

PATIMT-Bench: A Multi-Scenario Benchmark for Position-Aware Text Image Machine Translation in Large Vision-Language Models

Wanru Zhuang^{1*}, Wenbo Li^{1*}, Zhibin Lan¹, Xu Han², Peng Li², Jinsong Su^{1,3†}

¹School of Informatics, Xiamen University, China

²Tsinghua, Beijing, China

³Shanghai Artificial Intelligence Laboratory, China

{zhuangwanru, liwenbo}@stu.xmu.edu.cn jssu@xmu.edu.cn

Abstract

Text Image Machine Translation (TIMT) aims to translate texts embedded within an image into another language. Current TIMT studies primarily focus on providing translations for all the text within an image, while neglecting to provide bounding boxes and covering limited scenarios. In this work, we extend traditional TIMT into position-aware TIMT (PATIMT), aiming to support fine-grained and layout-preserving translation, which holds great practical value but remains largely unexplored. This task comprises two key sub-tasks: region-specific translation and full-image translation with grounding. To support existing models on PATIMT and conduct fair evaluation, we construct the PATIMT benchmark (PATIMT-Bench), which consists of 10 diverse real-world scenarios. Specifically, we introduce an Adaptive Image OCR Refinement Pipeline, which adaptively selects appropriate OCR tools based on scenario and refines the results of text-rich images. To ensure evaluation reliability, we further construct a test set, which contains 1,200 high-quality instances manually annotated and reviewed by human experts. After fine-tuning on our data, compact Large Vision-Language Models (LVLMs) achieve state-of-the-art performance on both sub-tasks. Experimental results also highlight the scalability and generalizability of our training data¹.

1 Introduction

Text Image Machine Translation (TIMT) is a challenging branch of Neural Machine Translation (NMT), offering broad application prospects in both academic research and commercial applications. Conventional TIMT methods (Jain et al., 2021; Zhu et al., 2023; Liang et al., 2024) typically focus on generating plain text or markdown-

* Equal contribution.

† Corresponding author.

¹Our benchmark and code are openly available at <https://github.com/XMUDeepLIT/PATIMT-Bench>

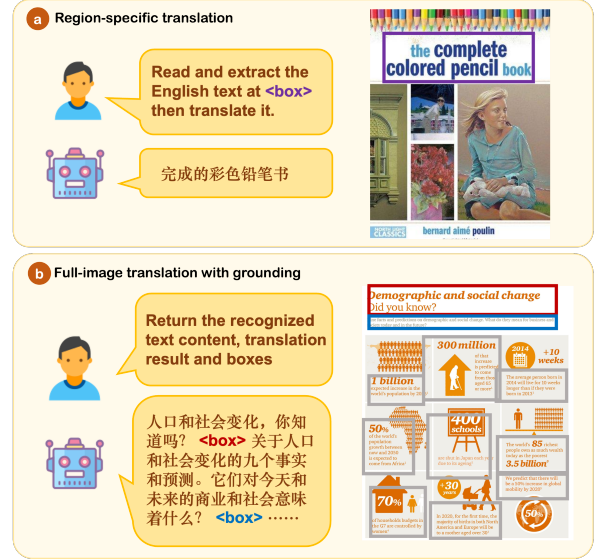


Figure 1: Two sub-tasks of PATIMT: region-specific translation and full-image translation with grounding.

formatted translations of all text within an image, failing to precisely preserve the original layout of source text in the image. This limitation gives rise to a critical position alignment problem in real-world applications, where users cannot reliably match translations to the corresponding source text. Additionally, all these methods overlook the localized translation requirements. This significantly limits their practical usability.

In this paper, we explore position-aware TIMT (PATIMT) which contains two core sub-tasks: region-specific translation and full-image translation with grounding as shown in Figure 1. Region-specific translation enables users to manually select one region of an image for translation, which allows fine-grained, user-controllable TIMT. Full-image translation with grounding ensures precise positional alignment between the translation and source text in the image, enabling seamless rendering of translated image version.

Recently, Large Vision-Language Models



Figure 2: Comparison between the Original LVLM and the LVLM fine-tuned on our data (Ours). In two types of fine-grained TIMT tasks, ours can correctly follow the translation instructions and conduct precise text referring and grounding within proper layout.

(LVLMs) (Gu et al., 2024; Wu et al., 2024; Chen et al., 2024; Bai et al., 2025) show remarkable performance across diverse multimodal benchmarks such as OCR (Liu et al., 2024b), image understanding (Mishra et al., 2019; Mathew et al., 2021, 2022; Masry et al., 2022; Lu et al., 2022) and visual grounding (Kazemzadeh et al., 2014; Mao et al., 2016; Li et al., 2022; Paiss et al., 2023). They appear to have great potential to perform region-specific translation and full-image translation with grounding. However, existing LVLMs usually fail to follow the above two types of translation instructions, as illustrated in Figure 2. This limitation primarily derives from data scarcity. Available TIMT datasets (Wang et al., 2021; Zhang et al., 2023b; Lan et al., 2023; Li et al., 2025) typically lack bounding box annotations or suffer from limited scenarios and scale, making them difficult to support accurate position-aware TIMT. Moreover, there is a lack of comprehensive benchmark. Existing TIMT benchmarks mainly focus on evaluating plain text or markdown translations and only specialize in a single scenario (Wang et al., 2021; Lan et al., 2023; Zhu et al., 2023; Liang et al., 2024). Although MIT-10M (Li et al., 2025) covers diverse image categories, it lacks bounding box annotations and is confined to simple scenarios, excluding document, infographic images and so on.

Nevertheless, constructing multi-scenario PATIMT datasets remains challenging for three main reasons. 1) General OCR tools typically provide line-by-line recognition results, leading to semanti-

cally incoherent annotations; 2) Document-specific OCR tools sometimes ignore text-containing areas and are not always optimal for other scenarios; 3) Manual annotation is labor-intensive and expensive. In this work, we address these issues by introducing an automated data processing pipeline to construct a **high-quality, multi-scenario** PATIMT dataset and a comprehensive benchmark: First, we introduce an **adaptive image OCR refinement pipeline** that combines a general EasyOCR² with a PDF-optimized MinerU (Wang et al., 2024) to adaptively process images from different scenarios and refine the results of text-rich samples. Second, we propose **PATIMT-Bench**, which is explicitly designed to evaluate region-specific translation and full-image translation with grounding for images from diverse domains. Specifically, we use our pipeline to construct training data, which provides fine-grained bounding boxes in proper layout for text within the images. As for the test set, we select 1200 images with high-quality manual annotations that are carefully reviewed by human experts.

