A Query-Response Framework for Whole-Page Complex-Layout Document Image Translation with Relevant Regional Concentration

Zhiyang Zhang[1,2], Yaping Zhang[1,2]*, Yupu Liang[1,2], Zhiyuan Chen[1,2], Lu Xiang[1,2], Yang Zhao[1,2], Yu Zhou[1,3], Chengqing Zong[1,2]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China

zhangzhiyang2020@ia.ac.cn, {yaping.zhang, lu.xiang, yang.zhao, yzhou, cqzong}@nlpr.ia.ac.cn

Abstract

Document Image Translation (DIT), which aims at translating documents in images from source language to the target, plays an important role in Document Intelligence. It requires a comprehensive understanding of document multi-modalities and a focused concentration on relevant textual regions during translation. However, most existing methods usually rely on the vanilla encoder-decoder paradigm, severely losing concentration on key regions that are especially crucial for complex-layout document translation. To tackle this issue, in this paper, we propose a new Query-Response DIT framework (QRDIT). QRDIT reformulates the DIT task into a parallel response/translation process of the multiple queries (*i.e.*, relevant source texts), explicitly centralizing its focus toward the most relevant textual regions to ensure translation accuracy. A novel dynamic aggregation mechanism is also designed to enhance the text semantics in query features toward translation. Extensive experiments in four translation directions on three benchmarks demonstrate its state-of-the-art performance, showing significant translation quality improvements toward whole-page complex-layout document images.

1 Introduction

Document image translation is a fundamental task in which model translates documents in images from source language to the target (Zhang et al., 2023). It plays a critical role in various applications such as cross-lingual document retrieval, summarization, and information extraction (Cui et al., 2021). However, performing DIT in real-world scenarios faces many difficulties including intricate layouts, complex multi-modality semantics, and cross-lingualities, *etc*.

Basically, DIT is formulated as an image/text-to-text transformation task. Early works (Afli and Way, 2016; Hinami et al., 2021; Shekar et al., 2021) directly use text contents extracted by optical character recognition (OCR) engine as the inputs of text translation model, and have been proven adequate for simple-layout text images such as single-line movie subtitle or multi-line, single-paragraph text. To better exploit multi-modal information of documents, recent methods propose to incorporate visual layout by encoding image pixels (Liang et al., 2024; Tian et al., 2023; Zhu et al., 2023) or layout positions (Zhang et al., 2023), and combine them with text features, achieving good results in DIT. However, as shown in Fig. 1, when translating whole-page documents with complex layouts, contents of each sentence change greatly in their textual regions, often taking up a relevant but local area of the whole document space. Therefore, a favorable DIT framework should first be aware and capable of concentrating on the most relevant region for each sentence and then generating the translation conditioning on these key regional texts. Nevertheless, these vital steps are severely overlooked in prior methods since they lack ex-

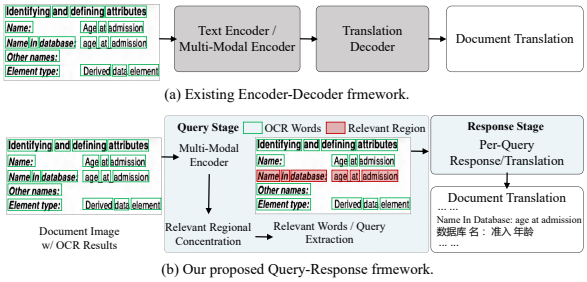


Figure 1: Comparison of different DIT frameworks. (a) Existing methods directly input the image and OCR results to a text-only or multi-modal encoder-decoder, overlooking the focus on key regions during translation. (b) In contrast, our proposed QRDIT based on Query-Response framework is capable of allocating dominant focus on key regions and extracting the most relevant texts to ensure accurate source texts & translations.

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 7138–7149

plicit objectives or modules targeting the relevant regional concentration and thus suffer degraded performance when translating complex-layout documents. *Therefore, how to incorporate relevant regional identification and concentration capabilities into DIT models to centralize their focus toward the most relevant regional texts, has become a vital step to improve whole-page complex-layout DIT.*

To address this problem, this paper presents the Query-Response Document Image Translation (QRDIT) network. It reformulates DIT as a novel query-response procedure where specific modules & objectives are introduced for relevant regional concentration. Specifically, QRDIT converts DIT into a response process of queries (*i.e.*, relevant regional texts), which includes a query stage and a response stage. First, the query stage extracts document multi-modal features, and then carries out sequence labeling to identify each query's prefix word. All prefixes interact with the multi-modal feature through a DETR-like (Carion et al., 2020) cross-attention to form query embeddings. Then, the text area relevant to each source sentence is identified by adaptively gathering the most relevant words, of which word-level relevance scores for each query are computed and employed as the gathering measure. Then, at the response stage, model first utilizes a dynamic gate-based aggregation to enhance the text semantics in query features. This mechanism alleviates the feature deviation issue (caused by prior query tasks), keeping features oriented at the translation goal. Finally, these enhanced query features are fed into a translation decoder, which employs a per-query strategy to simultaneously generate translation result as response for each query/source sentence. The query and response stages are integrated as an end-to-end framework with feature flowing across them consecutively. *Such a query-response framework successfully incorporates regional concentration capability into DIT and therefore leads to significant improvements.* Extensive experiments in four translation directions on three public benchmarks demonstrate QRDIT's SOTA performance. Our contributions are:

- We propose a new end-to-end DIT framework (QRDIT). It reformulates DIT into a parallel response procedure for the multi-queries, each composed of texts from most relevant region.
- We introduce a dynamic aggregation mechanism to enhance the query feature's text semantics toward the final translation goal.
- Experiment results on multiple directions and datasets show that with intrinsic relevant regional concentration ability, our model significantly outperforms prior SOTAs.

2 Task Formulation

The inputs of DIT include document image I and its OCR results - text words $T = \{t_i\}_{i=1}^{L}$ (L is # words) and word bounding boxes $B = \{b_i\}_{i=1}^{L}$. Each box $b_i = (x_{tl}, y_{tl}, x_{br}, y_{br})_i$ reports the top-left layout position $(x_{tl}, y_{tl})_i$ and bottom-right layout position $(x_{br}, y_{br})_i$ of i-th word t_i on the image. Given these multi-modal inputs (I, T, and B) for a document image, DIT aims to generate its target-language document-level translation $\hat{Y}$:

$$\hat{Y} = \arg\max_{Y} \prod_{j=1}^{|Y|} P_\theta \left(y_j \mid y_{<j}, I, T, B \right) \quad (1)$$

where $Y = \{y_j\}_{j=1}^{|Y|}$ is target sequence. y_j/$y_{<j}$ is current/previous target tokens. $P_\theta(\cdot)$ is the model.

