Less is More: Controlled Visual Evidence Routing and Redundancy Compression for Key Information Extraction

Yang Li[1,3] **Yajiao Wang**[1,3] **Wenhao Hu**[2]
Mengting Zhang[1,3] Zhixiong Zhang[1,3*]

[1]National Science Library, Chinese Academy of Sciences
[2]University of Electronic Science and Technology of China
[3]Department of Information Resources Management,
School of Economics and Management, University of Chinese Academy of Sciences

Abstract

Key Information Extraction (KIE) in visually-rich documents is inherently token-centric, yet prevailing multimodal encoders often fuse dense visual patches with text tokens indiscriminately, which can introduce low-density visual noise, intensify modality competition, and cause robustness collapse under distribution shifts. We propose **OTCR**, a lightweight and architecture-agnostic framework that turns vision from a competitor into a selective supporter for extraction. OTCR learns sparse, interpretable cross-modal coupling via optimal transport to route local visual evidence to the most relevant text tokens, applies token-level gating to control injection strength, and further suppresses spurious correlations through a variational information bottleneck. Experiments on FUNSD, CORD, and SROIE show consistent gains when OTCR is plugged into LayoutLMv3 and GeoLayoutLM, and ablations verify the complementary contributions of coupling, gating, and bottlenecking. Under distribution shifts from Do-GOOD(He et al., 2023a) and EC-FUNSD (Zhang et al., 2024), OTCR markedly mitigates performance degradation, indicating that controlled visual evidence can effectively compensate when text/layout shortcuts become unreliable.

1 Introduction

In visually rich document understanding, key information extraction (Cui et al., 2021) aims to recover structured semantics from unstructured document images, and has demonstrated substantial practical value in scenarios such as invoices, receipts, and reports or forms (Liu et al., 2019; Park et al., 2019; Jaume et al., 2019). Unlike pure text sequence modeling, visually rich documents contain complex two dimensional layouts, diverse typographic styles, and cross modal nonlinear alignment. To capture these heterogeneous signals, Transformer based multimodal pre-trained models (Xu et al., 2020b) introduce two dimensional spatial position

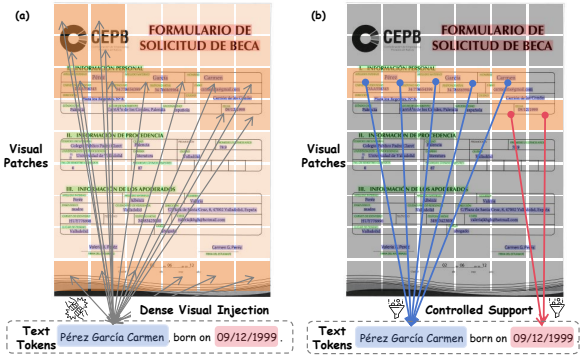

Figure 1: Dense multimodal fusion causes modality competition (a), whereas OTCR performs selective visual support via controlled cross-modal routing (b).

embeddings that explicitly associate text tokens with their bounding box coordinates, enabling the modeling of spatial topology in documents and substantially advancing related tasks. In recent years, research has further improved the upstream pretraining process to learn more general and more unified document representations (Li et al., 2024), thereby supporting a broad range of intelligent visual document tasks, including document classification, layout analysis, and document understanding.

However, what has received limited attention is that, for downstream key information extraction, the task is essentially a fine grained sequence labeling problem that focuses on local regions of a visually rich document. Its prediction ultimately relies on the final layer text representations of the encoder to perform per token classification, where vision serves as supporting evidence rather than the primary semantic carrier (Zhang et al., 2024). As a result, directly introducing page level image patches often incurs a drawback. For field level extraction, the visual evidence truly relevant to a given text token typically comes from a very small local region. In contrast, discretizing the entire page image into a large number of visual tokens inevitably introduces extensive irrelevant background

Proceedings of the 2nd Workshop on
Multimodal Augmented Generation via Multimodal Retrieval, pages 42–53

and spurious cues, including blank areas, textured backgrounds, decorative elements, table rules, and scanning artifacts, causing the visual modality to exhibit a pathological profile of low information density and high noise scale. More critically, visual tokens are dominant in quantity and often highly salient spatially, which can consume the limited attention budget during self attention interactions, thereby triggering modality competition and diluting text to text semantic interactions (Xie et al., 2026).

Furthermore, the aforementioned issues are significantly amplified under distribution shifts and real-world noise scenarios. The Do-GOOD (He et al., 2023a) study indicates that pre-trained visual document understanding models often exhibit a substantial performance gap between in-distribution and out-of-distribution settings. Fine-grained shift analyses reveal the vulnerability of existing models to factors such as template variations, scanning quality degradation, and error accumulation in optical character recognition. The root cause lies in the fact that when templates and layouts change and text or layout shortcuts become unreliable, the model increasingly needs to rely on visual cues for compensation. Concurrently, however, out-of-distribution scenarios introduce a significant increase in blurriness, compression artifacts, background textures, and noise tokens. This makes the visual modality much more susceptible to encompassing pseudo-correlated signals that are irrelevant to the extraction target. Consequently, this phenomenon amplifies the reliance on spurious cues and triggers a collapse in robustness.

To resolve these contradictions, we propose OTCR, a lightweight and architecture-agnostic controlled visual evidence injection framework tailored for Key Information Extraction. It is designed to transform the visual modality from an attention competitor into a selective supporter for text extraction. Our core strategy involves learning sparse and interpretable coupling relationships between text tokens and image patches to achieve structured routing of cross-modal evidence. This mechanism explicitly depicts which visual evidence should serve specific text tokens. Subsequently, we introduce a token-level gating mechanism to dynamically control the intensity of visual injection and suppress the interference of irrelevant visual signals on textual semantic modeling. Finally, we employ a Variational Information Bottleneck to filter and retain complementary information that truly contributes to the task. This enhances robustness without disrupting the text-dominant discriminative structure. Extensive experiments demonstrate that OTCR yields stable gains across multiple mainstream benchmarks and significantly mitigates performance degradation in stress tests involving distribution shifts and layout degradation.

