From Chaotic OCR Words to Coherent Document: A Fine-to-Coarse Zoom-Out Network for Complex-Layout Document Image Translation

Zhiyang Zhang[1,2], **Yaping Zhang**[1,2*], **Yupu Liang**[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) aims to translate documents in images from one language to another. It requires visual layouts and textual contents understanding, as well as document coherence capturing. However, current methods often rely on the quality of OCR output, which, particularly in complex-layout scenarios, frequently loses the crucial document coherence, leading to chaotic text. To overcome this problem, we introduce a novel end-to-end network, named Zoom-out DIT (ZoomDIT), inspired by human translation procedures. It jointly accomplishes the multi-level tasks including word positioning, sentence recognition & translation, and document organization, based on a fine-to-coarse zoom-out framework, to progressively realize "chaotic words → coherent document" and improve translation. We further contribute a new large-scale DIT dataset with multi-level fine-grained labels. Extensive experiments on public and our new dataset demonstrate significant improvements in translation quality towards complex-layout document images, offering a robust solution for reorganizing the chaotic OCR outputs to a coherent document translation.

1 Introduction

Document images such as scans, PDF renderings are important carriers of human knowledge. Document Image Translation (DIT), which is a crucial task of digital transformation, aims to generate the target-language translation for a document image based on its visual cues and textual contents (Zhang et al., 2023). However, DIT is a challenging task in practical applications and is faced with numerous difficulties (Cui et al., 2021): various document types, complex layouts, semantic understanding and cross-lingual translation, *etc.*

Currently, two groups of studies have been devoted to DIT task. The first group, vision-based methods (Lan et al., 2023; Liang et al., 2024; Mansimov et al., 2020; Tian et al., 2023; Zhu et al., 2023), directly input the visual features encoded by a vision encoder (*e.g.*, ViT (Dosovitskiy et al., 2021)) to a translation decoder. The second group, text-based methods, use the words extracted by Optical Character Recognition (OCR) for translation. They either use the sole text modality (Afli and Way, 2016; Hinami et al., 2021) or combine additional visual layouts with textual contents to leverage multi-modalities (Zhang et al., 2023), and achieve state-of-the-art (SOTA) performance. However, as shown in Fig. 1, when dealing with complex document images (Fig. 1 (a)), the translation should rely on coherent document, where words are grouped as semantically complete and logically organized sentences (Fig. 1 (c)), rather than chaotic OCR words (Fig. 1 (b)). Accordingly, a more favorable DIT framework should involve tasks spanning multiple levels (from word to sentence, and to document), including word positioning, sentence

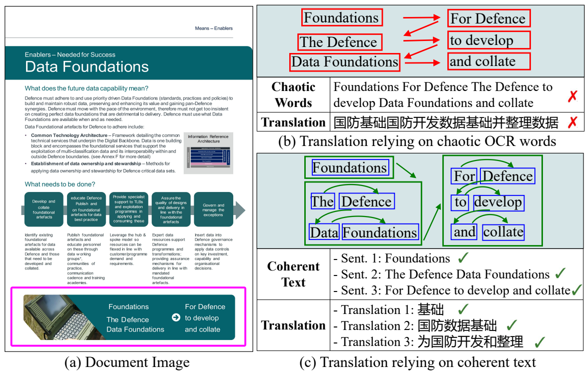


Figure 1: The critical multi-level tasks in DIT. (a) Document image. Red box indicates the text to translate. (b) DIT relying on solely word-level information from chaotic OCR words causes false translation. (c) Accomplishing the multi-level tasks including word positioning, sentence recognition, and document organization rearranges chaotic words to coherent text, thereby obtaining correct and well-formalized translation results.

* Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 10877–10890
January 19–24, 2025.

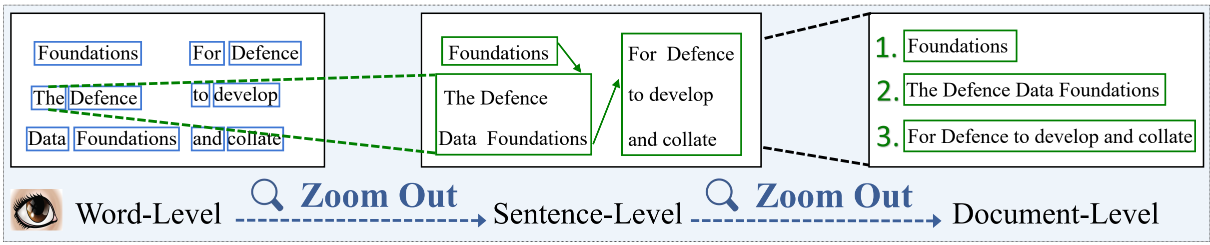


Figure 2: Overview of Zoom-Out Network. It first attends to the finest word-level text and visual layout, then zooms out to accomplish coarse sentence-level tasks, finally formalizes and outputs the global document-level translation. By fulfilling these multi-level tasks, it realizes "chaotic words → coherent document", thereby improving translation.

recognition & translation, and document organization. Nevertheless, in current methods, since there are no special modules and objectives to guide the multi-level tasks modeling, only the word-level information is used, and we consider their DIT capabilities are limited. *Therefore, how to effectively model and unify the multi-level tasks into DIT for a coherent document translation, is the vital step to improve the performance of DIT.*

To this problem, this paper introduces a novel end-to-end framework, named Zoom-out DIT (ZoomDIT), to model the multi-level tasks for DIT. In this framework, model's focus progressively "zooms out" from the finest word level to coarse sentence level, and finally reaches the global document level, resembling human translation patterns. Specifically, 1) First, at **word level**, model focuses on capturing each word's text and visual layout. 2) Second, at **sentence level**, model progressively locates, completes, translates and organizes sentences, aiming to reorganize the original chaotic OCR words to semantically intact, logically ordered sentences and generate their translations. 3) Third, at **document level**, model associates source and target sentences, formalizing them to a coherent document translation as DIT results. Each level deploys task-specific modules. A consecutive feature flows across them to unify modules as an end-to-end whole. *By modeling and integrating the multi-level tasks for DIT, ZoomDIT effectively realizes "chaotic words → coherent document" and improves translation quality.*

In addition, to facilitate DIT's further advancement, we propose a data pipeline that enables automatic web document extracting and fine-grained labels annotating. With this pipeline, we contribute the DIT700K dataset. Compared with prior DITrans (Zhang et al., 2023) and M3T (Hsu et al., 2024) datasets, DIT700K contains more document images (>700K) of various disciplines and provides multi-level fine-grained labels for DIT. Extensive experiments on DIT700K and the public DITrans in three translation directions show the SOTA performance of ZoomDIT. Our contributions are:

- A novel end-to-end DIT framework is proposed. It integrates the multi-level tasks, that have been largely overlooked, into DIT. With intrinsic document coherence capturing abilities, it relieves the reliance on chaotic OCR outputs and improves translation qualities.
- A new automatic data pipeline and benchmark DIT700K, which is the most large-scale and fine-grainedly labeled dataset, will be released to community[1].
- Experiments show the proposed ZoomDIT significantly outperforms prior SOTAs.

2 Zoom-Out DIT Network

Fig. 2 is the overview of our proposed Zoom-out DIT (ZoomDIT) Network. Its focus gradually zooms out from the finest word level, to coarse sentence level, and then to global document level. 1) At word level, model combines each word's textual, layout, and visual features as multi-modality features. 2) At sentence level, model accomplishes sentence prefix identification, completion, translation, and organization tasks to derive semantically intact, logically organized sentences, and generate their translations. 3) At document level, model formalizes these sentences and their paired translations to coherent document translation as DIT results. The internal structure of ZoomDIT is shown in Fig. 3.

