NEURAL INFORMATION PROCESSING SYSTEMS

Unified Pretraining Framework for Document Understanding

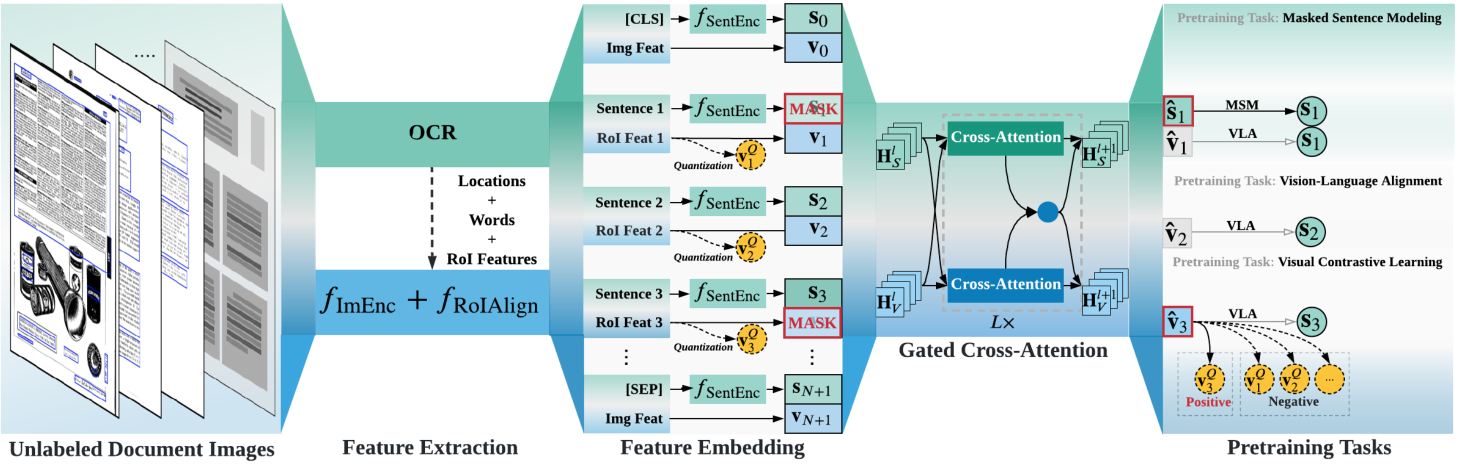
Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Nikolaos Barmpalios, Rajiv Jain, Ani Nenkova, Tong Sun

Adobe Research,
Adobe Document Cloud

Adobe®

Adobe Document Cloud

Problem and Contribution



1. We introduce ***UniDoc***, a powerful pretraining ***framework*** for document understanding. ***UniDoc*** is capable of learning contextual ***textual and visual*** information and ***cross-modal correlations*** within a ***single framework***, which leads to better performance.
2. We present ***Masked Sentence Modeling*** for language modeling, ***Visual Contrastive Learning*** for vision modeling, and ***Vision-Language Alignment*** for pretraining.
3. Extensive experiments and analysis provide useful insights on the effectiveness of the pretraining tasks and show outstanding performance on various downstream tasks.