The main contributions of this paper are as follows: (i)We propose OTCR, a lightweight and architecture-agnostic controlled visual evidence injection framework. Starting from the fine-grained nature of the KIE task, this framework successfully reshapes the visual modality from an attention competitor into a selective supporter of text semantics. (ii)We design a multi-level visual control and purification mechanism. By establishing sparse and interpretable cross-modal coupling, we achieve structured routing of visual evidence. Combined with a token-level gating mechanism to dynamically control injection intensity and a Variational Information Bottleneck (VIB) to deeply filter pseudo-correlated noise, we achieve precise retention of complementary information without disrupting the text-dominant structure. (iii)We conduct extensive experimental validation across multiple mainstream KIE benchmarks and backbones. Further out of distribution (OOD) / layout degradation stress tests and case analyses demonstrate that OTCR not only consistently improves task accuracy but also exhibits outstanding generalization capability and stable robustness when confronting complex noise and distribution shift scenarios.

2 Related Works

2.1 Key Information Extraction Method in Visually Rich Documents

Existing KIE methods can roughly be divided into four lines. Early **grid-based approaches** (Katti et al., 2018; Denk and Reisswig, 2019; Kerroumi et al., 2021; Dang et al., 2021) attempted to embed textual semantics directly into a 2D layout space, preserving both content and structure at the input level. LiuGraph (Liu et al., 2019) offered another perspective by modeling documents as **node–edge graphs**, shifting research attention toward more effective graph designs (Biescas et al., 2024; Zhang et al., 2022a; Tang et al., 2021). Subsequently, **large-scale pre-trained models** such as the LayoutLM (Xu et al., 2020b,a; Huang et al., 2022) series, together with more recent **instruction-driven**

MLLM methods (He et al., 2023b; Ye et al., 2023), have unified text, layout, and vision within a single framework, further advancing cross-task generalization. Despite methodological differences, these studies commonly aim to learn a strong multimodal representation to support downstream document intelligence. ViBERTgrid (Lin et al., 2021) integrates BERTgrid (Denk and Reisswig, 2019) with intermediate CNN layers to enable cross-modal interaction; GraphRevisedIE (Cao and Wu, 2023) employs graph revision techniques to combine multimodal embeddings with global contextual information; DocFormer (Appalaraju et al., 2021) leverages carefully designed multi-task unsupervised pre-training to enhance cross-modal alignment; and DocReL (Li et al., 2022) introduces relation consistency modeling to generate more effective relational representations.

2.2 Modality Interference Between Textual and Visual Tokens

Recent VrDU studies have noted that Transformer-style multimodal encoders can suffer from cross-modal interference when heterogeneous token streams, including text, layout, and visual patches, are processed jointly (Nguyen et al., 2021). In such settings, modalities do not contribute symmetrically, and visually salient yet semantically weak regions can disproportionately influence the shared representation space, making token-level linguistic cues that are crucial for extraction-oriented tasks harder to preserve (Zhai et al., 2023). Prior analyses describe this effect through token heterogeneity, visually dominant tokens, and multimodal sequence imbalance, and they consistently characterize the resulting modality competition as a practical bottleneck for stable document IE, particularly in scanned forms and receipts where fine textual distinctions are essential (Zhang et al., 2025).

A complementary line of work studies this issue from an efficiency and robustness perspective (He et al., 2023a). As image resolution increases, the number of visual tokens can grow rapidly, which lengthens multimodal sequences, amplifies interference, and makes cross-modal routing less reliable. Empirical findings further suggest that competition can persist even when visual tokens are not overwhelmingly more numerous, indicating that the issue is not purely a sequence-length effect but also stems from the heterogeneous semantics and salience of multimodal tokens (Toker et al., 2025). Representative directions include sparsifying document structures by pruning graph edges or restricting cross-modal connections, introducing contrastive or consistency objectives to suppress noisy correlations, and designing efficiency-aware tokenization to avoid excessive visual token accumulation (Rombach and Fettke, 2025). While these strategies improve stability in practice, they typically act as implicit regularization rather than providing an explicit mechanism to assign visual evidence to the most relevant textual units and to control how much visual information is retained, leaving room for lightweight and architecture-agnostic frameworks that support controlled evidence transmission and compression for KIE.

3 Proposed Method

3.1 Problem Formulation

Given a visually rich document D, the multimodal inputs consist of a sequence of textual tokens $T = \{t_i\}_{i=1}^{N}$ with corresponding spatial bounding boxes $B = \{b_i\}_{i=1}^{N}$, and a set of discrete visual patches $V = \{v_j\}_{j=1}^{M}$ extracted from the document image. Each bounding box $b_i = (x_i^0, y_i^0, x_i^1, y_i^1)$ provides the 2D geometric coordinates of t_i. The Key Information Extraction (KIE) task is fundamentally formulated as a fine-grained, token-level sequence labeling problem. For each text token t_i, the objective is to predict its corresponding entity label $y_i \in \mathcal{C}$, where $\mathcal{C}$ is a predefined semantic label space. Let $\mathcal{T} \in \mathbb{R}^{N\times d}$ and $\mathcal{V} \in \mathbb{R}^{M\times d}$ denote the initial textual and visual embeddings mapped into a shared feature space.

Standard unconstrained multimodal frameworks directly optimize a dense mapping $\hat{Y} = \Phi(\mathcal{T}, \mathcal{V})$, where textual tokens interact with visual patches in a largely unrestricted manner. However, page-level visual patches often contain low-density and task-irrelevant signals, including blank regions, background textures, decorative elements, table rules, and scanning artifacts. As a result, dense fusion can introduce noisy or spurious visual cues into token representations, increasing modality competition and weakening the fine-grained textual semantics required for token-level KIE.