2.1 Word-Level: Multi-Modal Feature Extraction

As shown in Fig. 3, the input is OCR words of a document image, and words have been serialized

[1]https://huggingface.co/datasets/zhangzhiyang/DIT700K

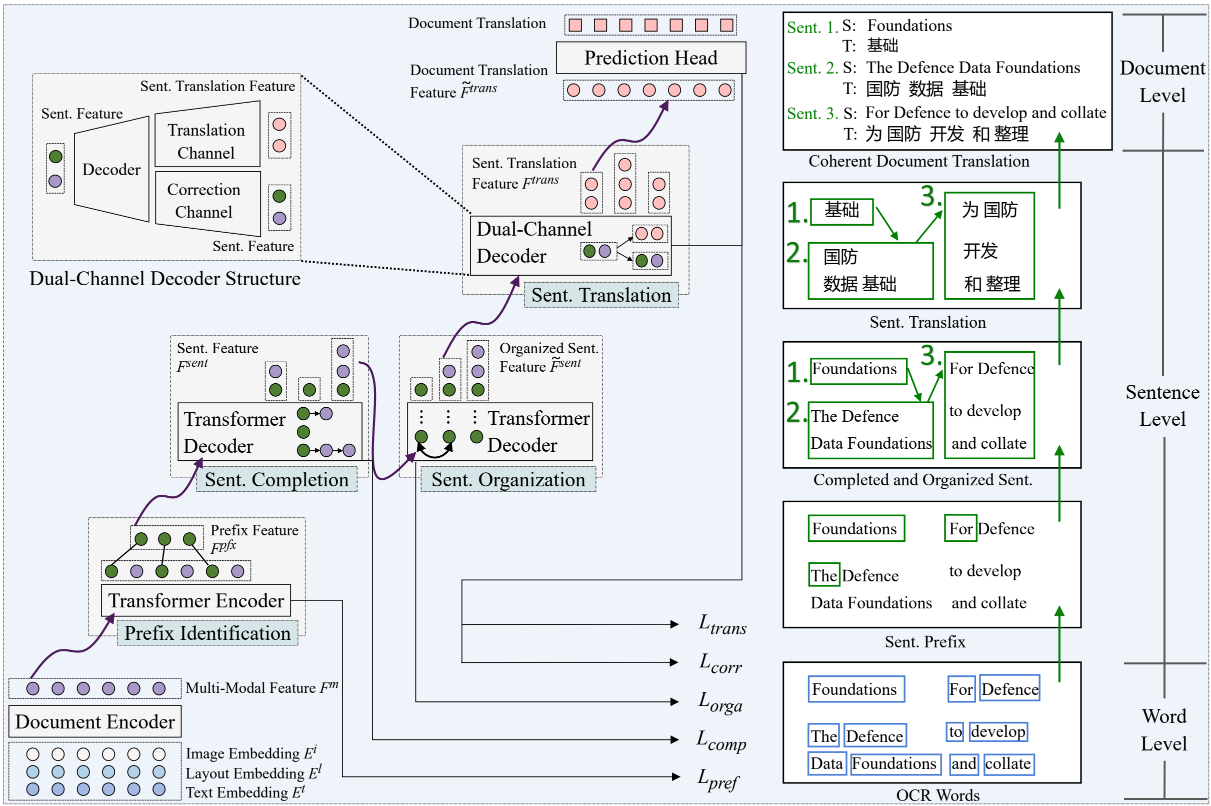


Figure 3: ZoomDIT's internal structure. Model climbs from bottom to top to fulfill tasks at each level, from 1) multi-modal feature extraction at word level, to 2) sentence prefix identification, completion, organization and translation at sentence level, to 3) formalizing coherent translation at document level. Each level deploys task-specific transformer-based modules. Feature flows across them consecutively to unify modules as an end-to-end whole.

according to a top-left to bottom-right order. Given the image and OCR words, each word's text embeddings E^t, layout embeddings E^l, and image embeddings E^i are extracted following previous literature (Huang et al., 2022). Take layout embeddings as an example, given a word bounding box $b = (x_{tl}, y_{tl}, x_{br}, y_{br})$, its top-left and bottom-right coordinates are encoded by looking-up two learnable embedding tables Emb_x and Emb_y respectively for x/y-direction:

$$E^l = \text{Lin}([\text{Emb}_x(x_{tl}, x_{br}); \text{Emb}_y(y_{tl}, y_{br})]) \quad (1)$$

where $[\cdot]$ is concatenation and $\text{Lin}(\cdot)$ is linear projection for dimension compatibility.

Embeddings of all modalities are aggregated via addition and fed into a document transformer to obtain contextualized multi-modal feature F^m.

2.2 Sentence-Level: Coherence Capturing and Translation

Conditioning on multi-modal word feature F^m, four tasks are accomplished at the sentence level: 1) A sentence prefix identification task identifies each sentence's prefix word; 2) A sentence completion task predicts suffix words to complete a sentence given prefix; 3) A sentence organization task organizes sentences in logical order; 4) A translation task generates each source sentence's translation.

Sentence Prefix Identification (SPI): It aims to identify the prefix word of each sentence. Specifically, given the word feature sequence F^m, a transformer encoder is employed for feature refinement, then a linear projection head classifies each word as *Prefix* or *Non-Prefix*. Loss function for SPI is:

$$\mathcal{L}_{pref} = \sum_{i=1}^{L} \text{CE}(pref_i, P_i^{pref})/L \quad (2)$$

where $pref_i$ is ground-truth prefix label for i-th word. P_i^{pref} is i-th word's classification probability. $\text{CE}(\cdot)$ is $\text{CrossEntropy}(\cdot)$. L is sequence length (*i.e.*, the number of words).

Sentence Completion (SC): It aims to complete the suffix words given a sentence's prefix. Specifically, with the prefix word as the beginning and F_m

as the context for cross-attention, SC employs a transformer decoder to auto-regressively calculate the hidden state F_i^{sent} for a given timestep i. Then, F_i^{sent} is used to calculate cosine similarity with each word in sequence F^m:

$$P_i^{sent} = \frac{\exp((F_i^{sent})^\intercal F_j^m + b_j)}{\Sigma_k \exp((F_i^{sent})^\intercal F_k^m + b_k)} \tag{3}$$

where P_i^{sent} is the normalized similarity score between i-th word and each word from F^m, b is learnable bias. The word with highest score is retrieved as i-th word of current sentence. This process continues until it meets the prefix of another sentence. After the completion of all sentences, we obtain the feature sequence $\{F_k^{sent}\}_{k=1}^M$ of M sentences, each feature F_k^{sent} corresponding to a sentence. Loss function for SC is:

$$\mathcal{L}_{comp} = \sum_{i=1}^{L_k} \mathrm{CE}(s_i, P_i^{sent}) / \sum_{i=1}^{L_k} i \tag{4}$$

where s_i is i-th word's ground-truth one-hot similarity score distribution over F_m. L_k is k-th sentence's sequence length. Note that above loss is for a single sentence. The total loss for the SC task should be further averaged over all sentences for a document.