3 Methodology

As shown in Fig. 2, by reformulating the DIT task as a novel query-response procedure, our proposed QRDIT consists of 1) an OCR stage, 2) a query stage for query extraction, and 3) a response stage for response/translation generation. Firstly, the multi-modal information including text, layout, and image are obtained from preconditioned OCR results. The words and bounding boxes are sorted with a "top-left to bottom-right" rule to keep a fixed input format. Afterward, in query stage, inputs are deeply aggregated as multi-modal document representation, based on which the prefix word of each query (*i.e.*, source sentence) is identified. Then, with each prefix acting as the "anchor", the most relevant regions are allocated predominant concentrations and words within that region are picked up (with an adaptive extraction strategy) to complement the query. Finally, all translation queries are semantically enhanced and then prompt response modules to generate their translations in parallel, which are concatenated as the final document translation results while also preserving each translation's layout position.

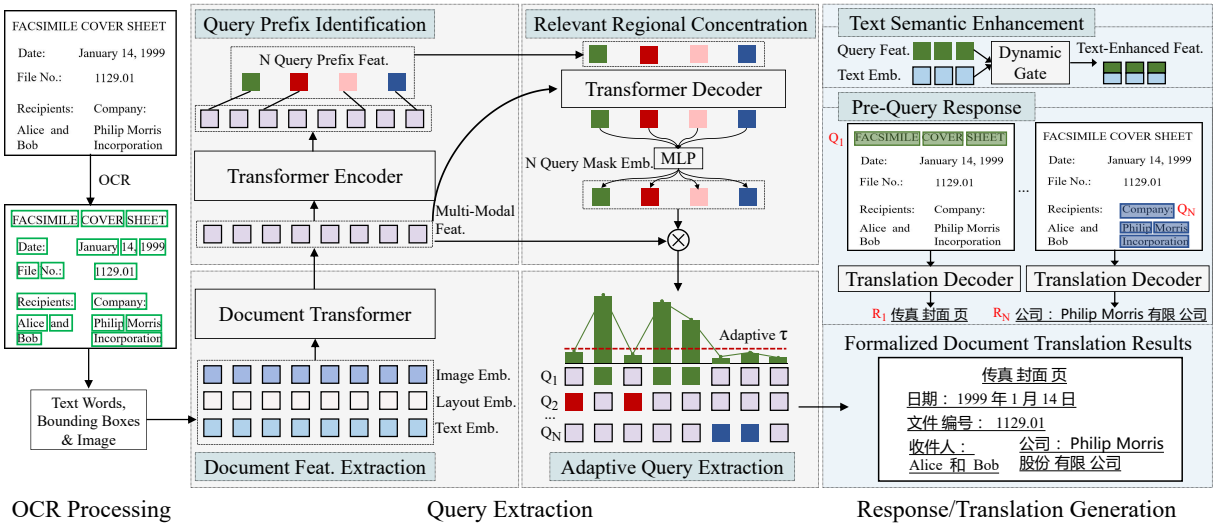


Figure 2: System overview of QRDIT. It realizes the whole-page complex-layout DIT by reformulating the DIT task as a novel Query-Response procedure: 1) In Query stage, model aggregates multi-modal document features, identifies query prefixes as "anchors", concentrates on each query's relevant regions and adaptively extracts regional texts as translation queries. 2) In Response stage, model responds to each query (after text semantics enhancement) to generate translations in parallel, and finally formalizes the document-level translation as task results.

3.1 Query Extraction with Relevant Regional Concentration

The query stage aims to prompt model to allocate its concentration predominantly on relevant textual regions, thereby locating the complete while non-redundant source texts to guarantee correct translation. This goal is progressively achieved through the following four sub-processes.

Multi-Modal Document Feature Extraction: We leverage advanced document transformers (*e.g.*, LayoutLMv3), which have been pre-trained on massive document images, for feature extraction. Specifically, given the image, text words, and bounding boxes, embedding for each modality is obtained following Huang et al., 2022. For example, given a word bounding box $b = (x_{tl}, y_{tl}, x_{br}, y_{br})$, its layout embedding is derived by looking up layout embedding tables according to layout positions:

$$E^l = \mathrm{Lin}([\mathrm{Emb}_x(x_{tl}, x_{br}); \mathrm{Emb}_y(y_{tl}, y_{br})]) \quad (2)$$

where $\mathrm{Emb}_{x/y}(\cdot)$ is the layout embedding table for x/y direction, $\mathrm{Lin}([\cdot;\cdot])$ is concatenation and linear mapping. All modality embeddings including text, layout, and image are fed into the pre-trained document transformer for feature contextualization and deep fusion. The output F^m is employed as the multi-modal document feature.

Query Prefix Identification: Given F^m, model first identifies each query's prefix word, which will further act as the "anchor" word to locate relevant regions for complete query extraction. Specifically, the query prefix identification is modeled as a sequence tagging problem over F^m:

$$F^{pref} = \mathrm{MHSA}(F^m); P^{pref} = \mathrm{Sigmoid}(F^{pref}) \quad (3)$$

where $\mathrm{MHSA}(\cdot)$ is multi-head self-attention (Vaswani et al., 2017). P_i^{pref} is i-th word's classification probability. Those words with $P_i^{pref} \geq 0.5$ are determined as positive predictions (*i.e.*, prefix words), each corresponding to a query.

Relevant Regional Concentration: All query prefix's features are fed to the relevant regional concentration module to compute each prefix's concentration scores, which will serve as the criterion for complete query extraction. Specifically, we adopt a DETR-style (Carion et al., 2020) transformer decoder where multi-head self-attention first models the interactions among query prefixes (with attention mask excluded to enable bidirectional attention) and then a multi-head cross-attention models interactions between query prefixes and document feature sequence F^m. This process yields N per-query embeddings in parallel. Then, a Multi-Layer Perceptron (MLP) with 2 hidden layers converts per-query embeddings to N query mask embeddings ε_{mask}. Finally, we obtain i-th query's concentration scores $s_i \in \mathbb{R}^L$ over F^m via dot product between query mask embeddings $\varepsilon_{mask} \in \mathbb{R}^{d \times N}$

and document feature $F^m \in \mathbb{R}^{d \times L}$:

$$s_i^\top = \text{Sigmoid}(\varepsilon_{mask}[:, i]^\top \cdot F^m) \quad (4)$$

where $\text{Sigmoid}(\cdot)$ is employed to normalize score values to the range [0, 1].