To formalize a mathematically rigorous filtering process, our OTCR framework defines the extraction pipeline as a constrained latent variable model. Specifically, rather than directly fusing modalities, we aim to learn an intermediate, token-wise latent representation $\mathcal{Z} = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_N\}$ through a con-

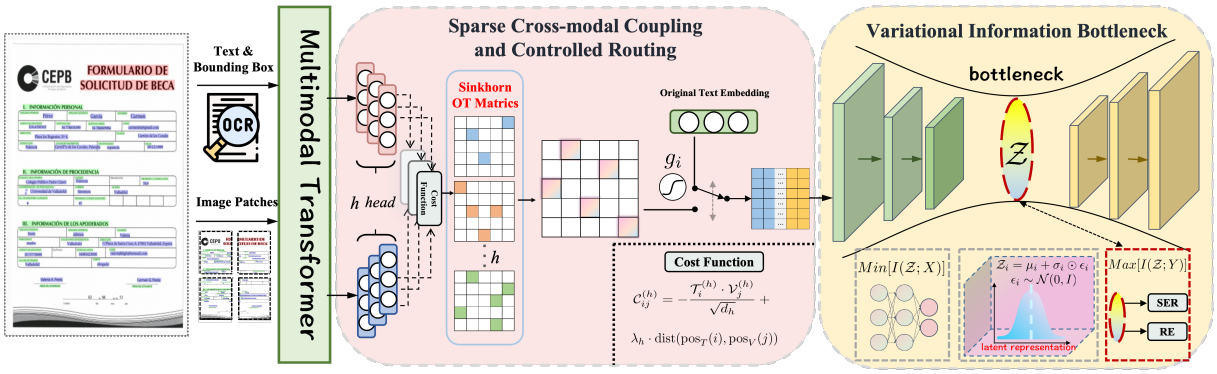

Figure 2: The overall framework of OTCR, which integrates sparse Optimal Transport coupling for controllable visual-to-text injection and a Variational Information Bottleneck for redundancy filtering and task-relevant representation learning.

trolled cross-modal routing function $g(\cdot)$:

$$\mathcal{Z} = g(\mathcal{T}, \mathcal{V}), \quad \hat{y}_i = f(\mathcal{Z}_i; \Theta), \tag{1}$$

where $f(\cdot)$ is the prediction network parameterized by Θ.

3.2 Overview

Our OTCR framework is illustrated in Figure 2. To explicitly prevent visual tokens from cannibalizing the attention budget, Section 3.3 introduces a **Sparse Cross-modal Coupling and Controlled Routing** mechanism. Based on Optimal Transport (OT), this module establishes interpretable alignment paths, guiding visual patches to be selectively injected into textual representations. A dynamic token-level gate is then employed to control the injection intensity, effectively shielding the text semantics from irrelevant visual noise. To further ensure robustness against distribution shifts (OOD), Section 3.4 proposes a **Variational Information Bottleneck (VIB)** training strategy. This module compresses the fused representation, filtering out out-of-distribution pseudo-features (compression artifacts) and retaining only the minimal, task-relevant complementary semantics.

3.3 Sparse Cross-modal Coupling and Controlled Routing

To obtain initial multimodal representations, we first employ an OCR system to extract textual tokens and their 2D bounding boxes. The tokens, layout coordinates, and document image are then fed into a multimodal Transformer, producing text-layout representations $\mathcal{T} \in \mathbb{R}^{N\times d}$ and visual patch representations $\mathcal{V} \in \mathbb{R}^{M\times d}$. Instead of injecting visual patches into text tokens through unconstrained dense fusion, we formulate visual-to-text routing as an entropy-regularized Optimal Transport (OT) problem, which provides a structured and interpretable coupling between textual tokens and visual patches. For the h-th head, we first project the two modalities into a shared subspace:

$$\mathcal{T}^{(h)} = \mathcal{T}\mathbf{W}_h^T, \quad \mathcal{V}^{(h)} = \mathcal{V}\mathbf{W}_h^V. \tag{2}$$

We then define the text-patch matching cost as:

$$\mathcal{C}_{ij}^{(h)} = -\frac{\mathcal{T}_i^{(h)}(\mathcal{V}_j^{(h)})^\top}{\sqrt{d_h}} + \lambda_h \operatorname{dist}\big(\operatorname{pos}_T(i), \operatorname{pos}_V(j)\big). \tag{3}$$

where the first term measures semantic affinity and the second term encourages spatially local coupling.

Given the text and visual marginal distributions $r \in \Delta^N$ and $c \in \Delta^M$, where r_i denotes the transport mass assigned to text token t_i and c_j denotes the transport mass assigned to visual patch v_j, we obtain the OT plan $\pi^{(h)} \in \mathbb{R}_+^{N\times M}$ by Sinkhorn normalization. In our implementation, we use uniform marginals by default, $r_i = 1/N$ and $c_j = 1/M$, so that each text token and visual patch contributes equally before task-driven routing. The transport plan satisfies the marginal constraints

$$\pi^{(h)}\mathbf{1}_M = r, \qquad (\pi^{(h)})^\top \mathbf{1}_N = c. \tag{4}$$

Specifically, the OT plan is computed as

$$\pi^{(h)} = \operatorname{diag}(\mathbf{u}) \exp\left(-\frac{\mathcal{C}^{(h)}}{\tau}\right) \operatorname{diag}(\mathbf{v}), \tag{5}$$

where $\mathbf{u}$ and $\mathbf{v}$ are Sinkhorn scaling vectors chosen to satisfy the above marginal constraints. The resulting plan is a soft coupling matrix rather than a

hard sparse assignment. A smaller τ encourages a more concentrated routing distribution.

