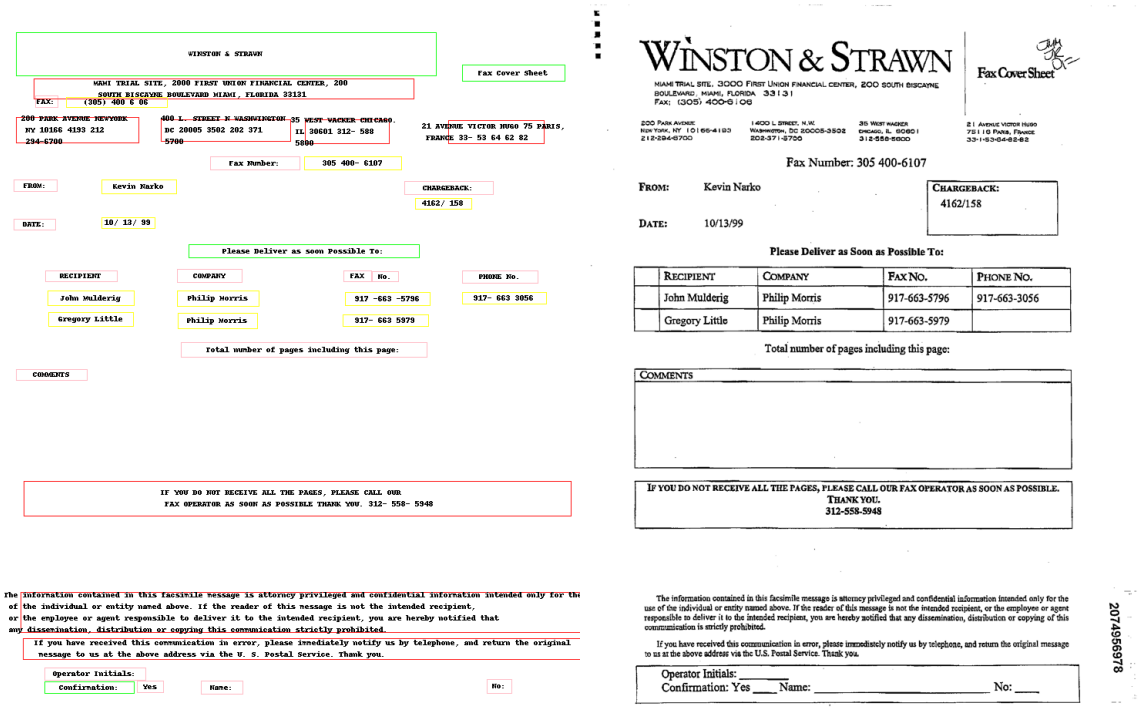


(a) MLM (b) MLM and 1-LOP (c) MLM, 1-LOP and 2-TSC

Figure 5: Visualization of representations learned under different pre-training tasks.

