ICDAR 2025 Competition on End-to-End Document Image Machine Translation Towards Complex Layouts

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

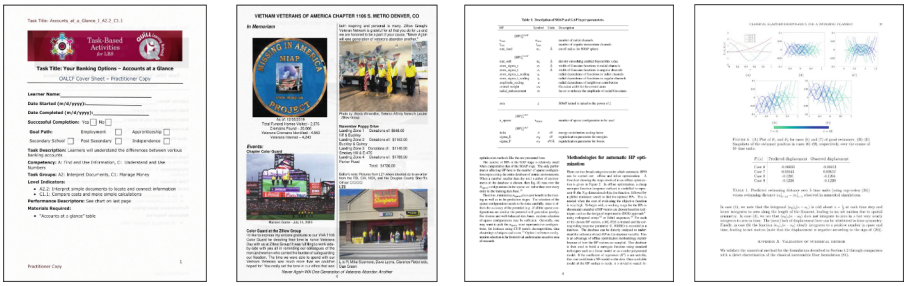
[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{yaping.zhang,lu.xiang,yang.zhao,cqzong}@nlpr.ia.ac,
{liangyupu2021,zhangzhiyang2020,chenzhiyuan2023}@ia.ac.
[2] University of Chinese Academy of Sciences, Beijing, China
yzhou@nlpr.ia.ac.cn
[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Company Ltd., Beijing, China

Abstract. Document Image Machine Translation (DIMT) seeks to translate text embedded in document images from one language to another by jointly modeling both textual content and page layout—bridging optical character recognition (OCR) and natural language processing (NLP). The DIMT 2025 Challenge advances research on end-to-end document image translation, a rapidly evolving area within multimodal document understanding. The competition features two tracks—OCR-free and OCR-based—each with two subtasks for small ($\leq$ 1B parameters) and large ($>$ 1B parameters) models. Participants submit a single unified DIMT system, with the option to incorporate provided OCR transcripts. Running from December 10, 2024 to April 20, 2025, the competition attracted 69 teams and 27 valid submissions in total. Track 1 had 34 teams and 13 valid submissions, while Track 2 had 35 teams and 14 valid submissions. In this report, we present the challenge motivation, dataset construction, task definitions, evaluation protocol, and a summary of results. Our analysis shows that large-model approaches establish a promising new paradigm for translating complex-layout document images and highlight substantial opportunities for future research.

1 Introduction

Recent advancements in large-scale language models (LLMs) have led to significant breakthroughs in optical character recognition (OCR) [1,2] and machine translation of plain text [3]. However, the translation of real-world document images, characterized by intricate layout structures—including mixed column formats, tables, and footnotes—continues to pose substantial challenges for current LLM architectures [4–7]. This inherent difficulty significantly curtails the effective application of LLMs for comprehensive knowledge extraction and limits their broader utility in diverse document-centric applications.

X.-C. Yin et al. (Eds.): ICDAR 2025, LNCS 16027, pp. 505–522, 2026.
https://doi.org/10.1007/978-3-032-04630-7_29



(a) (b) (c) (d)

Fig. 1. Document image examples of the DIMT 2025 challenge. (a) (b): Web document examples in Track 1; (c) (d): arXiv document examples in Track 2.

The task of Document Image Machine Translation (DIMT) addresses this by aiming to translate textual content embedded within document images from a source language to a target language. DIMT is fundamentally a multi-modal problem, necessitating the sophisticated integration of visual layout understanding with natural language processing (NLP) capabilities, thereby bridging the traditional divide between OCR and downstream textual analysis. While recent years have witnessed notable progress in DIMT technologies [8–12], existing solutions often fall short of the robustness and accuracy required for practical, real-world deployment. The primary impediments can be attributed to:

- Multi-modality and cross-linguality: Due to the inherent multi-modality nature, real-world document images often involve a intricate combination of complex layout, dense text, and visually-rich elements, resulting in difficulties in their comprehensive understanding and the cross-lingual translation.
- Image and text noise: Many factors such as image defect or OCR error may cause noise to the image or text as model input, leading to additional challenges to a DIMT system.
- Lack of samples and a unified benchmark: Due to the high annotation cost and different annotation protocals, existing datasets often suffer insufficient samples, inconsistent labels and evaluation metrics, resulting in model performance that are not directly comparable.

To address these gaps, we introduce the end-to-end Document Image Machine Translation Challenge at ICDAR 2025 (DIMT 2025), the first comprehensive benchmark for document image translation. This challenge offers standardized, large-scale datasets with detailed annotations and a unified evaluation protocol, featuring document images with complex layouts (see Fig. 1). Participants are invited to develop models across two complementary tracks:

- **OCR-Based:** This track focuses on translating text from OCR-extracted segments. The core task involves accurately reordering these potentially chaotic word-level outputs and translating them into a coherent, structurally sound

target-language text that faithfully preserves the original document's semantic content and layout integrity.
– **OCR-Free:** This track focuses on end-to-end DIMT without any OCR assistance, where participants translate source document images directly into markdown-formatted target translations while handling complex layouts and contextual information.