We aggregate the plans from all heads and row-normalize them to obtain the final token-wise routing matrix:

$$\mathcal{P} = \mathrm{RowNorm}\left(\frac{1}{H}\sum_{h=1}^{H}\pi^{(h)}\right). \tag{6}$$

The OT-routed visual evidence for token i is computed as:

$$\mathcal{F}_{ot}(i) = \sum_{j=1}^{M}\mathcal{P}_{ij}\mathcal{V}_j. \tag{7}$$

To prevent unreliable visual evidence from overwhelming text semantics, we introduce an entropy-aware token-level gate. We first compute the normalized routing entropy:

$$H_i = -\frac{1}{\log M}\sum_{j=1}^{M}\mathcal{P}_{ij}\log(\mathcal{P}_{ij}+\epsilon). \tag{8}$$

A larger H_i indicates that the visual evidence is more dispersed and therefore less reliable. The gate is then defined as:

$$g_i = \sigma(\mathbf{W}_g[\mathcal{T}_i;\mathcal{F}_{ot}(i)] + b_g - \alpha H_i), \tag{9}$$

where $\alpha \geq 0$ controls the strength of entropy-based suppression.

Finally, the controlled representation is obtained by interpolating between the original text-layout representation and the routed visual evidence:

$$\mathcal{T}_i' = (1-g_i)\mathcal{T}_i + g_i\mathcal{F}_{ot}(i). \tag{10}$$

When the routed visual evidence is concentrated and reliable, the gate allows it to complement the textual representation. When the evidence is scattered or ambiguous, the gate suppresses visual injection and preserves the text-dominant representation.

3.4 Variational Information Bottleneck for Spurious Noise Suppression

While the OT-based gating effectively restricts the intensity of visual injection, the routed representations may still entangle with spurious cues (e.g., scanning artifacts, domain-specific background textures), which are the primary culprits for robustness collapse in OOD scenarios. To guarantee that the final representation is strictly task-oriented, we introduce an intermediate latent representation $\mathcal{Z}$ governed by the Information Bottleneck principle. It aims to be sufficient and minimal: maximizing task-relevant information $I(\mathcal{Z};Y)$ while strictly compressing redundant modality inputs $I(\mathcal{Z};X)$.

Taking the gated representation $\mathcal{T}_i'$ as input, a variational encoder parameterizes a Gaussian distribution for each token:

$$\mu_i = W_\mu\mathcal{T}_i' + b_\mu, \quad \log\sigma_i^2 = W_\sigma\mathcal{T}_i' + b_\sigma. \tag{11}$$

Using the reparameterization trick, we draw the stochastic latent representation:

$$\mathcal{Z}_i = \mu_i + \sigma_i \odot \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,I). \tag{12}$$

To enforce the bottleneck, the overall optimization process is driven by two components. First, the task supervision loss $\mathcal{L}_{task}$ acts as the variational lower bound to maximize $I(\mathcal{Z};Y)$, formulated as the cross-entropy:

$$\mathcal{L}_{task} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in\mathcal{C}} y_{i,c}\,\log\hat{y}_{i,c}. \tag{13}$$

Second, to explicitly formalize the compression of redundancy ($I(\mathcal{Z};X)$), we calculate the Information Bottleneck penalty $\mathcal{L}_{VIB}$ as the Kullback-Leibler (KL) divergence between the posterior distribution and an isotropic Gaussian prior $\mathcal{N}(0,I)$:

$$\mathcal{L}_{VIB} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{KL}\big(q(\mathcal{Z}_i|\mathcal{T}_i')\,\|\,\mathcal{N}(0,I)\big). \tag{14}$$

Finally, the overall optimization objective seamlessly incorporates both terms:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \beta\mathcal{L}_{VIB}, \tag{15}$$

where β is a hyperparameter balancing discriminative power and information compression. See section 4.4.3 for the analysis of β.

By forcing $\mu \to 0$ and $\sigma^2 \to 1$ via $\mathcal{L}_{VIB}$ for task-irrelevant dimensions, the network explicitly penalizes and discards the low-density visual noise and pseudo-features introduced during scanning or template variations. Consequently, the remaining dimensions in $\mathcal{Z}_i$ securely preserve the highly discriminative, text-dominant complementary information, thereby yielding stable KIE performance across both ID and challenging OOD environments.

3.5 Task-Specific Prediction and Inference

After obtaining the purified latent representation $\mathcal{Z}_i$ for each token, we employ lightweight task-specific heads to perform downstream Key Information Extraction, which typically comprises Semantic Entity Recognition (SER) and Relation Extraction (RE).

For the SER task, we apply a multi-layer perceptron (MLP) over the latent representation $\mathcal{Z}_i$ to project it into the predefined label space $\mathcal{C}$, followed by a Softmax activation to obtain the entity probability distribution $\hat{y}_i$:

$$\hat{y}_i = \text{Softmax}(\text{MLP}_{ser}(\mathcal{Z}_i)). \tag{16}$$

For the RE task, which aims to predict the directed linkage between a pair of tokens (key-value pairs), we construct a pairwise representation. Given the latent variables $\mathcal{Z}_i$ and $\mathcal{Z}_j$ of two candidate tokens, we concatenate them and feed them into a relation classifier to predict the linkage probability $\hat{r}_{i,j}$:

$$\hat{r}_{i,j} = \sigma(\text{MLP}_{re}\left([\mathcal{Z}_i; \mathcal{Z}_j]\right)), \tag{17}$$

where σ denotes the sigmoid function. The corresponding relation loss is jointly optimized with the entity classification loss $\mathcal{L}_{task}$ defined in Section 3.4.

