Unveiling the Deficiencies of Pre-trained Text-and-Layout Models in Real-world Visually-rich Document Information Extraction

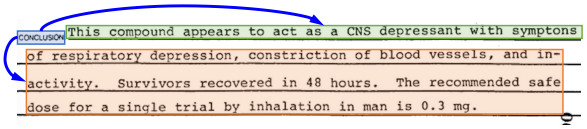
Chong Zhang[1], **Yixi Zhao**[2], **Yulu Xie**[3], **Chenshu Yuan**[4], **Yi Tu**[5], **Ya Guo**[5], **Mingxu Chai**[1], **Ziyu Shen**[1], **Yue Zhang**[1], **Qi Zhang**[1]
[1]Fudan University [2]Carnegie Mellon University [3]Tsinghua University [4]Nankai University
[5]Ant Tiansuan Security Lab, Ant Group
{chongzhang20, qz}@fudan.edu.cn

Abstract

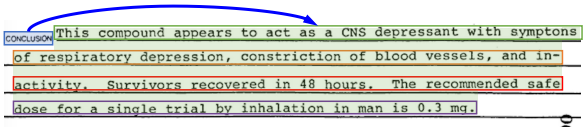
Recently developed pre-trained text-and-layout models (PTLMs) have shown remarkable success in multiple information extraction tasks on visually-rich documents (VrDs). However, despite achieving extremely high performance on benchmarks, their real-world performance falls short of expectations. Owing to this issue, we investigate the prevailing evaluation pipeline to reveal that: (1) The inadequate annotations within benchmark datasets introduce spurious correlations between task inputs and labels, which would lead to overly-optimistic estimation of model performance. (2) The evaluation solely relies on the performance on benchmarks and is insufficient to comprehensively explore the capabilities of methods in real-world scenarios. These problems impede the prevailing evaluation pipeline from reflecting the real-world performance of methods, misleading the design choices of method optimization. In this work, we introduce EC-FUNSD, an entity-centric dataset crafted for benchmarking information extraction from visually-rich documents (VrD-IE). This dataset disentangles the falsely-coupled segment and entity annotations that arises from the block-level annotation of FUNSD. Using the proposed dataset, we evaluate the real-world VrD-IE capabilities of PTLMs from multiple aspects, including their absolute performance, as well as generalization, robustness and fairness. The results indicate that prevalent PTLMs do not perform as well as anticipated in real-world VrD-IE scenarios. We hope that our study can inspire reflection on the directions of PTLM development.

1 Introduction

The research field of document AI is becoming more popular with the increase in industrial demands (Cui et al., 2021; Sassioui et al., 2023; Yang and Hsu, 2022). One of the primary objectives in this field is to extract useful information from visually-rich documents (VrDs), given the texts with their xy-coordinates on the document layout. In recent years, the advent of pre-trained text-and-layout models (PTLMs) has promoted the understanding of the semantic and spatial relations within the document layouts, and demonstrated great success in multiple visually-rich document information extraction (VrD-IE) tasks(Hong et al., 2022; Huang et al., 2022; Gu et al., 2022; Tu et al., 2023; Luo et al., 2023; Liao et al., 2023). It is consensually acknowledged by the document AI community that the recent success of VrD-IE methods is mainly attributed to the advancement in PTLMs, thus the research focus of this field has shifted to the improvement of PTLMs. Therefore, reliable evaluation of the real-world performance of PTLMs when they are applied in downstream tasks have become essential.



(a) A document image with its layout, entity and linking annotations in FUNSD. Each color of shades indicates an annotated entity. Each arrow indicates an entity pair.



(b) The corresponding annotations in EC-FUNSD.

Figure 1: Motivation of revising the annotations of FUNSD. (a) Block-level annotations within FUNSD do not correspond to semantic entities, making it unsuitable for entity-centric evaluations. Their linking relationships are also confused. (b) EC-FUNSD decouples the annotation of layouts and entities for proper evaluation of layout-aware models.

Conventionally, PTLMs serve as the embedding model for document layouts, just as the role contextualized pre-trained language models played in NLP tasks (Devlin et al., 2019; Liu et al., 2019; Sun et al., 2019). Their capabilities for adapting to

Findings of the Association for Computational Linguistics: EACL 2026, pages 1–16
March 24-29, 2026

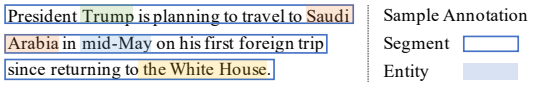


Figure 2: Illustration of necessity of disentangling the falsely-coupled visual segment and semantic entity annotations by introducing word/char-level entity annotations. In this sample, the entity "Saudi Arabia" spans across segments and also shares the same segment with the entities "Trump" and "mid-May".

VrD-IE tasks are usually evaluated by the semantic entity recognition (SER) and entity linking (EL) tasks, where PTLMs typically handle SER in a sequence-labeling manner, and tackle EL with simple classification heads. The requirements for these two tasks are in line with the capabilities required in real-world VrD-IE applications, and the performance of PTLMs on these two tasks reflects the capacity of their multimodal embeddings to facilitate the VrD-IE tasks. Currently, state-of-the-art PTLMs have achieved exceedingly good performance on prevailing VrD-IE benchmarks(Jaume et al., 2019; Park et al., 2019; Huang et al., 2019). As an extreme case, GeoLayoutLM (Luo et al., 2023) has achieved 100% test accuracy at the key-value entity linking task of CORD. We hope that these seemingly good performance of PTLMs on benchmarks could also be replicated in real-world VrD-IE applications.

However, the current prevalent benchmarks do not perfectly serve the evaluation of the VrD-IE capabilities of PTLMs, owing to their inherent drawbacks. For example, FUNSD (Jaume et al., 2019) and XFUND (Xu et al., 2022) are the most popular benchmarks in SER evaluation. Nevertheless, these datasets were originally designed for visual information extraction tasks and were transformed into the SER format merely for the purpose of evaluation. The layout semantic annotations within these datasets are based on visual blocks which do not have independent semantic meanings. As shown in Fig. 2, there are real-world cases that a block contain multiple entities, or an entity spans across multiple blocks. As a result, some of the blocks in these datasets do not correspond to semantic entities, making these datasets unsuitable for SER evaluation. Besides, the other benchmark datasets suffer from (1) the lack of diversity in layout patterns and entity semantics; (2) the insufficiency of complex semantic entities [1]; (3) the granularity-deficiency and low-quality of layout annotations; and (4) the mismatch of task paradigm. These issues are further demonstrated in §B, which illustrates the limitation of prevalent datasets for VrD-IE evaluation of PTLMs.

When these aforementioned unsuitable datasets are utilized in the evaluation pipeline, several potential risks will arise, significantly reducing the reliability of the evaluation. One of the most serious potential risks is the spurious correlation bias introduced by the block-level layout semantic annotations. The annotations within several prevailing benchmarks, including FUNSD, XFUND and CORD (Park et al., 2019), are organized by visual regions (i.e., blocks) of the document layout, in which the semantic category labels and the association relationship between regions are annotated on the block-level. In SER evaluation, segments and entities are simultaneously represented by these semantic blocks. Therefore, models would predict entity boundaries specified by block annotations with leveraging the segment-level layout inputs from block annotations. Layout elements within the same entity would have exactly the same xy-coordinate inputs, leading to label leakage during training, and resulting in inflated test scores which could not reflect the real model performance. Besides, the existing evaluation pipeline is solely centered around the test performance on benchmarks, and fails to comprehensively reflect the real-world performance of methods. These issues impede the accurate evaluation of the real-world performance of VrD-IE methods, which may negatively influence the direction of method optimization.