Adaptive Query Extraction: Given the concentration score for a query, one could deterministically select the top-k words with highest scores from the input sequence to find the complete query. Nevertheless, since the number of words varies for different queries, this fixed strategy might suffer incomplete recall issue for some queries while excessive recall issue for the others. However, in our experiments, we find that the concentration scores exhibit remarkable discrimination property between query words (with considerably high scores) and non-query words (with quite low scores). In view of this, we propose the adaptive query words extraction strategy. It simply conducts the k-means ($k = 2$) algorithm on the concentration scores, partitioning them into two clusters to discriminate high concentration scores from low scores. Through this process, this strategy derives an adaptive threshold τ to filter out non-query words, only preserving and concentrating on words with significantly high scores. By collecting each query words' features from F^m, we derive N feature sequences $Q = \{q_k\}_{k=1}^N$, each corresponding to a query.

3.2 Response Generation for Translation

The above sub-processes in query stage convert the OCR outputs to structured and semantically intact source sentences. They serve as translation queries, prompting the response modules to generate translations through the following two sub-processes.

Text Semantics Enhancement: While the query stage requires document multi-modalities, the translation stage deserves more textual information. For this, we propose to enhance text semantics in query features to alleviate their deviation from the translation goal caused by prior query stage. Specifically, the text embeddings E^t (from document feature extraction module) are assumed to contain rich text semantics and thereby employed to enhance the query features via a dynamic gate aggregation:

$$\eta = \text{Sigmoid}(w_q\, q + w_t E_q^t + b) \quad (5)$$

$$q' = \eta \odot q + (1 - \eta) \odot E_q^t \quad (6)$$

where q is a given query's feature (subscript k is omitted for brevity), E_q^t is the textual embeddings for words in this query. w_q, w_t, b are learnable weights and bias. η is the gate factor, which dynamically controls the proportion of q and E_q^t. q' is the required text-enhanced query feature.

Per-Query Response: Given the k-th query feature q'_k, a transformer decoder deems it as the feature of source sentence for cross-attention to generate its translation as the response for this query. Note that all queries are responded to in parallel to realize translating all source sentences in one pass and promote inference efficiency. Finally, all query-response pairs could be organized with layout position of query prefixes indicating their locations on the image (as in Fig. 2), or could be concatenated to formalize the document-level source text and translation as model output.

3.3 Loss Function

Query Prefix Identification Loss: The prefix identification process is supervised by a sequential position-wise binary classification loss:

$$\mathcal{L}_{pref} = \frac{1}{L} \sum_{i=1}^{L} \text{CE}(P_i^{pref}, pref_i) \quad (7)$$

where P_i^{pref} and $pref_i$ are model probability and golden label for i-th word. $\text{CE}(\cdot)$ is cross-entropy.

Relevant Regional Concentration Loss: The loss function for regional concentration process is:

$$\mathcal{L}_{reg} = \frac{1}{NL} \sum_{i=1}^{N} \sum_{j=1}^{L} \text{CE}(s_{i,j}, s_{i,j}^*) \quad (8)$$

where $s_{i,j}$ is model predicted score (calculated by Eq. 4) for i-th query at position j, $s_{i,j}^* \in \{0, 1\}$ is golden score. $s_{i,j}^*$ is 1 only if the word at position j belongs to the i-th query, otherwise it is 0.

Response Loss: Loss function for response process is essentially a translation loss ($\mathcal{L}_{trans}$).

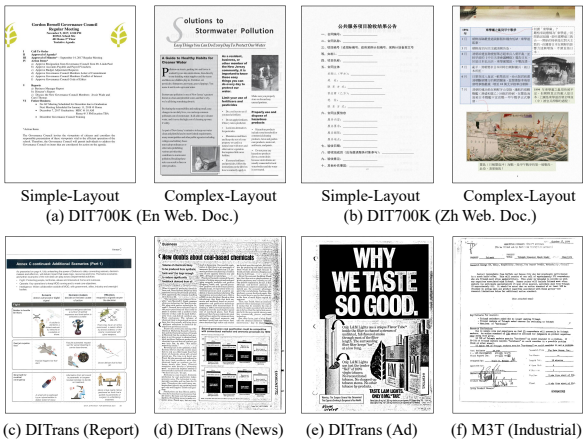
Overall Loss: Above loss functions are summed for our model's end-to-end joint training:

$$\mathcal{L} = \mathcal{L}_{pref} + \mathcal{L}_{reg} + \mathcal{L}_{trans} \quad (9)$$

4 Experiments

4.1 Experiment Settings

Datasets: We comprehensively conduct model evaluations and comparisons on three public DIT benchmarks across five domains and four translation directions. The three benchmarks include DIT700K, DITrans, and M3T.



Simple-Layout Complex-Layout Simple-Layout Complex-Layout
(a) DIT700K (En Web. Doc.) (b) DIT700K (Zh Web. Doc.)
(c) DITrans (Report) (d) DITrans (News) (e) DITrans (Ad) (f) M3T (Industrial)

Figure 3: Image examples from the three used datasets.

Dataset	DIT700K		DITrans			M3T
Domain/Subset	En Web Doc.	Zh Web Doc.	Report	News	Ad.	Industrial, *etc*
Trans. Direction	En→Zh/De	Zh→En/Fr	En→Zh			En→Zh/De
# Image	619K	99K	902	396	377	143
# Word/Image	237	431	245	219	123	263
# Sent./Image	30	28	24	16	16	31
Trainset						
# Images	618K	98K	722	316	302	-
Layout Score	82.16	87.44	72.97	76.16	61.18	-
Testset - Simple Layout						
# Images	512	128	-	-	-	-
Layout Score	86.72	91.88	-	-	-	-
Testset - Complex Layout						
# Images	512	128	180	80	75	143
Layout Score	74.29	73.62	69.35	74.92	55.31	51.77

Table 1: Statistics of the three used datasets. #Word/Image is counted on English (En) words and Chinese (Zh) characters respectively for En and Zh documents.

- DIT700K (Zhang et al., 2025)**:** The most large-scale dataset that contains 718K (619K in English, 99K in Chinese) document images crawled from the websites. The samples in DIT700K are parsed and annotated automatically with an elaborate document analysis pipeline, providing translation references in multiple target languages which support four translation directions: English-to-Chinese/German (En→Zh/De), Chinese-to-English/French (Zh→En/Fr).