Crucially, the behavior of our OTCR framework differs between the training and inference phases. During training, $\mathcal{Z}_i$ is stochastically sampled via the reparameterization trick to enforce the information bottleneck constraint and explore the latent space. However, during the deterministic inference phase, we disable the stochastic noise ϵ and directly utilize the predicted mean μ_i as the definitive latent representation ($\mathcal{Z}_i = \mu_i$). This ensures that the evaluation is stable and strictly relies on the highly selective, text-dominant semantics purified by our controlled routing mechanism.

4 Experiments

4.1 Experiment Settings

4.1.1 Datasets and Baselines

We evaluate OTCR on three standard KIE benchmarks: FUNSD (Jaume et al., 2019), CORD (Park et al., 2019), and SROIE (Huang et al., 2019), covering diverse scanned forms and real-world receipts. To demonstrate the architecture-agnostic nature of our framework, we integrate OTCR into two representative Transformer-based VrDU backbones: **LayoutLMv3** (Huang et al., 2022) (based on ViT-style patch processing) and **GeoLayoutLM** (Luo et al., 2023) (emphasizing geometric pre-training). We fine-tune these models for semantic entity recognition across all datasets and additionally report relation extraction results for FUNSD. Besides the base backbones, we compare OTCR with other representative methods, including graph-based **DocGraphLM** (Wang et al., 2023), structure-aware **LiLT** (Wang et al., 2022a). Following standard protocols, we report the entity-level F1 score for all experiments.

4.1.2 Implementation Details

We employ `LayoutLMv3-base` and `GeoLayoutLM-base` as our primary encoders. Input document images are resized to 224×224 for LayoutLMv3 consistent with its pre-training, while maintaining the original resolution for GeoLayoutLM's geometric extraction. To accommodate document length variations, we set the sequence length to 512 for LayoutLMv3 and 1024 for GeoLayoutLM. For the Sparse Cross-modal Coupling module, the number of optimal transport heads is set to $H = 12/16$, aligned with the backbone's attention heads. We set the entropy regularization coefficient $\tau = 0.1$ for the Sinkhorn algorithm to encourage sparse alignment, and the learnable spatial bias parameter λ is initialized to 1.0. For the Variational Information Bottleneck (VIB), the trade-off hyperparameter β is determined via sensitivity analysis (see Section 4.4.3). All models are optimized using AdamW with a weight decay of 1×10^{-2}. The learning rate is initialized at 2×10^{-5} for LayoutLMv3-based experiments and 1×10^{-5} for GeoLayoutLM-based experiments. We employ a linear warmup for the first 5% of training steps, followed by a linear decay schedule. The batch size is set to 16. Training is conducted for 100 epochs on FUNSD and 50 epochs on CORD and SROIE to ensure full convergence. All experiments are executed on a single NVIDIA A100 (80GB) GPU, with fixed random seeds to ensure reproducibility.

4.2 Main Results

We evaluated OTCR on three KIE benchmarks: FUNSD, CORD, and SROIE, using LayoutLMv3 and GeoLayoutLM as backbones. Table 1 reports the performance of OTCR on three standard benchmarks. The data explicitly shows that integrating

Table 1: Main results on KIE benchmarks (FUNSD, CORD, and SROIE). OTCR is plugged into two representative VrDU backbones, LayoutLMv3 and GeoLayoutLM, and evaluated under the standard SER/RE setting. Best results are marked in **bold** and second-best results are underlined. Results reproduced by us are marked with *.

Type	Method	#Params	FUNSD		CORD	SROIE
			SER	RE	SER	SER
Graph	DocGraphLM (Wang et al., 2023)	base	88.77	-	96.93	-
	GraphDoc (Zhang et al., 2022b)	265M	87.77	-	96.93	98.02*
	mmLayout (Wang et al., 2022b)	large	86.49	-	97.38	97.91
	FormNet (Lee et al., 2022)	large	84.69	-	97.28	-
	DocFormer (Appalaraju et al., 2021)	502M	84.55	-	96.99	-
	MatchVIE (Tang et al., 2021)	base	81.33	-	-	96.57
	RE^2 (Ramu et al., 2024)	base	-	71.76	-	-
	Doc2Graph (Gemelli et al., 2022)	base	-	53.36	-	-
	GraphLayoutLM (Li et al., 2023)	372M	-	-	97.86*	-
Attention	LayoutLMv3 (Huang et al., 2022)	368M	91.81*	79.67*	97.02*	96.10*
	LayoutLMv3 (Huang et al., 2022)	133M	90.85	69.80	95.95*	94.73*
	BROS (Hong et al., 2022)	340M	84.52	77.01	97.28	96.62
	DocTr (Liao et al., 2023)	153M	84.0	73.9	98.2	-
	LiLT (Wang et al., 2022a)	base	88.41	62.76	96.07	-
	LAGaBi (Zhu et al., 2023)	133M	91.00	-	97.05	-
	SERA (Zhang et al., 2021)	base	-	65.96	-	-
	SPADE (Hwang et al., 2021)	base	72.0	41.3	-	-
Pre-trained	GeoLayoutLM (Luo et al., 2023)	399M	91.10	88.06	98.23*	96.93*
	Wukong-Reader (Bai et al., 2023)	470M	**93.62**	-	97.27	**98.15**
	LayoutMask (Tu et al., 2023)	404M	93.20	-	97.19	97.27
	Bi-VLDoc (Luo et al., 2022)	409M	93.44	-	97.84	-
	ERNIE-Layout (Peng et al., 2022)	large	93.12	-	97.21	97.55
	DocReL (Li et al., 2022)	142M	-	46.1	97.0	-
	StrucTexT (Li et al., 2021)	107M	83.09	44.1	-	96.88
Ours	OTCR-LayoutLMv3	133M+30	91.95 (↑1.10)	72.18 (↑2.38)	97.01 (↑1.06)	95.18 (↑0.45)
		368M+30	92.33 (↑0.52)	81.17 (↑1.50)	97.35 (↑0.33)	96.93 (↑0.83)
	OTCR-GeoLayoutLM	399M+30	93.12 (↑2.02)	**88.75** (↑0.69)	**98.63** (↑0.40)	98.08 (↑1.15)

OTCR into both LayoutLMv3 and GeoLayoutLM consistently improves results across all datasets.