Sentence Organization (SO): Since sentences are spatially placed onto the 2-dimensional image, the SO task aims to derive sentences' logical order to guarantee document coherence for translation. Specifically, given the sentence feature sequence $\{F_k^{sent}\}_{k=1}^M$ and the prefix word feature of each sentence $\{F_k^{pref}\}_{k=1}^M$, SO employs a transformer decoder to predict the prefixes' logical order (which is also equivalent to sentences' order). It employs $\{F_k^{pref}\}_{k=1}^M$ as the context for cross-attention and uses a "[CLS]" special token to prompt the decoding process to auto-regressively decide which sentence's prefix in $\{F_k^{pref}\}_{k=1}^M$ is logically adjacent to current sentence prefix. The decoding continues until all prefixes have been selected, after which sentences are organized in correct logical order. The reordered sentence feature is denoted as $\{\widetilde{F}_k^{sent}\}_{k=1}^M$. Loss function for SO is:

$$\mathcal{L}_{orga} = \sum_{k=1}^{M} \mathrm{CE}(ord_k, P_k^{ord}) / M \tag{5}$$

where ord_k is k-th sentence's ground-truth order, P_k^{ord} is the classification probability over [1, M].

Sentence Translation (ST): It is in charge of sentence translation. Considering that the text semantics in $\{\widetilde{F}_k^{sent}\}_{k=1}^M$ may be deviated due to the noisy OCR input and preceding translation-agnostic tasks, we employ a dual-channel decoder following Passban et al., 2021. It comprises a correction channel to generate the denoised source sentence and a translation channel to generate each sentence's translation. Take the translation channel as an example, given k-th sentence's feature $\widetilde{F}_k^{sent}$, translation channel calculates the hidden states as follows (subscript k is omitted for simplicity):

$$H_{n,\le j}^{trans} = \mathrm{MHCA}(\mathrm{MHSA}(H_{n-1,\le j}^{trans}, O_j), \widetilde{F}^{sent}) \tag{6}$$

where $H_{n,\le j}^{trans}$ is the hidden states output by the n-th layer, MHSA/MHCA denotes multi-head self/cross attention (Vaswani et al., 2017), O_j is causal attention mask. $H_{N,\le j}^{trans}$ from the top layer is employed as the translation features $\{F_k^{trans}\}_{k=1}^M$. Features of the correction channel are calculated similarly.

2.3 Document-Level: Formalize and Output Coherent Document Translation

At the document level, translation features of all sentences $\{F_k^{trans}\}_{k=1}^M$ are sequentially concatenated as document translation feature, based on which a translation head predicts the target-language token to generate the document translation. Note that to promote training and inference efficiency, during implementation, the translation head is applied to sentence features $\{F_k^{trans}\}_{k=1}^M$ to generate all sentence translations in parallel, which is equivalent to translating document features:

$$P_{k,j}^{trans} = \mathrm{Softmax}\left(\mathrm{Linear}\left(F_{k,j}^{trans}\right)\right) \tag{7}$$

where $P_{k,j}^{trans}$ is the classification probability over target-language vocabulary. Based on $P_{k,j}^{trans}$, the target token is predicted via beam search. Loss function for the translation channel and head is:

$$\mathcal{L}_{trans} = \sum_{k=1}^{M} \sum_{j=1}^{|Y_k|} \mathrm{CE}(Y_{k,j}, P_{k,j}^{trans}) / \sum_{k=1}^{M} \sum_{j=1}^{|Y_k|} j \tag{8}$$

where $Y_{k,j}$ is the ground-truth target token of the k-th sentence at timestep j. Loss function for the correction channel is similar and is denoted as $\mathcal{L}_{corr}$.

By associating translations with the organized source sentences from SO task results, we derive

Dataset	# Images	Trans. Direction	Document Domain	Word Text	Word Box	Sent. Prefix	Sent. Order	Sent. Translation	Doc. Translation
DITrans	1,796	En→Zh	Report, News, *etc.*	✓	✓	✓	✓	✓	✓
M3T	1,016	En→Zh/De, *etc.*	Report, Legal, *etc.*	✓	-	-	-	-	✓
DIT700K *ours*	619K	En→Zh/De	General Web Doc.	✓	✓	✓	✓	✓	✓
	99K	Zh→En							

Table 1: Comparisons with prior datasets. DIT700K offers multi-level fine-grained labels and a lot more images.

the coherent document translation, where (source sentence, translation) pairs are organized in logical order and their layout positions on image are preserved according to prefix word bounding boxes.

ZoomDIT is trained with all above multi-level tasks to optimize all its modules jointly:

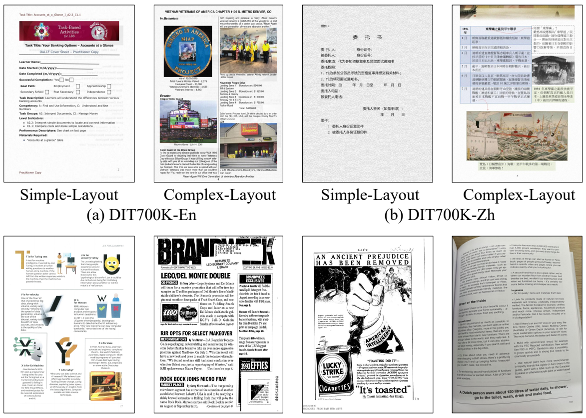
$$\mathcal{L} = \mathcal{L}_{pref} + \mathcal{L}_{comp} + \mathcal{L}_{orga} + \mathcal{L}_{trans} + \mathcal{L}_{corr} \quad (9)$$

3 Large-Scale Multi-Level Dataset

Along with ZoomDIT, we propose an automatic data pipeline and a new dataset to facilitate DIT.

Data Pipeline: It automatically extracts and annotates *Word* documents from the web. 1) It crawls *Word* file URLs and downloads .docx and XML source files. 2) A coloring scheme (Li et al., 2020) assigns a unique color to each word in XML. 3) The colored XML is rendered to PDF. 4) Document text is extracted from XML, acquiring the (word, color) pairs list that preserves logical order from XML. PDF is parsed, acquiring (word box, color) pairs list. 5) The two lists are merged by color, acquiring (word, word box) pairs list. 6) Finally, each PDF page is converted to .jpg image format. A SOTA model (Nguyen et al., 2021) with high F1 (90%) labels sentence prefixes and Google API provides sentence translations. Refer to App. A for details.

DIT700K Dataset: With this pipeline, we contribute a new DIT700K dataset. It contains 718K images (619K in English, 99K in Chinese) with multi-level fine-grained labels including word text and box; sentence prefix, order, and translation; and document translation in three directions. As shown in Tab. 1, compared with prior dataset M3T (Hsu et al., 2024), DIT700K provides more fine-grained labels that support multi-level tasks of DIT. Document images in DIT700K are also more large-scale and diverse-disciplinary than prior M3T and DITrans (Zhang et al., 2023). These properties constitute a more comprehensive benchmark for DIT.



Simple-Layout Complex-Layout (a) DIT700K-En Simple-Layout Complex-Layout (b) DIT700K-Zh

(c) DITrans-Report (d) DITrans-News (e) DITrans-Ad (f) DITrans-Book

Figure 4: Image examples of the used datasets.