- DITrans (Zhang et al., 2023)**:** A multi-domain dataset that contains 1,675 document images with fine-grained labels for document layout analysis and translation tasks. The samples in DIT700K are manually collected to have complex and diverse layouts spanning three domains (Report, News, and Ad.). Translation labels are also manually annotated. This dataset supports the En→Zh translation task.

- M3T (Hsu et al., 2024)**:** A multi-domain DIT testset whose document images are sourced from multiple document banks (*e.g.*, RVL-CDIP industrial documents, *etc*). The samples in M3T are manually annotated in multiple target languages and we select its En→Zh/De directions for evaluation.

Tab. 1 shows the statistics and data split details of the three datasets. Fig. 3 present some document examples.

Setups: Models are evaluated in both simple and complex-layout setups. For this, we follow Wang et al., 2021 to compute a BLEU score between the layout-agnostic document texts serialized via simple rule (top-left to bottom-right) and the layout-preserving document texts parsed via ground-truth layouts to obtain a criterion (termed layout score) that reflects layout complexity (lower scores indicate more complex layouts that are difficult to be depicted via a simple parsing rule). Then, examples in DIT700K with the highest/lowest layout scores are selected as simple/complex-layout testsets, each having 512 examples. As for DITrans and M3T, their examples are manually selected and have complex layouts, so we split DITrans with the ratio *Train: Test = 4: 1* and keep all examples in M3T as test examples given that M3T is essentially a testset. Following Zhang et al., 2023, page-level BLEU and chrF++ are employed as evaluation metrics.

Baselines: Prior text/vision/layout-based methods are comprehensively compared with QRDIT.

- Text-Based Method - TextMT: It is a text-only encoder-decoder translation model that only utilizes the text modality of document images.

- Vision-Based Methods: a) **DonutTrans**: It is based on the document understanding model Donut (Kim et al., 2022), which utilizes a vision transformer (ViT) (Dosovitskiy et al., 2020) for image encoding and a text decoder for document content parsing/generation. We use DIT datasets to fine-tune the pre-trained Donut to be a DIT expert. b) **DIMTDA** (Liang et al., 2024): It is the SOTA vision-based DIT model that dynamically assembles visual and layout features from two pre-trained ViTs, followed by a translation decoder.

- Layout-Based Methods: a) **LayoutEnc-Dec** model series: They employ the pre-trained layout-aware transformers (*e.g.*, LayoutLMv3) as encoders to leverage document multi-modalities (text and visual layout) for DIT, and employ a text decoder to generate translation. We experiment with

Method	Modality	DIT700K-En (En→Zh) Simple-Layout BLEU	chrF	Complex-Layout BLEU	chrF	Average BLEU	chrF	DIT700K-En (En→De) Simple-Layout BLEU	chrF	Complex-Layout BLEU	chrF	Average BLEU	chrF	# Params
†DonutTrans[1]	V	32.87	44.54	24.28	33.77	28.58	39.16	34.27	56.96	25.81	49.97	30.04	53.47	131M
DIMTDA[2]	V	34.39	46.03	25.16	34.41	29.78	40.22	36.32	58.99	26.85	51.27	31.59	55.13	216M
TextMT[Roberta[3]]	T	33.24	46.06	25.96	37.55	29.60	41.81	36.28	60.53	26.47	56.19	31.38	58.36	149M
LayoutLM[4]-Dec	T+L	36.18	48.99	27.54	38.04	31.86	43.52	38.02	61.90	29.77	57.65	33.90	59.78	136M
†LayoutXLM[5]-Dec	T+L+V	36.50	49.68	28.63	39.42	32.57	44.55	39.75	63.34	30.04	58.06	34.90	60.70	387M
LiLT[Roberta][6]-Dec	T+L	36.32	48.97	30.21	41.02	33.27	44.99	38.44	61.76	31.23	58.00	34.84	59.88	152M
LayoutLMv3[7]-Dec	T+L+V	37.76	48.23	28.81	39.74	33.29	43.99	39.79	63.46	30.59	58.26	35.19	60.86	149M
LayoutDIT[8]	T+L	38.09	48.68	29.15	40.57	33.62	44.63	39.88	64.25	31.54	60.70	35.71	62.48	141M
QRDIT[LayoutLMv3[7]]	T+L+V	**39.19**	**50.24**	**31.72**	**42.82**	**35.46**	**46.53**	**41.39**	**65.61**	**34.26**	**63.13**	**37.83**	**64.37**	154M
		DIT700K-Zh (Zh→En)						DIT700K-Zh (Zh→Fr)						
†DonutTrans[1]	V	30.73	57.37	18.84	40.32	24.79	48.85	22.93	43.96	15.57	38.06	19.25	41.01	137M
DIMTDA[2]	V	32.68	59.27	21.79	46.77	27.24	53.02	26.12	48.97	18.80	44.40	22.46	46.69	221M
†TextMT[InfoXLM[9]]	T	29.45	57.17	17.22	41.04	23.34	49.11	24.51	51.35	15.58	38.31	20.05	44.83	293M
†TextMT[XLM-Roberta[10]]	T	29.91	58.67	18.19	41.64	24.05	50.16	26.29	52.37	15.57	37.83	20.88	45.10	301M
†LiLT[XLM][6]-Dec	T+L	35.57	60.85	26.23	49.56	30.90	55.21	30.41	56.46	23.36	45.95	26.88	51.21	304M
†LayoutXLM[5]-Dec	T+L+V	37.61	63.10	27.91	51.07	32.76	57.09	32.18	57.46	23.41	45.00	27.80	51.23	393M
†QRDIT[LayoutXLM[5]]	T+L+V	**41.79**	**64.35**	**37.29**	**61.81**	**39.54**	**63.08**	**35.14**	**58.80**	**31.57**	**56.32**	**33.36**	**57.56**	410M

Table 2: Results of DIT700K's four directions (En→Zh/De, Zh→En/Fr). **Bold**/underline indicate the best/second best. V, T, L respectively denote vision, text, and layout information that are leveraged as model input. †Multi-lingual model. []: Pre-trained weights used for model initialization. # model params for directions En→Zh and Zh→En are reported. [1](Kim et al., 2022); [2](Liang et al., 2024); [3](Liu et al., 2019); [4](Xu et al., 2020); [5](Xu et al., 2021); [6](Wang et al., 2022); [7](Huang et al., 2022); [8](Zhang et al., 2023); [9](Chi et al., 2021); [10](Conneau et al., 2019).

