

# SUPPLEMENTARY MATERIAL

## 1 STRUCTEXTV2 FOR FORM UNDERSTANDING

To more accurately verify the capability of StrucTexTv2 on semantic learning, we evaluate the performance of our pre-trained model on FUNSD for form understanding. We build a classification layer to predict four categories (question, answer, header, or other) for each semantic entity. As shown in Tab. 1, the entity-level F1-score of StrucTexTv2<sub>Small</sub> is 89.23% and the entity-level F1-Score of StrucTexTv2<sub>Large</sub> is 91.82%, which achieve comparable performance with the latest state-of-the-art methods. StrucTexTv2 achieves great semantic learning while the ground-truth OCR is utilized on FUNSD. However, it needs to be explained that our method can present better end-to-end information extraction results than the two-stage methods with the actual OCR engine, which is more referential for practical applications.

Table 1: Experimental results and performance comparison of the form understanding task on FUNSD. The entity-level F1-Score is used to measure model accuracy.

Methods	#Params.	FUNSD
DocFormer <sub>Base</sub> Appalarajuet al. (2021)	183M	83.34%
SelfDoc Li et al. (2021b)	-	83.36%
LayoutLMv2 <sub>Large</sub> Xu et al. (2021a)	426M	84.20%
DocFormer <sub>Large</sub> Appalarajuet al. (2021)	536M	84.55%
StructuralLM <sub>Large</sub> Liet al. (2021a)	355M	85.14%
UDoc Gu et al. (2021)	272M	87.93%
StrucTexTv2 <sub>Small</sub>	<b>28M</b>	89.23%
XDoc Chen et al. (2022)	146M	89.40%
LayoutLMv3 <sub>Base</sub> Huang et al. (2022)	133M	90.29%
StrucTexTv2 <sub>Large</sub>	238M	91.82%
LayoutLMv3 <sub>Large</sub> Huang et al. (2022)	368M	<b>92.08%</b>

## 2 STRUCTEXTV2 IN CHINESE

To further analyze StrucTexTv2 in the Chinese language, we pre-train StrucTexTv2<sub>Small</sub><sup>†</sup> with extra collected Chinese document images and evaluate the model on XFUND-ZH Xu et al. (2022). XFUND is a multi-lingual extended dataset of the FUNSD Jaume et al. (2019) in 7 languages. StrucTexTv2<sub>Small</sub><sup>†</sup> directly adopts the text recognition model from PaddleOCR since only 149 train images of XFUND-ZH can not fulfill the recognition branch training for the end-to-end information extraction task. Experiment results in Tab. 2 show that StrucTexTv2 can benefit the task of document image understanding in Chinese.

Table 2: Experimental results on XFUND-ZH. We re-implement LayoutXLM<sub>Base</sub><sup>\*</sup> with PaddleOCR to provide OCR results. The Normalized Edit Distance (1-NED) is used to evaluate for the end-to-end information extraction task.

Methods	OCR	1-NED
LayoutXLM <sub>Base</sub> <sup>*</sup> Xu et al. (2021b)	PaddleOCR	60.40%
StrucTexTv2 <sub>Small</sub> <sup>†</sup>	Text Recognition Only	67.46%

### 3 ABLATION STUDY FOR IMAGE RESOLUTION

In this ablation, we investigate the impact of training with different image resolutions. As shown in Tab.3, by replacing the resolution of input images with  $224 * 224$ ,  $512 * 512$ , and  $960 * 960$ , the accuracy of RVL-CDIP is 89.5%, 90.4%, and 92.5%, respectively. We also report the results on PubLayNet, the F1-Score of PubLayNet is 94.9%, 92.1%, and 85.9%, respectively. It suggests that the best image resolution of our pre-training tasks is  $960 * 960$ . The performance of image-only vision tasks is more sensitive to the selection of resolution of the input.

Table 3: Ablation study for Different Image Resolution.

Image Resolution	Pre-train Tasks	RVL-CDIP	PubLayNet
224*224	MLM+MIM	89.5%	85.9%
512*512	MLM+MIM	90.4%	92.1%
960*960	MLM+MIM	92.5%	94.9%

### 4 ABLATION STUDY FOR OCR SELECTION

In this ablation, we look at the impact of various thresholds used in OCR selection for pre-training. We choose a commercial OCR engine to get the OCR results of the pre-training data. Besides the coordinates and text content, we keep the confidence score for each OCR word. For our pre-training, only the high-confidence word with a score above 0.8 was retained. Then the masked words are selected within the preserved text. We conduct experiments to analyze the impacts of the quality of OCR results on our pre-trained model via performing different score thresholds. As shown in Tab.4, the performance decreases from 55.0% to 54.9% and 54.8% by replacing the score threshold from 0.8 to 0.6 and 0.4 on FUNSD, respectively. Similarly, on RVL-CDIP, only slight modifications in accuracy by replacing the score threshold from 0.8 to 0.6 and 0.4. The subtle performance changes in downstream tasks suggest that StrucTexTv2 is robust against the selection of word’s thresholds.