4 Experiments

4.1 Experiment Settings

Datasets: We do experiments on DIT700K and DITrans datasets. 1) For DIT700K, testsets are divided according to layout complexity. Specifically, following Wang et al., 2021, document words are serialized via "top-left to bottom-right" rule to discard layouts and are calculated a BLEU score with ground-truth layout-preserving document text. Lower BLEU means a more complex layout that is not well-captured by the rule. With this metric (termed Layout Score), documents with the lowest/highest scores are selected as complex/simple-layout testset, each having 1024 examples. 2) For DITrans, it provides complex-layout documents carefully selected from four specific domains, each domain is split with the ratio `Train: Test` ≈ `4:1`. Fig. 4 shows examples from DIT700K and DITrans. Tab. 2 gives their detailed statistics.

Setups: Four setups with increasing difficulty are conducted for comprehensive evaluations. 1) Setup-Simple.GT: Evaluation of simple-layout documents with ground truth as input; 2) Setup-Simple.OCR: Evaluation of simple-layout documents with OCR results as input; likewise, we have more difficult 3) Setup-Complex.GT and 4) Setup-Complex.OCR. SAR (Li et al., 2019) is employed as OCR engine. Following Zhang et al. 2023, page-level BLEU and

Dataset	Domain	Acquisition	Direction	# Image	# Word/ Image	Trainset			Testset (Simple Layout)			Testset (Complex Layout)		
						# Image	# Sent.	Lay. Score	# Image	# Sent.	Lay. Score	# Image	# Sent.	Lay. Score
DIT700K-En	General	Digit-Born	En→Zh/De	619K	237	617K	20M	81.22	1024	25,477	87.99	1024	63,629	74.49
DIT700K-Zh	General	Digit-Born	Zh→En	99K	431	98K	2.7M	88.65	256	5,966	92.71	256	11,715	74.12
DITrans-Report	Report	Scan	En→Zh	902	245	722	17,030	72.97	-	-	-	180	4,878	69.35
DITrans-News	News	Scan	En→Zh	396	219	316	4,589	76.16	-	-	-	80	1,841	74.92
DITrans-Ad	Ad.	Scan	En→Zh	377	123	302	4,416	61.18	-	-	-	75	1,702	55.31
DITrans-Book	Book	Camera	En→Zh	121	247	91	1,635	44.96	-	-	-	30	678	40.40

Table 2: Statistics of the used datasets.

Method	Modality	Setup-Simple.GT		Setup-Simple.OCR		Setup-Complex.GT		Setup-Complex.OCR		Average		Params
		BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	
DIT700K-En (En→Zh)												
DIMTDA[1]	V	34.65	45.93	34.65	45.93	25.49	35.11	25.49	35.11	30.07	40.52	216M
TextMT[BERT[2]]	T	38.94	48.43	33.49	46.33	30.93	42.56	26.19	37.90	32.39	43.81	134M
LayoutLM[3]-Dec	T+L	40.27	49.84	35.58	48.42	32.66	43.46	28.03	38.84	34.14	45.14	136M
BROS[4]-Dec	T+L	41.36	51.68	36.57	47.48	33.31	44.53	28.37	39.74	34.90	45.86	134M
*LayoutXLM[5]-Dec	T+L+V	41.97	51.46	38.12	48.05	32.07	43.06	28.51	39.56	35.17	45.54	387M
LayoutLMv3[6]-Dec	T+L+V	41.66	51.85	37.54	48.00	32.58	43.37	28.90	39.87	35.17	45.77	149M
LiLT-Roberta[7]-Dec	T+L	40.28	50.82	35.80	48.64	34.57	45.02	30.60	41.50	35.31	46.50	152M
LayoutDIT[8]	T+L	42.47	52.67	38.35	48.96	34.40	44.90	30.59	41.51	36.45	47.01	141M
ZoomDIT[LayoutLMv3[6]] ***ours***	T+L+V	**44.45**	**54.52**	**40.24**	**50.85**	**37.07**	**47.42**	**33.13**	**43.86**	**38.72**	**49.16**	159M
DIT700K-Zh (Zh→En)												
*TextMT[XLM-Roberta[9]]	T	31.63	59.20	29.66	57.63	19.30	44.35	18.32	42.85	24.73	51.01	301M
*TextMT[InfoXLM[10]]	T	34.52	61.05	29.64	56.71	19.50	45.42	18.32	43.20	25.49	51.60	293M
*LiLT-XLM[7]-Dec	T+L	37.05	61.74	35.43	60.28	29.04	51.58	27.18	50.36	32.18	55.99	304M
*LayoutXLM[5]-Dec	T+L+V	42.83	67.23	38.52	63.48	31.53	55.17	28.70	52.77	35.39	59.66	387M
ZoomDIT**[LayoutXLM[5]] ***ours	T+L+V	**44.45**	**67.25**	**41.14**	**65.12**	**39.86**	**62.59**	**37.34**	**60.61**	**40.70**	**63.89**	415M
DIT700K-En (En→De)												
DIMTDA[1]	V	37.21	59.20	37.21	59.20	28.60	52.51	28.60	52.51	32.91	55.86	221M
TextMT[BERT[2]]	T	41.95	65.01	37.12	61.13	31.73	60.23	26.87	56.30	34.42	60.67	142M
LayoutLMv3[6]-Dec	T+L+V	44.27	66.75	40.42	63.70	34.10	60.79	30.63	58.10	37.36	62.33	157M
LiLT-Roberta[7]-Dec	T+L	43.69	65.65	38.95	61.81	35.52	61.16	31.34	58.21	37.38	61.71	159M
LayoutDIT[8]	T+L	44.88	68.11	40.43	64.58	35.82	63.32	31.47	60.12	38.15	64.03	149M
ZoomDIT[LayoutLMv3[6]] ***ours***	T+L+V	**47.05**	**69.60**	**42.82**	**66.33**	**39.67**	**66.38**	**35.28**	**63.33**	**41.21**	**66.41**	167M

Table 3: Results of En→Zh/De task on DIT700K-En dataset and Zh→En task on DIT700K-Zh dataset. T, L, V denote text, layout, vision modality of model input. *The multilingual model. []: Pre-trained weights for initialization. [1](Liang et al., 2024); [2](Devlin et al., 2019); [3](Xu et al., 2020); [4](Hong et al., 2022); [5](Xu et al., 2021); [6](Huang et al., 2022); [7](Wang et al., 2022); [8](Zhang et al., 2023); [9](Conneau and Lample, 2019); [10](Chi et al., 2021).

chrF++ are employed as evaluation metrics.

Baselines: Baselines include 1) The vision-based SOTA model **DIMTDA**; 2) Text-based models, including: **TextMT** based on text-only encoder-decoder to use only text modality; DocEnc-Dec model series based on document encoder-decoder to incorporate text and visual layout multi-modalities, *e.g.*, **LayoutLM-Dec**, **LiLT-Dec**, and the SOTA **LayoutDIT**, ***etc***. We ensure all baselines' and our model's parameter numbers are comparable when implementation. All models are first pre-trained on the large-scale DIT700K and then continually trained for DITrans experiments. Refer to App. B for more baseline and implementation details.

4.2 Comparison with Prior State-of-the-Arts

We evaluate the performance of ZoomDIT on the public DITrans and our proposed DIT700K.