Each track consists of 2 subtasks: small-model (with $\leq$ 1B parameters) and large-model (with $>$ 1B parameters), encouraging innovation in both resource-constrained and large-scale settings. Submissions are evaluated using an end-to-end document-level translation metric, referred to as document-level BLEU [13]. By deeply fusing OCR and NLP into a unified multimodal framework and rigorously evaluating on complex layouts, the DIMT 2025 Challenge aims to eliminate brittle preprocessing pipelines and catalyze breakthroughs in multilingual document understanding and translation.

2 Related Works

LLMs have significantly impacted Document AI, with developments classifiable by their reliance on OCR: OCR-based methods using text from OCR engines, and OCR-free methods directly processing document images.

OCR-Based Document AI. OCR-based methods first extract text using an OCR engine, then feed this, often with layout data, to an LLM. The LayoutLM series [14,15] was foundational, pre-training models by integrating text, 2D position, and visual embeddings from OCR outputs. LiLT [16] extended this with a language-independent layout transformer.

OCR-Free Document AI. OCR-free methods bypass OCR error propagation by directly interpreting document images. Donut [11] pioneered this with an end-to-end transformer converting images to structured text. LayoutLMv3 [17] advanced multimodal pre-training, capable of OCR-free understanding by treating text detection as a pre-training task using image patch embeddings. More recently, Nougat [18] focuses on converting scientific document images into structured Markdown, performing OCR and parsing simultaneously. General-purpose Vision-Language Models (VLMs) like GPT-4o [19,20], InternVL [1], LLaVA [12], and MiniCPM [2] also show promise in zero-shot OCR-free document understanding by reasoning about text in images. These methods leverage visual and spatial cues but can be computationally intensive and sensitive to visual variations.

While these approaches have significantly advanced the field, existing visually-rich document benchmarks predominantly focus on evaluating coarse-grained understanding. Common tasks include document visual question answering (e.g., DocVQA [21], RDVQA [22], CS-DVQA [23]) or structured/key information extraction (e.g., VisualMRC [24], DUDE [25]). Even recent, large-scale benchmarks like MMVQA [26], which provides extensive QA pairs over scientific documents including tables and figures, and DocILE [27], offering numerous real

and synthetic documents for key information and line item recognition in financial/legal domains, primarily assess the ability to locate and interpret specific pieces of information rather than demanding a comprehensive understanding of all textual content.

Additionally, we provide a multilingual dataset with a broader range of categories. An overview of these datasets is presented in Table 1. Notably, DIT700K [28] is a large-scale document image translation dataset containing 718,000 images (619,000 in English, 99,000 in Chinese), with multi-level fine-grained labels including word text/box, sentence prefix/order/translation, and document translation, across three translation directions (En→Zh/De, Zh→En). Constructed through an automated pipeline (web crawling, XML/PDF processing, color-coded word-box matching), DIT700K spans diverse disciplines and layout complexities, which are categorized into simple and complex subsets based on the Layout Score.

Furthermore, DoTA [7] introduces a new challenge: Document Image Machine Translation to Markdown (DIMT2Markdown). This challenge aims to translate source document images with long-context and complex layout structures into markdown-formatted target translations. The accompanying dataset consists of 126,345 paired images, English markdown files, and their corresponding translations in Chinese, French, and German. The dataset is split into 124,338 images for training, 1,004 for validation, and 1,003 for testing.

Table 1. Existing Visually-rich Document Images Datasets.

Dataset	# Image number	Language	Granularity	Document type	Year
DocVQA [21]	12,767	English	Word	Industrial Reports	2021
VisualMRC [24]	10,197	English	Word	Website	2021
RDVQA [22]	8,514	English	Word	Data Analysis Report	2022
CS-DVQA [23]	600	English	Word	Industry Documents	2022
DUDE [25]	28,709	English	Word	Cross-Domain	2023
DITrans [8]	1,675	English	Word	Repor, News, Ad.	2023
MMVQA [26]	30,239	English	Word	Academic Paper	2024
DIT700K [28]	718,000	English, Chinese	Word	Web Documents	2025
DoTA [7]	126,345	4 languages	Page	Academic Paper	2024

3 Benchmark Description

3.1 Organization

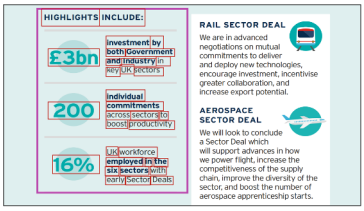
ICDAR 2025 competition on DIMT is organized by a team from the Institute of Automation, Chinese Academy of Sciences (CASIA). The DIMT2025 competition is presented on the official website[1], which provides datasets download links, baseline code download links, and submission pages for uploading results via CodaLab. We received substantial support from the CodaLab web team.

[1] https://cip-documentai.github.io/.

Based on the input type, we release two tracks: 1) OCR-based DIMT, where model inputs include the image and its OCR results (word and word bounding box); 2) OCR-free DIMT, where model input is the image. Both tracks are given English document images and are required to translate them to Chinese.

3.2 Track 1. OCR-Based DIMT

This track focuses on translating text from document images where OCR results (containing words and word bounding boxes) are provided, as shown in Fig. 2. The task aims to generate an intact, ordered target translation from the raw, often chaotic OCR-extracted words in an end-to-end manner. Participants are required to handle misordered, fragmented, or disjointed OCR outputs and accurately reorder them while translating the text into the target language. The goal is to produce a coherent and structured target-language text that reflects the original document's content and layout, ensuring the translation maintains both meaning and order.