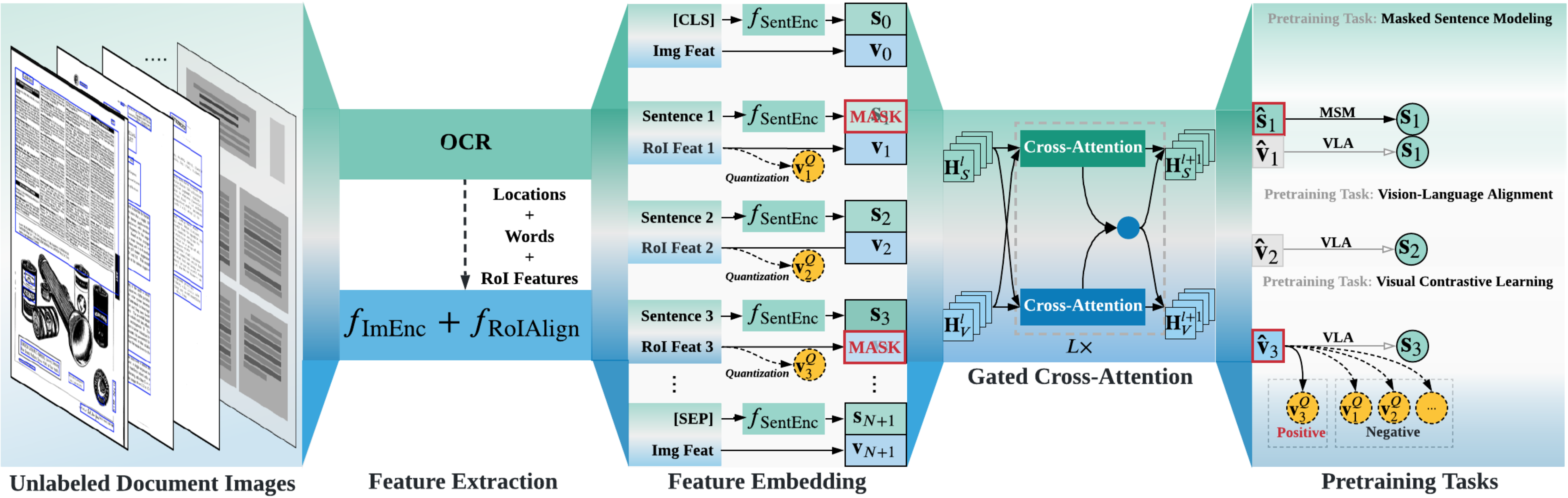
Motivation

1. Documents are composed of semantic regions

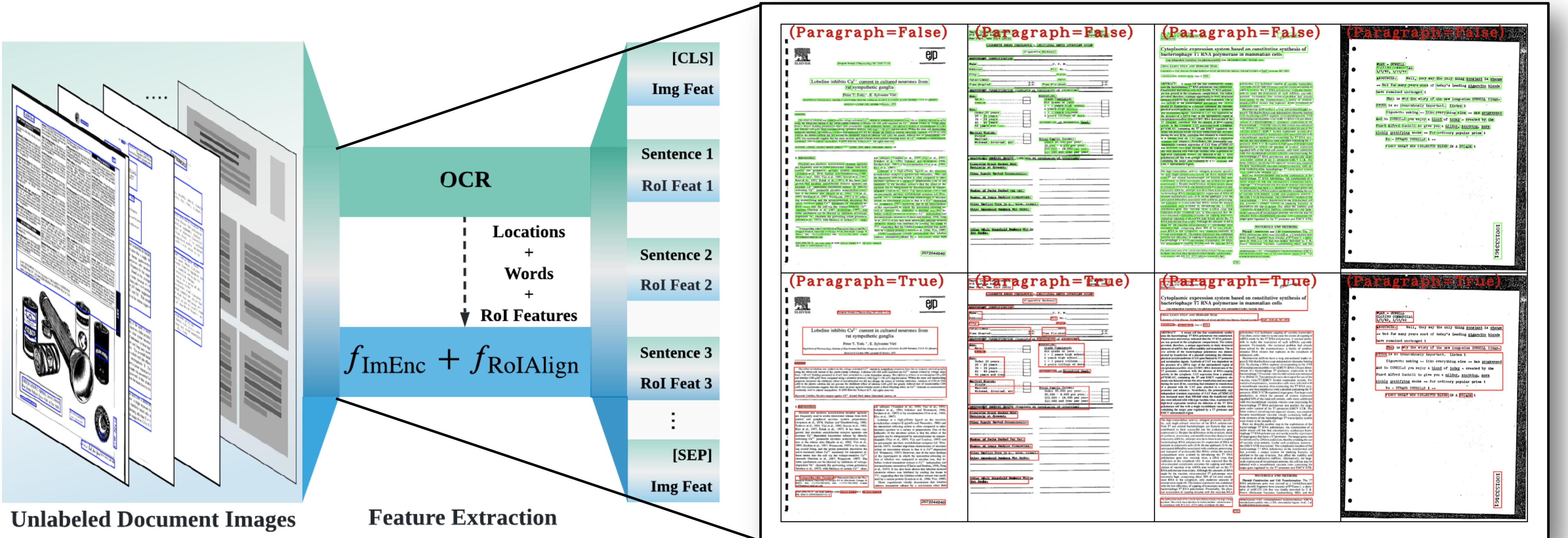
2. Documents are more than words

3. Documents have spatial layout

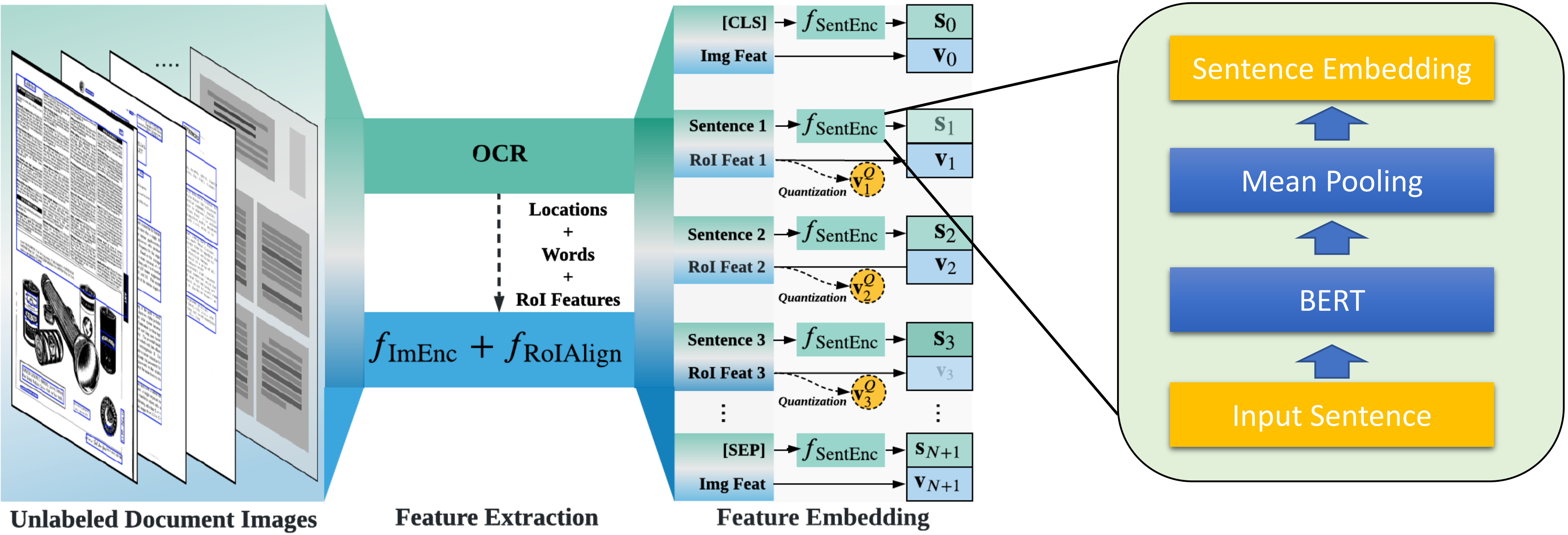
Framework


Unlabeled Document Images
OCR
Locations + Words + RoI Features
$f_{\text{ImEnc}} + f_{\text{RoIAlign}}$
Feature Extraction
[CLS]
Img Feat
Sentence 1
RoI Feat 1
Sentence 2
RoI Feat 2
Sentence 3
RoI Feat 3
[SEP]
Img Feat
f_{SentEnc}
Quantization
MASK
Feature Embedding
Cross-Attention
Cross-Attention
$L\times$
Gated Cross-Attention
Pretraining Task: Masked Sentence Modeling
MSM
VLA
Pretraining Task: Vision-Language Alignment
VLA
Pretraining Task: Visual Contrastive Learning
VLA
Positive
Negative
Pretraining Tasks