In this paper, our aim is to establish a valid pipeline for evaluating the comprehensive performance of PTLMs in real-world VrD-IE applications. The primary step is to introduce a more appropriate benchmark to evaluate the capabilities of PTLMs in VrD-IE tasks. Based on the specification of requirements, we propose **EC-FUNSD**, an **E**ntity-**C**entric benchmark derived from FUNSD (Jaume et al., 2019) that aims to provide a fair and unbiased evaluation for VrD-IE capabilities of PTLMs. This benchmark is constructed by manually revising the annotations in FUNSD, and is designed to be used for SER and EL tasks. Based on the newly proposed unbiased benchmark, we conduct a comprehensive evaluation of prevalent baseline PTLMs on real-world VrD-IE tasks. The evaluation aspects include not only the test performance of the model, but

[1] Complex entities refer to those entities that span multiple rows and columns and overlap with other entities in region.

Table 1: The requirements of benchmarks being suitable for evaluating the VrD-IE capabilities of PTLMs.

Dimension	Requirement	FUNSD	SROIE	CORD	EPHOIE	FUNSD-r	RFUND	EC-FUNSD
Layout Annotation Quality	Segment- and word/char-level annotations	✓	✗	✓	✓	✓	✓	✓
	Conventional distribution	✗	✓	✗	✗	✓	✓	✓
	High annotation quality	✓	✓	✗	✓	✗	✓	✓
Entity Annotation Quality	Decoupled from layout annotation	✗	✓	✗	✓	✓	✗	✓
	Annotated in word/char-level	✗	✗	✗	✓	✓	✗	✓
	Diverse and complex entities	✓	✓	✗	✗	✓	✓	✓
Suitable to PTLM	Following sequence-labeling paradigm	✓	✓	✓	✓	✗	✓	✓

also: (1) generalization to unknown distributions; (2) robustness under natural perturbations, and (3) fairness to difficult test sample subsets. Experimental results indicate that the true performance of the baseline models is not as good as they have claimed. Although EC-FUNSD and FUNSD are very similar in almost all aspects, the baseline models suffer from a significant performance drop on the same task settings, particularly a 7.48-8.55 decrease of F1 on SER. This reveals the potential risk of current PTLMs that they may develop to excessively overfit the biased benchmarks, but the actual benefits brought by the advancements are suspicious. Moreover, in the real-world performance evaluation focusing on the aforementioned aspects, the baseline models have exhibited notable deficiencies, revealing the potential risks of using them in practical applications. In the evaluation, the newly proposed unbiased benchmark allows these shortcomings to be identified more clearly, emphasizing the necessity for its introduction.

The contribution of this paper are as follows:

1. We point out that the capability of PTLMs on real-world VrD-IE cannot be adequately evaluated using the existing pipeline, since current benchmarks are not properly tailored for the evaluation purpose, and the existing pipeline is constrained by a monotonous evaluation standard.
2. We introduce EC-FUNSD, a entity-centric dataset of SER and EL that serves as a precise benchmark to evaluate the capabilities of PTLMs. This dataset has eliminated the spurious correlation bias in the previous dataset to support the real-world VrD-IE evaluation of PTLMs.
3. Our experiments with prevalent PTLMs show that these models may not perform as good as they claimed in real-world VrD-IE tasks, and are facing challenges related to generalization, robustness and fairness in practical applications.

We anticipate that this research will inspire the community to carry out more appropriate evaluation for existing methods, and develop novel methods that are more adaptable to real-world applications[2].

2 Related Work

Driven by the success of contextualized pre-trained language models in diverse NLP tasks and the growing demands of document AI, extensive studies have integrated multimodal features-such as layout and image information—into pre-trained language models, leading to the emergence of PTLMs. The effectiveness of these models in document representation has been successfully validated across various downstream VrD tasks.

The LayoutLM series (Xu et al., 2020, 2021a; Huang et al., 2022; Xu et al., 2021b) established the foundation for the prevailing paradigm of PTLMs, drawing inspiration from the common practice of representative pre-trained language models (Devlin et al., 2019; Liu et al., 2019; Sun et al., 2019) to further pre-train on text, layout, and visual modalities. Following the paradigm, many PTLMs have been developed, employing diverse pre-training strategies to enhance the interaction between multimodal features (Li et al., 2021; Lyu et al., 2024; Zhai et al., 2023; Tu et al., 2023; Appalaraju et al., 2021, 2024; Hong et al., 2022). Moreover, to better tackle specific downstream tasks, certain PTLMs have carefully crafted their pre-training strategies to align with targeted distributions of various tasks, including entity linking (Li et al., 2022; Krishnan et al., 2024; Peng et al., 2022; Luo et al., 2023; Yu et al., 2023b), open-set SER (Wang et al., 2022; Yang et al., 2023; Tu et al., 2024), and long document understanding (Yao et al., 2023; Jiang et al., 2024).

While these approaches demonstrate strong performance on downstream tasks, their effectiveness in real-world applications remains to be fully validated. Our work is fully motivated by generalization, robustness, and fairness evaluation studies for deep learning systems (see Appendix A), aiming to provide a comprehensive evaluation for VrD-IE capabilities of PTLMs.

[2]Code at https://github.com/chongzhangFDU/ROOR.

Table 2: Statistics of the proposed dataset. # is short for "Number of". The statistics of FUNSD is also listed in comparison, with invalid entity linking pairs removed.

Dataset	# Segments	# Words	# Segs. per Sample	# Words per Sample	Avg. Len. of Segment
FUNSD	9,743	31,485	48.95	158.21	3.23
EC-FUNSD	10,662	31,297	53.57	157.27	2.93
Dataset	**# Entities**	**# Ents. per Sample**	**Avg. Len. of Entity**	**# Relation Triplets**	**# Rel. Triplets per Sample**
FUNSD	8,529	42.85	2.92	3,966	19.92
EC-FUNSD	8,398	42.20	2.96	3,912	19.65

Table 3: The proportion of complex entities in each benchmark dataset.

Dataset	The proportion of complex entities
EPHOIE (Wang et al., 2021a)	$0/9,823 = 0.00\%$
CORD (Park et al., 2019)	$266/13,515 = 1.96\%$
FUNSD (Jaume et al., 2019)	$576/8,529 = 6.75\%$
RFUND (Lin et al., 2024)	$7,098/10,4184 = 6.81\%$
SROIE (Huang et al., 2019)	$925/3,780 = 24.47\%$
EC-FUNSD	$682/8,398 = 8.12\%$

3 EC-FUNSD: An Entity-Centric VrD-IE Benchmark Dataset

In this section, we introduce EC-FUNSD, a **E**ntity-**C**entric version of **FUNSD** for quantifying the negative impacts of using existing PTLMs in real-world VrD-IE applications. In previous works, popular datasets (Jaume et al., 2019; Xu et al., 2022; Park et al., 2019; Wang et al., 2021a; Huang et al., 2019; Zhang et al., 2023; Lin et al., 2024) have been adopted in the evaluation of PTLMs. Despite the widespread use of these datasets, they still have several limitations that hinder them from being suitable and reliable for evaluating the VrD-IE capabilities of PTLMs. According to Tab. 1, current benchmarks fall short of meeting the requirements for the evaluation (a more detailed defeat analysis can be found in Appendix B). It is highlighted the necessity of establishing new VrD-IE benchmarks to enable reliable and comprehensive evaluation.