(a)

Reordering Results	HIGHLIGHTS INCLUDE: £3bn Investment by both Government and Industry in key UK sectors 200 Individual commitments across sectors to boost productivity 16% UK workforce employed in the six sectors with early Sector Deals
Translation Results	亮点 包括： 30 亿 英镑 政府 和 工业 对 英国 主要 行业 的 投资 200 个 跨 部门 的 个人 承诺 以 提高 生产力 6% 受雇于 早期 有 行业 协议 的 六个 部门 的 英国 劳动力

(b)

Fig. 2. An example of input and output for Track 1 is shown in (a). The purple rectangular box highlights the text areas that need to be reordered and translated, with the words and their corresponding bounding boxes (indicated by the red boxes) serving as the model's input. For clarity, only the area within the rectangular box is illustrated, but in practice, the entire page is considered. (b) The results of reordering and translating the text in the rectangular box area.

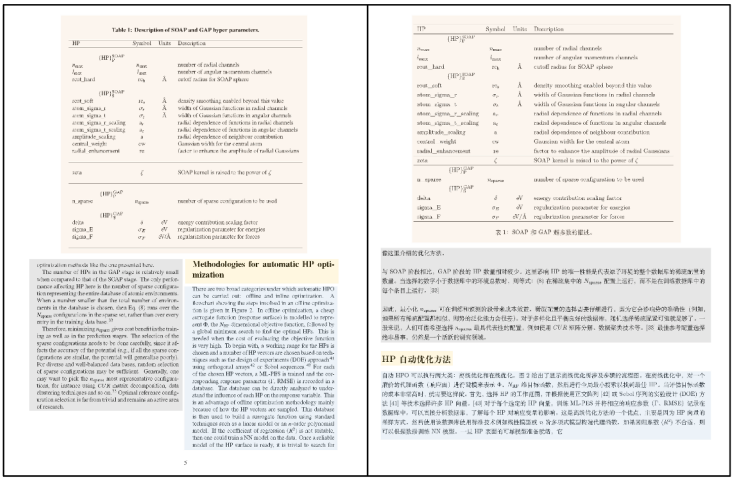
Track 1.1 OCR-Based DIMT-LLM: aims to evaluate the performance of LLM-based methods on the DIMT task. In this sub-track, participants must use large language models (LLMs) with over 1 billion parameters to achieve the OCR-based DIMT task. Open-source LLMs can be utilized, and participants are allowed to fine-tune these models to improve performance. The number of parameters in the model must be specified in the submitted reports used during inference.

Track 1.2 OCR-Based DIMT-Small: aims to evaluate the performance of small model-based methods on the DIMT task. In this sub-track, participants are only allowed to use small models, where the number of parameters is fewer than 1 billion, to achieve the OCR-based DIMT task. Participants must focus on optimizing these smaller models for accurate translation and reordering. The

number of parameters in the model must be specified in the submitted reports used during inference.

3.3 Track 2. OCR-Free DIMT

This track focuses on an OCR-free approach to DIMT, where participants are tasked with translating a source document image directly into a markdown-formatted target translation without any OCR assistance in an end-to-end way, as shown in Fig. 3. In the OCR-free scene, participants must handle the complex layout and contextual information directly from the image, delivering a coherent and accurate translation in the target language. The final submission must follow markdown format reflecting the content and layout of the original document.



(a) (b)

Fig. 3. An input and output example of Track 2. (a) is the original document image. (b) is the translated text in markdown format after rendering. The blocks with the same color represent the corresponding original text and translated text.

Track 2.1 OCR-Free DIMT-LLM: aims to evaluate the performance of LLM-based methods on the OCR-free DIMT task. In this sub-track, participants must use large language models (LLMs) with over 1 billion parameters to achieve the OCR-free DIMT task. Open-source LLMs can be utilized and fine-tuned to handle complex layouts and long context translation. Participants must specify the number of parameters in the model used during inference in their submitted reports.

Track 2.2 OCR-Free DIMT-Small: aims to evaluate the performance of small model-based methods on the OCR-free DIMT task. In this sub-track,

participants are only permitted to use small models with fewer than 1 billion parameters to achieve the OCR-free DIMT task. Participants must work within these constraints, focusing on optimizing smaller models to handle complex layouts and long context. The number of parameters in the model must be specified in the submitted reports used during inference.

4 Dataset

In this competition, we have assembled a dataset of over 42,400 document pages drawn from two domains: web documents and scientific articles, with a held-out test set of 1,000 pages. Detailed statistics of the competition dataset are summarized in Table 2. The following provides a brief description of the datasets used in this competition. For more comprehensive details, please refer to our previous work [7,28].

- **DIMT-WebDoc-300K** dataset for Track 1: This dataset includes a training set of 300,000 images and a validation set of 1,000 images, all derived from publicly available web documents. Each image is accompanied by the corresponding OCR results, which include word-level text and bounding boxes, as well as the word-level reading order index, sentence-level translations, and document-level translations.
- **DIMT-arXiv-124K** dataset for Track 2: This dataset includes a training set of 124,000 images and a validation set of 1,000 images. These images are extracted from PDF and LaTeX files crawled from Arxiv. Each image is paired with corresponding source language text and target language text in markdown format.