Dataset	DITrans (En→Zh)								M3T (En→Zh)						M3T (En→De)					
Subset/Domain	Report		News		Ad.		Average		RVL-CDIP		DocLayNet		Average		RVL-CDIP		DocLayNet		Average	
Method	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
TextMT[Roberta]	23.16	37.55	20.39	32.20	15.61	26.74	19.72	32.16	9.85	22.26	13.80	27.28	14.79	24.77	10.54	32.81	12.73	38.29	11.64	35.55
LayoutLM-Dec	25.05	38.69	21.10	33.50	18.24	29.94	21.46	34.04	11.77	22.76	15.77	29.00	13.77	25.88	12.34	35.24	15.94	41.66	14.14	38.45
LayoutLMv3-Dec	26.74	39.67	22.52	35.04	21.32	32.52	23.53	35.74	12.69	21.39	18.91	31.77	15.80	26.58	12.43	37.13	18.56	44.59	15.50	40.86
LayoutDIT	28.25	39.98	23.29	34.44	22.55	32.56	24.70	35.66	14.44	23.80	19.68	31.97	17.06	27.88	13.44	36.90	18.37	43.18	15.91	40.04
QRDIT[LayoutLMv3]	**29.41**	**41.86**	**24.62**	**37.76**	**26.49**	**37.90**	**26.84**	**39.17**	**15.94**	**26.02**	**20.04**	**32.53**	**17.99**	**29.28**	**16.10**	**38.58**	**20.00**	**47.43**	**18.05**	**43.00**

Table 3: Results of DITrans' three domains (En→Zh) and M3T's two subsets (En→Zh and En→De).

several widely-used layout-aware transformers including the representative LayoutLM (Xu et al., 2021), the two-stream LiLT (Wang et al., 2022), the models further include vision information beyond layout - LayoutXLM (Xu et al., 2021) and layoutLMv3 (Huang et al., 2022). Baselines are denoted as LayoutEnc-Dec (*e.g.*, LayoutLMv3-Dec). b) **LayoutDIT** (Zhang et al., 2023): It is an improved variant of LayoutLM-Dec which decomposes LayoutLM-Dec decoder to a three-step decoder to alleviate long-context and wrong-order problems and shows advanced DIT performance.

Refer to App. A for more details about model implementation.

4.2 Comparisons with Existing Methods

QRDIT is evaluated on above three public benchmarks including DIT700K, DITrans, and M3T.

DIT700K: Results on DIT700K are reported in Tab. 2. Methods all perform better under simple-layout setups than complex-layout setups, demonstrating layout complexity effects on DIT. In En-Zh direction, text-only or vision-only methods (*e.g.*, TextMT, DIMTDA) perform worst due to their overlook or inadequate leverage of the crucial layout information. Compared with them, methods from LayoutLM-Dec to LayoutDIT achieve better results since they have incorporated more comprehensive multi-modality information. Notably, by concentrating on local regions and extracting the most relevant regional texts as translation queries, QRDIT achieves the best results. Specifically, compared with its backbone LayoutLMv3-Dec, QRDIT brings 1.43/2.91 BLEU improvement under Simple/Complex-Layout setups. Its SOTA performances are also consistently observed in En→De and Zh→En/Fr directions.

DITrans and M3T: Results on DITrans and M3T

Setup	Query Prefix Identification			Query Words Extraction					
				Query Words			Non-Query Words		
	P	R	F_1	P_{25}	M	P_{75}	P_{25}	M	P_{75}
DIT700K-En (En→Zh)									
Simple	93.98	92.33	92.52	96.66	97.10	98.89	1.71	2.72	5.28
Complex	93.46	89.73	90.62	88.95	95.17	96.91	2.00	3.84	7.97
DIT700K-Zh (Zh→En)									
Simple	94.80	88.72	90.67	89.65	94.94	96.59	4.69	7.52	13.09
Complex	90.98	89.64	88.50	87.12	91.27	93.60	7.71	12.02	15.97
DIT700K-En (En→De)									
Simple	93.79	92.19	92.38	94.73	96.80	97.80	1.68	2.67	4.89
Complex	93.34	89.63	90.57	87.82	95.61	96.37	2.01	3.77	7.71
DITrans-Report (En→Zh)									
Complex	95.81	93.93	94.42	83.72	92.90	97.26	1.88	3.92	9.41
DITrans-News (En→Zh)									
Complex	93.46	93.81	93.24	86.88	91.78	94.21	2.06	4.03	8.32

Table 4: Evaluation results of query extraction ability.

Query Extraction Strategy		Simple-Layout BLEU	Simple-Layout chrF	Complex-Layout BLEU	Complex-Layout chrF
Top-k [$k = k' \times L_{sent.}$] Deterministic Extraction Strategy ($L_{sent.}$: Avg. Sent. Length)	k'=0.5	16.88	30.01	16.33	28.44
	k'=1.0	21.86	35.79	18.76	31.38
	k'=1.5	29.34	42.51	25.37	37.34
	k'=2.0	31.19	44.10	26.70	38.57
	k'=2.5	29.86	43.16	24.10	36.23
	k'=3.0	28.70	42.35	23.54	34.94
Adaptive Extraction Strategy *ours*		**39.19**	**50.24**	**31.72**	**42.82**

Table 5: Comparing different query extraction strategies.

are summarized in Tab. 3. Note that all models are continually trained with DITrans training data using their pre-trained checkpoints on the large-scale DIT700K for warming-up. As for M3T, since it is essentially a testset without sufficient training data, all models are tested using their checkpoints trained on DIT700K. As Tab. 3 shows, models' performances are relatively lower than those on DIT700K due to the more complex layouts (as Tab. 1 shows) and much fewer or unavailable training data in DITrans and M3T. Similarly, LayoutLMv3-Dec outperforms TextMT due to its multi-modality pre-training on massive documents. Our QRDIT still achieves the best results in all test subsets/domains, achieving new SOTA performance.

4.3 Analysis on Query Modules

Our model's improvements stem from the concentrations of relevant regions and accurately extracted regional words. Therefore, this section deeply investigates modules and processes in query stage.

Quantitative Evaluation of Query Extraction:

We first provide quantitative results of the two key sub-processes - query prefix identification and query words extraction. Specifically, 1) for query prefix identification, we evaluate model's prefix/non-prefix classification capability, of which the metrics are Precision (P), Recall (R), and F_1. 2) For query words extraction, we inspect model's concentration scores allocated on the golden relevant regions (*i.e.*, ground-truth words of each sentence) and irrelevant regions. The statistics for each region's concentration scores (averaged over the entire testset), including 25/75-th percentile (P_{25}, P_{75}) and Median (M), are reported.