On the FUNSD dataset, OTCR-GeoLayoutLM achieved the highest performance in RE 88.75% and competitive SER performance, surpassing other methods, including GeoLayoutLM 91.10% and LAGaBi 91.00%. Similarly, OTCR-LayoutLMv3 improved the SER score to 91.95%, outperforming LayoutLMv3 90.85% and showing the framework's ability to enhance performance even with a smaller model size (133M parameters). For CORD, OTCR-GeoLayoutLM achieved 98.63% in SER, outperforming GeoLayoutLM 98.23% and LayoutLMv3 97.02%. OTCR-LayoutLMv3 (368M) also showed notable improvement, reaching 97.35% in SER, a 0.33% point increase over the baseline. These results indicate that OTCR enhances the ability of the model to extract relevant visual information while minimizing noise, even in complex document layouts such as receipts. In SROIE, OTCR-GeoLayoutLM achieved a SER score of 98.08%, which is very competitive compared to other methods such as LayoutLMv3 96.10% and Wukong-Reader 98.15%. The results confirm OTCR's effectiveness in improving KIE performance, particularly in noisy environments and complex document structures like receipts.

4.3 Ablation Study

The ablation study in Table 2 systematically evaluates the impact of key components in the OTCR framework: Optimal Transport (OT), Gate mechanism, and Variational Information Bottleneck (VIB). For both LayoutLMv3-large and GeoLayoutLM, the results reveal that excluding OT significantly impairs performance, particularly in SER and RE tasks, underscoring the critical role of cross-modal coupling for effective text-visual alignment. The Gate mechanism also contributes notably to performance, especially in reducing visual noise and ensuring relevant visual features are integrated effectively. Its removal leads to further declines, particularly in RE on FUNSD.

While the VIB mechanism provides additional performance benefits by filtering redundant visual features, its absence causes a smaller drop in performance compared to OT and Gate, especially on tasks like SROIE. These findings highlight that OT,

Table 2: Ablation study of OTCR components on two backbones: LayoutLMv3-large and GeoLayoutLM. We report SER on FUNSD/CORD/SROIE and RE on FUNSD. Best results are marked in **bold** and second-best results are underlined.

Backbone	Components			FUNSD		CORD	SROIE
	OT	Gate	VIB	SER	RE	SER	SER
LayoutLMv3-large	–	–	–	91.81	79.67	97.02	96.10
	–	–	✓	91.76	79.69	97.09	96.22
	✓	–	–	91.99	79.94	97.13	96.31
	✓	✓	–	92.21	80.95	97.19	96.55
	✓	–	✓	92.12	80.46	97.10	96.70
	✓	✓	✓	**92.33**	**81.17**	**97.35**	**96.93**
GeoLayoutLM	–	–	–	91.10	88.06	98.23	96.93
	–	–	✓	91.40	88.16	98.22	96.98
	✓	–	–	91.52	88.13	98.35	97.22
	✓	✓	–	92.70	88.69	98.58	97.91
	✓	–	✓	92.43	88.23	98.46	97.63
	✓	✓	✓	**93.12**	**88.75**	**98.63**	**98.08**

Gate, and VIB all play vital roles in enhancing KIE performance, with OT being the most influential in enabling effective cross-modal interaction. The full OTCR model consistently outperforms all ablated versions, demonstrating that the combination of these components is essential for maximizing model accuracy, particularly in complex and noisy document layouts.

4.4 Further Analysis

4.4.1 Robustness Evaluation

In the main experiments, OTCR consistently improves SER and RE on standard KIE benchmarks, showing that controlled visual evidence injection is beneficial under conventional in-distribution evaluation. However, real-world document extraction often involves corrupted text, manual edits, template variation, or different annotation protocols. We therefore conduct two complementary evaluations in Table 3. For distribution-shift robustness, models are fine-tuned on FUNSD and evaluated on FUNSD, OOD_H (human-intervened documents), and OOD_T (text corruption) from Do-GOOD (He et al., 2023a). For annotation-protocol evaluation, we report results under the supervised EC-FUNSD setting (Zhang et al., 2024), which re-annotates FUNSD from an entity-centric perspective.

The results show that OTCR brings clear gains under Do-GOOD shifts, especially when text/layout cues are disrupted. For LayoutLMv3-base, OTCR improves SER F1 by 5.88 points on OOD_H and 3.23 points on OOD_T. For GeoLayoutLM, it improves OOD_H by 5.25 points. These gains

Table 3: **Robustness and entity-centric evaluation.** Models are fine-tuned on FUNSD and evaluated on FUNSD, OOD_H, and OOD_T for distribution-shift evaluation. For EC-FUNSD, models are evaluated under the supervised entity-centric annotation setting. We report SER F1, with values in parentheses denoting absolute gains over the corresponding baseline.

Backbone	Model	FUNSD	OOD_H	OOD_T	EC-FUNSD
LayoutLMv3 (base)	Baseline	90.85	73.25	86.82	82.30
	OTCR	91.95 (+1.10↑)	79.13 (+5.88↑)	90.05 (+3.23↑)	83.56 (+1.26↑)
LayoutLMv3 (large)	Baseline	91.81	80.16	87.95	83.88
	OTCR	92.33 (+0.52↑)	84.33 (+4.17↑)	87.98 (+0.03↑)	83.27 (-0.61↓)
GeoLayoutLM	Baseline	91.10	84.26	89.37	83.62
	OTCR	93.12 (+2.02↑)	89.51 (+5.25↑)	90.14 (+0.77↑)	85.30 (+1.68↑)

support our central claim: selectively routed visual evidence can act as a useful complement when textual or layout shortcuts become unreliable. Under the supervised EC-FUNSD setting, OTCR also improves LayoutLMv3-base and GeoLayoutLM, but slightly decreases LayoutLMv3-large. This suggests that controlled visual injection is generally helpful, but its benefit depends on the backbone and the type of shift. Overall, OTCR provides a lightweight mechanism for making visual evidence supportive rather than competitive in KIE.