Experimental results demonstrate that all compact LVLMs achieve state-of-the-art performance on PATIMT-Bench after fine-tuning on our training data, outperforming larger models such as Qwen2.5-VL-72B and closed-source models like GPT-4o. A series of systematic analyses are conducted which demonstrate the scalability and generalizability of our dataset.

2 Related Work

2.1 TIMT models

Early approaches predominantly rely on cascaded systems (Sable et al., 2023; Zhang et al., 2023b; Lan et al., 2023), where OCR and NMT models are separately optimized and pipelined. Such methods suffer from error propagation, and they only provide plain text translation. Recent advancements explore end-to-end frameworks (Jain et al., 2021; Ma et al., 2022; Zhu et al., 2023; Liang et al., 2024) to mitigate these issues. A representative approach developed by Liang et al. (2024) enables markdown format translations for document-style images. While this approach achieves layout-aware translation, it remains limitations in handling other scenarios such as infographic, chart, natural scene where markdown is inadequate to establish accurate localization correspondence.

²<https://github.com/JaidedAI/EasyOCR>

Dataset	Source	Bounding box	Scenario
OCRMT30K (Lan et al., 2023)	realistic	✓	street-view
DITrans (Zhang et al., 2023b)	realistic	✓	document
DoTA (Liang et al., 2024)	realistic	×	document
UMTIT (Niu et al., 2024)	synthetic	×	document
MIT-10M (Li et al., 2025)	realistic	×	multi-scene
PATIMT-Bench (Ours)	realistic	✓	multi-scene

Table 1: Comparison of PATIMT-Bench with other TIMT datasets.

2.2 TIMT Datasets

Dataset scarcity remains a critical challenge in TIMT research (Shen et al., 2024). Early studies (Mansimov et al., 2020; Jain et al., 2021; Ma et al., 2022; Niu et al., 2024) primarily rely on synthetic data, which are generated by rendering source language text onto background images. However, synthetic data are unable to capture the nuanced complexity of text in real-world translation applications (e.g., occlusions, irregular layouts), leading to an inevitable performance gap.

Recent efforts aim to construct real-world TIMT datasets. Lan et al. (2023) develops OCRMT30K, derived from street view images and their OCR annotations; Zhang et al. (2023b) constructs DITrans, which considers reading order in document images; Liang et al. (2024) introduces DoTA, a document image machine translation dataset in markdown format. Afterwards, Li et al. (2025) constructs MIT-10M, a large-scale, real-world dataset with diverse image categories. However, it lacks bounding box annotations and omits more complex scenarios such as infographic and document. In this work, we propose the Adaptive Image OCR Refinement Pipeline, an automated and cost-effective solution for processing text within images. Our pipeline provides bounding box labels in proper layouts for images from varying scenarios. Table 1 shows the comparison of our dataset with existing TIMT datasets, image examples and output format comparison are listed in Appendix A.1.

2.3 Large Vision-Language Models

Recent advances in LVLMs (Gu et al., 2024; Wu et al., 2024; Chen et al., 2024; Bai et al., 2025) demonstrate remarkable performance across diverse multimodal benchmarks, including visual question answering (Mathew et al., 2021; Singh et al., 2019; Mishra et al., 2019; Mathew et al., 2022; Lu et al., 2022), OCR (Liu et al., 2024b; Fu et al., 2025), and visual grounding (Kazemzadeh

et al., 2014; Mao et al., 2016). The prevailing architecture integrates a powerful visual encoder with a large language model (LLM) via cross-modal adapters. This unified framework exhibits two strengths: (1) superior translation quality with the powerful LLM, and (2) precise text grounding that enabling position-aware TIMT for diverse images. Despite these potentials, no existing work has systematically explored position-aware TIMT capability for LVLMs. In this work, we present PATIMT-Bench, which is designed to evaluate PATIMT through region-specific translation and full-image translation with grounding tasks.

3 Adaptive Image OCR Refinement Pipeline

To develop a high-quality PATIMT dataset for diverse real-world scenarios, we first extensively collect existing open-source image-text datasets and classify these images into corresponding scenarios. Secondly, we introduce an adaptive processing with refinement strategy to adaptively process images from different scenarios. Finally, we prompt GPT-4o (Hurst et al., 2024) to generate the instruction tuning data. Figure 3 illustrates the overall pipeline.

3.1 Data Collection and Preprocessing

We collect data from the following sources: MIT-10M (Li et al., 2025), CC12M (Changpinyo et al., 2021), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), Wukong (Gu et al., 2022), WTW (Rujiao et al., 2021), LSVT (Sun et al., 2019), CDLA³, pdfa-eng-wds⁴, an English hand-written OCR dataset from Nexdata⁵, and some online data.

After collecting data from various sources, we categorize the collected data using CLIP (Radford et al., 2021) following Zhang et al. (2023a), resulting in 10 different scenarios: *advertisement*, *poster*, *book cover*, *natural scene*, *street view*, *chart*, *table*, *hand-written*, *infographic*, and *document*. Appendix A.2 lists the detailed implementation. To ensure the classification accuracy, we randomly sample 200 images from the easy and hard categories separately for verification, which achieves an accuracy rate of 98.5%. Figure 4 shows the proportion of different scenarios.