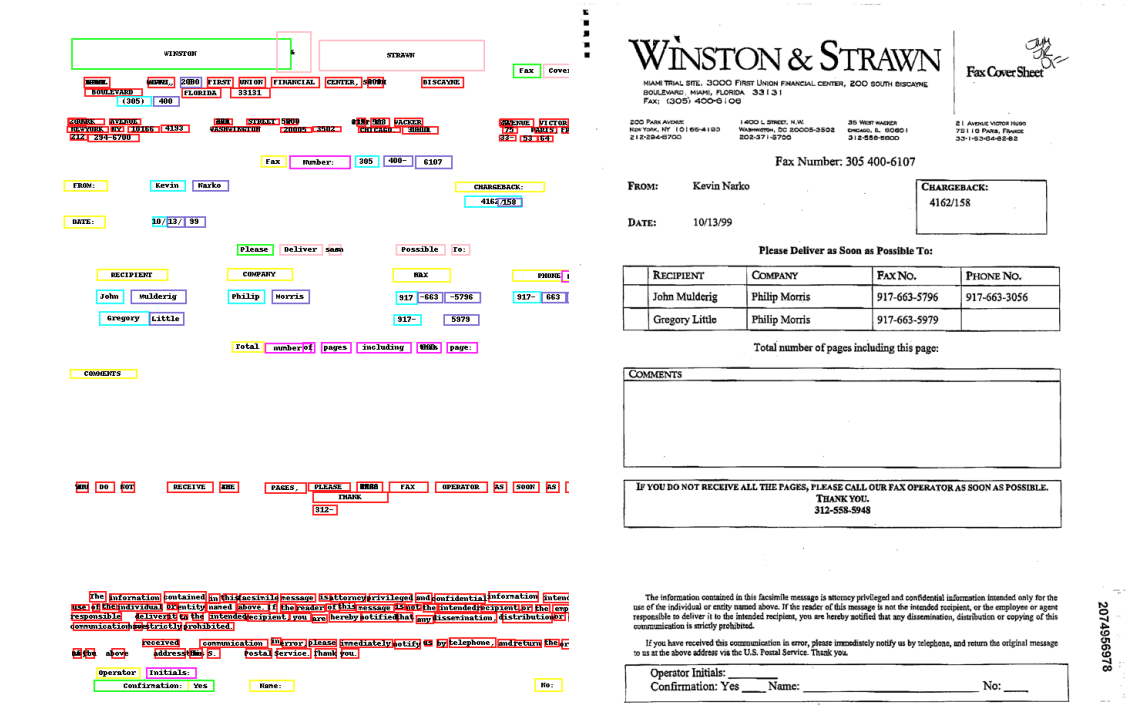


(a) Manually annotated words and word-wise bounding boxes.

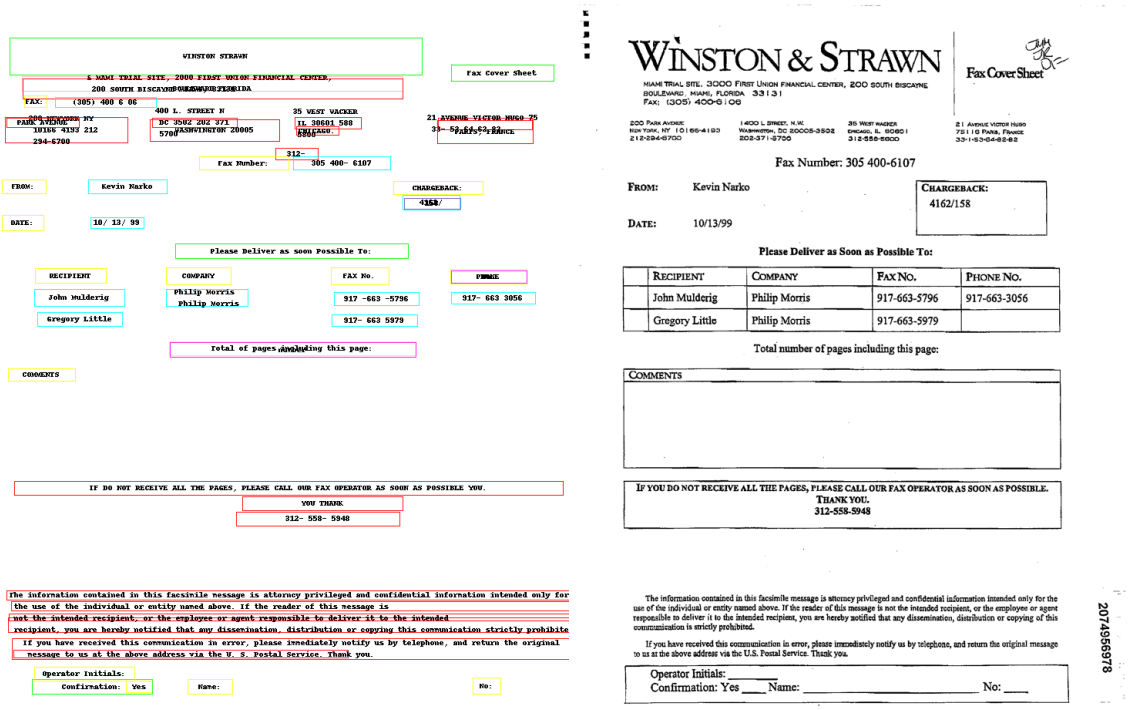


(b) Manually annotated words and segment-wise bounding boxes.

Figure 6: Visualization of the VrDU Task on the FUNSD Sample.