Table 2. Statistical data information of the DIMT 2025 challenge.

Track	Dataset	# of Examples		
		Train	Valid	Test
Track 1	DIMT-WebDoc-300K	300K	1K	1K
Track 2	DIMT-arXiv-124K	124K	1K	1K

5 Evaluation Protocol

We adopt the evaluation metric used in DoTA [7] and DIT700K [28], employing document-level BLEU [13] for both tracks. Participants are asked to treat the final translation output of an entire image as a single text string. Chinese segmentation is performed using Jieba[2], followed by BLEU-4 score computation between the predicted translation and the reference text using the NLTK toolkit[3].

[2] https://github.com/fxsjy/jieba.
[3] https://github.com/nltk/nltk.

6 Submissions and Results

The competition attracted 69 participants and 27 valid submissions from both academia and industry. Participant statistics are provided in Table 3. During the competition, we established an evaluation platform on CodaLab, which automatically computed BLEU scores for participants' submissions in real-time and updated the leaderboard accordingly. As each new submission would overwrite the previous one, we reminded participants to upload their best final results before the deadline. The winners for each track were determined based on the highest BLEU scores. The complete leaderboard for all tracks is available on the official website (See footnote 1).

Table 3. The statistics of the competition submission

Track	Subtrack	Participate	Submission
Track 1 OCR-based DIMT	LLM	18	6
	Small	16	7
	Total	34	13
Track 2 OCR-free DIMT	LLM	20	9
	Small	15	5
	Total	35	14

After the competition, we required participants to specify the number of model parameters and whether any additional training data was used in their technical reports. This was in accordance with the competition rules, which permitted the use of pre-trained models but restricted training to the provided dataset. The number of model parameters determined each team's subtrack classification. Ultimately, all submitted technical reports met these requirements.

6.1 Track 1.1 OCR-Based DIMT-LLM

In this section, we present the results and methods employed by the submissions for Track 1.1 OCR-based DIMT-LLM. Models with more than 1 billion parameters are used to achieve OCR-based end-to-end document image machine translation. In this setup, the input consists of the raw OCR results from a document, and the output is the document-level translation, including the reading order. The reported results are selected from the top five entries in the leaderboard, based on the intersection with the submitted method reports, as shown in Table 4.

Baseline Method. We provide a baseline in which Qwen2-VL-7B [29] is employed to directly generate the complete translated text based on the input document image and its corresponding OCR results.

Table 4. The results of Track 1.1 OCR-based DIMT-LLM

Rank	Team Name	Team Members	Institute	BLEU
Baseline	–	–	–	26.34
1	Hw-tsc	Zhanglin Wu, Tengfei Song, Ning Xie, Weidong Zhang, Pengfei Li, Shuang Wu, Chong Li, Junhao Zhu, and Hao Yang	Huawei Translation Service Center	70.48
2	Lucky Star	Zhentao Guo, Zining Wang, Tongkun Guan, Hao Sun, Chen Duan, and Kai Zhou	Beijing Institute of Technology, Meituan, Shanghai Jiao Tong University, Chinese Academy of Sciences	34.35
3	Abantika Bose	Abantika Bose	Friedrich-Alexander-Universität Erlangen-Nürnberg	10.64

1st Ranking Method. "Hw-tsc" team primarily utilizes the InternVL2.5-8B-MPO framework [30], which integrates multi-task learning with a perceptual chain-of-thought training approach. This enables the model to process both visual layouts and linguistic content concurrently. During inference, the system further applies minimum Bayesian decoding and post-processing techniques to optimize the quality of the generated output.
2nd Ranking Method. "Lucky Star" team's model architecture employs an Encoder-Decoder framework, comprising a LayoutLM-based [14] Encoder and a Qwen2.5-7B-based [31] Decoder. The LayoutLM Encoder processes both textual content and the coordinates of text boxes from document images. By leveraging the Transformer mechanism, it integrates textual content with visual layout features to produce rich contextual and layout-aware representations. These representations are subsequently passed through a latent space to the Qwen2.5-7b Decoder. The Qwen2.5-7b Decoder receives the context-rich representations generated by the Encoder and, based on its advanced Transformer-based language modeling capabilities, is fine-tuned to generate target-language text sequences. This process ensures that the translated output is coherent, fluent, and natural.
3rd Ranking Method. "Abantika Bose" team splits the DIMT task into two sequential stages: Stage 1: Reordering: Llama-4-Maverick-17B-128E (402 B) [32] via few-shot in-context prompting, no fine-tuning. Stage 2: Translation. Reordered English text from either expert is translated into Chinese using Helsinki-NLP/opus-mt-en-zh (77.9 M) [33], followed by Jieba segmentation.

6.2 Track 1.2 OCR-Based DIMT-Small

In this section, we present the results and methods employed by the submissions for Track 1.2 OCR-based DIMT-Small. In this track, the input and output are consistent with Track 1.1, with the only difference being that the models used in this track must have fewer than 1 billion parameters. The aim of this track is to explore the potential of smaller models in comparison to large models. The reported results are selected from the top five entries in the leaderboard, based on the intersection with the submitted method reports, as shown in Table 5.
Baseline Method. We construct an encoder-decoder model based on LayoutLM [14] and a Transformer decoder, and train it on the provided training data for 10,000 steps.