³<https://github.com/buptlihang/CDLA.git>

⁴<https://huggingface.co/datasets/pixparse/pdfa-eng-wds>

⁵from <https://www.nexdata.ai>

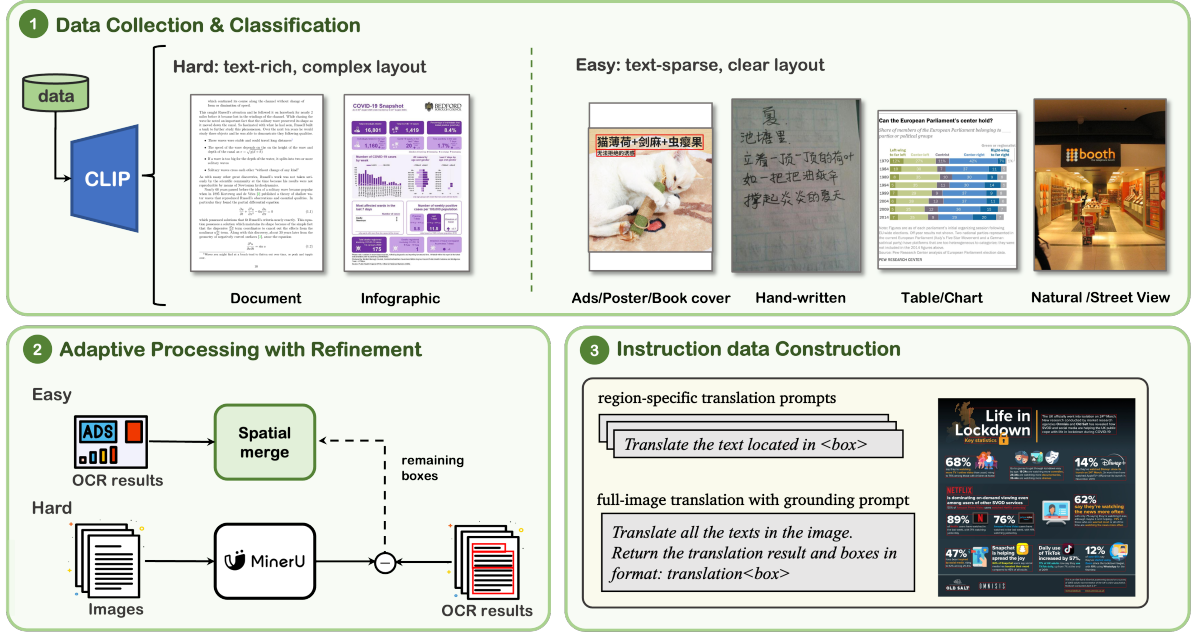


Figure 3: Our pipeline includes three steps: (1) collecting images from open-source datasets with initial OCR filtering, classifying images into easy/hard categories using CLIP; (2) adaptive processing with refinement strategy that generates accurate annotations for both easy and hard categories, and (3) constructing instruction dataset using GPT-4o.

We then implement EasyOCR to generate the OCR results for the collected images and conduct coarse-grained filtering. Specifically, images are excluded if: (1) their OCR results are empty; (2) their OCR results contain repetitive character sequences of length ≥ 3 ; or (3) the average character pixels are less than 3% of the total image pixels.

3.2 Adaptive Processing with Refinement

This process handles images from different scenarios and generates box annotations. To begin with, we categorize the aforementioned ten scenarios into two groups based on the level of difficulty:

- **Easy:** Characterized by images containing sparse text, clean layouts, and typically low resolutions with modest aspect ratios.
- **Hard:** Characterized by text-rich images with small font sizes, complex layouts, and high resolutions with potentially extreme aspect ratios. These scenarios present intricate spatial arrangements where text, graphics, and tables are sometimes interleaved.

We classify document and infographic as the hard scenarios, and the others as the easy scenarios. For images belonging to easy scenarios, we directly merge the OCR results based on their spatial relevance, the algorithm is shown in Table 2.

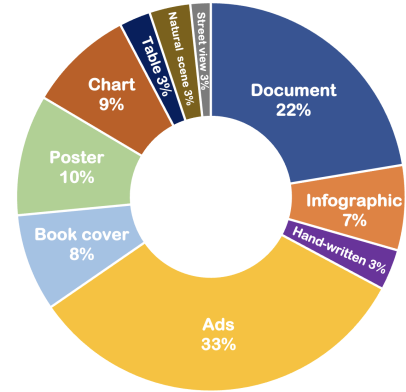


Figure 4: The proportion of different scenarios in our dataset.

As for images belonging to hard scenarios, we first employ MinerU (Wang et al., 2024) to process the original images. MinerU is a specialized cascaded system designed for document-type pdf, which organizes the recognized content into blocks attached by corresponding bounding box and block type labels such as text, image or table. Nevertheless, it sometimes overlooks some text regions or misclassifies text-containing blocks as images. To mitigate this issue, we refine MinerU’s output by leveraging the initial OCR results. Specifically, we extract a subset of initial OCR results that are omitted by MinerU through analysis of bounding box over-

Algorithm 1 Spatial merge

Input: OCR results $\mathcal{O}$, thresholds x_{ths} , y_{ths}
Output: Merged text boxes $\mathcal{R}$ 1: $\mathcal{B} \leftarrow \emptyset$
2: **for** $b \in \mathcal{O}$:
3: Store $(b_{\text{text}}, x_{\min}, x_{\max}, y_{\min}, y_{\max}, h, \bar{y}_c, 0)$
4: where $h = y_{\max} - y_{\min}$, $\bar{y}_c = \frac{y_{\min} + y_{\max}}{2}$
5: $g \leftarrow 1$
6: **while** ungrouped boxes exist:
7: **if** group g empty:
8: Assign first ungrouped box to g
9: **else**:
10: Compute group bounds using $\pm x_{\text{ths}}\bar{h}$, $\pm y_{\text{ths}}\bar{h}$
11: **for** ungrouped box u :
12: **if** u overlaps bounds:
13: Assign u to g ; **break**
14: **if** no assignment: $g \leftarrow g + 1$
15: $\mathcal{R} \leftarrow \emptyset$
16: **for** each group k :
17: text $\leftarrow ""$
18: box $\leftarrow [0, 0, 0, 0]$
19: **while** group boxes remain:
20: Find highest candidate row S
21: Select leftmost box b^* in S
22: Append b^*_{text} to text
23: Combine box b^*_{box} and box
24: Store merged text and group bbox
25: $\mathcal{R} \leftarrow \mathcal{R} \cup \{(\text{text} : \text{text.strip}(), \text{box} : \text{box})\}$
26: **return** $\mathcal{R}$

Table 2: Algorithm of spatial merge, which merges the OCR results based on their spatial relevance.

laps, and merge them based on spatial relevance. These recovered OCR results are then extended into MinerU’s output to supplement its original results. As shown in Figure 5, utilizing adaptive processing with refinement strategy can accurately handle images from different scenarios.

3.2.1 Instruction Tuning Dataset

Based on the processed results, we leverage GPT-4o (Hurst et al., 2024) to generate translations. Specifically, we prompt GPT-4o to generate 100 diverse questions for region-specific translation and full-image translation with grounding tasks, which are randomly sampled as questions for each instance. For each image, we construct a region-specific translation question-answer pair for each bounding box, and one full-image translation with grounding question-answer pair. The details of the templates are shown in Appendix A.3.