3.1 Construction of EC-FUNSD

Based on the review of prevailing datasets, we claim that evaluating PTLMs with these benchmarks may be imprecise to reflect their real VrD-IE capabilities. In order to establish a high-quality dataset that is suitable to the evaluation, we propose five essential requirements for an appropriate benchmark: (1) The layout annotations should follow the conventional distribution, including segment-level text regions (as text lines) and fine-grained character (or word) bounding boxes, corresponding to typical outputs generated by OCR engines; (2) The layout annotations should be of high quality, with minimal omission of words and segments; (3) The annotation of semantic entities should confirm to unified semantic-driven definitions to ensure consistency across samples; (4) Each sample should be a single-page document that contains rich layout and diversified entities and relations to increase the validness of evaluation on the benchmark; and (5) The layout annotation of the text of semantic entities needs to be continuous to meet the requirements for the sequence-labeling paradigm in the SER task.

To construct a dataset that satisfies these requirements and serves as a proper benchmark for the evaluation, we choose to revise the existing annotations within the FUNSD dataset due to the following reasons: (1) Compared with other prevalent benchmarks, FUNSD stands out by offering a diverse range of layouts among its samples, rendering it a comprehensive benchmark. (2) FUNSD contains extensive block annotations with various semantic types and their relations. These existing annotations can be revised into semantic entity and relation annotations with little modification.

Therefore, based on the original dataset, we executed the revision of the layout and IE annotations to create a **E**ntity-**C**entric version of **FUNSD**, namely **EC-FUNSD**. The details for manual annotation is demonstrated in Appendix C. In two steps, we manually corrected the word and segment-level layout annotations of each sample, and revised the entity and relation annotations.

3.2 Statistics of EC-FUNSD

The statistics of EC-FUNSD are displayed in Tab. 2. The main differences are: (1) EC-FUNSD has a smaller total word count than FUNSD, primarily due to the removal of low-quality word annotations; (2) The number of segments increases since multiple-row blocks are split to several segments; (3) The average length of entities slightly increases as the entities that were split into multiple semantic blocks were combined; (4) Among 8,529 entities from FUNSD, 8,207 of them are kept in EC-FUNSD, while 322 are discarded or corrected,

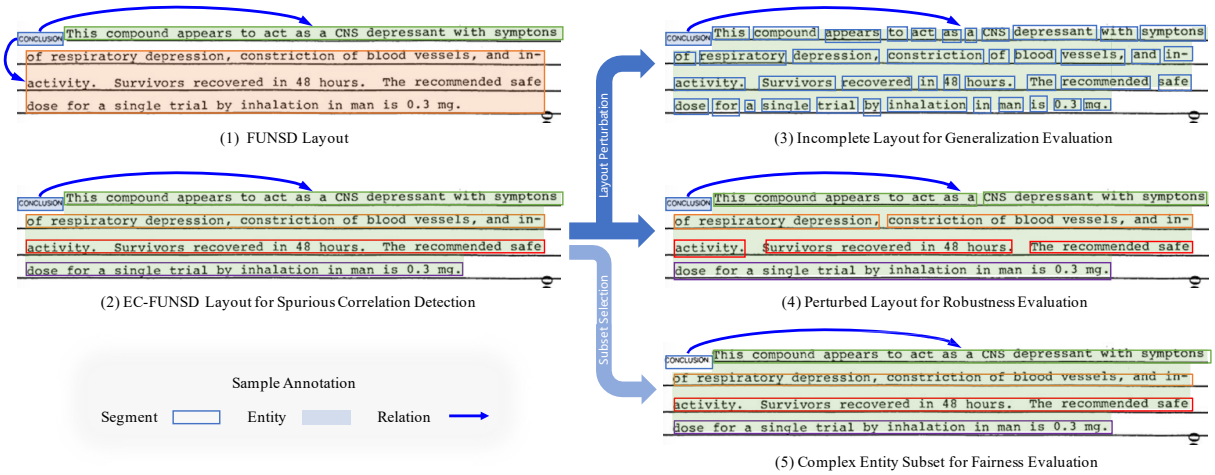


Figure 3: Document layout samples used in each evaluation.

resulting in 191 newly emerged entities in EC-FUNSD. The F1 between FUNSD and EC-FUNSD entity set among all samples is 96.96%, ensuring the task requirements be unchanged.

Compared with the aforementioned benchmarks, these key factors of EC-FUNSD making it well-suited for the evaluation of the real-world VrD-IE capabilities of PTLMs: (1) EC-FUNSD offers high-quality annotations. These include word- and segment-level layout annotations, along with manually-corrected entities and linking relationships, laying a solid foundation for evaluation. (2) The entity annotations in EC-FUNSD follow the sequence labeling paradigm, facilitating the evaluation. (3) EC-FUNSD features diverse layout patterns, mirroring real-world document layouts. As shown in Tab.3, it also includes a sufficient number of complex entities, enabling comprehensive evaluation of model generalization, robustness, and fairness in real-world scenarios.

4 Real-world VrD-IE Evaluation for Pre-trained Text-and-Layout Models

Based on the proposed dataset, we are now able to conduct a comprehensive evaluation of the capabilities of PTLMs in real-world VrD-IE. In this section, based on the specific VrD-IE task settings we focus on (see Appendix D), we illustrate the motivation and details of our evaluation methods.

4.1 Spurious Correlation Detection

This evaluation aims to identify potential spurious correlations between input features and output labels in the prevailing datasets and quantify its influence to model evaluation. As mentioned in §B, when evaluating on SER with FUNSD-like datasets (Jaume et al., 2019; Xu et al., 2022; Yang et al., 2023; Sun et al., 2021; Šimsa et al., 2023), the block-level annotations simultaneously represent both segments and entities. Therefore, models would predict entity boundaries and types specified by block annotations while leveraging the segment-level layout inputs from the same block annotations, which leads to a potential risk that models would overly rely on correlations between input layout features and entity labels. During training and testing, input units (e.g., tokens in a Transformer model) within the same block share the same layout feature (the block's bounding box), which leaks the block boundary information that need to be predicted. Therefore, models quickly converge during training by fitting these identical layout input features. However, the correlation between input layout features and entity labels is spurious and specific to certain datasets. In real-world SER, there is typically no inherent strong link between visual-based semantic block layout information and semantic-based entity boundaries. Thus, this correlation should not be leveraged as a reliable task feature.

In this evaluation, we measure the real performance of PTLMs on VrD-IE by removing the spurious input-label correlations. We apply PTLMs on EC-FUNSD for SER and EL tasks and compare the results with FUNSD, to determine whether current PTLMs overfit the spurious input-label correlations and quantify the corresponding negative impact.