Table 4: Ablation study for the impacts of different thresholds used in OCR selection. The model is pre-trained with only MIM task.

OCR	RVL-CDIP	FUNSD
threshold=0.8	93.4%	55.0%
threshold=0.6	93.4%	54.9%
threshold=0.4	93.3%	54.8%

### 5 QUALITATIVE RESULTS OF MASK IMAGE MODELING

The qualitative results of Mask Image Modeling, which are shown in Fig. 1, express that StrucTexTv2 is capable of recovering the RGB value of the randomly selected masked text regions according to the surrounding context.

### REFERENCES

- Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding. In *ICCV*, pp. 973–983, 2021.
- Jingye Chen, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. Xdoc: Unified pre-training for cross-format document understanding. *CoRR*, 2022.
- Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova, and Tong Sun. Unified pretraining framework for document understanding. In *NeurIPS*, pp. 39–50, 2021.

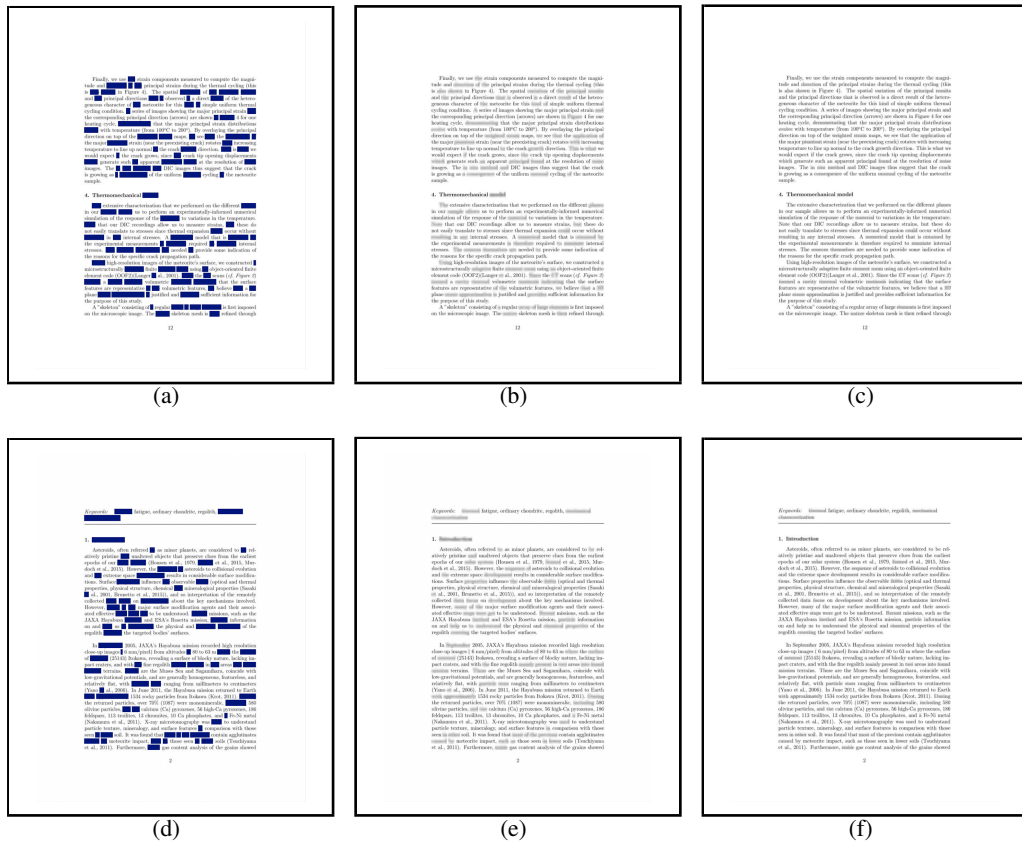


Figure 1: The qualitative results of Mask Image Modeling. From left to right: the masked document image, document reconstruction without content information from MLM, and document reconstruction with content information from MLM.

- Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document AI with unified text and image masking. In *ACM Multimedia*, 2022.
- Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. FUNSD: A dataset for form understanding in noisy scanned documents. In *ICDAR*, pp. 1–6, 2019.
- Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, and Luo Si. Structurallm: Structural pre-training for form understanding. In *ACL/IJCNLP*, pp. 6309–6318, 2021a.
- Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *CVPR*, pp. 5652–5660, 2021b.
- Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. pp. 2579–2591, 2021a.
- Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florêncio, Cha Zhang, and Furu Wei. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *CoRR*, abs/2104.08836, 2021b.
- Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, and Furu Wei. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *ACL*, pp. 3214–3224, 2022.