4 PATIMT Benchmark

In this section, we present a detailed description of our PATIMT-Bench. Firstly, we formally define the

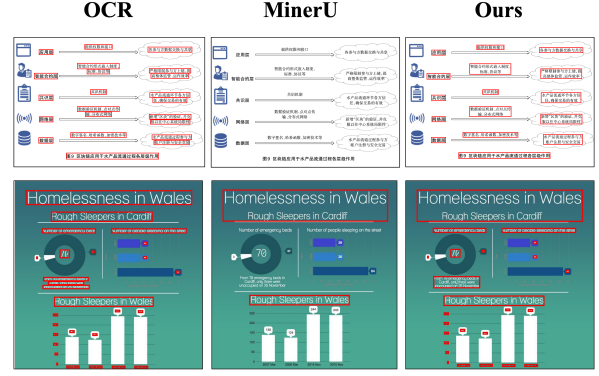


Figure 5: Comparison of utilizing different strategies. EasyOCR offer line-by-line results ignoring semantic coherence, while MinerU often fail to identify some text-containing areas. Our proposed adaptive processing with refinement strategy can accurately handle images from different scenarios.

PATIMT	Images	OCR boxes	Boxes	Src Words	Tgt Words
Train	48,884	1,307,516	417,066	24,827,252	30,437,907
Test	1,200	-	11,102	564,656	685,375

Table 3: Data statistics of PATIMT-Bench. OCR boxes, boxes refer to raw OCR box count and box count utilizing our pipeline, src words and tgt words refer to total number of words in source text and target text. Typically, test set is manually labeled.

two sub-tasks mentioned above. Secondly, we conduct a comprehensive analysis of our datasets. The specific details of these two aspects are presented as follows.

4.1 Task Definition

Our PATIMT-Bench focus on two sub-tasks:

- **Region-specific translation:** Given an input image with specified bounding box coordinates in the prompt, the model needs to generate its accurate translation.
- **Full-image translation with grounding:** Given an input image, the model needs to generate the text translation and the corresponding bounding box for each layout. This supports spatial correspondence between target text and source text and within the input image in practical applications.

4.2 Analysis of PATIMT-Bench

4.2.1 Training Dataset

We quantify key dataset metrics including image numbers, original OCR detection box numbers,

Model	Region-Specific Translation				Full-Image Translation with Grounding					
	EN-ZH		ZH-EN		EN-ZH			ZH-EN		
	BLEU	COMET	BLEU	COMET	BLEU	COMET	IoU	BLEU	COMET	IoU
<i>Proprietary LVLMs</i>										
Qwen2.5-VL-72B	<u>45.0</u>	76.6	37.3	75.2	13.2	48.7	0.185	<u>11.3</u>	53.0	0.251
GPT-4o	22.8	60.6	16.3	58.7	6.9	47.4	0.068	8.2	48.5	0.094
<i>Compact LVLMs</i>										
Aquila-VL-2B	3.1	45.9	2.2	44.0	1.0	15.7	0.037	0.3	20.8	0.056
Aquila-VL-2B*	40.3	79.7	17.3	63.5	19.6	<u>65.0</u>	0.359	7.4	53.1	0.332
InternVL2.5-2B	17.6	59.8	12.3	58.1	6.6	47.5	0.057	5.3	47.2	0.047
InternVL2.5-2B*	44.0	79.8	29.2	74.0	<u>20.0</u>	59.6	<u>0.411</u>	10.9	54.6	<u>0.426</u>
DeepseekVL2-Tiny	3.2	46.1	3.1	50.9	0.0	0.0	0.000	0.0	0.0	0.000
DeepseekVL2-Tiny*	<u>43.8</u>	<u>85.4</u>	<u>29.8</u>	<u>77.8</u>	16.3	61.0	0.199	<u>11.2</u>	<u>57.3</u>	0.313
SmoVLM2-2.2B	2.1	41.2	1.1	37.3	0.0	0.0	0.000	0.0	0.0	0.000
SmoVLM2-2.2B*	22.0	66.4	11.6	53.1	12.3	57.5	0.257	10.2	51.2	0.235
PaliGemma2-3B	0.1	34.8	1.1	37.6	0.0	0.0	0.000	0.0	0.0	0.000
PaliGemma2-3B*	13.4	55.4	24.8	65.4	14.7	54.6	0.106	10.9	51.7	0.157
Qwen2.5-VL-3B	19.5	63.0	10.5	58.3	3.3	19.6	0.073	2.2	17.8	0.068
Qwen2.5-VL-3B*	53.6	87.7	36.8	80.5	26.4	67.0	0.457	17.5	59.4	0.427
<i>Cascade Pipelines</i>										
EasyOCR + LLM	21.3	58.0	19.1	63.0	5.5	47.0	0.223	5.4	49.0	0.305
GOT-OCR + LLM	38.2	75.5	27.3	71.7	11.6	45.9	-	6.3	47.4	-

Table 4: Evaluation results of proprietary, compact LVLMs and cascade systems on PATIMT-Bench across two sub-tasks: region-specific translation and full-image translation with grounding, evaluated on both English → Chinese (EN-ZH) and Chinese → English (ZH-EN) using BLEU, COMET, and IoU metrics. Models marked with * indicate fine-tuning on our PATIMT train set. Best results are marked in **bold** and second-best results are underlined.

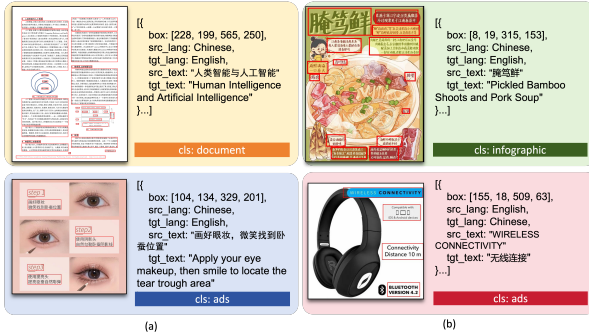


Figure 6: samples from train set (left) and test set (right).

box numbers after processing by our pipeline, and source/target word numbers, as presented in Table 3. A significant drop can be observed in the number of bounding boxes after processing, indicating that our pipeline effectively merges line-level OCR detection boxes to mitigate potential semantic fragmentation. To ensure the quality of our dataset, we further conduct manual validation on a randomly sampled subset of 1,000 training images and verify both bounding box annotations and translations, which achieves a 92% approval rate. We also ex-

hibit some examples as shown in Figure 6.

4.2.2 Test set

Specifically, considering stylistic similarities in text across certain image categories, during the evaluation, we group advertisement/poster/book cover, table/chart, and natural scene/street view into three categories and remain the other scenarios unchanged. Therefore, the final evaluation includes images of six categories. The test set consists of 1,200 images including English → Chinese and Chinese → English, with 100 images manually selected and annotated for each category. Similar to the training set, the statistical results of the indicators and the sample demonstrations are shown in Table 3 and Figure 6.

