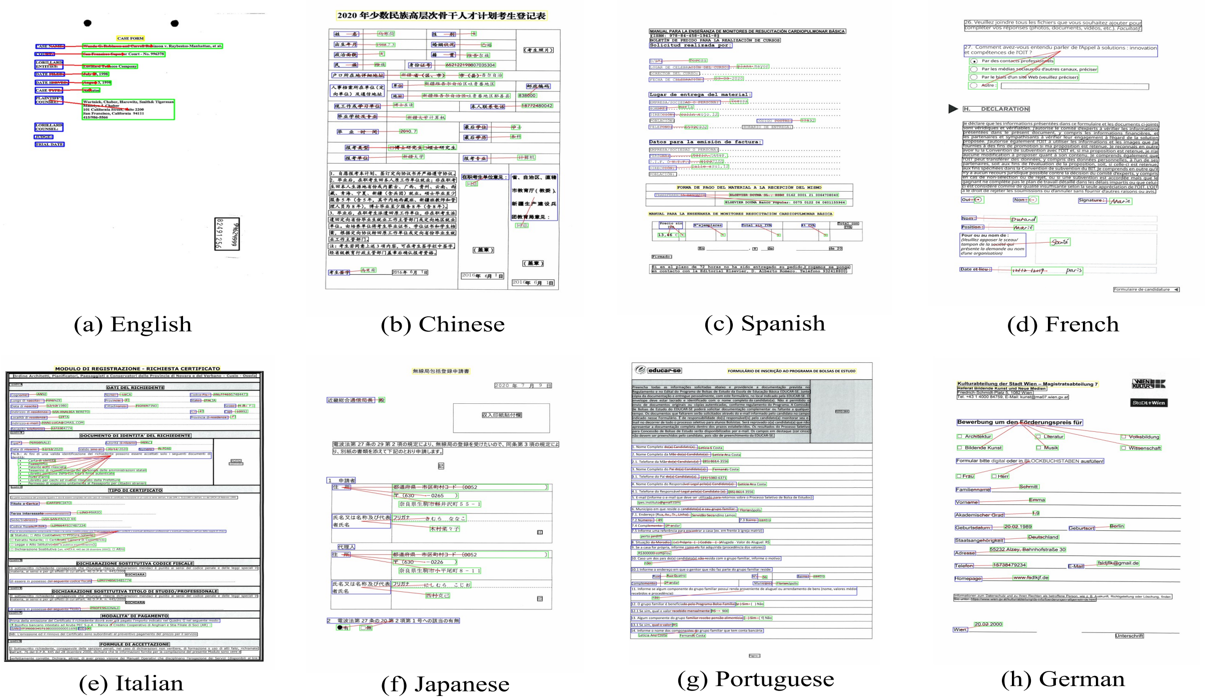

(a) English (b) Chinese (c) Spanish (d) French

(e) Italian (f) Japanese (g) Portuguese (h) German

Figure 3: Qualitative results of our KVPFormer on FUNSD and XFUND datasets. Bule boxes, green boxes, yellow boxes and black boxes represent Question, Answer, Header and Other entities, respectively. Red arrows stand for the predicted key-value relationships pointing from key entities to value entities. Best viewed in color.

Model	EN	ZH	JA	ES	FR	IT	DE	PT	Avg.
XLM-RoBERTa$_{\text{BASE}}$	0.3638	0.6797	0.6829	0.6828	0.6727	0.6937	0.6887	0.6082	0.6341
InfoXLM$_{\text{BASE}}$	0.3699	0.6493	0.6473	0.6828	0.6831	0.6690	0.6384	0.5763	0.6145
LayoutXLM$_{\text{BASE}}$	0.6671	0.8241	0.8142	0.8104	0.8221	0.8310	0.7854	0.7044	0.7823
LiLT[InfoXLM]$_{\text{BASE}}$	0.7407	0.8471	0.8345	0.8335	0.8466	0.8458	0.7878	0.7643	0.8125
Baseline#1 (LayoutXLM$_{\text{BASE}}$ + SCF)	0.8627	0.8971	0.8486	0.8721	0.8904	0.8563	0.8302	0.8196	0.8596
Baseline#2 (LayoutXLM$_{\text{BASE}}$ + SCF + DP)	0.9557	0.9373	0.9122	0.9349	0.9366	0.9240	0.9175	0.8874	0.9257
KVPFormer	**0.9570**	**0.9427**	**0.9423**	**0.9523**	**0.9719**	**0.9411**	**0.9241**	**0.9219**	**0.9442**