4.4.2 Comparison with Large-Parameter Models.

Table 4 compares OTCR with representative large-parameter models, including zero-shot multimodal

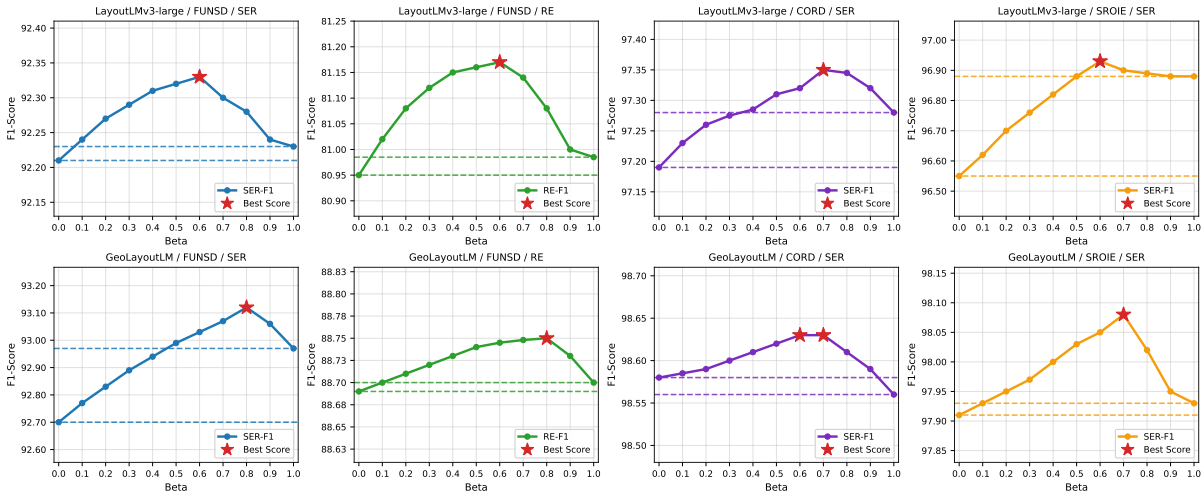

Figure 3: Sensitivity Analysis of β.

Table 4: **Main comparison with large-parameter models.**

Model	FUNSD	CORD	Params	Setting
GPT-5.1	86.78	94.12	–	Zero-shot
Qwen3-VL	79.64	89.93	235B	Zero-shot
Gemini 2.5 Flash	85.52	93.75	–	Zero-shot
LayoutLLM	95.30	98.64	6914.38M	Fine-tune
OTCR-GeoLayoutLM	**93.12**	**98.63**	**399M**+30	**Fine-tune**

LLMs and a finetuned large document model. Despite being built on a lightweight 399M-parameter backbone, **OTCR-GeoLayoutLM** achieves strong finetuned performance, reaching 93.12% on FUNSD and 98.63% on CORD, which is competitive with much larger specialized models (LayoutLLM at 95.30%/98.64% with 6.9B parameters) and substantially exceeds the zero-shot results of MLLMs such as GPT-5.1 (86.78%/94.12%), Gemini 2.5 flash (85.52%/93.75%), and Qwen3-VL (79.64%/89.93%). This highlights that, for extraction-centric VrDU, controlled and task-aligned visual evidence routing can be more effective than scaling alone, enabling small-to-mid scale backbones to approach the performance of multi-billion-parameter systems under supervised finetuning.

4.4.3 Sensitivity Analysis of β

We further analyze the sensitivity of OTCR to the VIB trade-off coefficient β by sweeping $\beta \in [0, 1]$ and reporting the resulting SER/RE curves for LayoutLMv3-large and GeoLayoutLM on FUNSD, CORD, and SROIE (Figure 3). Overall, the performance follows a consistent rise–plateau–slight-drop pattern: increasing β from 0 (no bottleneck) yields steady gains, suggesting that moderate information compression effectively filters redundant multimodal signals and stabilizes token representations, whereas overly large β starts to degrade accuracy due to over-compression. Notably, the optimal β varies slightly across backbones and datasets, reflecting different noise levels and modality redundancy in forms versus receipts; nevertheless, the best-performing region is consistently concentrated around $\beta \in [0.6, 0.8]$ for most settings (peaks marked by $\star$), indicating that OTCR is not overly sensitive to precise tuning and admits a broad, transferable operating range in practice.

5 Conclusion

We propose OTCR, a lightweight and architecture-agnostic framework for controlled visual evidence injection in Key Information Extraction. By introducing sparse optimal transport-based cross-modal routing, token-level gating, and variational information bottleneck compression, OTCR explicitly models modality asymmetry and reshapes vision into a selective supporter of text semantics. Extensive experiments show consistent gains across multiple benchmarks and backbones, with notably improved robustness under distribution shifts.

Limitation

While OTCR demonstrates consistent improvements across multiple benchmarks and backbones, the present study still has several limitations. Our work mainly focuses on extraction-oriented document understanding, and the applicability of con-

trolled visual evidence routing to other visually rich document understanding tasks remains for future investigation. In addition, although OTCR is lightweight and architecture-agnostic, it still introduces several tunable components, and further simplification of the framework may improve its practicality in deployment.

Acknowledgments

The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

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.

Haoli Bai, Zhiguang Liu, Xiaojun Meng, Li Wentao, Shuang Liu, Yifeng Luo, Nian Xie, Rongfu Zheng, Liangwei Wang, Lu Hou, et al. 2023. Wukong-reader: Multi-modal pre-training for fine-grained visual document understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13386–13401.