5 Experiments

Section 5.1 outlines our experimental setup, including evaluation metrics, baseline models and implementation details. Section 5.2 presents the main results, demonstrating performance improvements obtained by training on our dataset. In Section 5.3, we conduct an ablation study to evaluate the effec-

tiveness of our data construction pipeline. Section 5.4 assesses the scalability of the training dataset, while Section 5.5 examines its generalizability on the relevant benchmark. Finally, Section 5.6 provides an analysis of the trade-off between speed and performance across varying image-compression ratios.

5.1 Experimental Setting

5.1.1 Metrics

We report case-sensitive detokenized BLEU using SacreBLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) to evaluate translation quality, and assess the grounding capability in full-image translation with grounding task using the Intersection over Union (IoU) metric.

5.1.2 Baselines

Compact LVLMS. We select six LVLMS as our baselines: Aquila-VL-2B (Gu et al., 2024), InternVL-2.5-2B (Chen et al., 2024), Deepseek-VL2-Tiny (Wu et al., 2024), SMOVLMS-2.2B⁶, PaliGemma2-3B (Steiner et al., 2024) and Qwen2.5-VL-3B (Bai et al., 2025). Each model is evaluated under two conditions: zero-shot inference and fine-tuning on our proposed training data. This dual evaluation provides a reliable assessment of the quality of our constructed dataset. Detailed introduction of these baseline models are listed in Appendix B.1.

Proprietary LVLMS. To benchmark our approach against more advanced vision-language models, we establish proprietary models using two state-of-the-art LVLMS: Qwen2.5-VL-72B (Bai et al., 2025) and GPT-4o (Hurst et al., 2024), both evaluated without fine-tuning on our dataset.

Cascade Pipelines. Additionally, we implement cascade baselines that integrate EasyOCR or GOT-OCR (Wei et al., 2024) for text recognition with Qwen2.5-VL-3B (Yang et al., 2024) for translation to facilitate comparison with the aforementioned end-to-end image machine translation methods.

5.1.3 Implementation Details

Through our experiments, proprietary LVLMS, Qwen2.5-VL-3B and InternVL-2.5-2B can consistently generate outputs in JSON format, which yield superior performance, we fine-tune Qwen2.5-VL-3B and InternVL-2.5-2B and evaluate all these four models in JSON format. The

⁶<https://huggingface.co/blog/smolvlm2>

Data Process	EN-ZH		ZH-EN	
	BLEU	COMET	BLEU	COMET
OCR only	42.0	79.4	29.8	75.6
MinerU only	46.7	81.4	31.7	74.2
Ours	49.3	84.0	34.0	76.7

(a) Ablation on region-specific translation.

Data process	EN-ZH			ZH-EN		
	BLEU	COMET	IoU	BLEU	COMET	IoU
OCR only	16.0	58.8	0.319	10.9	50.8	0.326
MinerU only	21.0	63.9	0.343	13.0	54.4	0.357
Ours	22.6	66.0	0.414	14.3	57.9	0.367

(b) Ablation on full-image translation with grounding.

Table 5: Ablation on different data processing methods. *OCR only* and *MinerU only* denote using EasyOCR and MinerU to generate OCR results without spatial merge and refinement. Ours denotes using our data construction pipeline. Best results are marked in **bold**.

training of compact LVLMS is conducted with a batch size of 128 on four A6000 GPUs. Complete inference and training settings are provided in Appendix B.2.

5.2 Main Results

Table 4 reports the average performance across different image scenarios in the PATIMT-Bench. Compact LVLMS struggle to follow PATIMT instructions under zero-shot settings, resulting in low BLEU, COMET, and IoU scores. After fine-tuning on our proposed dataset, all compact LVLMS achieve competitive performance, with some results even surpassing Qwen2.5-VL-72B and GPT-4o. For instance, the BLEU score of Aquila-VL-2B in the EN-ZH region-specific translation task increases from 3.1 to 40.3, and COMET improves from 45.9 to 79.7. Remarkably, Qwen2.5-VL-3B stands out among the baselines, outperforming both cascade pipelines and proprietary LVLMS by wide margins in most metrics. Detailed results for each scenario are provided in Appendix C.

For clear visualization, we render the results of full-image translation with grounding based on the predicted grounding information, as shown in Figure 7, which demonstrates that our model generates translations with accurate grounding.

Overall, these results validate the effectiveness of our dataset in enhancing translation quality and text spatial grounding to handle PATIMT.

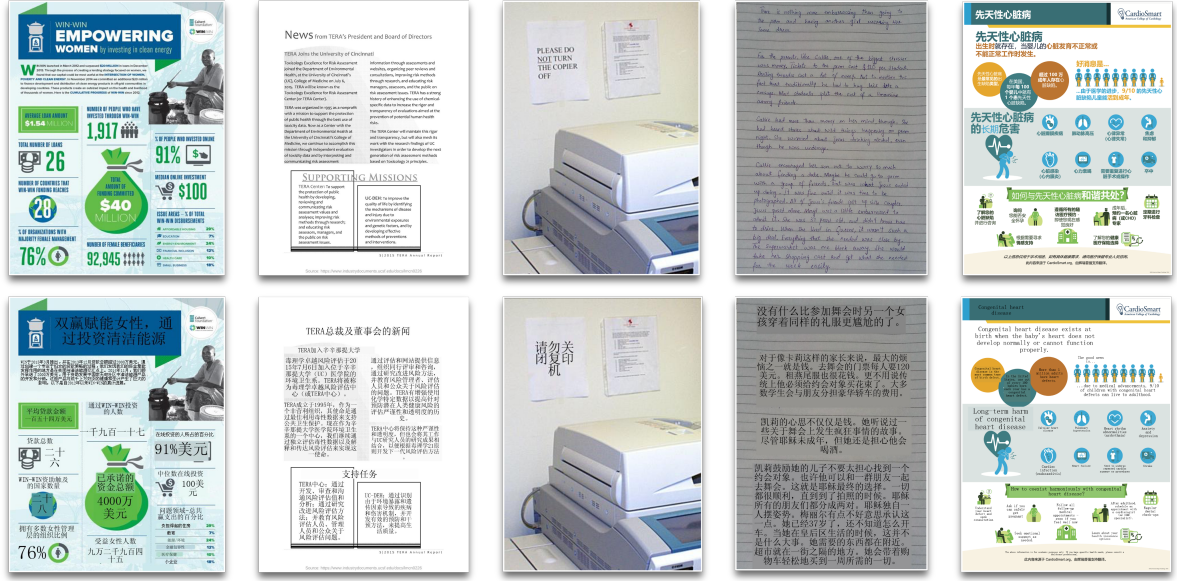


Figure 7: Visualization of the full-image translation with grounding results by rendering the model outputs onto the corresponding source images based on their grounding information. The top row shows the source images, while the bottom row displays the rendered outputs.