Table 2: Multitask learning F1-score for key-value pair extraction on XFUND dataset, where "SCF" means Spatial Compatibility Feature and "DP" stands for Dependency Parsing.

Model	EN	ZH	JA	ES	FR	IT	DE	PT	Avg.
XLM-RoBERTa$_{\text{BASE}}$	0.2659	0.1601	0.2611	0.2440	0.2240	0.2374	0.2288	0.1996	0.2276
InfoXLM$_{\text{BASE}}$	0.2920	0.2405	0.2851	0.2481	0.2454	0.2193	0.2027	0.2049	0.2423
LayoutXLM$_{\text{BASE}}$	0.5483	0.4494	0.4408	0.4708	0.4416	0.4090	0.3820	0.3685	0.4388
LiLT[InfoXLM]$_{\text{BASE}}$	0.6276	0.4764	0.5081	0.4968	0.5209	0.4697	0.4169	0.4272	0.4930
Baseline#1 (LayoutXLM$_{\text{BASE}}$ + SCF)	0.8533	0.7633	0.7350	0.7503	0.7888	0.7572	0.6781	0.7155	0.7552
Baseline#2 (LayoutXLM$_{\text{BASE}}$ + SCF + DP)	0.9385	0.9131	0.8864	0.9007	0.9161	0.8808	0.8771	0.8506	0.8954
KVPFormer	**0.9555**	**0.9223**	**0.8907**	**0.9047**	**0.9366**	**0.8848**	**0.8743**	**0.8642**	**0.9041**

Table 3: Cross-lingual zero-shot transfer learning F1-score for key-value pair extraction on XFUND dataset, where "SCF" means Spatial Compatibility Feature and "DP" stands for Dependency Parsing.

#	Encoder	Decoder	SCAB	C2F	F1
1					0.7498
2a	✓				0.7632
2b	✓		✓		0.7817
3a		✓			0.7767
3b		✓	✓		0.7906
4a	✓	✓			0.7825
4b	✓	✓	✓		0.8017
4c	✓	✓	✓	✓	**0.8223**

Table 4: Ablation studies of different component in KVPFormer on FUNSD dataset, where "SCAB" means Spatial Compatibility Attention Bias and "C2F" means Coarse-to-Fine answer prediction module.

achieved a new state-of-the-art result, i.e., 78.84% in F1-score. Compared with this strong Baseline#2, our proposed KVPFormer is still significantly better by improving F1-score from 78.84% to 82.23%, which can demonstrate the advantage of our approach. To evaluate the upper bound performance of our approach, following SERA (**?**), we further concatenate the entity label embeddings (gold labels) to entity representations. The performance of our KVPFormer with gold labels can be improved to 90.86% in F1-score. This performance gap indicates that there is still much room for improvement of our model if we can achieve better accuracy on question identification with semantic entity labels.

XFUND. In this dataset, we perform two experimental settings designed in (**?**) to verify the effectiveness of our approach: 1) Multi-task learning; 2) Zero-shot transfer learning. Moreover, following the implementation of (**?**), all models here have leveraged gold labels for fair comparisons.

1) Multi-task learning. In this setting, all models are trained on all languages and then tested on each language. As shown in Table 2, our proposed KVPFormer achieves the best result of 94.42% in average F1-score, outperforming previous methods and our strong baselines by a substantial margin.