4.2 Generalization Evaluation

The purpose of this evaluation is to examine the generalization ability of PTLMs to the distribution

shift of real-world datasets. Existing PTLMs are designed for and pre-trained on structured documents (contracts, forms, etc.) and leverage both char/word- and segment-level layouts as input features. Ideally, each user should apply a model with its *expected* input data format, as constrained by standardized benchmark datasets. However, distribution shifts may occur in real-world applications as real-world processing demands often extend beyond standardized textual formats (invoices, tickets, cards, WordArts, and UI screenshots). For example, auto-extracted layouts from OCR engines may exhibit incompleteness as segment-level layouts are often absent or too noisy to be effectively utilized. For example, being the most popular benchmark for document reading order prediction, ReadingBank (Wang et al., 2021d) only offers word-level layout annotations; typical Chinese VrD benchmarks, such as CTW (Yuan et al., 2018), provide only char-level layout annotations; and synthesized datasets like EATEN (Guo et al., 2019; Kim et al., 2022) also rely solely on single-level auto-generated layout annotations. Therefore, in generalization evaluation, we simulate the distribution shift in real-world scenarios by masking segment-level layout information and keeping only word-level layout information as input, evaluating the generalization ability of PTLMs by measuring their performance changes when adapting to the shifted distribution.

4.3 Robustness Evaluation

This evaluation is designed to assess the robustness of PTLMs against the real-world input perturbations. The processing of real-world documents may suffer from the printing distortion with overlapped, blurred or misaligned texts. Scanning-related issues like skew, jitter, or inconsistent lighting can further degrade the quality of scanned document pages, negatively influencing the model performance. In this evaluation, we apply random shifting, clipping, and rotation to the input bounding boxes to simulate real-world perturbations, measuring the extent to which current PTLMs are affected by real-world layout perturbations.

4.4 Fairness Evaluation

The aim of this evaluation is to test the consistency of model performance on specific subset of samples. In real-world applications, we expect that the model performance on any subset of samples should not be significantly lower than its average performance, ensuring the reliability across diverse data distributions that the model remains robust and fair even in less frequent or challenging cases.

In tackling SER task with PTLMs, we focus on complex entities. As shown in Fig. 1(2), the entire paragraph corresponds to a single entity, whose complex layout patterns like line wrapping and indentation making it extremely difficult for the model to accurately predict the boundaries of it. Additionally, due to the nature of the sequence-labeling paradigm, false prediction on any single token within the long entity can result in the failure to correctly recognize it. Therefore, this evaluation focuses on the capability of PTLMs on recognizing complex entities, and we take it as a key dimension for measuring the real-world fairness of PTLMs.

Table 4: Performance of baseline models on FUNSD and EC-FUNSD.

Task	Model	FUNSD	EC-FUNSD
SER	LayoutLMv3-base	90.85	82.30 (↓8.55)
	LayoutLMv3-large	91.70	83.88 (↓7.82)
	GeoLayoutLM	91.10	83.62 (↓7.48)
EL	LayoutLMv3-base	69.80	67.47 (↓2.33)
	LayoutLMv3-large	79.37	78.14 (↓1.23)
	GeoLayoutLM	88.06	86.18 (↓1.88)

5 Experiments

In this section, we report the results for the aforementioned evaluation on various PTLMs , together with detailed analyses. The implementation details are elaborated in Appendix E.

5.1 Analysis on Spurious Correlation of FUNSD

Tab. 4 shows the standard fine-tuning results of baseline models on FUNSD and EC-FUNSD, demonstrating than **spurious correlations between input features and labels have drawn significantly negative impact to the evaluation**. Although EC-FUNSD and FUNSD are very similar in almost all aspects, three models perform significantly worse on EC-FUNSD than on FUNSD, especially on the SER task, with a 7.48-8.55 drop in F1 scores on EC-FUNSD. We attribute the performance degradation to false overfitting on the block-level text region layout feature for training and inference. As illustrated in §B, FUNSD derives its segment and entity annotations directly from blocks, which leads to considerable bias when fine-tuning on this dataset. For SER, tokens within the same entity have exactly the same xy-

Table 5: Performance of baseline models in generalization evaluation.

(a) Performance on FUNSD.

Task	Model	Ori.	Generalization	Adaptation
SER	LayoutLMv3-base	90.85	22.75 (↓68.10)	81.06 (↓9.79)
	LayoutLMv3-large	91.70	19.92 (↓71.78)	83.23 (↓8.47)
	GeoLayoutLM	91.10	19.47 (↓71.63)	81.70 (↓9.40)
EL	LayoutLMv3-base	69.80	53.83 (↓15.97)	63.29 (↓6.51)
	LayoutLMv3-large	79.37	69.43 (↓9.94)	77.92 (↓1.45)
	GeoLayoutLM	88.06	59.66 (↓28.40)	84.98 (↓3.08)

(b) Performance on EC-FUNSD.

Task	Model	Ori.	Generalization	Adaptation
SER	LayoutLMv3-base	82.30	50.53 (↓31.77)	79.74 (↓2.56)
	LayoutLMv3-large	83.88	42.48 (↓41.40)	83.25 (↓0.63)
	GeoLayoutLM	83.62	38.04 (↓45.58)	82.38 (↓1.24)
EL	LayoutLMv3-base	67.47	61.97 (↓5.50)	64.80 (↓2.67)
	LayoutLMv3-large	78.14	72.47 (↓5.67)	77.34 (↓0.80)
	GeoLayoutLM	86.18	78.11 (↓8.07)	84.66 (↓1.52)

Table 6: Performance of baseline models in robustness evaluation.

(a) Performance on FUNSD.

Task	Model	Ori.	Robustness	AFT
SER	LayoutLMv3-base	90.85	61.41 (↓29.44)	83.22 (↓7.63)
	LayoutLMv3-large	91.70	60.67 (↓31.03)	85.58 (↓6.12)
	GeoLayoutLM	91.10	90.12 (↓0.98)	90.53 (↓0.57)
EL	LayoutLMv3-base	69.80	54.13 (↓15.67)	62.02 (↓7.78)
	LayoutLMv3-large	79.37	68.56 (↓10.81)	72.25 (↓7.12)
	GeoLayoutLM	88.06	87.13 (↓0.93)	86.07 (↓1.99)

(b) Performance on EC-FUNSD.

Task	Model	Ori.	Robustness	AFT
SER	LayoutLMv3-base	82.30	64.03 (↓18.27)	77.84 (↓4.46)
	LayoutLMv3-large	83.88	69.36 (↓14.52)	81.19 (↓2.69)
	GeoLayoutLM	83.62	82.93 (↓0.69)	82.75 (↓0.87)
EL	LayoutLMv3-base	67.47	60.29 (↓7.18)	61.39 (↓6.08)
	LayoutLMv3-large	78.14	71.77 (↓6.37)	72.35 (↓5.79)
	GeoLayoutLM	86.18	85.80 (↓0.38)	84.12 (↓2.06)

coordinate inputs, inducing the models to learn entity boundaries simply by determining whether the layout features between tokens are consistent or not. For EL, the entity representations are heavily dominated by the layout features, as all tokens within an entity share the same layout features, resulting in less attention to entity semantics. This experiment reveals the potential risk of current PTLMs that they may develop to excessively overfit the biased benchmarks, but the actual benefits brought by the advancements are suspicious.

Additionally, according to the experiment, **EC-FUNSD serves as a better indicator of model performance than FUNSD**. The performance gap between FUNSD and EC-FUNSD illustrates that the bias in FUNSD significantly influences its trustworthiness as an evaluation benchmark. In evaluation on FUNSD, since spurious correlations exist in both the training and validation sets, this issue is not reflected in the performance metrics. In contrast, the evaluation on EC-FUNSD effectively reveals the shortcomings: once the spurious correlations in the dataset are removed, the model performance would drop sharply.