Scale	EN-ZH		ZH-EN	
	BLEU	COMET	BLEU	COMET
5K	49.3	84.0	34.0	76.7
10K	51.0	84.3	35.2	77.9
all	53.7	87.7	36.8	80.5

(a) Results on region-specific translation.

Scale	EN-ZH			ZH-EN		
	BLEU	COMET	IoU	BLEU	COMET	IoU
5K	22.6	63.0	0.414	14.3	57.9	0.367
10K	24.3	64.7	0.432	15.8	58.7	0.405
all	26.4	67.0	0.457	17.5	59.4	0.427

(b) Results on full-image translation.

Table 6: Results of scalability of our data based on Qwen2.5-VL-3B, dK denotes the base model is fine-tuned on dK subset from our train set, *all* represents training on the entire dataset. The best results are marked in **bold**.

5.3 Ablation Study

To evaluate the effectiveness of our data construction pipeline, we assess the performance of Qwen2.5-VL-3B using EasyOCR or MinerU annotations without implementing our adaptive processing with refinement strategy. Given the high cost of GPT-based labeling, we randomly sample 10 % of instances from each scenario in our training dataset, resulting in a subset of 5,000 examples. As shown in Table 5, the first row and second row

Models	BLEU	COMET
Fox	13.8	36.6
Qwen2.5-VL-3B	9.4	89.2
Qwen2.5-VL-3B*	47.9	91.7

Table 7: Comparison of our fine-tuned model on Fox benchmark (Liu et al., 2024a). The best results are marked in **bold**.

represent training on the subset annotated by EasyOCR and MinerU without spatial merge and refinement. The last row corresponds to training on the subset processed by our pipeline, demonstrating a clear performance improvement.

5.4 Scalability

To assess the scalability of our dataset, we construct two additional training subsets of 5,000 and 10,000 instances. As illustrated in Table 6, we observe a steady improvement in performance, thereby demonstrating the scalability of our dataset.

5.5 Extending to Other Benchmarks

To further assess the generalizability of our training data, we evaluate our baseline models on the Fox benchmark (Liu et al., 2024a), which contains document region-specific text image machine translation. As shown in Table 7, our fine-tuned model substantially outperform both the Fox model and the baseline model by approximately 400% in

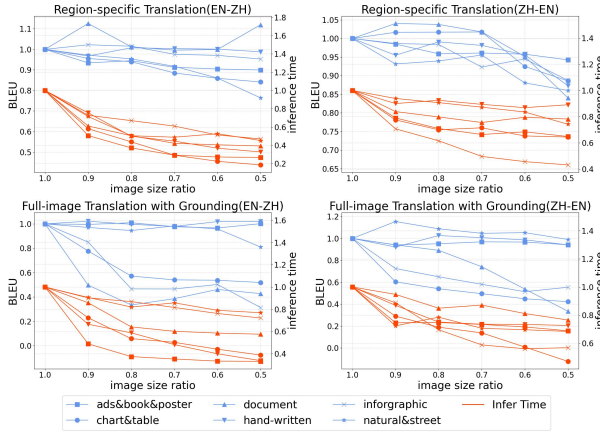


Figure 8: Illustration of the BLEU score (left y-axis) and inference time (right y-axis) across varying compression ratios (x-axis).

BLEU score and 250% in COMET, demonstrating the broad applicability of our training data and benchmark.

5.6 Tradeoff Between Speed and Performance Across Image Compression Ratios.

High-resolution images provide fine-grained visual information facilitating great performance, while generate excessive visual tokens which significantly increase the inference time (Zhang et al., 2024). To explore this influence in PATIMT task, we conduct a series of experiments. Specifically, we here compress the images in our test set to different ratios and measure the change of both inference time and BLEU score, as shown in Figure 8. For clarity, We report BLEU scores and inference times as decimal fractions of the uncompressed baseline performance. From the resulting plot, we draw the following conclusions:

- In region-specific translation task, the performance remains comparable even when images are significantly compressed. It indicates that this task holds the great potential to accelerate through the compression of visual features.
- In full-image translation with grounding task, the performance of some scenarios such as chart&table, document, and infographic are sensitive to image compression ratio. These scenarios may be adversely affected by limited image size, whereas other categories maintain stable performance.

6 Conclusion

In this paper, we extend the conventional TIMT task into PATIMT task, which encompasses two sub-tasks: region-specific translation and full-image translation with grounding. Confronted with data scarcity, we construct the PATIMT-Bench, a benchmark featuring 10 distinct image scenarios. We introduce an Adaptive Image OCR Refinement Pipeline to construct training data, which adaptively selects suitable OCR tools according to different image scenarios and refines the results for text-rich images to ensure high-quality annotations. Notably, to ensure the accuracy of evaluation, we manually annotate bounding boxes and review the translation results of 1,200 instances to construct the test set. LVLMs fine-tuned on our data achieve state-of-the-art performance on PATIMT-Bench, and demonstrate the scaling ability of our training data. In the future, we will explore the application of our dataset to domains including in-image machine translation (Tian et al., 2023; Lan et al., 2024) and visual text generation (Tuo et al., 2023; Liu et al., 2023; Li et al., 2024b; Esser et al., 2024). Additionally, we aim to expand our benchmark into a large-scale multilingual version.

Limitations

Despite the contributions of our benchmark in advancing PATIMT and achieving impressive performance, several limitations still remain. Our benchmark predominantly focuses on bounding boxes for region annotation. However, in practical applications, users may prefer or require other formats such as polygons, points, or free-form shapes. Besides, multilingual translation is not explored in our benchmark.