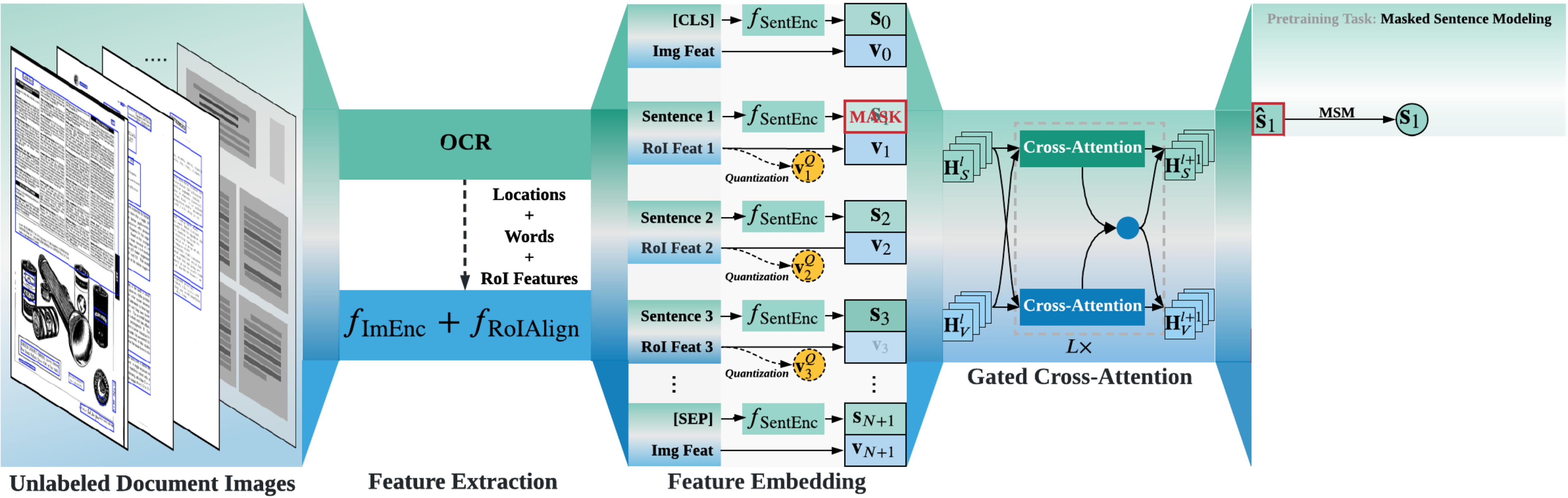
Feature Extraction



Feature Embedding

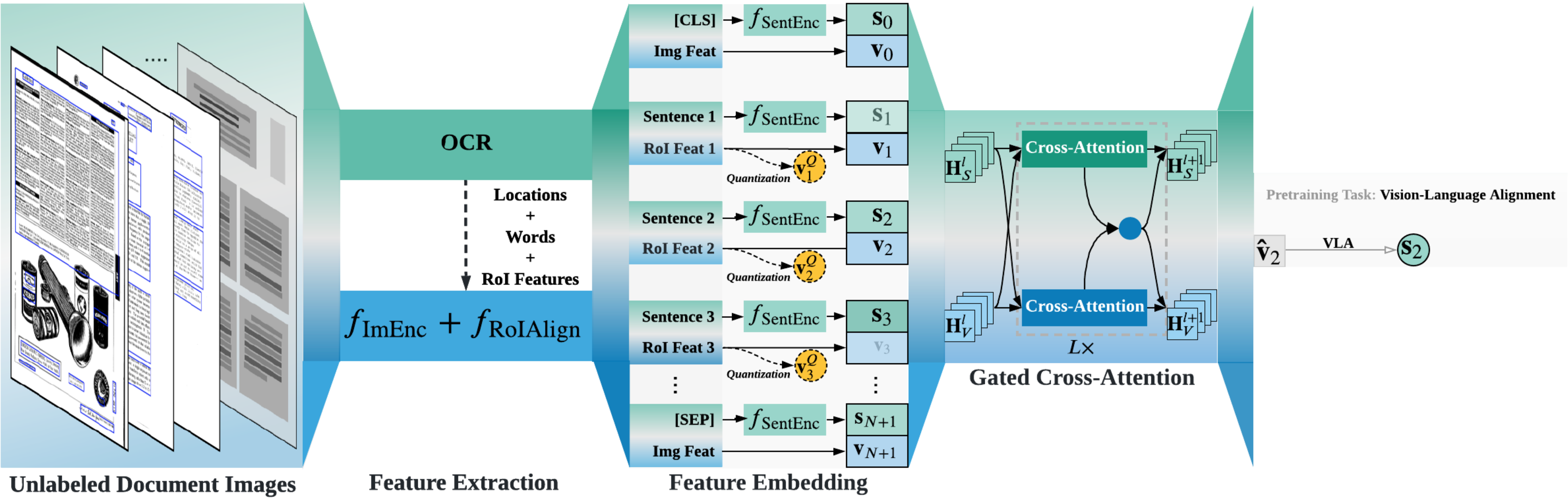


Pretraining Task: *Masked Sentence Modeling*



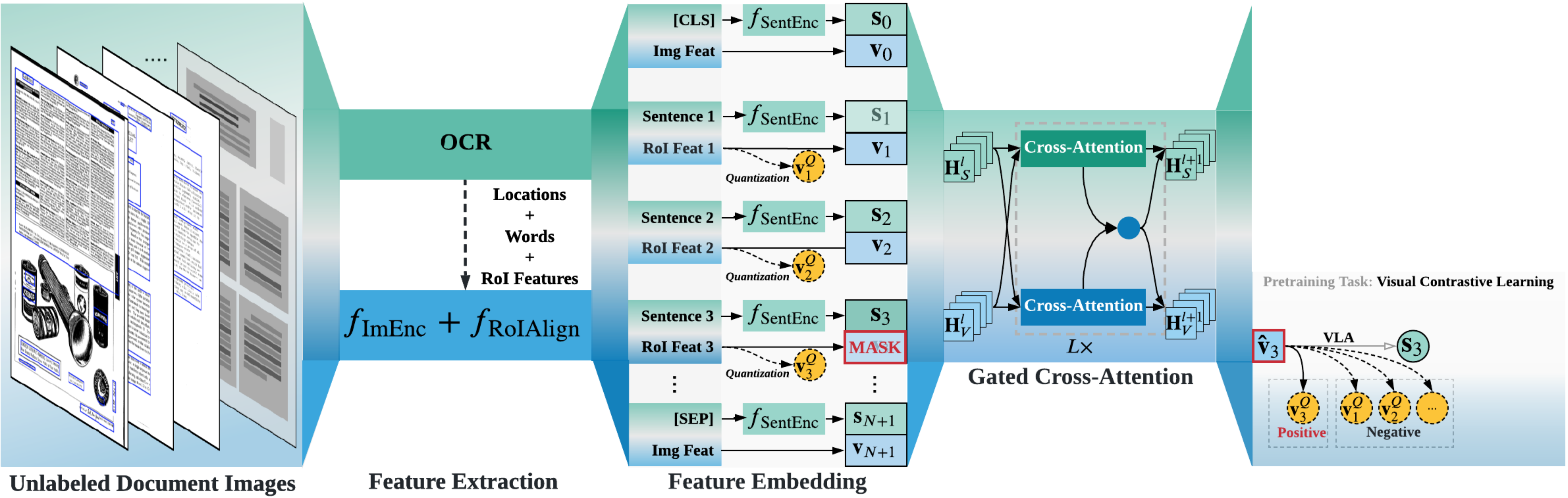
$$\mathcal{L}_{\text{MSM}}(\Theta) = \sum_i \text{smooth}_{L_1}(\boldsymbol{s}_i - f_{\text{UniDoc}}(\boldsymbol{s}_i | \boldsymbol{s}_{\setminus i}, \tilde{\mathbf{V}}))$$