5.2 Analysis on Generalization of PTLMs

Tab. 5 presents the generalization evaluation results of baseline PTLMs. Real-world generalization simulates scenarios where a deployed PTLM must handle input samples lacking segment-level layout features, thereby evaluating the ability of PTLMs to generalize to unseen input distributions. Real-world adaptation mimics scenarios where PTLMs must adjust to specific downstream tasks that lack segment-level features, evaluating the adaptability of PTLMs to real-world tasks with distributions differing from those seen during pre-training and fine-tuning.

According to the results, (1) **Despite achieving strong performance on prevailing benchmarks, current PTLMs fall short in real-world scenarios.** In generalization adaptation, the performance of PTLMs decline on both datasets. Notably, for the SER task on FUNSD, the F1 scores of three models drop significantly by 8-10%. (2) **EC-FUNSD exhibits its superiority as an evaluation benchmark.** In generalization adaptation, the performance of GeoLayoutLM drops significantly on FUNSD by over 9% and 3% on SER and EL, despite using both word- and segment-level layout features. We hypothesis that GeoLayoutLM overly relies on segment-level layout features after fine-tuning on FUNSD, while failing to fully leverage word-level layout and semantic features. This further highlights the negative impact of the spurious correlations in FUNSD for evaluation, casting doubt on the previous results. In contrast, the results on EC-FUNSD exhibit smaller drops under both settings, indicating that PTLMs trained on EC-FUNSD rely less on segment-level layout features. The results also verify that the raw performance of PTLMs on EC-FUNSD is not inflated by spurious correlations, and thus can serve as a more credible metric to reflect the real-world VrD-IE capabilities of PTLMs.

5.3 Analysis on Robustness of PTLMs

Tab. 6 presents the performance of baseline models in robustness evaluation. In real-world robustness evaluation, layout inputs are randomly perturbed during inference to simulate naturally distorted

samples, mimicking real-world challenging scenarios. Adversarial fine-tuning (AFT) perturbs layout features during both training and inference to evaluate the adaptability of PTLMs in real-world applications, where layout perturbations would be commonly encountered.

The result shows that **conducting adversarial training in pre-training phase brings more robustness improvement, rather than fine-tuning**. According to the results, LayoutLMv3 is significantly affected in both settings, while GeoLayoutLM exhibits little change. The performance degradation of GeoLayoutLM under robustness evaluation is even better than LayoutLM under adversarial fine-tuning. The robust performance of GeoLayoutLM can be attribute to its noise-tolerant pre-training, which introduces noise into the input layout features through the Poisson Line Segmentation algorithm – closely resembling our perturbation settings. Another notable observation is that GeoLayoutLM in EL task consistently performs better under robustness evaluation than under adversarial fine-tuning. Based on the results, we hypothesize that GeoLayoutLM has learned to prioritize semantic features while remaining robust to noise in layout features during pre-training. When handling the EL task which relies more on semantic features, GeoLayoutLM is not significantly affected by layout perturbations in robustness evaluation. However, during adversarial fine-tuning, the model may overfit the perturbed, low-quality training data, resulting in performance degradation. On the other hand, **EC-FUNSD serves as a clear and robust indicator for evaluating model performance**. We observe that models achieving better original performance on EC-FUNSD also exhibit less degradation in robustness evaluation. Additionally, the extent of performance drop on EC-FUNSD is consistently smaller than that on FUNSD. These two observations suggest that EC-FUNSD is a reliable benchmark for evaluation purpose.

5.4 Analysis on Fairness of PTLMs

Tab. 7 shows the performance of baseline models in fairness evaluation, from which we conclude that (1) **The fairness issue cannot be overlooked**, as the recall of complex entities is notably lower than the total recall among all baseline models. The inherent task difficulty and imbalance data distribution may attribute to this issue. (2) **The negative impact of spurious correlations in FUNSD on evaluation is also revealed in the fairness evaluation**, whereas **EC-FUNSD successfully addresses this issue, highlighting its advantages for evaluation**. Due to the spurious correlations, the performance of PTLMs on the complex entity subset of FUNSD does not decline significantly, failing to expose the fairness issue. In contrast, the performance on EC-FUNSD drop by nearly 30%, clearly revealing the deficiencies of PTLMs in recognizing complex entities.

Table 7: Performance of baseline models in fairness evaluation. "Total." and "Comp." denote the recall of all entities and the subset of complex entities, respectively.

Model	FUNSD		EC-FUNSD	
	Total.	Comp.	Total.	Comp.
LayoutLMv3-base	91.18	83.97 (↓7.21)	84.05	53.29 (↓30.76)
LayoutLMv3-large	93.76	89.10 (↓4.66)	86.65	59.53 (↓27.12)
GeoLayoutLM	93.31	88.46 (↓4.85)	86.84	58.79 (↓28.05)

6 Conclusion

In this paper, our aim is to provide the first comprehensive view on evaluating the capabilities of PTLMs to address real-world VrD-IE tasks. We question the real-world performance of prevalent PTLMs, even though they have achieved excellent performance on prevailing VrD-IE benchmarks, they may be working on flawed targets which hindered the development of this field. Consider that these flaws are only introduced from issues in dataset construction and evaluation methodology, we propose the new dataset EC-FUNSD and a multi-aspect evaluation to fill this critical gap, enabling a more accurate and realistic assessment of PTLM performance in practical applications. From the analysis of extensive experiments, we draw two key conclusions: (1) Spurious correlations between input features and labels exist in prevalent benchmarks, drawing negative impact to the evaluation. In particular, the strong performance achieved by previous models on these benchmarks may be unreliable and fail to reflect their true capabilities. The proposed EC-FUNSD dataset addresses the issues in previous benchmarks, and is proven suited to enable a fair and accurate evaluation. (2) Prevalent PTLMs tend to suffer from generalization, robustness, and fairness issues, hindering their effectiveness in real-world VrD-IE applications. We anticipate our work would suggest improvement directions for future models, and inspire future works to establish better evaluation pipelines to advance document AI.

Limitations

There are several aspects which may undermine the reliability of the main claims of our work. We are cautious on each aspect and manage to provide reasonable explanations for each of them.