Nil Biescas, Carlos Boned, Josep Lladós, and Sanket Biswas. 2024. Geocontrastnet: Contrastive key-value edge learning for language-agnostic document understanding. In *International Conference on Document Analysis and Recognition*, pages 294–310. Springer.

Panfeng Cao and Jian Wu. 2023. Graphrevisedie: Multimodal information extraction with graph-revised network. *Pattern Recognition*, 140:109542.

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

Tuan Anh Nguyen Dang, Duc Thanh Hoang, Quang Bach Tran, Chih-Wei Pan, and Thanh Dat Nguyen. 2021. End-to-end hierarchical relation extraction for generic form understanding. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5238–5245. IEEE.

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

Andrea Gemelli, Sanket Biswas, Enrico Civitelli, Josep Lladós, and Simone Marinai. 2022. Doc2graph: a task agnostic document understanding framework based on graph neural networks. *arXiv preprint arXiv:2208.11168*.

Jiabang He, Yi Hu, Lei Wang, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. 2023a. Do-good: towards distribution shift evaluation for pre-trained visual document understanding models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 569–579.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023b. Icl-d3ie: In-context learning with diverse demonstrations updating for document information extraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19485–19494.

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.

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.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021. Spatial dependency parsing for semi-structured document information extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343.

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.

Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2d documents. *arXiv preprint arXiv:1809.08799*.

Mohamed Kerroumi, Othmane Sayem, and Aymen Shabou. 2021. Visualwordgrid: information extraction from scanned documents using a multimodal approach. In *International Conference on Document Analysis and Recognition*, pages 389–402. Springer.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3735–3754, Dublin, Ireland. Association for Computational Linguistics.

Qiwei Li, Zuchao Li, Xiantao Cai, Bo Du, and Hai Zhao. 2023. Enhancing visually-rich document understanding via layout structure modeling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4513–4523.

Qiwei Li, Zuchao Li, Ping Wang, Haojun Ai, and Hai Zhao. 2024. Hypergraph based understanding for document semantic entity recognition. *arXiv preprint arXiv:2407.06904*.

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. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4614–4624.

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.

Weihong Lin, Qifang Gao, Lei Sun, Zhuoyao Zhong, Kai Hu, Qin Ren, and Qiang Huo. 2021. Vibertgrid: a jointly trained multi-modal 2d document representation for key information extraction from documents. In *International Conference on Document Analysis and Recognition*, pages 548–563. Springer.

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

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, Guozhi Tang, Qi Zheng, Cong Yao, Lianwen Jin, Chenliang Li, Yang Xue, and Luo Si. 2022. Bi-vldoc: Bidirectional vision-language modeling for visually-rich document understanding. *arXiv preprint arXiv:2206.13155*.

Laura Nguyen, Thomas Scialom, Jacopo Staiano, and Benjamin Piwowarski. 2021. Skim-attention: Learning to focus via document layout. *arXiv preprint arXiv:2109.01078*.

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*.

Pritika Ramu, Sijia Wang, Lalla Mouatadid, Joy Rimchala, and Lifu Huang. 2024. Re2: Region-aware relation extraction from visually rich documents. 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 8723–8739.

Alexander Michael Rombach and Peter Fettke. 2025. Deep learning based key information extraction from business documents: Systematic literature review. *ACM Computing Surveys*, 58(2):1–37.

Guozhi Tang, Lele Xie, Lianwen Jin, Jiapeng Wang, Jingdong Chen, Zhen Xu, Qianying Wang, Yaqiang Wu, and Hui Li. 2021. Matchvie: Exploiting match relevancy between entities for visual information extraction. *arXiv preprint arXiv:2106.12940*.

Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik, and Yonatan Belinkov. 2025. Padding tone: A mechanistic analysis of padding tokens in t2i models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7618–7632.

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*.

Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, Kang Gu, and Sameena Shah. 2023. Docgraphlm: documental graph language model for information extraction. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 1944–1948.

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

Wenjin Wang, Zhengjie Huang, Bin Luo, Qianglong Chen, Qiming Peng, Yinxu Pan, Weichong Yin, Shikun Feng, Yu Sun, Dianhai Yu, et al. 2022b. mmlayout: Multi-grained multimodal transformer for document understanding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4877–4886.

Tingwei Xie, Jinxin He, and Yonghong Song. 2026. Roap: A reading-order and attention-prior pipeline for optimizing layout transformers in key information extraction. *arXiv preprint arXiv:2601.05470*.

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.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

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, Yixi Zhao, Yulu Xie, Chenshu Yuan, Yi Tu, Ya Guo, Mingxu Chai, Ziyu Shen, Yue Zhang, and Qi Zhang. 2024. Unveiling the deficiencies of pre-trained text-and-layout models in real-world visually-rich document information extraction. *arXiv preprint arXiv:2402.02379*.

Jiaxin Zhang, Wentao Yang, Songxuan Lai, Zecheng Xie, and Lianwen Jin. 2025. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9923–9932.

Yue Zhang, Zhang Bo, Rui Wang, Junjie Cao, Chen Li, and Zuyi Bao. 2021. Entity relation extraction as dependency parsing in visually rich documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2759–2768.

Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. 2022a. Multimodal pre-training based on graph attention network for document understanding. *IEEE Transactions on Multimedia*, 25:6743–6755.

Zhenrong Zhang, Jiefeng Ma, Jun Du, Licheng Wang, and Jianshu Zhang. 2022b. Multimodal pre-training based on graph attention network for document understanding. *Trans. Multi.*, 25:6743–6755.

Xi Zhu, Xue Han, Shuyuan Peng, Shuo Lei, Chao Deng, and Junlan Feng. 2023. Beyond layout embedding: Layout attention with gaussian biases for structured document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7773–7784.