Table 5. The results of Track 1.2 OCR-based DIMT-Small

Rank	Team Name	Team Members	Institute	BLEU
Baseline	–	–	–	38.81
1	Hw-tsc	Zhanglin Wu, Tengfei Song, Ning Xie, Weidong Zhang, Pengfei Li, Shuang Wu, Chong Li, Junhao Zhu, and Hao Yang	Huawei Translation Service Center	66.16
2	CMC ATI	Doan Kien Thai, An Nguyen Thai, and Vu Son Lam Nguyen	CMC ATI	39.73
3	sunmk	Mai Sun, Jiazi Wang, and Long Yuan	Beijing Jiaotong University	37.06
4	Lucky Star	Zhentao Guo, Zining Wang, Tongkun Guan, Hao Sun, Chen Duan, and Kai Zhou	Beijing Institute of Technology, Meituan, Shanghai Jiao Tong University, Chinese Academy of Sciences	33.61
5	Abantika Bose	Abantika Bose	Friedrich-Alexander-Universität Erlangen-Nürnberg	9.40

1st Ranking Method. The Hw-tsc team utilized the same framework as in the Track 2.1 OCR-based DIMT-LLM track, but with a smaller pretrained version, InternVL2.5-1B [30]. This approach enabled the team to secure first place in the competition.

2nd Ranking Method. "CMC ATI" team's pipeline includes 2 models for distinct tasks: Text Ordering and Machine Translation. The outputs from Text Ordering Model will be served as the input to the Machine Translation model. In detail, they used the LayoutLMv3 [34] model for the reordering model due to its pre-trained English ability and the Qwen2.5-0.5B-Instruct [31] translation model due to the fact that the predominant languages trained on this model are reported to be English and Chinese. In both models, they performed supervised fine-tuning on the competition dataset to enhance the specialized ability in each task.

3rd Ranking Method. "sunmk" team proposes a layout-aware encoder-decoder architecture for document-level sequence generation tasks, where the input consists of recognized words along with their corresponding bounding boxes. The encoder leverages LayoutLM [14] to incorporate spatial layout information, while the decoder is based on BERT-base [35] and generates textual sequences in an auto-regressive manner. The encoder and decoder are connected through a Transformer-based cross-attention fusion module.

6.3 Track 2.1 OCR-Free DIMT-LLM

In this section, we present the results and methods employed by the submissions for Track 2.1 OCR-free DIMT-LLM. In this setup, the model is fully end-to-end, with the input being a document image and the output in markdown format, including structural information. Models with more than 1 billion parameters are used to achieve the task of Track 2.1. The reported results are selected from the top five entries in the leaderboard, based on the intersection with the submitted method reports, as shown in Table 7.

Table 6. The results of Track 2.1 OCR-free DIMT-LLM

Rank	Team Name	Team Members	Institute	BLEU
Baseline	–	–	–	48.63
1	Hw-tsc	Zhanglin Wu, Tengfei Song, Ning Xie, Weidong Zhang, Pengfei Li, Shuang Wu, Chong Li, Junhao Zhu, and Hao Yang	Huawei Translation Service Center	60.78
2	360ailab	Junhui Yu	360 AI Research Institute	57.78
3	Lucky Star	Zhentao Guo, Zining Wang, Tongkun Guan, Hao Sun, Chen Duan, and Kai Zhou	Beijing Institute of Technology, Meituan, Shanghai Jiao Tong University, Chinese Academy of Sciences	44.41

Baseline Method. We provide participants with a baseline model, where LoRA fine-tuning [36] is applied to adapt Qwen2-VL-7B [29] on the provided training data for 1,000 steps.

1st Ranking Method. "Hw-tsc" team's OCR-free-LLM approach integrates OCR-based and OCR-free translation tasks within a single framework, eliminating the need for separate pipelines. This approach is built upon the InternVL2.5-8B-MPO framework [30].

2nd Ranking Method. "360ailab" team proposes a robust end-to-end document translation method based on the Qwen2.5-VL multimodal large language model [37]. Additionally, through adversarial training, they apply adversarial training algorithms related to the visual encoder of the large model, such as Projected Gradient Descent and Fast Gradient Method, to enhance the model's robustness and improve its generalization ability.

3rd Ranking Method. "Lucky Star" team's methods are mainly based on CLIP [38] and Qwen2.5-7B [31]. The CLIP encoder is responsible for processing the input document images, extracting both visual and textual content features. This encoding phase converts complex page layouts into a structured set of representations, bridging the gap between visual appearance and semantic content. Subsequently, these representations are fed into the Qwen2.5-7B decoder. This decoder utilizes its expansive language model to transform the encoded data into target language translations, ensuring context-aware and semantically accurate outputs. The integration between these components is facilitated through a tailored interface that harmonizes visual and textual data flow, optimizing the translation process.