Pretraining Task: *Vision-Language Alignment*



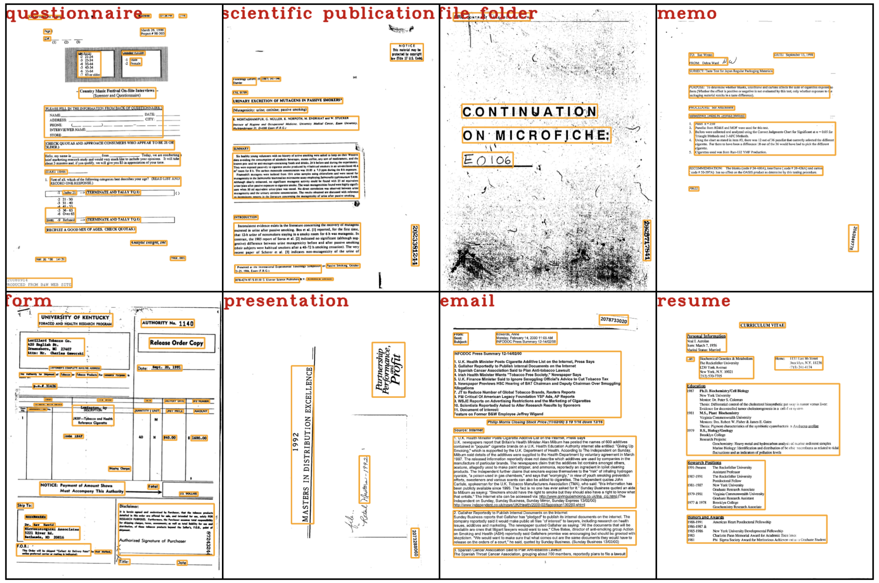
$$\mathcal{L}_{\text{VLA}}(\Theta) = \frac{1}{N \times N} ||f_{\text{Norm}}(\mathbf{S} \cdot \mathbf{S}^\top) - f_{\text{Norm}}(\mathbf{H}_V^L \cdot \mathbf{H}_V^{L\top})||_F^2$$

Pretraining Task: *Visual Contrastive Learning*



$$\mathcal{L}_{\text{VCL}}(\Theta) = -\sum_{\tilde{\boldsymbol{v}}_i \in \tilde{\mathbf{V}}} \left(\log \frac{\exp(\text{sim}(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_i^Q)/\kappa)}{\sum_{\boldsymbol{v}_j^Q} \exp(\text{sim}(\hat{\boldsymbol{v}}_i, \boldsymbol{v}_j^Q)/\kappa)} \right) + \lambda \frac{1}{CE} \sum_{c=1}^{C} \sum_{e=1}^{E} p_{c,e} \log p_{c,e}$$

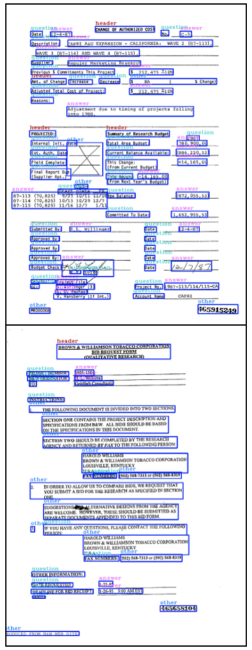
Finetuning Task: *Document Classification*