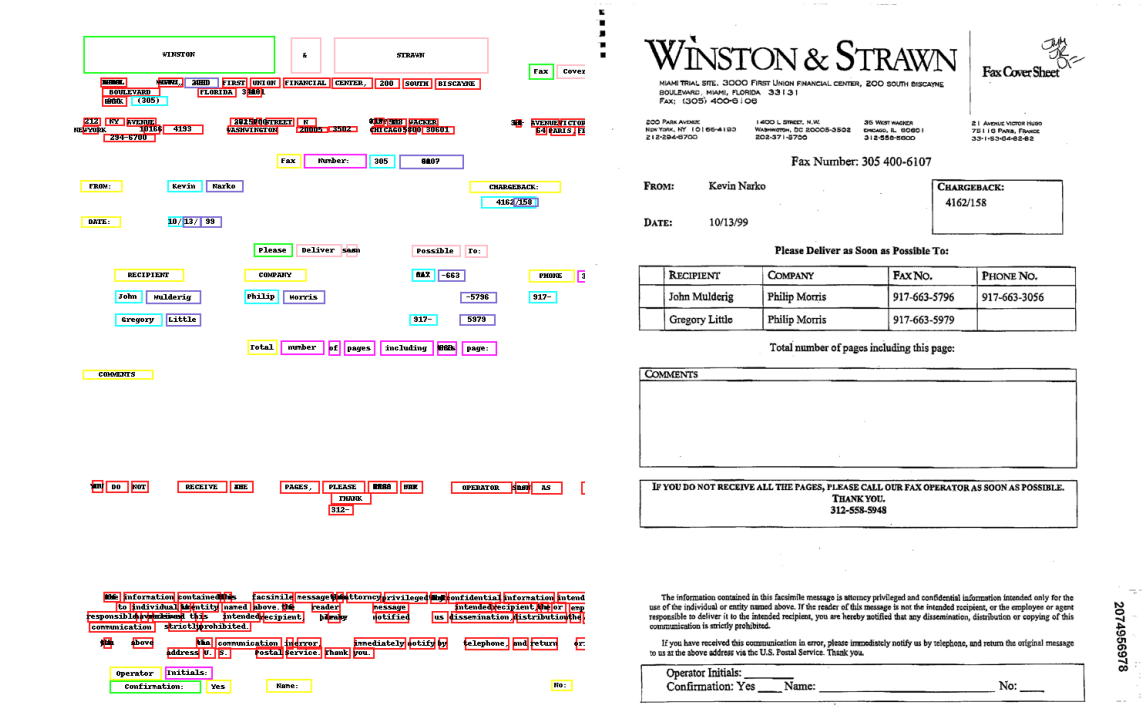


(a) MSR-automatically annotated words and word-wise bounding boxes.

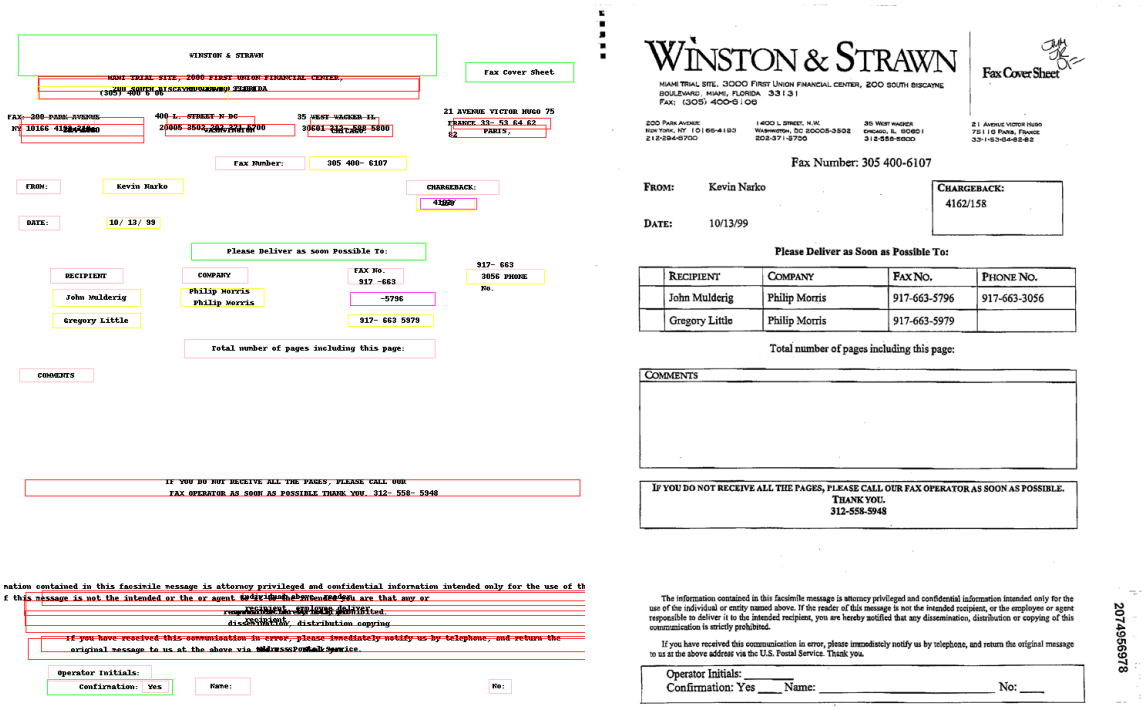


(b) MSR-automatically annotated words and segment-wise bounding boxes.

Figure 7: Visualization of the ReVrDU Task on the FUNSD Sample with MSR.



(a) PPOCR-automatically annotated words and word-wise bounding boxes.