6.4 Track 2.2 OCR-Free DIMT-Small

In this section, we present the methods employed by the submissions for Track 2.2 OCR-free DIMT-Small, as shown in Table 7, highlighting the diverse approaches used to address the challenges of OCR-free document image machine translation. The reported results are selected from the top six entries in the leaderboard, based on the intersection with the submitted method reports.

Baseline Method. We provide the OCR-free DIMT small baseline model, which uses Nougat's encoder [18] along with a pre-trained translation decoder.

Table 7. The results of Track 2.2 OCR-free DIMT-Small

Rank	Team Name	Team Members	Institute	BLEU
Baseline	–	–	–	27.03
1	Intime & HY	Gengluo Li, Huawen Shen, Chengquan Zhang, Xingyu Wan, Pengyuan Lyu, Liang Wu, Zhuohao Chen, Han Hu, and Yu Zhou	Chinese Academy of Sciences, Tencent, Nankai University	59.96
2	Hw-tsc	Zhanglin Wu, Tengfei Song, Ning Xie, Weidong Zhang, Pengfei Li, Shuang Wu, Chong Li, Junhao Zhu, and Hao Yang	Huawei Translation Service Center	59.56
3	Blue Jays	Cameron Carpenter, David Etter, Matt Lee, Paul McNamee, Kevin Duh, and Kenton Murray	Johns Hopkins University	58.41
4	Lucky Star	Zhentao Guo, Zining Wang, Tongkun Guan, Hao Sun, Chen Duan, and Kai Zhou	Beijing Institute of Technology, Meituan, Shanghai Jiao Tong University, Chinese Academy of Sciences	44.41

A Transformer-based model is pre-trained on the DoTA dataset for the text machine translation task. Subsequently, an encoder-decoder model is constructed using Nougat's encoder [18] and the pre-trained translation decoder. The model is further fine-tuned for 20,000 steps on the provided dataset.

1st Ranking Method. "Intime & HY" team combines document parsing-based self-filtering with reinforcement learning to enhance the performance of small model-based OCR-free document image machine translation. The team employs a self-developed model named HYOCR-1B as the backbone and further trains it on English-Chinese document data to enhance structured document parsing capabilities. They also use the DPO reinforcement learning strategy to optimize the model's outputs, reducing hallucinations and improving translation accuracy.

2nd Ranking Method. "Hw-tsc" team's methods are mainly based on InternVL2.5-1B-MPO [30]. This framework integrates multi-task learning with perceptual chain-of-thought training approach, enabling the model to simultaneously comprehend visual layouts and linguistic content. During the inference phase, the system further employs minimum Bayesian decoding and post-processing strategies to optimize output quality.

3rd Ranking Method. "Blue Jays" team trained a small end-to-end vision-language model, which follows a three-part design consisting of a Vision Transformer image encoder [39], an MLP projector, and an auto-regressive LLM decoder [40].

7 Discussion

In this section, we analyze participation trends, performance gaps, and the methods used in the submission reports, as shown in Table 8.

7.1 Participation Trends and Performance Gap

Participation Trends. The competition provided a unique opportunity to explore the interest in both OCR-based and OCR-free paradigms. Participa-

tion in the OCR-based and OCR-free tracks was nearly identical, with 34 teams in the OCR-based track and 35 in the OCR-free track, reflecting similar levels of interest in both approaches. However, a notable trend emerged: the LLM tracks (both OCR-based and OCR-free) consistently had more participants than their small-model counterparts. Among them, the OCR-free-LLM track attracted the highest number of submissions, highlighting the growing confidence in using large-scale models for OCR-free document image translation, a historically more challenging task.

Performance Gap. The performance gap between OCR-based and OCR-free models remains substantial. OCR-based models consistently delivered better results, underscoring the reliability of OCR methods in extracting text from document images. In contrast, OCR-free models are still catching up and face greater challenges when translating documents without OCR support. However, the progress made by OCR-free models, especially in large-scale tracks, suggests that these models will eventually close the performance gap with OCR-based systems as they continue to improve.

Table 8. The method reports submitted by the participating teams

OCR-based							
Model Type	Team Name	Method		SFT	DPO	Parameters	Extra Data
LLMs	hw-tsc	InternVL2.5-8B-MPO	70.48	✓	✓	8B	
	Lucky Star	LayoutLM+qwen2.5-7B	34.35	✓	ĀŮ	7.3B	
	Abantika Bose	LLaMA-402B	10.64	ĀŮ	ĀŮ	402B	
Small Models	hw-tsc	InternVL2.5-1B-MPO	66.16	✓	✓	1B	
	tadkt	LayoutLMv3+Qwen2.5-0.5B-Instruct	39.73	✓	ĀŮ	633M	
	sunmk	LayoutLM+BERT	37.06	✓	ĀŮ	211M	
	Lucky Star	LayoutLM+RobeERTa-largge	33.61	✓	ĀŮ	674M	✓
	Abantika Bose	LayoutT5-Reorder	9.4	ĀŮ	ĀŮ	264M	
OCR-free							
LLMs	hw-tsc	InternVL2.5-8B-MPO	60.78	✓	✓	8B	
	yujunhuinlp	Qwen2.5-VL-7B-Instruct+PGD	57.78	✓	ĀŮ	7B	
	Lucky Star	CLIP+Qwen2.5-7b	44.41	✓	ĀŮ	8B	
Small Models	Intime & HY	HYOCR-1B	59.96	✓	✓	1B	
	hw-tsc	InternVL2.5-1B-MPO	59.56	✓	✓	1B	✓
	ccarpe18	SigLip-2+Qwen2.5-0.5B-Instruct	58.41	✓	ĀŮ	927M	✓
	Lucky Star	Donut+Nougat	44.41	✓	ĀŮ	500M	