Acknowledgments

The project is supported by National Key R&D Program of China (No. 2022ZD0160501), Natural Science Foundation of Fujian Province of China (No. 2024J011001), and the Public Technology Service Platform Project of Xiamen (No.3502Z20231043). We also thank the reviewers for their insightful comments.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie

- Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. [Qwen2.5-vl technical report](#). *CoRR*, abs/2502.13923.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. [Internlm2 technical report](#).
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. [Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts](#). In *CVPR*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 21 others. 2024. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *CoRR*, abs/2412.05271.
- Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. [Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models](#).
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. [Scaling rectified flow transformers for high-resolution image synthesis](#). *Preprint*, arXiv:2403.03206.
- Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2025. [Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning](#). *CoRR*, abs/2501.00321.
- Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. 2022. [Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark](#). In *NeurIPS*.
- Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, Zhenchong Hu, Bo-Wen Zhang, Jijie Li, Dong Liang, Yingli Zhao, Yulong Ao, Yaoqi Liu, Fangxiang Feng, and Guang Liu. 2024. [Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data](#). *CoRR*, abs/2410.18558.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. [Gpt-4o system card](#). *CoRR*, abs/2410.21276.
- Puneet Jain, Orhan Firat, Qi Ge, and Sihang Liang. 2021. [Image translation network](#). In *Image Translation Model*.
- Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. [ReferItGame: Referring to objects in photographs of natural scenes](#). In *EMNLP*.
- Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, Min Zhang, and Jinsong Su. 2024. [Translatotron-V\(ision\): An end-to-end model for in-image machine translation](#). In *ACL*.
- 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 *ACL*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. [Llava-onevision: Easy visual task transfer](#). *CoRR*, abs/2408.03326.
- Bo Li, Shaolin Zhu, and Lijie Wen. 2025. [MIT-10M: A large scale parallel corpus of multilingual image translation](#). In *COLING*.
- Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022. [Grounded language-image pre-training](#). In *CVPR*.
- Wenbo Li, Guohao Li, Zhibin Lan, Xue Xu, Wanru Zhuang, Jiachen Liu, Xinyan Xiao, and Jinsong Su. 2024b. [Empowering backbone models for visual text generation with input granularity control and glyph-aware training](#). In *EMNLP*.
- 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 *NAACL*.
- Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023. [On the cultural gap in text-to-image generation](#). *Preprint*, arXiv:2307.02971.
- Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. [Focus anywhere for fine-grained multi-page document understanding](#). *CoRR*, abs/2405.14295.

- Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. [Mminstruct: a high-quality multi-modal instruction tuning dataset with extensive diversity](#). *Science China Information Sciences*.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Cong Ma, Yaping Zhang, Mei Tu, Xu Han, Linghui Wu, Yang Zhao, and Yu Zhou. 2022. [Improving end-to-end text image translation from the auxiliary text translation task](#). In *ICPR*.
- Elman Mansimov, Mitchell Stern, Mia Xu Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. [Towards end-to-end in-image neural machine translation](#). *CoRR*, abs/2010.10648.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. [Generation and comprehension of unambiguous object descriptions](#). In *CVPR*.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. [Chartqa: A benchmark for question answering about charts with visual and logical reasoning](#). In *ACL*.
- Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. [Infographicvqa](#). In *WACV*.
- Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. [Docvqa: A dataset for VQA on document images](#). In *WACV*.
- Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. [OCR-VQA: visual question answering by reading text in images](#). In *ICDAR*.
- Liqiang Niu, Fandong Meng, and Jie Zhou. 2024. [UMTIT: unifying recognition, translation, and generation for multimodal text image translation](#). In *COLING*.
- Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. 2023. [Teaching CLIP to count to ten](#). In *ICCV*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *ACL*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *ICML*, Proceedings of Machine Learning Research.
- Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *EMNLP*.
- Long Rujiao, Wang Wen, Xue Nan, Gao Feiyu, Yang Zhibo, Wang Yongpan, and Xia Gui-Song. 2021. Parsing table structures in the wild. In *ICCV*.
- Nilesh P. Sable, Priya Shelke, Ninad Deogaonkar, Nachiket Joshi, Rudra Kabad, and Tushar Joshi. 2023. Doc-handler: Document scanner, manipulator, and translator based on image and natural language processing. In *ESCI*.
- Huangjun Shen, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024. [A survey on multi-modal machine translation: Tasks, methods and challenges](#). *Preprint*, arXiv:2405.12669.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. [Towards VQA models that can read](#). In *CVPR*.
- Andreas Steiner, André Susano Pinto, Michael Tschanen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey A. Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, R. Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. [Paligemma 2: A family of versatile vlms for transfer](#). *CoRR*, abs/2412.03555.
- Yipeng Sun, Dimosthenis Karatzas, Chee Seng Chan, Lianwen Jin, Zihan Ni, Chee Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, and Jingtuo Liu. 2019. [ICDAR 2019 competition on large-scale street view text with partial labeling - RRC-LSVT](#). In *ICDAR*.
- 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 *EMNLP*.
- Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. 2023. [Anytext: Multilingual visual text generation and editing](#).
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liquan Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. [Mineru: An open-source solution for precise document content extraction](#). *Preprint*, arXiv:2409.18839.
- Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. [Layoutreader: Pre-training of text and layout for reading order detection](#). *Preprint*, arXiv:2108.11591.
- Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao,

- Jianjian Sun, Yuang Peng, and 1 others. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. [Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding](#). *CoRR*, abs/2412.10302.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. [Vision-language models for vision tasks: A survey](#). *Preprint*, arXiv:2304.00685.
- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023a. [Llavar: Enhanced visual instruction tuning for text-rich image understanding](#). *CoRR*.
- Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023b. [Layoutdit: Layout-aware end-to-end document image translation with multi-step conductive decoder](#). In *EMNLP*.
- 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 *ACL*.

A Adaptive Image OCR Refinement Pipeline

A.1 Data Comparison

For more clear comparison, we show the example input images and the output format of our dataset comparing to existing TIMT datasets, as shown in Figure 8.

A.2 CLIP-based Categorization

Following (Zhang et al., 2023a), we divide the images in our training data into 10 distinct classes. Each class is associated with one or more descriptive labels.