As Tab. 4 shows, QRDIT achieves high F_1 ($\geq$ 90%) for query prefix identification. Conditioning on these accurate prefixes, the subsequent query words extraction is also well accomplished, exhibiting high median ($\geq$ 90%) for query words and low median ($\leq$ 15%) for non-query words, demonstrating that model effectively distributes the majority ($\geq$ 90%) of its concentrations to relevant regional words, thereby obtaining the intact and non-redundant source texts.

Visualizing Relevant Regional Concentration: We further analyze our model's relevant regional concentration results as well as its translation results against the baseline model LayoutLMv3-Dec. As Fig. 4 shows, in the first case, LayoutLMv3-Dec ignores the separation between the two columns, leading to falsely mixed, cluttered source text and false translation. In contrast, QRDIT accurately locates the relevant regional words by allocating remarkable concentration scores to them, thereby obtaining correct queries/source texts and translations. Likewise, in the second case, LayoutLMv3-Dec overlooks the table structure while QRDIT effectively captures table cells. The effectiveness of query modules in our framework is notably demonstrated with these examples.

Ablation on Adaptive Extraction Strategy: The proposed adaptive extraction strategy aims to accurately select the query words while filtering out non-query words for a given query prefix, relying on their concentration scores. We also compare it with the deterministic strategy (on DIT700K En→Zh task), which selects the words with top-k highest scores. Here the parameter k is a multiple (the multiplier factor is denoted as k') of the average sentence length $L_{sent.}$ in each document. *E.g.*, $k' = 1$ means that the # words for each query is equivalent to $L_{sent.}$. As Tab. 5 shows, performance rapidly increases as k' improves, but reaches a saturation at $k' = 2.0$ and then a slight decrease. These results align with our motivation, since deterministically setting a fixed k for all queries violates

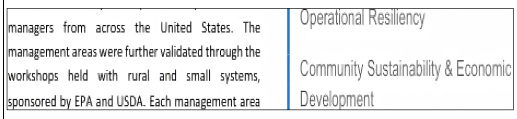
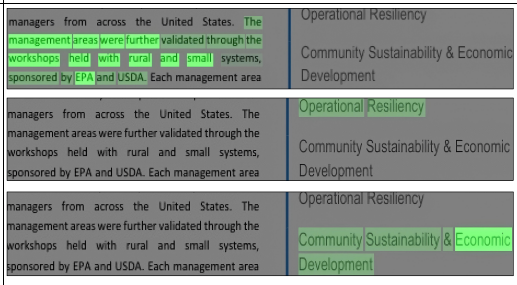
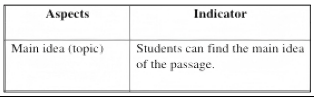
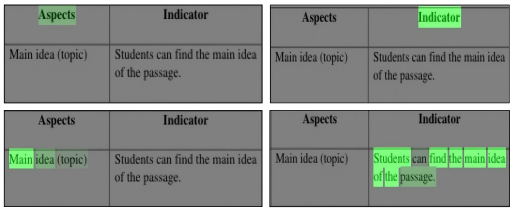
Model	Document Image	Extracted Query / Source Text	Translation Results
Layout LMv3-Dec		**Source Text:** Operational Resiliency The management areas were further validated through the Community Sustainability & Economic workshops held with rural and small systems, Development sponsored by EPA and USDA. ✗	**Translation:** 运营 弹性 管理 领域 通过 由 环保 局 和 美国 农业 部 赞助 的 农村 和 小型 系统 发展 举办 的 社区 可持续性 和 经济 研讨会 得到 了 进一步 的 验证 。 ✗
QRDIT (Ours)		**Query #1:** The management areas were further validated through the workshops held with rural and small systems, sponsored by EPA and USDA. ✓ **Query #2:** Operational Resiliency ✓ **Query #3:** Community Sustainability & Economic Development ✓	**Response #1:** 通过 由 环保 局 和 美国 农业 部 赞助 的 农村 和 小型 系统 举办 的 研讨会 ， 进一步 验证 了 管理 区域 。 ✓ **Response #2:** 操作 弹性 ✓ **Response #3:** 社区 可持续性 & 经济 发展 ✓
Layout LMv3-Dec		**Source Text:** Aspects Indicator Main idea (topic) Students can find the main idea of the passage. ✗	**Translation:** 方面 指标 主旨 （主题） 学生 可以 找到 文章 的 主旨。 ✗
QRDIT (Ours)		**Query #1:** Aspects ✓ **Query #2:** Indicator ✓ **Query #3:** Main idea (topic) ✓ **Query #4:** Students can find the main idea of the passage. ✓	**Response #1:** 方面 ✓ **Response #2:** 指示 ✓ **Response #3:** 主要 思想 （主题） ✓ **Response #4:** 学生 可以 找到 文章 的 主要 思想 。✓

Figure 4: Two visualization DIT cases on document images with two columns (top) and a table (bottom). For brevity, the images have been cropped to only showcase text regions of interest. For each case, we present the extracted queries/source texts and translations from our model QRDIT and baseline LayoutLMv3-Dec. Particularly, QRDIT's concentration scores on relevant regional words are plotted in green color (darker colors symbolize greater scores).

the prior that # words varies for different queries, thereby probably suffering incomplete or excessive recall issue for some queries. In contrast, our adaptive strategy significantly outperforms deterministic strategy by selecting query words through adaptive threshold.

4.4 Analysis on Response Modules

Ablation on Semantics Enhancement: The proposal of text semantics enhancement in response stage aims to enhance the textual information by aggregating text embedding and query feature through dynamic gate. To investigate its effects, we compare it with other options on DIT700K En→Zh task. As Tab. 6 shows: 1) Dynamic enhancement (model a) significantly outperforms primitive query feature without enhancement (model b), verifying its positive effect. 2) Model (a) also outperforms model (c) which discards the query feature but only uses text embedding from encoder. We infer that although text embedding contains rich textual information, discarding its aggregation with query feature destroys the query-response modules' joint coordination, therefore deteriorating performance. 3)

Tag	Model	Simple-Layout BLEU	Simple-Layout chrF	Complex-Layout BLEU	Complex-Layout chrF
(a)	w/ Dynamic Gate Agg. *ours*	**39.19**	**50.24**	**31.72**	**42.82**
(b)	w/o Agg. (Use Query Feat.)	37.34	48.02	29.86	40.57
(c)	w/o Agg. (Use Text Emb.)	32.82	44.32	25.13	36.80
(d)	w/ Mean Pooling Agg.	33.05	44.25	25.20	36.51
(e)	w/ Concat Projection Agg.	38.41	49.60	30.98	42.25

Table 6: The effects of text semantics enhancement.