7.2 Analysis of Participation Methods

OCR-Free vs OCR-Based Approaches. OCR-based models consistently outperformed OCR-free models. OCR-based methods leverage well-established OCR technologies to extract text from document images, providing a solid foundation for subsequent translation. These models benefit from decades of

OCR algorithm development, which enables robust text extraction even in documents with complex layouts (e.g., multiple columns, tables, footnotes). This makes OCR-based solutions particularly reliable when translating documents that require precise text extraction.

Although OCR-free models did not outperform OCR-based models, they showed considerable progress, particularly in handling document images without OCR. Models like InternVL2 and Qwen2.5 performed well in OCR-free tracks, indicating that OCR-free models are advancing in their ability to handle multimodal document translation tasks. The performance gap between OCR-based and OCR-free models is narrowing, suggesting that as OCR-free models continue to evolve, they will become more competitive with traditional OCR-based systems.

Large vs Small Models. A key takeaway from the competition is the clear advantage of larger models in handling complex document image translation tasks. Models like InternVL2.5-8B-MPO consistently outperformed smaller models in both the OCR-based-LLM and OCR-free-LLM tracks. For example, InternVL2.5-8B-MPO achieved a score of 70.48 in the OCR-based track, significantly higher than the 66.16 achieved by the top small model. Similarly, in the OCR-free track, the top LLM submission reached 60.78, while the best small model reached 59.96. These results underscore the scalability and capacity of larger models to address complex, real-world document image translation tasks involving diverse text types and structural nuances.

Despite the dominance of LLMs in raw performance, smaller models still have value, especially in resource-constrained scenarios or when dealing with less complex document structures. Smaller models like InternVL2-1B and Qwen2.5-0.5B performed competitively, achieving solid results even with limited computational resources. The performance of smaller models suggests they can be highly optimized for specific tasks, achieving good results with a smaller computational footprint.

Interestingly, smaller models, including those with sizes like 1B and 211M, demonstrated that fine-tuning on domain-specific datasets, such as OCR data or translated documents, can significantly boost performance. While LLMs excel at handling the complexity of multimodal tasks, smaller models can still achieve competitive results when appropriately fine-tuned, making them valuable for applications with limited resources or real-time translation needs.

Fine-Tuning Strategies. The competition results underscore the critical role of fine-tuning strategies, particularly Supervised Fine-Tuning (SFT), in enhancing the performance of models for document image translation tasks. SFT was the predominant technique employed across most of the top-performing models, highlighting that fine-tuning on domain-specific data—such as OCR data or document images—is essential for refining model performance. This approach enables models to adapt to the intricacies of document structures and language pairs, leading to substantial improvements in translation accuracy. Notably, Direct Preference Optimization (DPO) [41] was utilized by both the 1st and 2nd ranked methods in both OCR-free and OCR-based tracks, indicating

that DPO holds promising potential in the context of Document Image Machine Translation (DIMT).

Pretraining with Specialized Models. Specialized models, such as LayoutLM, LayoutT5, and SigLip-2, were employed by several teams, demonstrating the continued value of layout-aware techniques in the pretraining process. These models excel at understanding and processing document structures, which is critical for accurate document translation. However, despite their ability to handle complex document layouts, these specialized models did not perform as competitively in the larger tracks as models like InternVL2 or Qwen2.5. This highlights a key challenge in document image translation: while layout-aware models are effective for documents with simpler or more structured layouts, they struggle to match the broader capabilities and performance of general-purpose LLMs that can handle diverse and complex document structures.

In contrast, models with broader pretraining, such as InternVL2 and Qwen2.5, demonstrated superior performance in both OCR-based and OCR-free tracks. These models are better equipped to handle the diversity of document layouts and content types, which are characteristic of complex document image translation tasks.

8 Conclusion

We hosted the DIMT competition, which focused on the complex task of translating text from document images, both with and without OCR preprocessing. The competition introduced new datasets and challenges, providing participants with opportunities to explore both OCR-based and OCR-free document image translation. Submissions highlighted significant interest from both academia and industry in the integration of multimodal technologies for document translation, but also revealed the substantial challenges in dealing with complex document structures, such as mixed layouts and non-textual elements, especially in OCR-free scenarios. The results indicate that LLMs show promising potential for overcoming these challenges, but achieving high performance across diverse and real-world document image scenarios remains difficult. Future iterations of this competition could expand on these challenges by introducing even more intricate datasets, encouraging innovation in model architectures and training strategies. This would not only attract further advancements in the intersection of OCR and NLP but also drive progress in the broader field of Document AI, with numerous applications in areas like automatic document processing and cross-lingual information extraction.