DIT700K: As shown in Tab. 3, generally, all methods perform best under Setup-Simple.GT and worst under Setup-Complex.OCR, revealing the significant impact of layout complexity and OCR noise on DIT. On En-Zh direction, DIMTDA and TextMT perform the worst since the single modality (vision or text) is insufficient for DIT. DIMTDA shows consistent results across GT/OCR setups since it is vision-based and OCR-free. Compared with them, methods (LayoutLM-Dec → LayoutDIT) incorporating text with visual layout show better results, *e.g.*, LayoutLMv3-Dec improves 5.10/2.78 avg. BLEU on DIMTDA/TextMT. By modeling multi-level tasks, ZoomDIT significantly improves on

DITrans (En→Zh)	DITrans-Report				DITrans-News				DITrans-Ad				DITrans-Book				Average	
Setup (Complex.GT/OCR)	GT		OCR		GT		OCR		GT		OCR		GT		OCR			
Method	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
TextMT[BERT]	23.16	37.55	20.79	35.12	20.39	32.20	15.12	27.66	15.61	26.74	11.66	21.76	10.10	21.77	8.08	19.72	15.61	27.81
LayoutLMv3-Dec	26.74	39.67	23.51	36.93	22.52	35.04	17.62	29.26	21.32	32.52	17.58	28.50	12.91	25.11	10.34	20.89	19.07	30.99
LiLT-Roberta-Dec	28.25	39.98	24.16	36.40	23.29	34.44	18.73	29.78	22.55	32.56	16.68	26.66	14.50	**26.85**	**12.37**	**24.49**	20.07	31.40
LayoutDIT	28.04	40.65	24.32	37.26	23.31	36.18	19.84	33.20	24.95	36.66	21.14	32.86	12.93	25.26	11.69	23.94	20.78	33.25
ZoomDIT[LayoutLMv3] *ours*	**30.00**	**42.37**	**25.68**	**38.45**	**25.14**	**37.98**	**20.59**	**33.71**	**26.75**	**38.21**	**23.48**	**34.93**	**14.52**	26.07	11.18	23.54	**22.17**	**34.41**

Table 4: Results of the En→Zh task on the four specific domains of DITrans dataset.

top of LayoutLMv3-Dec and achieves the best results under all setups. Its SOTA performance is also observed in Zh-En and En-De directions.

DITrans: Due to the more complex layouts as depicted in Tab. 2 and fewer training examples, DITrans results (Tab. 4) are relatively lower than DIT700K. Likewise, LayoutLmv3-Dec outperforms TextMT due to multi-modality utilization. LiLT-Roberta-Dec shows better results, especially on DITrans-Book. We attribute this to its dual-stream backbone for text-layout decoupling, which improves learning efficiency under low-resource scenarios (*e.g.*, DITrans-Book). Clearly, ZoomDIT still achieves the best results in most domains and on average. Its SOTA results on two datasets reveal the effectiveness of ZoomDIT's fine-to-coarse framework in unifying multi-level tasks into DIT.

4.3 Intermediate Results Evaluation

To investigate whether ZoomDIT accomplishes multi-level tasks well, except for the final translation results, all intermediate results are also thoroughly evaluated. Metrics for sent. prefix identification task are precision, recall, and F1. As for sent. completion and organization tasks, referring to prior format-preserving OCR task (Blecher et al., 2024; Sun et al., 2024), metrics are BLEU and chrF, which compute the similarity between model predicted document text and the ground-truth text.

As shown in Tab. 5, our model achieves high F1 values on sent. prefix identification task. Based on the accurately identified prefixes, subsequent completion and organization tasks are effectively fulfilled with BLEU scores approaching or surpassing 80 on most datasets and setups except for DITrans-Book, since DITrans-Book has very complex-layout document images (Layout Score ≈ 40 as depicted in Tab. 2) and extremely scarce training examples. These promising intermediate results demonstrate that our model successfully fulfills multi-level tasks and reorganizes chaotic words into a coherent document, thus improving translation results.

Setup	Prefix Identification			Sent. Completion		Sent. Organization	
	Prec.	Rec.	F1	BLEU	chrF	BLEU	chrF
DIT700K-En (En→Zh)							
Simple.GT	92.94	92.84	92.23	95.09	97.58	96.49	97.96
Simple.OCR	93.08	92.70	92.23	94.88	97.47	96.28	97.85
Complex.GT	92.89	90.44	90.83	89.26	93.56	91.47	94.21
Complex.OCR	92.77	90.00	90.54	89.07	93.60	91.29	94.25
DIT700K-Zh (Zh→En)							
Simple.GT	93.84	88.35	89.70	92.99	91.99	95.32	95.00
Simple.OCR	94.02	88.17	89.75	92.66	91.66	95.13	94.83
Complex.GT	91.45	89.42	89.59	76.63	74.81	82.78	82.18
Complex.OCR	91.47	88.78	89.26	74.28	72.63	80.53	80.05
DIT700K-En (En→De)							
Simple.GT	92.67	92.62	92.00	94.69	97.38	96.11	97.75
Simple.OCR	92.69	92.53	91.96	94.43	97.15	95.86	97.51
Complex.GT	92.43	90.07	90.35	89.52	93.72	91.73	94.38
Complex.OCR	92.54	89.64	90.18	89.28	93.79	91.51	94.44
DITrans-Report (En→Zh)							
Complex.GT	95.69	94.51	94.70	83.50	91.53	87.10	92.40
Complex.OCR	95.32	93.42	93.96	81.09	90.10	84.83	91.01
DITrans-News (En→Zh)							
Complex.GT	93.89	93.69	93.40	90.94	94.41	91.67	94.57
Complex.OCR	93.47	92.59	92.64	90.34	93.88	91.19	94.12
DITrans-Ad (En→Zh)							
Complex.GT	88.89	90.60	89.25	80.26	88.68	82.87	89.44
Complex.OCR	88.69	89.02	88.26	80.68	88.94	82.61	89.47
DITrans-Book (En→Zh)							
Complex.GT	89.76	86.77	87.88	66.71	78.09	70.01	78.83
Complex.OCR	88.66	84.55	86.10	63.40	76.44	65.75	77.20

Table 5: Detailed evaluation of intermediate results.

4.4 Discussions and Ablations

We deeply study each task module's effectiveness of our model on the DIT700K-En dataset in Tab. 6, where (e) is full model as the performance anchor.

1) First, to ablate **prefix identification task module**, we use only the first word of the serialized OCR words, instead of model-predicted whole prefix words, as the beginning for subsequent sent. completion (model (a)). This severely damages sentence completion (a vs. e) since it is more difficult to complete the whole document in one pass. Translation is also affected negatively. 2) Second, to ablate **completion task module**, we replace it with hard-code rule (model (b)). The rule simply takes

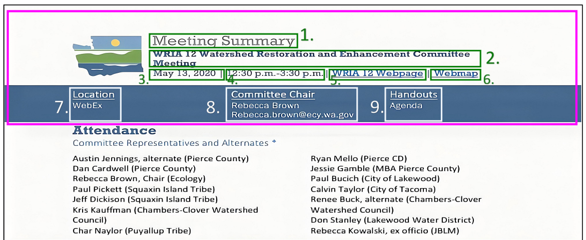
Tag	Different Task Modules				Setup-Complex.GT				Setup-Complex.OCR			
	Sent. Pref.	Sent. Comp.	Sent. Orga.	Sent. Trans.	Pref.	Comp.	Orga.	Trans.	Pref.	Comp.	Orga.	Trans.
(a)	First Word	Model Pred.	Model Pred.	Dual-Channel	-	81.30	81.30	27.12	-	80.95	80.95	24.09
(b)	Model Pred.	Rule-Based	Model Pred.	Dual-Channel	90.80	74.26	74.47	23.80	90.53	73.82	74.02	21.61
(c)	Model Pred.	Model Pred.	Rule-Based	Dual-Channel	90.83	89.26	89.26	36.24	90.54	89.07	89.07	32.31
(d)	Model Pred.	Model Pred.	Model Pred.	Single-Channel	90.61	88.97	91.20	36.01	90.40	88.27	90.46	31.60
(e)	Model Pred.	Model Pred.	Model Pred.	Dual-Channel	**90.83**	**89.26**	**91.47**	**37.07**	**90.54**	**89.07**	**91.29**	**33.13**
(f)	Ground-Truth	Model Pred.	Model Pred.	Dual-Channel	100.00	90.73	93.12	38.03	100.00	90.47	92.87	33.57
(g)		Ground-Truth	Model Pred.		100.00	95.75	97.37	38.94	100.00	95.67	96.93	34.64
(h)			Ground-Truth		100.00	95.75	100.00	40.52	100.00	95.67	100.00	36.07

Pref., *Comp.*, *Orga.*, and *Trans.* denote *prefix identification*, *completion*, *organization*, and *translation*. Metrics: Pref. - F1, Comp./Orga./Trans. - BLEU.