Two other aggregation techniques - stationary mean pooling (model d) and concat projection (model e) - also show worse results (mean pooling aggregation is even worse than using query feature without aggregation in model b), verifying the advantages of dynamic gate aggregation.

Evaluating the Efficiency Promotion from Per-Query Response: Benefiting from the per-query response strategy, QRDIT achieves the parallel response to all queries in one pass, leading to promoted efficiency. For quantitative comparison, we evaluate the average time (on NVIDIA A100) that model consumes for each training step (as training efficiency) and for each test image (as inference

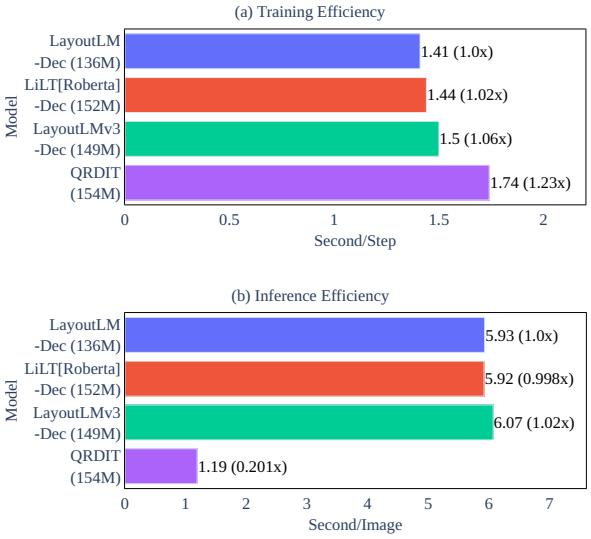


Figure 5: Quantitative evaluation of model efficiency.

efficiency). Results are shown in Fig. 5, which demonstrates that: 1) The training efficiency of QRDIT is on par with baseline models, causing only 16% more training time than LayoutLMv3-Dec. The extra time is consumed by some intermediate predictions (*e.g.*, the predictions for query prefixes and relevant regions). 2) The efficiency of inference (in auto-regressive mode) is lower than training (in teacher-forcing mode). However, our model notably shows $\sim 5\times$ higher inference efficiency than baseline models, demonstrating the significant efficiency promotion from the per-query response in our framework.

5 Related Work

Deep neural models have demonstrated significant effectiveness and driven the expansion of machine translation from plain text to multimodal domains (Liang et al., 2024; Ma et al., 2023a,b,c; Yu et al., 2024, 2025; Zhang et al., 2023; Zhao et al., 2023). As a multimodal machine translation task, Document Image Translation (DIT) involves the collaborative processing of document text and visual layout. There have been many works oriented at the simple-layout single-line/paragraph text image (*e.g.*, movie subtitle or street view text) translation. Most of them attempt to alleviate the image-text modality gap in feature space, with techniques such as multi-modal codebook (Lan et al., 2023), cross-modal knowledge transfer (Zhu et al., 2023) or distillation (Ma et al., 2023c), and multi-modal contrastive learning (Ma et al., 2023a). Some recent studies (Lan et al., 2024; Tian et al., 2023) begin exploring the image-format output for both text translation and visual style preserving. However, these methods may suffer deteriorated performance for whole-page complex-layout documents due to the lack of visual layout modeling toward complex layouts. For complex-layout documents, early cascade solutions rely on an OCR-translation pipeline (Afli and Way, 2016) or add an extra layout parser (Hinami et al., 2021) for layout analysis. Many recent works have verified incorporating visual layout into an end-to-end DIT model by encoding image pixels (Liang et al., 2024) or word positions (Zhang et al., 2023). Nevertheless, these methods lack special awareness of key text areas during translation. As a resolution, our proposed framework deploys specific modules and objectives toward relevant regional concentration to ensure more accurate source texts & translations.

6 Conclusion

This paper proposes the Query-Response framework. It integrates relevant regional concentration capabilities, achieving accurate relevant text extraction and improving translation. We also propose a dynamic aggregation mechanism to enhance query feature text semantics toward translation. Experiments in four directions on three datasets demonstrate QRDIT significantly outperforms prior methods and achieves new SOTA performance.

Limitations

Since our model relies on the LayoutXLM backbone model to process non-English (*e.g.*, Chinese) document images, its model size is accordingly enlarged (compared with its English counterpart that relies on LayoutLMv3) due to LayoutXLM's multi-lingual embedding table. In our future work, we will attempt to compress the vocabulary and embedding table of LayoutXLM toward a specific language via techniques such as vocabulary transfer learning or embedding table clipping to maintain a compact model without redundant parameters, which would be of practical value for realistic scenarios where translating massive documents in a specific language is required.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant numbers 62336008 and 62476275.

References

Haithem Afli and Andy Way. 2016. Integrating optical character recognition and machine translation of historical documents. In *Proc. of LT4DH*, pages 109–116.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proc. of ECCV*, page 213–229.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proc. of NAACL*, pages 3576–3588.

Alexis Conneau, Guillaume Lample, and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Proc. of NIPS*.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of ICLR*, pages 1–22.

Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. Towards fully automated manga translation. In *Proc. of AAAI*, pages 12998–13008.

Benjamin Hsu, Xiaoyu Liu, Huayang Li, Yoshinari Fujinuma, Maria Nadejde, Xing Niu, Ron Litman, Yair Kittenplon, and Raghavendra Pappagari. 2024. M3T: A new benchmark dataset for multi-modal document-level machine translation. In *Proc. of NAACL*, pages 499–507.

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 *Proc. of ACM MM*, pages 4083–4091.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *ArXiv*, abs/2301.0874.

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 *Proc. of ECCV*, pages 498–517.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, Min Zhang, and Jinsong Su. 2024. Translatotron-v(ison): An end-to-end model for in-image machine translation. In *Proc. of ACL Findings*.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook. In *Proc. of ACL*, pages 3479–3491.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. In *Proc. of NAACL*, pages 7077–7088.

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*, abs/1907.11692.

Cong Ma, Xu Han, Linghui Wu, Yaping Zhang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023a. Modal contrastive learning based end-to-end text image machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–13.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023b. E2timt: Efficient and effective modal adapter for text image machine translation. In *Proc. of ICDAR*, pages 70–88.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023c. Multi-teacher knowledge distillation for end-to-end text image machine translation. In *Proc. of ICDAR*, pages 484–501.

K Chandra Shekar, Maria Anisha Cross, and Vignesh Vasudevan. 2021. Optical character recognition and neural machine translation using deep learning techniques. In *Proc. of ICICSE*, pages 277–283.