Acknowledgements. This competition is supported by the National Natural Science Foundation of China (No. 62476275, No. 62336008) and the Young Scientists Fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS2024316). The organizers thank Binyao Xu and Zheng Lian for their significant efforts and support. We are also deeply grateful to the ICDAR 2025 competition chairs (Zhouhui Lian, Michael Coustaty, and Ron Litman), and the CodaLab web team

for their crucial assistance and tremendous support throughout the organization of this competition.

References

1. Dong, X., et al.: Internlm-xcomposer2-4khd: a pioneering large vision-language model handling resolutions from 336 pixels to 4k HD. Adv. Neural. Inf. Process. Syst. **37**, 42566–42592 (2024)
2. Yao, Y., et al.: MiniCPM-V: a GPT-4V level MLLM on your phone. arXiv preprint arXiv:2408.01800 (2024)
3. Yang, W., Wu, J., Wang, C., Zong, C., Zhang, J.: Implicit cross-lingual rewarding for efficient multilingual preference alignment. arXiv preprint arXiv:2503.04647 (2025)
4. Qian, Z., et al.: AnyTrans: translate AnyText in the image with large scale models. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 2024, pp. 2432–2444. Association for Computational Linguistics (2021)
5. Lan, Z., et al.. Exploring better text image translation with multimodal codebook. In: ACL (1) (2023)
6. Zhang, Z., et al.: Understand layout and translate text: unified feature-conductive end-to-end document image translation. IEEE Trans. Pattern Anal. Mach. Intell. **47**(5), 3358–3376 (2025)
7. Liang, Y., et al.: Document image machine translation with dynamic multi-pre-trained models assembling. In: Proceedings of NAACL, pp. 7077–7088 (2024)
8. Zhang, Z., et al.: LayoutDIT: layout-aware end-to-end document image translation with multi-step conductive decoder. In: Proceedings of EMNLP Findings, pp. 10043–10053 (2023)
9. Liang, Z., et al.:. A survey of multimodel large language models. In: Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering, pp. 405–409 (2024)
10. Liu, Y., et al.: Ocrbench: on the hidden mystery of OCR in large multimodal models. SCIENCE CHINA Inf. Sci. **67**(12), 220102 (2024)
11. Kim, G., et al.: OCR-free document understanding transformer. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13688, pp. 498–517. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19815-1_29
12. Zhang, R., Zhou, Y., Chen, J., Gu, J., Chen, C., Sun, T.: Llava-read: enhancing reading ability of multimodal language models. arXiv preprint arXiv:2407.19185 (2024)
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318. Association for Computational Linguistics (2002)
14. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1192–1200 (2020)

15. Xu, Y., et al.: Layoutlmv2: multi-modal pre-training for visually-rich document understanding. In: Proceedings of ACL, pp. 2579–2591 (2021)
16. Wang, J., Jin, L., Ding, K.: LILT: a simple yet effective language-independent layout transformer for structured document understanding. In: Proceedings of ACL, pp. 7747–7757 (2022)
17. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: pre-training for document AI with unified text and image masking. In: Proceedings of ACM MM, pp. 4083–4091 (2022)
18. Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.: Nougat: neural optical understanding for academic documents. In: The Twelfth International Conference on Learning Representations (2024)
19. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
20. Hurst, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
21. Mathew, D.K.M., Jawahar, C.V.: DocVQA: a dataset for VQA on document images. In: Proceedings of CVPR, pp. 2200–2209 (2021)
22. Wang, R., et al.: A region-based document VQA. In: Proceedings of ACM MM, pp. 4909–4920 (2022)
23. Li, K., Tian, J., Xiao, L., Du,Q., Wang, Q., Jin, Y.: Calm: commen-sense knowledge augmentation for document image understanding. In: Proceedings of ACM MM, pp. 3282–3290 (2022)
24. Nishida, K., Tanaka, R., Yoshida, S.: VisualMRC: machine reading comprehension on document images. In: Proceedings of AAAI, pp. 13878–13888 (2021)
25. Borchmann, Ł., et al.: Document understanding dataset and evaluation (dude). In: Proceedings of ICCV, pp. 19528–19540 (2023)
26. Huang, J., Luo, S., Ding, Y., Ren, K., Han, S.C.: A dataset for multimodal information retrieval in pdf-based visual question answering. In: Proceedings of ICCV, pp. 19528–19540 (2023)
27. Uřičář, M., et al.: Docile benchmark for document information localization and extraction. In: Proceedings of ICDAR, pp. 147–166 (2023)
28. Zhang, Z., et al.: From chaotic OCR words to coherent document: a fine-to-coarse zoom-out network for complex-layout document image translation. In: Proceedings of COLING, pp. 10877–10890 (2025)
29. Wang, p., et al.: Qwen2-vl: enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
30. Chen, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
31. Yang, A., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024)
32. Touvron, H., et al.: Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
33. Tiedemann, J., et al.: Democratizing neural machine translation with OPUS-MT. Lang. Resour. Eval. **58**, 713–755 (2023)
34. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: pre-training for document AI with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083–4091 (2022)
35. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers), pp. 4171–4186 (2019)

36. Hu, E.J., et al.: Lora: low-rank adaptation of large language models. ICLR, **1**(2), 3 (2022)
37. Bai, S., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
38. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PmLR (2021)
39. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
40. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
41. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Advances in Neural Information Processing Systems, vol. 36, pp. 53728–53741 (2023)