Table 6: Effects of different task modules in our model on DIT700K dataset.

the words between two adjacent prefix words for first sentence completion. This causes heavy degradation in completion and translation results (b vs. e). 3) Third, to ablate **organization task module**, we replace it with "top-left to bottom-right" rule to reorder prefix words for sent. organization (model (c)). Since all model input words (including prefixes) have already been sorted with this rule, it brings no improvements to organization task and causes worse translation results (c vs. e). 4) Finally, for **translation task module**, we disentangle correction channel's effect in model (d), which has almost no impact on intermediate results but causes 1.06/1.53 BLEU decline under GT/OCR setup. This verifies the auxiliary effectiveness of correction channel.

In addition, to explore performance upper bound, we gradually replace model predictions with **ground-truth labels** in (f) (g) (h). Intermediate results significantly improve or achieve 100.00 scores, continuously benefiting translation. This reveals an ideal coherent document facilitates translation and ZoomDIT provides promising results.



LayoutLMv3-Dec	会议 摘要 12 流域 恢复 和 增强 委员会 会议 2020 年 5 月 13 日 | 中午 12 : 30 - 下午 3 : 30 | 12 网页 | 网页 | 地点 委员会 主席 讲义 网页 Rebecca Brown Rebecca.brown@ecy.wa.gov 议程 ✗
ZoomDIT (Ours)	(1) 会议 记录 ✓ (2) WRIA 12 届 流域 恢复 和 增强 委员会 会议 ✓ (3) 2020 年 5 月 13 日 | ✓ (4) 中午 12 : 30 - 3 : 30 | ✓ (5) WRIA 12 网页 | ✓ (6) 网页 ✗ (7) 地点 网址 ✓ (8) 委员会 主席 Rebecca Brown Rebecca.brown@ecy.wa.gov ✓ (9) 讲义 议程 ✓
Ground-Truth S: Source T: Target	(1) S: Meeting Summary T: 会议 纪要 (2) S: WRIA 12 Watershed Restoration and Enhancement Committee Meeting T: 第 12 届 WRIA 流域 恢复 和 加强 委员会 会议 (3) S: May 13, 2020 | T: 2020 年 5 月 13 日 | (4) S: 12:30 p.m.-3:30 p.m. | T: 中午 12 : 30 - 下午 3 : 30 | (5) S: WRIA 12 Webpage | T: WRIA 12 网页 | (6) S: Webmap T: 网络 地图 (7) S: Location WebEx T: 位置 网讯 (8) S: Committee Chair Rebecca Brown Rebecca.brown@ecy.wa.gov T: 委员会 主席 Rebecca Brown Rebecca.brown@ecy.wa.gov (9) S: Handouts Agenda T: 讲义 议程

Figure 5: DIT case. Top: Document image. Red box indicates the text to translate. The recognized sentences and their logical order predicted by our model are visualized with green/white-color boxes and numbers. Bottom: Translation results from models and ground truth.

4.5 Visualization Cases

A visualized case is given in Fig. 5. Translation from LayoutLMv3-Dec is incoherent and semantically confused since it excessively depends on the chaotic OCR words. For example, the source word "Location" should be grouped with "WebEx" as a translation unit but is linked with "Committee Chair", causing false source text and translation. In contrast, ZoomDIT successfully predicted all sentences and their logical order, therefore producing a coherent source document and correct, well-formalized translation. Refer to App. C for more visualization cases.

5 Related Work

Deep neural models have proven to be very successful and have motivated machine translation from plain text to multi-modalities (Liang et al., 2024; Ma et al., 2023a,b,c; Yu et al., 2024; Zhang et al., 2023; Zhao et al., 2023). As a multi-modal machine translation task, DIT involves the cooperation of document text and visual layout. To this task, many efforts have been devoted to the simple-layout sentence/paragraph image (*e.g.*, movie subtitle) translation. They mostly model this task as image-to-text transformation based on vision-encoder text-decoder paradigm, with specially designed modules or tasks to bridge image-text modality gap, such as modal contrastive learning (Ma et al., 2023a), auxiliary text translation task (Zhu et al., 2023), multi-modal unified codebook (Lan et al., 2023), *etc*. Some other works (Lan et al., 2024; Mansimov et al., 2020; Tian et al., 2023) convert the source-text image to target-translation image to realize in-image text translation for higher efficiency. These methods have achieved impressive results on sentence/paragraph images. However, it is hard for them to generalize well to whole-page document images since they presuppose that sentences/paragraphs could be ideally cropped from the image, which is not always true in practice.

As for document image translation, early work (Afli and Way, 2016) directly translates OCR words with a text encoder-decoder. Considering the multi-modality nature of document images, recent studies incorporate extra visual layout information into DIT, with external layout parser (Hinami et al., 2021) or intrinsic layout-oriented encoders (Liang et al., 2024; Zhang et al., 2023). These methods are not restricted to sentence/paragraph images but can also tackle complex-layout document images. However, they still conduct translation based on chaotic OCR words, ignoring the document coherence which is crucial for DIT. As a remedy, this work models the multi-level tasks to recover a coherent document for DIT, thus improving translation quality and achieving new SOTA performance.

6 Conclusion

This paper proposes the Zoom-Out DIT framework. It combines multi-granularity, multi-level tasks in an end-to-end framework, thereby recovering a coherent document and achieving joint optimization. The information-rich intermediate results can also facilitate relevant document tasks. Besides, we construct a comprehensive benchmark with large scale and multi-level labels, which will prompt DIT community. Extensive experiments have demonstrated our model significantly outperforms prior methods, pushing DIT to a higher performance level.

Limitations

Although ZoomDIT achieves the best results in most domains, it slightly lags behind the LiLT-Roberta-Dec model in the DITrans-Book domain. We suppose this may be due to the distribution shift from DIT700K digit-born regular image to DITrans-Book camera deformed image, which causes performance degradation to our model. Referring to literature (Wang et al., 2022), in future work, we will consider incorporating the LiLT-Roberta-Dec model's text-layout dual-stream backbone into our framework to improve its domain transferring efficiency toward low-resource DIT scenarios.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China under Grant No. 62336008, and in part by the National Natural Science Foundation of China under Grant No.62106265.

References

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

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural optical understanding for academic documents. In *Proc. of ICLR*, pages 1–17.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv*, abs/2303.12712.