2) Zero-shot transfer learning. In this setting, all models are trained on English (i.e., FUNSD) only and tested on other languages. As shown in Table 3, our Baseline#1 can significantly improve the average F1-score of LayoutXLM$_{\mathrm{BASE}}$ from 43.88% to 75.52%, which indicates that rich spatial compatibility features play an important role in cross-lingual zero-shot transfer learning. Furthermore, KVPFormer achieves a new state-of-the-art result of 90.41% in average F1-score, demonstrating that our approach has a better capability of transferring knowledge from the seen language to unseen languages for key-value pair extraction.

Some qualitative results of our proposed KVPFormer on FUNSD and XFUND datasets are depicted in Fig. 3, which demonstrate that our approach can handle many challenging cases, e.g., large empty spaces between key and value entities, one key entity has multiple value entities. For failure cases, we observe that our approach cannot work equally well in extracting key-value pairs in tables, e.g., one column header is related to multiple table cells.

5.4 Ablation Study

We conduct a series of ablation experiments to evaluate the effectiveness of each component in KVPFormer. All experiments are conducted on FUNSD dataset and the results are presented in Table 4.

Transformer Encoder-Decoder Architecture. The results in #1, #2a-#4a, and #2b-#4b rows show that 1) Both encoder and decoder can improve the performance; 2) Decoder-only is slightly better than encoder-only models; 3) The combination of encoder and decoder leads to the best performance.

Spatial Compatibility Attention Bias. As shown in results #2a-#2b, #3a-#3b, and #4a-#4b, no matter which architecture is used (i.e., encoder-only, decoder-only and encoder-decoder), the proposed spatial compatibility attention bias can consistently improve the performance.

Coarse-to-Fine Answer Prediction. As shown in results #4b and #4c, the proposed coarse-to-fine answer prediction algorithm can lead to about 2.1% improvement in F1-score, which demonstrates its effectiveness.

6 Conclusion and Future Work

In this paper, we formulate key-value pair extraction as a QA problem and propose a new Transformer-based encoder-decoder model (namely KVPFormer) to extract key-value pairs, which makes the QA framework be more general on VDU tasks. Moreover, we propose three effective techniques to improve both the efficiency and accuracy of KVPFormer: 1) A DETR-style decoder to predict answers for all questions in parallel; 2) A coarse-to-fine answer prediction algorithm to improve answer prediction accuracy; 3) A new spatial compatibility attention bias to better model the spatial relationships in each pair of entities in self-attention/cross-attention layers. Experimental results on the public FUNSD and XFUND datasets have demonstrated the effectiveness of our KVPFormer on key-value pair extraction. In future work, we will explore how to train a unified QA model to solve various VDU tasks including key-value pair extraction. Moreover, we will also explore the effectiveness of our approach on more challenging but under-researched relation prediction problems in VDU area such as hierarchical key-value pair extraction and choice group extraction.

Doloremque id quas harum temporibus voluptatum minus, dolor cupiditate doloribus beatae mollitia molestias a quae dignissimos doloremque sapiente labore, nesciunt perferendis voluptate iure ipsa, dolorem molestias eveniet cumque?Itaque rem ipsa culpa sunt vel et at corrupti nulla harum magnam, debitis laborum illum perspiciatis magni, nisi doloremque pariatur officia voluptatibus laborum cum eveniet, itaque iusto deserunt neque?Veritatis eum ad inventore sapiente molestiae itaque autem perferendis vel, unde dolor nesciunt, incidunt nulla quidem tempora ratione facere cum molestiae suscipit.Fugit voluptatum eius nisi ratione dolores, veritatis nesciunt architecto similique, possimus repellendus sed commodi cum deserunt illum?Cum molestias eum facere, perspiciatis impedit laudantium eligendi delectus exercitationem quisquam quidem maiores porro architecto sequi, nemo quidem nihil voluptatibus soluta impedit doloremque modi ducimus libero porro aut, libero reiciendis quisquam excepturi quam?