1. The proposed dataset EC-FUNSD is not significantly larger than previous benchmarks. We construct EC-FUNSD based on the original images of FUNSD and therefore maintain its original scale as 199 samples. We believe the current scale of the dataset is sufficient to serve the evaluation aims of this study for the following reasons: (1) Prior work has widely adopted FUNSD as an evaluation benchmark, indicating a consensus that its dataset scale is sufficient to reflect diverse layout patterns in VrD-IE tasks. Similarly, our dataset features a comparable range of layout diversity at the same scale, ensuring the trustworthiness of experimental results. (2) Since our dataset is designed specifically for evaluation purposes, maintaining parity in data volume with FUNSD (which represents existing benchmarks) is essential to enable fair comparisons and guarantee the trustfulness of our conclusions.
2. There is less discussion of novel generalization or robustness improvement methods in this paper. In Sec. 5, we only attempt the vanilla adversarial training as the performance improvement method under generalization and robustness test. We believe that proposing new generalization or robustness improvement methods is not the main focus of this paper, as the contribution of proposing these methods alone may be limited. In the previous studies for addressing downstream tasks, the improvement achieved through merely enhancing the downstream task model is limited, due to the scarcity of domain-relevant data resources and the high cost of data annotation. In contrast, pre-trained models adapt to the tasks by designing corresponding pre-training tasks. These models are able to leverage massive document layout data covering various domains during pre-training to achieve a higher performance. As a result, it is acknowledged by the document AI community that the recent success of VrD-IE methods is mainly attributed to the advancement in PTLMs, thus the research focus of this field has shifted to the improvement of PTLMs, as stated in the introduction.
   We believe that a more promising direction is to design PTLMs with inherently stronger robustness and generalization capabilities, rather than incremental refinements to generalization or robustness techniques, since developing new methods on top of underperforming PTLMs is unlikely to fundamentally address the main concerns. With this in mind, the proposed comprehensive evaluation aims to inspire the community to focus on building more reliable and capable PTLMs.
3. Evaluation of generative models are not included. In this paper, we focus more on discriminative models to better meet the industrial demands. Although newly appeared layout-aware generative models (Luo et al., 2024; Lu et al., 2024) reveal an alternative roadmap to VrD-IE, PTLM-based methods remain the mainstream solution. Till now, generative methods still underperform PTLM-based methods on VrD-IE tasks (Zhang et al., 2024). Since our work focuses on application drawbacks in real-world scenarios, we prioritized PTLMs in our experiments.
4. Some of the evaluation settings are similar in form to previous works. For example, (Appalaraju et al., 2024; Zhang et al., 2023) conducted similar experiments to the proposed generalization evaluation setting. However, the objectives of the experiments differ. These works conducted the experiments as ablation studies to determine the best model configuration. However, our work aims to provide a holistic view on evaluating the comprehensive capabilities of PTLMs in real-world scenarios.

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.

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R Manmatha. 2024. Docformerv2: Local features for document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 709–718.

Rajas Bansal. 2022. A survey on bias and fairness in natural language processing. *arXiv preprint arXiv:2204.09591*.

Simon Caton and Christian Haas. 2024. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7).

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document ai: Benchmarks, models and applications. *arXiv preprint arXiv:2111.08609*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. 2023. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10583–10602.

Aparna Elangovan, Jiayuan He, Yuan Li, and Karin Verspoor. 2024. Principles from clinical research for nlp model generalization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2293–2309.

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.

He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2019. Eaten: Entity-aware attention for single shot visual text extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 254–259. IEEE.

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 *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4083–4091. ACM.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

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.

Feihu Jiang, Chuan Qin, Jingshuai Zhang, Kaichun Yao, Xi Chen, Dazhong Shen, Chen Zhu, Hengshu Zhu, and Hui Xiong. 2024. Towards efficient resume understanding: A multi-granularity multi-modal pre-training approach. *arXiv preprint arXiv:2404.13067*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Prashant Krishnan, Zilong Wang, Yangkun Wang, and Jingbo Shang. 2024. Towards few-shot entity recognition in document images: A graph neural network approach robust to image manipulation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16514–16526.

Xin Li, Yan Zheng, Yiqing Hu, Haoyu Cao, Yunfei Wu, Deqiang Jiang, Yinsong Liu, and Bo Ren. 2022. Relational representation learning in visually-rich documents. *arXiv preprint arXiv:2205.02411*.

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

Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R. Manmatha, and Vijay Mahadevan. 2023. Doctr: Document transformer for structured information extraction in documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19584–19594.

Zening Lin, Jiapeng Wang, Teng Li, Wenhui Liao, Dayi Huang, Longfei Xiong, and Lianwen Jin. 2024. Peneo: unifying line extraction, line grouping, and entity linking for end-to-end document pair extraction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5171–5180.

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

Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, and 1 others. 2024. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *arXiv preprint arXiv:2407.01976*.

Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023. Geolayoutlm: Geometric pre-training for visual information extraction. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640.

Pengyuan Lyu, Yulin Li, Hao Zhou, Weihong Ma, Xingyu Wan, Qunyi Xie, Liang Wu, Chengquan Zhang, Kun Yao, Errui Ding, and 1 others. 2024. Structextv3: An efficient vision-language model for text-rich image perception, comprehension, and beyond. *arXiv preprint arXiv:2405.21013*.

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, and 1 others. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *arXiv preprint arXiv:2210.06155*.

Abdellatif Sassioui, Rachid Benouini, Yasser El Ouargui, Mohamed El Kamili, Meriyem Chergui, and Mohammed Ouzzif. 2023. Visually-rich document understanding: Concepts, taxonomy and challenges. In *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pages 1–7.

Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalickỳ, Jiří Matas, Antoine Doucet, Mickaël Coustaty, and 1 others. 2023. Docile benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition*, pages 147–166. Springer.

Hongbin Sun, Zhanghui Kuang, Xiaoyu Yue, Chenhao Lin, and Wayne Zhang. 2021. Spatial dual-modality graph reasoning for key information extraction. *arXiv preprint arXiv:2103.14470*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

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*.

Yi Tu, Chong Zhang, Ya Guo, Huan Chen, Jinyang Tang, Huijia Zhu, and Qi Zhang. 2024. Uner: A unified prediction head for named entity recognition in visually-rich documents. *arXiv preprint arXiv:2408.01038*.

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021a. Towards robust visual information extraction in real world: new dataset and novel solution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2738–2745.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, and 1 others. 2021b. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2021c. Measure and improve robustness in nlp models: A survey. *arXiv preprint arXiv:2112.08313*.

Zifeng Wang, Zizhao Zhang, Jacob Devlin, Chen-Yu Lee, Guolong Su, Hao Zhang, Jennifer Dy, Vincent Perot, and Tomas Pfister. 2022. Queryform: A simple zero-shot form entity query framework. *arXiv preprint arXiv:2211.07730*.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021d. Layoutreader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, and 1 others. 2021a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of*

the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2579–2591.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. 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. 2021b. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *arXiv preprint arXiv:2104.08836*.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. 2022. Xfund: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3214–3224.

Huichen Yang and William Hsu. 2022. Transformer-based approach for document layout understanding. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 4043–4047.

Zhibo Yang, Rujiao Long, Pengfei Wang, Ro Song, Humen Zhong, Wenqing Cheng, Xiang Bai, and Cong Yao. 2023. Modeling entities as semantic points for visual information extraction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15358–15367.

Kaichun Yao, Jingshuai Zhang, Chuan Qin, Xin Song, Peng Wang, Hengshu Zhu, and Hui Xiong. 2023. Resuformer: Semantic structure understanding for resumes via multi-modal pre-training. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3154–3167. IEEE.

Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, and 1 others. 2023a. Icdar 2023 competition on structured text extraction from visually-rich document images. *arXiv preprint arXiv:2306.03287*.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023b. Structextv2: Masked visual-textual prediction for document image pre-training. *arXiv preprint arXiv:2303.00289*.

Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, and Shi-Min Hu. 2018. Chinese text in the wild. *arXiv preprint arXiv:1803.00085*.

Mingliang Zhai, Yulin Li, Xiameng Qin, Chen Yi, Qunyi Xie, Chengquan Zhang, Kun Yao, Yuwei Wu, and Yunde Jia. 2023. Fast-structext: An efficient hourglass transformer with modality-guided dynamic token merge for document understanding. *arXiv preprint arXiv:2305.11392*.

Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang, and Tao Gui. 2023. Reading order matters: Information extraction from visually-rich documents by token path prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13716–13730.

Chong Zhang, Yi Tu, Yixi Zhao, Chenshu Yuan, Huan Chen, Yue Zhang, Mingxu Chai, Ya Guo, Huijia Zhu, Qi Zhang, and Tao Gui. 2024. Modeling layout reading order as ordering relations for visually-rich document understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9658–9678, Miami, Florida, USA. Association for Computational Linguistics.

A Generalization, Robustness, and Fairness Evaluation for Deep Learning Systems

Considering that the focus of generalization, robustness and fairness may vary across different contexts, this section first specifies their concepts within this paper, and then introduces previous studies on the general principles and specific methods for the evaluation of these aspects.

Generalization Generalization refers to the ability of a trained model to make accurate predictions on new, unseen data. In real-world applications, the out-of-distribution (OOD) generalization issue leads to performance degradation as the data distribution in real-world applications may differ from the training distribution, highlighting the necessity of generalization evaluation. Hupkes et al. (2023) identified key dimensions of generalization evaluation, including compositional, structural, cross-task, cross-lingual, cross-domain, and robustness generalization. Elangovan et al. (2024) introduced internal and external validity as key measurements for generalization evaluation. Internal validity measures the capability of methods to capture cause-and-effect relationships, while external validity pertains to task-specific generalization. From this perspective, the OOD generalization challenge is categorized as an external validity challenge, and an evaluation framework for cause-and-effect modeling is proposed.

Robustness In this paper, we focus on the natural robustness of PTLMs, i.e. the robustness to natural distribution shifts. Ideally, the model performance should remain stable when the inputs are drawn from any distribution that is naturally occurring in real-world scenarios. Wang et al. (2021c) summarized the concept of natural robustness with two key assumptions: label-preserving and semantic-preserving. The label-preserving assumption implies that human predictions should remain unchanged despite perturbations in the input. The semantic-preserving assumption states that perturbed inputs should retain semantic consistency with the original inputs. Based on these assumptions, several studies have designed robustness evaluation falling into two categories: (1) those where the inputs change significantly while labels remain unchanged, and (2) those where the inputs remain relatively stable but the labels change. Wang et al. (2021b) introduced a unified multilingual robustness evaluation toolkit that integrates universal linguistically based text transformations, task-specific transformations, adversarial attacks, subpopulation analysis, and their combinations to provide comprehensive robustness assessment.

Fairness Our focus of fairness in this paper is on the equality of opportunity (group fairness) of PTLMs, which requires the model performance remains stable across any subgroup. (Chang et al., 2019; Bansal, 2022; Caton and Haas, 2024; Dong et al., 2023) have pointed out that the concerns of algorithmic fairness vary across different applications, influencing the design principles of fairness enhancement methods. According to these studies, common group fairness concepts include demographic parity, equality of odds, and equality of opportunity. Based on these concepts, these studies identified the potential drawbacks of current deep learning systems, scenarios that would induce bias, and introduced metrics to quantify these biases. They also proposed design principles and solutions to mitigate bias and improve fairness.

B Defect Analysis for Prevailing VrD-IE Benchmark Datasets

In previous works, popular datasets include FUNSD (Jaume et al., 2019), XFUND (Xu et al., 2022), CORD (Park et al., 2019), EPHOIE (Wang et al., 2021a), SROIE (Huang et al., 2019), FUNSD-r and CORD-r (Zhang et al., 2023) have been adopted in the evaluation of PTLMs. Despite the widespread use of these datasets, they still have several limitations that hinder them from being suitable and reliable for evaluating the VrD-IE capabilities of PTLMs. The defect analysis of these datasets highlights the necessity of establishing new VrD-IE benchmarks.

The most influential benchmarks are not well-suited to the VrD-IE evaluation due to their specific limitations. For FUNSD and XFUND, their annotation does not conform to the SER task settings. Specifically, these datasets are adapted from the visual information extraction task, and their annotations are organized by visual regions (i.e., blocks or segments) of the document layout, with the category labels and association relationships annotated on the block-level. However, the bounding box annotation of these blocks does not take the semantic of contents into account, resulting in massive cases where the contents within a block

may not correspond to an entity. In real-world scenarios, document layout annotations are usually generated by an off-the-shelf OCR engine, which focuses solely on visual features while ignoring semantic features. As a result, there are cases that a block contain multiple entities, or an entity spans across multiple blocks, as shown in Fig. 2. Nevertheless, current evaluation pipeline would use the block-level annotations to represent both segments and entities in SER and EL tasks, which introduces ambiguity and interferes with accurate performance evaluation. As shown in Fig. 1(1), the paragraph is split into two blocks, both labeled as "answer" and linked to the preceding "question" block, yet the proper annotation should treat the whole paragraph as a single semantic entity as the "answer" to the preceding "question". Due to this phenomenon, FUNSD and XFUND are not suitable for evaluation, as well as (Yang et al., 2023; Sun et al., 2021; Šimsa et al., 2023) which share the same issue. (Huang et al., 2019; Lin et al., 2024; Yu et al., 2023a) have tackled this issue by decoupling the annotation of blocks and entities. However, these datasets lack of word- or char-level layout annotations, and their entity annotations are still at block-level. According to Fig. 2, clearly indicating entities requires word-level annotations. Thus, these datasets are still unsuitable for reliably evaluating the VrD-IE capabilities of PTLMs. CORD and EPHOIE suffer from a lack of diversified layout formats and semantic entities. Most samples in CORD are receipts with similarly simple layouts, and the entities they contain are highly repetitive across samples – many samples even share identical entity content. Most samples in EPHOIE are long, narrow information columns from test papers with simple layouts and sparse text. Since EPHOIE was originally designed for key-value extraction tasks, each type of entity appears at most once in each sample, limiting the interactions between entities of the same type. Besides, as shown in Tab. 3, complex entities, such as those span multiple rows/columns or are interrupted by other contents, are lacking in these datasets. However, accessing the model's capability of recognizing such challenging entities is crucial for evaluation, indicating that CORD and EPHOIE are unsuitable for evaluating the VrD-IE capabilities of PTLMs. FUNSD-r and CORD-r acknowledge the importance of fine-grained annotations, offering decoupled char-level layout and entity annotations. However, these datasets adopted a special schema for entity annotation, which do not guarantee the continuity of entities contained in the word sequence. Therefore, these datasets cannot be completely tackled by sequence-labeling models and cannot directly be used to evaluate PTLMs. Besides, their layout annotations are generated by automated OCR engines and are of low quality.

To sum up, current benchmarks fall short of meeting the requirements for the evaluation. In this paper, our aim is to identify the problems of using existing PTLMs in real-world VrD-IE applications with quantifying the negative impacts. Thus, it is essential to propose new evaluation pipeline with new benchmarks to enable reliable and comprehensive evaluation.