(b) PPOCR-automatically annotated words and segment-wise bounding boxes.

Figure 8: Visualization of the ReVrDU Task on the FUNSD Sample with PPOCR.

Q: how many points are there in modifications to readout instrumentation?
GT: 5

2. The rotometer in the original design was replaced by a 0-100 SLPM mass flow meter (Hastings #200B) to provide more precise flow settings. As in the original design, flow is set before each smoke run by manual adjustment of a 3/4" needle valve (Whitey SS65XTF12-F16) and the flowmeter is bypassed during smoke measurements to minimize contamination. In the new design the mass flowmeter is bypassed during smoke runs by the parallel combination of a 3/4" needle valve (Whitey SS65XTF12-F16) and a length of unrestricted 1/2" piping. The latter valve is adjusted to give an identical pressure drop (as indicated by a 0-25 cm H_2O Dwyer Magnehelic® gauge) to that of the mass flowmeter, thus maintaining the flow when the meter is removed from the circuit.

Modifications to Readout Instrumentation:

1. The single turn sensitivity potentiometer supplied on the Oriel #7072 amplifier was replaced with a ten turn potentiometer to allow more precise settings of the zero attenuation level.

2. To reduce noise and drift in the computer data acquisition system, the output voltage divider internal to the Oriel amplifier was altered to increase the instrument's output by a factor of 100. The sensitivities of the computer data system and the parallel strip chart recorder were correspondingly reduced.

3. Signal leads were re-dressed and shielded to minimize ground loops and RF pickup. The exposed terminal board was eliminated.

4. A running display of output voltage was added to the computer screen as a visual aid to the operator.

5. Changes to the computer program were made by Randy Greene to eliminate minor operational problems. The changes did not affect the calculation of any reported parameters.

cc: P. N. Gauvin
B. L. Goodman
A. C. Lilly
D. D. McRae
E. B. Sanders
J. E. Wickham
Central Files

-2-

2022156306

Source: https://www.industrydocuments.ucsf.edu/docs/sxxj0037

Q: what is the name of the tobacco company?
GT: rj reynolds tobacco company

RLS

R.J. Reynolds Tobacco Company

Interoffice Memorandum

RJR SECRET ADDRESSEE ONLY

Subject: Monthly Highlights - June **Date:** June 18, 1992

No. 515 By

To: J. D. Phillips **From:** M. D. Shannon

PROJECT XC:

CS (Carbon Scrubbing) Filter:

Significance: A previous aging study showed that a contaminant from the tobacco rod caused carbonyl yield to increase ~25% after 6 weeks of aging at 98°F. Propylene glycol was implicated as the contaminant. A new study was designed in which the level of propylene glycol was varied. **Status:** Levels of propylene glycol of 0.44%, 0.90%, and 1.15% (level currently used in Camel Light) were used in this study. After 6 weeks at 98°F, the carbonyls of the 1.15% sample increased ~25% as previously observed. At the level of 0.44%, no increase in the carbonyl yield occurred. These results define an acceptable upper limit for propylene glycol and suggest that any problems with de-activation of the carbon can be resolved. **Next Steps:** The experiment will be carried out to eighteen weeks for further evaluation.

Even Puff-by-Puff Filter Development (EPPCAT):

Significance: A designed study has been completed to gain an understanding of variables that influence the performance of the EPPCAT filter. This work will lead to the development of a model to permit accurate control over performance characteristics. **Status:** A designed experiment was conducted to quantify the importance of three filter characteristics: filter pressure drop, cigarette air dilution, and vent location. A Box-Behnken design was selected in order to evaluate each of the three factors at three different levels each. This type of design allowed for the estimation of all main effects, quadratic effects, and all linear two-way interactions. Filter pressure drop was set at 69, 79 and 89 mm water. Cigarette air dilution was set at 0, 20 and 40 per cent. Vent position was fixed at 19, 24, and 29 mm from the filter end of the cigarette. Cigarettes were smoked by FTC conditions with puff-by-puff pad WTPM collected. The mean values for TOTAL PAD WTPM and PUFF RATIO were the initial response variables studied. PUFF RATIO is determined by summing the WTPM for the first two puffs and dividing it by the sum of the WTPM for the last two puffs. Observations for these two response variables showed that 1) TOTAL PAD WTPM decreases as filter pressure drop increases and decreases as cigarette air dilution increases. The location of the vent holes has little effect. 2) PUFF RATIO increases as filter pressure drop increases and shows some increase as dilution increases for higher pressure drop filters. Again, the location of the vent holes had little effect. Average TOTAL PAD WTPM values ranged from

51210 9589

Source: https://www.industrydocuments.ucsf.edu/docs/qqbl0037

Figure 9: Visualization of DocVQA samples