Yanzhi Tian, Xiang Li, Zeming Liu, Yuhang Guo, and Bin Wang. 2023. In-image neural machine translation with segmented pixel sequence-to-sequence model. In *Proc. of EMNLP Findings*, pages 15046–15057.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*, pages 5998–6008.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. Lilt: A simple yet effective language-independent layout transformer for structured document understanding. In *Proc. of ACL*, pages 7747–7757.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. Layoutreader: Pre-training of text and layout for reading order detection. In *Proc. of EMNLP*, pages 4735–4744.

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 *Proc. of KDD*, pages 1192–1200.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, and Furu Wei. 2021. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *ArXiv*, abs/2104.08836.

Donglei Yu, Xiaomian Kang, Yuchen Liu, Yu Zhou, and Chengqing Zong. 2024. Self-modifying state modeling for simultaneous machine translation. In *Proc. of ACL*, pages 9781–9795.

Donglei Yu, Yang Zhao, Jie Zhu, Yangyifan Xu, Yu Zhou, and Chengqing Zong. 2025. Simulpl: Aligning human preferences in simultaneous machine translation. In *Proc. of ICLR*, pages 1–23.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023. LayoutDIT: Layout-aware end-to-end document image translation with multi-step conductive decoder. In *Proc. of EMNLP Findings*, pages 10043–10053.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2025. From chaotic ocr words to coherent document: A fine-to-coarse zoom-out network for complex-layout document image translation. In *Proc. of COLING*, pages 10877–10890.

Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2023. Transformer: A general framework from machine translation to others. *Machine Intelligence Research*, 20:514–538.

Shaolin Zhu, Shangjie Li, Yikun Lei, and Deyi Xiong. 2023. PEIT: Bridging the modality gap with pre-trained models for end-to-end image translation. In *Proc. of ACL*, pages 13433–13447.

A Implementation Details

Model Architecture: The four key modules including the document feature extraction module, query prefix module, regional concentration module, and response/translation module all utilize the stacked transformer encoder or decoder layers as the structure. Specifically, we set their layer numbers to 6, 1, 2, and 4, respectively. Following prior literature (Wang et al., 2022; Xu et al., 2020), the resolution of the input image is normalized to 1000×1000 pixels. Baseline models and ours are initialized with corresponding document transformers that have been pre-trained on massive documents. E.g., our model employs the pre-trained LayoutLMv3 (for En→Zh/De) or LayoutXLM (for Zh→En/Fr) to initialize its feature extraction module. Other modules are randomly initialized for training.

Pre-training and Fine-tuning: As described in Sec. 4.2, to improve performance, we first pre-train QRDIT on the large-scale DIT700K. After pre-training, model is further fine-tuned on the DITrans dataset for adaptation or zero-shot tested on the M3T dataset. Both pre-training and fine-tuning employ the multi-task losses as described in Sec. 3.3. We utilize the Adam optimizer with an initial learning rate of 1×10^{-4} for pre-training and 4×10^{-5} for fine-tuning, and incorporate a linear decay schedule to stabilize convergence. Model is pre-trained on DIT700K for 2 epochs with a batch size of 10. As for fine-tuning, model is tuned for 20 epochs with a smaller batch size of 6 to ensure optimal performance and sufficient training steps for convergence. Following prior literature (Zhang et al., 2023), during inference, beam search (beam size is set to 4) is leveraged by the response/translation module to improve translation quality.

B Comparison with Advanced Large Models

Since large models have exhibited inspiring abilities across various text and multi-modality tasks, this section aims to evaluate their DIT performance to provide knowledge of their DIT capabilities as well as to conduct a comparison with our model. To this end, we choose the DIT700K and DITrans datasets as the benchmark. Specifically, the English/Chinese complex-layout subsets from DIT700K and the Report/News-domain subsets from DITrans are included, with a total of 256 image examples (64 examples are randomly picked up from each subset). As for the large mod-

Model	DIT700K-En En→Zh		DIT700K-Zh Zh→En		DITrans-Report En→Zh		DITrans-News En→Zh		Average	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
GPT4-o1	42.35	42.65	46.58	42.00	32.90	29.52	32.89	33.28	38.68	36.86
Gemini-1.5-Pro	44.07	43.71	43.68	41.43	30.90	30.59	30.98	32.17	37.41	36.98
QRDIT$_{Base,154M}$	36.79	47.56	39.65	62.31	28.19	39.65	23.57	36.97	32.05	46.62
QRDIT$_{Large,345M}$	**46.30**	**55.59**	**49.90**	**73.56**	**35.44**	**44.74**	**32.42**	**46.87**	**41.02**	**55.19**

Table 7: Comparison with advanced large models for document image translation.

els, we choose two representative and most advanced[1] large models - GPT4-o1[2] and Gemini-1.5-Pro[3] - as the test models. Models are instructed to generate the translation of the input document image, with the prompt "You are a professional translator, please accurately and fluently translate the text on the image from English/Chinese to Chinese/English: {Document Image}", referring to literature (Jiao et al., 2023). As for our model, we scale its model size (within a reasonable range) to stimulate and improve its performance while keeping it a small-scale DIT expert to compare it with these general-purpose large models. Specifically, we mainly scale the feature extraction, regional concentration, and response/translation modules' layer numbers (from their primitive 6, 2, and 4 layers to 12, 6, and 12 layers, respectively). This scaled version is denoted as QRDIT$_{Large}$.

The evaluation results are shown in Tab. 7. 1) Although being devised and trained as a multi-task generalist, both GPT4-o1 and Gemini can adequately tackle the challenging DIT task, especially on DIT700K web documents (*e.g.*, Gemini's 44.07 BLEU on DIT700K-En, GPT4-o1's 46.58 BLEU on DIT700K-Zh). 2) With appropriate scaling, our task-specific DIT model QRDIT$_{Large}$ consistently shows the best results on all test subsets, resulting in the SOTA average overall performance. These results demonstrate that the task-specific training and the explicit modeling of relevant regional concentration with our query-response framework can lead to a small-size expertise DIT model with superior parameter efficiency and DIT capability compared with the large models. However, we also approve of large models' simple and unified architectures and their fluent translation generation abilities. Therefore, in our future study, we will investigate the cooperation between small expertise models (like our QRDIT) and the large generalist models, such as prioritizing our model before a large model for document sentence positioning and extraction to provide well-formalized document structure for the large model-based document understanding or translation.

[1]Up to our test date December 2024.

[2]https://chatgpt.com/

[3]https://gemini.google.com/app