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 and Guillaume Lample. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. of ICLR*, pages 1–21.

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

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 *Proc. of AAAI*, pages 10767–10775.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*, pages 1–15.

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.

Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. 2019. Show, attend and read: A simple and strong baseline for irregular text recognition. In *Proc. of AAAI*, pages 8610–8617.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020. Docbank: A benchmark dataset for document layout analysis. In *Proc. of COLING*, pages 949–960.

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.

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.

Elman Mansimov, Mitchell Stern, Mia Xu Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. Towards end-to-end in-image neural machine translation. In *Proc. of NLPBT*, pages 70–74.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proc. of EACL*, pages 80–90.

Peyman Passban, Puneeth Saladi, and Qun Liu. 2021. Revisiting robust neural machine translation: A transformer case study. In *Proc. of EMNLP Findings*, pages 3831–3840.

Yu Sun, Dongzhan Zhou, Chen Lin, Conghui He, Wanli Ouyang, and Han sen Zhong. 2024. Locr: Location-guided transformer for optical character recognition. *ArXiv*, abs/2403.02127.

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.

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.

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.

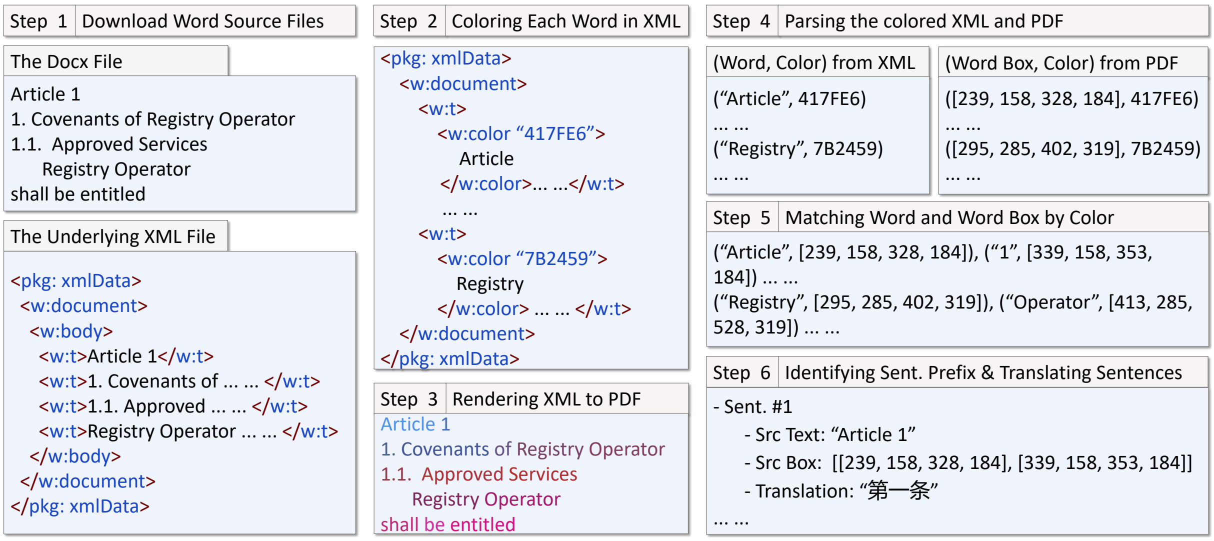


Figure 6: The automatic data pipeline for processing a document file. 1) Step 1: *Word* source files (*Word* .docx and the corresponding XML file) are crawled from websites. 2) Step 2: Each word in XML file is assigned a unique color code as the identifier. 3) Step 3: The colored XML is rendered to PDF. 4) Step 4: Extracting (word, color) pairs from XML, (word box, color) pairs from PDF. 5) Step 5: Associating each word with its bounding box using color as the key. 6) Step 6: Labeling sentence prefixes and sentence translations with model and Google API.

A Automatic Data Pipeline Details

The pipeline is shown in Fig. 6. The first step is to crawl websites to extract URLs that point to *Word* files and download source files (.docx file and the underlying XML file) from these URLs. Then, similar to Li et al. 2020, we use a coloring scheme to assign a unique color code to each word in the XML file, the colored document is then rendered to a PDF file with LibreOffice[2] library. Next, we extract the full document text from the XML file using python-docx[3], acquiring a sequence of (word, color) pairs. The text is in logical order due to the internal well-organized XML structure. At the same time, we parse the PDF file with PyMuPDF library[4] to extract word bounding boxes and word color, acquiring a sequence of (word box, color) pairs. Next, the two sequences are merged with color code as the key, resulting in the sequence of (word, word box) pairs. In the final step, each PDF page is converted to .jpg image format. An advanced prefix detection tool (F1 $\geq$ 90%) supplied by NLP toolkit (Nguyen et al., 2021) is used to annotate the sentence prefixes, and Google Translation API is used to produce sentence translations.

[2]https://www.libreoffice.org/
[3]https://github.com/python-openxml/python-docx
[4]https://github.com/pymupdf/PyMuPDF

B Experiment Setting Details

B.1 Baselines

Vision-Based Method:

- **DIMTDA** (Liang et al., 2024): It is the SOTA vision-based OCR-free model for academic document translation. It employs two separate pre-trained ViT encoders to extract visual and layout features from the image and a text decoder to generate translation.

Text-Based Methods:

- **TextMT**: A standard transformer encoder-decoder model (Vaswani et al., 2017) for translation based on solely text modality.

- **DocEnc-Dec**: Models of this series employ advanced pre-trained document encoders such as LayoutLMv3 as the encoder to incorporate the multi-modality feature of a document, and employ a text decoder to generate translation. In our experiments, several representative document encoders are experimented with, including 1) the canonical LayoutLM (Xu et al., 2020), 2) BROS (Hong et al., 2022) that considers relative spatial positions, 3) the dual-stream document encoder – LiLT (Wang et al., 2022), and 4) models that further incorporate visual features beyond layout and text – LayoutXLM (Xu et al., 2021) and LayoutLMv3 (Huang et al., 2022). Baselines of this class are denoted as DocEnc-Dec (*e.g.*, LayoutLM-Dec).

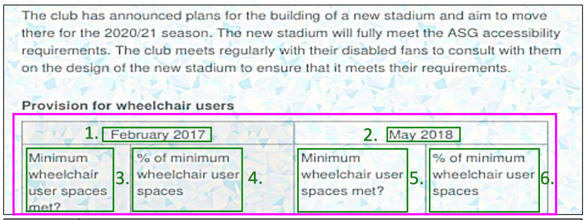
- **LayoutDIT** (Zhang et al., 2023): This model resembles LayoutLM-Dec with a multi-modal encoder for document feature extraction but decomposes the one-step decoding in LayoutLM-Dec into three-step decoding to alleviate the long context and text order problems in document image translation.

B.2 Implementation Details

Model Configurations: As described in Sec. 2, ZoomDIT's modules are all based on transformer encoder/decoder layers. Specifically, its document transformer encoder employs 6 encoder layers. Its sentence prefix identification/completion/organization/translation modules employ 1/3/1/3 encoder/decoder layers, respectively. Following previous literature (Devlin et al., 2019), each encoder/decode layer has 768-dimensional hidden sizes, 12 attention heads, and 3,072 feed-forward hidden units. Baseline models' hyper-parameters are consistent with our model. *e.g.*, the DocEnc-Dec models employ 6/6 layers for their encoder/decoder to have comparable parameter numbers as our model.