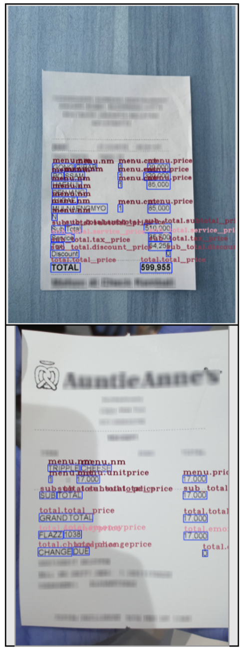
Samples from RVL-CDIP

Method	Pretraining						FUNSD	CORD	RVL-CDIP
	Source	#Data	Scale	Max #Words	Modality	#Param.	F1	F1	Accuracy
$\text{BERT}_{\text{BASE}}$ [5]	–	–	Word	512	L	110M	60.26	89.68	89.81
$\text{BERT}_{\text{LARGE}}$ [5]	–	–	Word	512	L	340M	65.63	90.25	89.92
$\text{LayoutLM}_{\text{BASE}}$ [5]	IIT-CDIP	11M	Word	512	L	113M	78.66	94.72	94.42
$\text{LayoutLM}_{\text{LARGE}}$ [5]	IIT-CDIP	11M	Word	512	L	343M	78.95	94.93	94.43
$\text{LayoutLMv2}_{\text{BASE}}$ [5]	IIT-CDIP	11M	Word	512	V+L	200M	82.76	94.95	95.25
$\text{LayoutLMv2}_{\text{LARGE}}$ [5]	IIT-CDIP	11M	Word	512	V+L	426M	84.20	96.01	95.64
SelfDoc [6]	RVL-CDIP	320K	Region	50×512	V+L	–	83.36	–	92.81
SelfDoc+VGG-16 [6]	RVL-CDIP	320K	Region	50×512	V+L	–	–	–	93.81
TILT-Base [34]	RVL-CDIP+	1.1M	Word	512	V+L	230M	–	95.11	95.25
TILT-Large [34]	RVL-CDIP+	1.1M	Word	512	V+L	780M	–	96.33	95.52
UDoc	IIT-CDIP	1M	Region	64×512	V+L	272M	87.96	98.85	93.96
UDoc*	IIT-CDIP	1M	Region	64×512	V+L	272M	87.93	98.94	95.05‡

Finetuning Task: *Document Entity Recognition*



FUNSD



CORD

Method	Pretraining						FUNSD	CORD	RVL-CDIP
	Source	#Data	Scale	Max #Words	Modality	#Param.	F1	F1	Accuracy
$\text{BERT}_{\text{BASE}}$ [5]	–	–	Word	512	L	110M	60.26	89.68	89.81
$\text{BERT}_{\text{LARGE}}$ [5]	–	–	Word	512	L	340M	65.63	90.25	89.92
$\text{LayoutLM}_{\text{BASE}}$ [5]	IIT-CDIP	11M	Word	512	L	113M	78.66	94.72	94.42
$\text{LayoutLM}_{\text{LARGE}}$ [5]	IIT-CDIP	11M	Word	512	L	343M	78.95	94.93	94.43
$\text{LayoutLMv2}_{\text{BASE}}$ [5]	IIT-CDIP	11M	Word	512	V+L	200M	82.76	94.95	95.25
$\text{LayoutLMv2}_{\text{LARGE}}$ [5]	IIT-CDIP	11M	Word	512	V+L	426M	84.20	96.01	95.64
SelfDoc [6]	RVL-CDIP	320K	Region	50×512	V+L	–	83.36	–	92.81
SelfDoc+VGG-16 [6]	RVL-CDIP	320K	Region	50×512	V+L	–	–	–	93.81
TILT-Base [34]	RVL-CDIP+	1.1M	Word	512	V+L	230M	–	95.11	95.25
TILT-Large [34]	RVL-CDIP+	1.1M	Word	512	V+L	780M	–	96.33	95.52
UDoc	IIT-CDIP	1M	Region	64×512	V+L	272M	87.96	98.85	93.96
UDoc*	IIT-CDIP	1M	Region	64×512	V+L	272M	87.93	98.94	95.05‡

At the Poster:

- **Additional details**
- **Quantitative results**
- **Discussion**