C Human Annotation Process for EC-FUNSD

The two-step annotations for EC-FUNSD is illustrated as follows. In the first step, we constructed the word and segment-level layout annotations by manually revising the original layout annotations of each sample. We removed empty words in the original layout annotations, and marked unrecognizable handwritten words as "<unk>". We appended omitted words, associating them to appropriate existing segments or creating new segments when necessary. We rectified all errors on texts or bounding boxes of words and segments. We removed words that are of low resolution and deemed unimportant to the remaining contents, because their original annotations are erroneous and we are unable to correct these illegible words. We manually split multiple-row blocks within the original annotations into multiple segments, each segment confined in one row. After row-splitting, we combined segments that are tightly adjacent to each other into one segment. Throughout this process, we preserved the sequential order of words within each segment and also the mapping of words and segments from revised annotations to the old ones, to ensure the mapping of entity and relation annotations is preserved. After correcting layout annotations, in the second step, we revised the entity and relation annotations by manually transforming the block annotations to semantic entity annotations. We mapped the original block annotations to form the preliminary entity annotations for revision. After that, we corrected the annotation of entities being annotated

across multiple semantic blocks, and entities with missing words in their annotations. Additionally, we removed the invalid linking pairs and modified the corresponding linking pairs following the modifications made to the entity annotations. We generally preserved the segment order within the original annotations to ensure that each entity span is continuous in the annotations, making the form of this dataset suitable for sequence-labeling models. The annotating procedures above are carried out by two qualified annotators who are familiar with document AI.

D VrD-IE Task Formulation

The SER and EL tasks on document layouts are formalized as follows. A document layout with $N_{\mathcal{D}}$ words is represented as $\mathcal{D} = \{(w_i, \mathbf{b}_i)\}_{i=1,\ldots,N_{\mathcal{D}}}$, where w_i denotes the i-th word in document and $\mathbf{b}_i = (x_i^0, y_i^0, x_i^1, y_i^1)$ denotes the position of w_i in the document layout. The coordinates (x_i^0, y_i^0) and (x_i^1, y_i^1) correspond to the bottom-left and top-right vertex of w_i's bounding box, respectively. Given the predefined semantic entity types $\mathcal{E} = \{e_i\}_{i=1,\ldots,N_{\mathcal{E}}}$ and relation types $\mathcal{R} = \{r_i\}_{i=1,\ldots,N_{\mathcal{R}}}$, the semantic entities within document $\mathcal{D}$ is denoted as $s_{\mathcal{D}} = \{s_1, \ldots, s_J\}$, where the j-th entity $s_j = \{e_j, (j_1, j_2)\}$ is identified by its entity type $e_j \in \mathcal{E}$ and an index span (j_1, j_2) indicating the position of words in the inputs, satisfying $1 \le j_1 \le j_2 \le N_{\mathcal{D}}$. The set of relationships between entities of $s_{\mathcal{D}}$ is denoted as $t_{\mathcal{D}} = \{t_1, \ldots, t_K\}$, where the k-th relation triplet $t_k = \{r_k, (ks, ko)\}$ indicates that the relation between the subject s_{ks} and object s_{ko} is $r_k \in \mathcal{R}$ $(1 \le ks, ko \le J)$. It is guaranteed that there would be at most one relation triplet from one entity to another. The aim of SER is to recognize all the entity that spans together with their semantic categories, while EL aims to identify the possible relationship between two arbitrary entities, as well as type of relationship in the document.

E Implementation Details

We use two baseline PTLMs to be evaluated: LayoutLMv3 (Huang et al., 2022) as the most popular PTLM, and GeoLayoutLM (Luo et al., 2023) as the current state-of-the-art PTLM. LayoutLMv3 enhances the perception of visual signals by pre-training to align the inputs of modalities. GeoLayoutLM introduces multi-level geometric pre-training tasks to model spacial relationships between layout elements.

We fine-tune the two baseline models for SER and EL tasks on FUNSD and EC-FUNSD, using F1 score metrics for spurious correlations evaluation, generalization evaluation and robustness evaluation while using the recall for fairness evaluation. In spurious correlation detection, we fine-tune models in their normal way. In generalization evaluation, we use word-level layout as input instead of segment-level layout to simulate the distribution offset of real-world datasets. We measure the generalization ability of PTLMs on two settings, namely real-world generalization, where the distribution shifts of inputs are applied only during inference; and real-world adaptation, where the shifts are applied during both training and inference. In robustness evaluation, the layout input is perturbed to simulate the random deviation in the real-world scanning results. Specifically, we first simulate the falsely split text regions by sampling cutoff length with $Exp(0.1)$ distribution and splitting each segment with the sampled value. Next, each text region is centrally rotated $U(-5, 5)$ degrees to simulate rotational misalignment during scanning. Finally, an offset of $N(0, U(5, 20)^2)$ pixels is added to the upper, lower, left and right boundaries of each text region to simulate the deviation in the recognition result. Similar to generalization evaluation, we also measure the robustness of PTLMs on two settings, namely real-world robustness where the perturbations of inputs are applied only during inference, and adversarial fine-tuning (AFT) where the perturbations are applied during both training and inference. In the fairness experiment, we also fine-tune the models in the normal way, but evaluate them with their recall on the whole set of entities and subset of complex entities.

The detailed configuration of fine-tuning these models are further illustrated as follows. In experiments, we use the official implementation and pre-trained weights of LayoutLMv3-base [3] (Huang et al., 2022) and GeoLayoutLM [4] provided by their official GitHub repositories. It is important to note that the predefined maximum sequence length of textual tokens in both models is limited to 512. Therefore, when processing long documents that surpass this limit, LayoutLMv3 divides the

[3] https://github.com/microsoft/unilm/tree/master/layoutlmv3

[4] https://github.com/AlibabaResearch/AdvancedLiterateMachinery/tree/main/DocumentUnderstanding/GeoLayoutLM

document into several segments, whereas GeoLayoutLM truncates the content beyond the maximum length. Both means inevitably disrupt the integrity of EL labels in the documents, resulting in unfair comparison with other methods. To address this issue, we increase the maximum sequence length to 1024 by initializing the positional embedding of index 512-1023 by those of index 0-511 before fine-tuning. This adjustment guarantees none of the training or validation samples exceed the maximum length, and results in a slight difference compared to the results proposed in the original releases of the baselines.

The hyperparameters of fine-tuned the baseline models on FUNSD and EC-FUNSD for SER and EL tasks are reported as follows. In fine-tuning LayoutLMv3 for SER, we follow all the original setting of (Huang et al., 2022). In fine-tuning LayoutLMv3 for EL, we generally follow all the original setting with 400 epochs of fine-tuning. In fine-tuning GeoLayoutLM for SER and EL, for better performance, instead of following the original settings proposed in (Huang et al., 2022), we use an AdamW optimizer with 2% linear warming-up steps and a 1e-2 weight decay with a cosine scheduler. The learning rate and batch size were 1e-5 and 16 as the optimal configure searching from lr={8e-6, 1e-5, 1.5e-5, 2e-5} and bs={6, 16}. All models are fine-tuned by 500 epochs and the checkpoints with the best performance on SER and EL are kept. Generally, we ensure the consistency in fine-tuning and evaluating on FUNSD and EC-FUNSD, with the sole exception that we disabled the vision branch of GeoLayoutLM when fine-tuning on EC-FUNSD. In specific, we notice that the vision feature was only available in block-level in GeoLayoutLM, which directly contributes to the entity feature when fine-tuning FUNSD. However, a corresponding block-level vision representation is not available for every entity in EC-FUNSD, e.g. for the entities that span multiple rows and overlap with other entities in region. Therefore, the vision inputs are disabled in fine-tuning GeoLayoutLM on EC-FUNSD.