Training and Inference Configurations: Models are first pre-trained on the large-scale DIT700K dataset and then continually trained for DITrans experiments. During training, Adam optimizer (Kingma and Ba, 2015) is applied with $(\beta_1 = 0.9, \beta_2 = 0.98)$. Both dropout rate and label smoothing are set to 0.1. For training runs on DIT700K, the learning rate is $1e^{-4}$ with a warm-up on 5% training steps and then a linear schedule strategy. Models are trained for 80K steps with a batch size of 8. Before training, all models are initialized with pre-trained weights from their corresponding pre-trained models to improve performance, *e.g.*, pre-trained BERT (Devlin et al., 2019) for TextMT, pre-trained LayoutLMv3 for LayoutLMv3-Dec. In particular, for En-Zh/De tasks, our model is initialized with LayoutLMv3 which has been pre-trained on English documents; for Zh-En task, our model is initialized with LayoutXLM which has been pre-trained on multi-lingual documents. For training runs on DITrans, the learning rate is reduced to $2e^{-5}$. Models are trained for 20 epochs with a batch size of 6. During inference, beam search is applied for translation with a beam size of 4.

C More Visualization Cases



LayoutLMv3-Dec	2017 年 2 月 2018 年 5 月 最低 轮椅 使用者 空间 的 最低 轮椅 使用者 空间 百分比 满足 最低 轮椅 使用者 空间 的 最低 轮椅 使用者 空间 的 百分比 ✗
ZoomDIT (Ours)	(1) 2017 年 2 月 ✓ (2) 2018 年 5 月 ✓ (3) 最低 轮椅 使用者 空间 满足 吗？ ✓ (4) 最低 轮椅 使用者 空间 的 百分比 ✓ (5) 最低 轮椅 使用者 空间 满足 吗？ ✓ (6) 最低 轮椅 使用者 空间 的 百分比 ✓
Ground-Truth S: Source T: Target	(1) S: February 2017 T: 2017 年 2 月 (2) S: May 2018 T: 2018 年 5 月 (3) S: Minimum wheelchair user spaces met? T: 满足 最小 轮椅 使用者 空间 要求 吗？ (4) S: % of minimum wheelchair user spaces T: 最小 轮椅 使用者 空间 的 百分比 (5) S: Minimum wheelchair user spaces met? T: 满足 最小 轮椅 使用者 空间 要求 吗？ (6) S: % of minimum wheelchair user spaces T: 最小 轮椅 使用者 空间 的 百分比

Figure 7: DIT case study of table.

LayoutLMv3-Dec	圣经 文本 1 . 比尔 士王 的 宴会 。 2 . 比尔 士王 的 绝望 。 3 . 比尔 士王 的 俘虏 。 4 . 比尔 士王 的 死亡 。 关键词 巴尔 士 - 巴尔都 是 保护者 。 ✗
ZoomDIT (Ours)	(1) 圣经 文本 丹尼尔 第 5 章 ✓ (2) 关键 诗篇 丹尼尔 5 : 25 ✓ (3) 关键词 贝尔法 萨尔 - 所以 是 保护者 。 ✗ (4) 1 . 贝尔法 萨尔 王 的 宴会 。 ✓ (5) 2 . 贝尔法 萨尔 王 之谜 。 ✓ (6) 3 . 贝尔法 萨尔 王 的 俘虏 。 ✓ (7) 4 . 贝尔法 萨尔 王 的 灭亡 。 ✓
Ground-Truth S: Source T: Target	(1) S: Bible Text Daniel chapter 5 T: 圣经 但 以理书 第 5 章 (2) S: Key Verse Daniel 5:25 T: 关键 经文 但 以理书 5 : 25 (3) S: Key Word Belshazzar- bal is the protector. T: 关键词 伯沙 撒 - 保护者 。 (4) S: 1. The banquet of King Belshazzar. T: 1 . 伯沙撒王 的 宴会 。 (5) S: 2. The enigma of King Belshazzar. T: 2 . 伯沙撒王 之谜 。 (6) S: 3. The captive of King Belshazzar. T: 3 . 伯沙撒王 的 俘虏 。 (7) S: 4. The demise of King Belshazzar. T: 4 . 伯沙撒王 的 灭亡 。

Figure 8: DIT case study of item list.

Our motivation behind ZoomDIT is to jointly model multi-level tasks including sentence recognition and organization and unify them into DIT. In Fig. 7, ZoomDIT successfully predicts sentences in table cells and their logical order, thereby giving correct translations. However, LayoutLMv3-Dec treats all table cells as one sentence, which is counter-intuitive and the translation is also incorrect. A similar comparison can be also observed in Fig. 8, where LayoutLMv3-Dec ignores the separation between the two item lists and mingles them as one paragraph, while our model exactly translates items in the left list and then those of the right list.

D Comparision with Large VLMs

Recently, large vision language models (VLMs) have shown remarkable success on various multi-modal tasks (Bubeck et al., 2023). In view of this, we evaluate their DIT capabilities for comparison with our model. Specifically, we randomly sampled 64 document images from the complex-layout testsets of DIT700K-En/Zh as test examples and conducted evaluations in En→Zh and Zh→En directions. Two advanced VLMs - OpenAI's ChatGPT4-o[5] and Google's Gemini-Pro[6] - are evaluated. VLMs are instructed with the document image and a user prompt *Above is an English/Chinese document image. Translate its text content from English/Chinese to Chinese/English.* As for our model, it is further enhanced by expanding its document encoder to 12 layers, and expanding its sentence prefix identification/completion/organization/translation module to 1/6/1/12 layers (350M parameter numbers in total). The enhanced model is denoted as ZoomDIT*.

Evaluation results are presented in Tab. 7, from which we observe: 1) Both ChatGPT4-o and Gemini-Pro show promising DIT results although they have not been specially trained on our datasets. Another advantage is their robustness against OCR noise due to their reliance on only image inputs, which leads to consistent results across the GT and OCR setups. However, both VLMs might suffer from the under-translation issue and tend to have lower chrF scores. 2) Our best model ZoomDIT* still shows superior performances in both directions, especially in Zh→En direction. Despite the negative effects of OCR noise, our model still outperforms the two VLMs under Setup.OCR significantly, *e.g.*, 2.17/4.40 BLEU improvements compared with Gemini-Pro/ChatGPT4-o in En→Zh. The margin is further enlarged under Setup.GT, which means our model has more advantages if using better OCR engines such as commercial OCR APIs. Considering that DIT is always confronted with diverse document domains and translation directions, developing and training a task-specific model like our ZoomDIT is still a more reliable solution. In future work, we will pay more attention to enhancing ZoomDIT's noise resistance as well as exploring the collaboration of large VLMs and small task-specific models for better DIT systems.

Model	DIT700K-En (En→Zh)				DIT700K-Zh (Zh→En)				Avg.	
	Setup.GT		Setup.OCR		Setup.GT		Setup.OCR			
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
ChatGPT4-o	43.48	43.17	43.48	43.17	46.77	36.71	46.77	36.71	45.13	39.94
Gemini-Pro	45.71	44.23	45.71	44.23	43.99	35.42	43.99	35.42	44.85	39.83
ZoomDIT* *ours*	**57.05**	**64.66**	**47.88**	**56.83**	**57.16**	**76.09**	**55.67**	**75.02**	**54.44**	**68.15**

Evaluation time: August 2024.

Table 7: Comparison with SOTA large VLMs.

[5] https://chatgpt.com/

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