- **ads:** "advertisement"
- **book:** "book cover", "magazine cover", "comic book cover"
- **poster:** "movie poster", "podcast poster", "TV show poster", "event poster", "poster", "concert poster", "conference poster", "travel poster", "art poster"
- **natural:** "natural scene", "landscape", "nature background", "wildlife scene", "Trail sign", "Park map", "Info board", "Gate sign", "Stone plaque", "Wood post", "Kiosk sign", "Exhibit panel"
- **street:** "street view", "urban scene", "city street", "suburban neighborhood", "rural road", "traffic scene", "billboard", "shop front"
- **hand-written:** "hand-written", "handwriting letter"
- **infographic:** "infographic", "diagram", "mind map", "statistical graph"
- **document:** "document", "contract"
- **chart:** "chart", "bar chart", "pie chart", "scatter plot", "line chart", "Histogram", "area chart", "bubble chart",
- **table:** "table", "spreadsheet", "matrix", "grid"

For each word, we apply the same textual templates used in Zhang et al. (2023a) to achieve embedding-space ensembling (Radford et al., 2021):

- "a photo of a {}.",
- "a blurry photo of a {}.",
- "a black and white photo of a {}.",
- "a low contrast photo of a {}.",

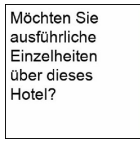
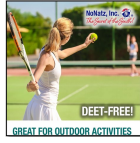
Dataset	Input Image	Output Format
OCRMT30K		Plain Text
DiTrans		Plain Text
DoTA		Markdown Text
UMTIT		Image
MIT-10M		Plain Text
PATIMT(Ours)		Plain Text & Bounding Box

Table 8: Image and output format comparison of PATIMT with other popular image translation datasets.

- "a high contrast photo of a {}.",
- "a bad photo of a {}.",
- "a good photo of a {}.",
- "a photo of a small {}.",
- "a photo of a big {}."

Using CLIP-ViT-L/14, we compute the similarity between each image and all associated labels. Each image is then assigned to the corresponding superclass (e.g., book) of the label (e.g., "book cover") with the highest similarity score.

A.3 Instruction Tuning Data

In this section, we detail the question and label formats used for each baseline model during training, as illustrated in Table 9.

Image	Question Format	Response Format
	First pinpoint the words in <code>Box([0.13, 0.09, 0.28, 0.15])</code>, then express them in Chinese. Output only the recognized text content and translation result in format: text <translation> translation.	step1 <translation> 步骤1
	Extract all visible text and offer its English meaning. Return the recognized text content, translation result and boxes in format: text <translation> translation Box([x1, y1, x2, y2]).	画好眼妆微笑找到卧蚕位置 <translation> Apply your eye makeup, then smile to locate the tear trough area Box([0.14, 0.17, 0.43, 0.26])\n ...

(a) Example of instruction tuning data with plain-text format. Aquila-VL-2B and DeepseekVL2-Tiny utilize this format.

Image	Question Format	Response Format
	First read the English snippet at <code>Box([40, 553, 730, 596])</code>, then provide its Chinese version.. Output result in the following JSON format (note xxx is placeholder for text, x1,y1,x2,y2 are placeholders for coordinate).{"bbox_2d": Box([x1,y1,x2,y2]), "text_content": xxx, "translation": xxx}	json{"bbox_2d": "Box([40, 553, 730, 596])","text_content": "GIVE THEM A SAFE CUD-DLE SPACE","translation": "给他们一个安全的拥抱空间"}
	Can you do text detection and translation from English to Chinese?. Output result in the following JSON format (note xxx is placeholder for text, x1,y1,x2,y2 are placeholders for coordinate, ... means there may be more contents in the image).["bbox_2d": Box([x1,y1,x2,y2]), "text_content": xxx, "translation": xxx,...].	json[{"bbox_2d": "Box([40, 553, 730, 596])","text_content": "GIVE THEM A SAFE CUD-DLE SPACE","translation": "给他们一个安全的拥抱空间"}, ...]

(b) Example of instruction tuning data with JSON format. InternVL2.5-2B and Qwen2.5VL-3B utilize this format.

Table 9: Example of instruction tuning data with different format. texts marked in **bold** refer to diverse question generated by GPT-4o, `Box(·)` denotes converting bounding box to the format utilized by each baseline model, such that `Box([10,20,30,40])` is `[10,20,30,40]` for Qwen2.5-VL-3B and `<box>[[10,20,30,40]]</box>` for InternVL2.5-3-2B.

B Detailed Experiment Settings

B.1 Details of baseline models.

We introduce the trainable parameters, bounding box format and other settings of our selected LVLM baseline models as the following:

- *Aquila-VL-2B* (Gu et al., 2024). This model is developed based on the LLaVA-One-Vision framework (Li et al., 2024a), utilizing the Qwen2.5-1.5B-Instruct (Yang et al., 2024) as the language model and SigLIP-SO400M-Patch14-384⁷ as the vision tower. It contains a total of 2.18 billion trainable parameters. The bounding box format is `[x1, y1, x2, y2]`, where each coordinate represents a normalized ratio in the range `[0,1]`.
- *InternVL-2.5-2B* (Chen et al., 2024). This model employs InternLM2.5-1.8B-Chat (Cai et al., 2024) as the large language model and InternViT-300M-448px-V2.5⁸ as the vision

tower, with a randomly initialized MLP projector. It has 2.21 billion trainable parameters. The bounding box format is `<box>[x1, y1, x2, y2]</box>`, where the coordinates are normalized to the range `[0,1000]`.

- *DeepSeek-VL2-Tiny* (Wu et al., 2024). This model is based on DeepSeekMoE-3B (Dai et al., 2024), comprising 3.37 billion trainable parameters and 1.0 billion activated parameters during inference. The bounding box format is `<det>[x1, y1, x2, y2]</det>`, where coordinates are normalized to the range `[0,999]`.
- *SmoVLM2-2.2B*. This model is designed for efficient video understanding across various devices, offering strong visual understanding and localization capabilities. The bounding box format is `[x1, y1, x2, y2]`, where each coordinate represents a normalized ratio in the range `[0,1]`.
- *PaliGemma2-3B* (Steiner et al., 2024). This model connects the SigLIP image encoder with the Gemma2 language model, supporting

⁷<https://huggingface.co/google/siglip-so400m-patch14-384>

⁸https://huggingface.co/OpenGVLab/InternViT-300M-448px-V2_5

various input resolutions (224x224, 448x448, and 896x896) for different use cases. The bounding box format is $[x1, y1, x2, y2]$, where each coordinate represents a normalized ratio in the range $[0, 1]$.

- *Qwen2.5-VL-3B* (Bai et al., 2025). This model demonstrates strong visual understanding and localization capabilities. It has 3.75 billion trainable parameters. The bounding box format is $[x1, y1, x2, y2]$, using absolute position coordinates.

B.2 Details of Training and Inference Configuration

We list the detailed training settings as the following:

- **Optimization Settings:**
 - Learning rate: $1e-5$ with cosine scheduling
 - Warmup ratio: 0.1
 - Weight decay: 0.0
 - Batch size: 128
 - Training epoch: 1
 - Optimizer: AdamW
- **Computational Environment:**
 - Precision: bfloat16 (bf16)
 - Acceleration framework: DeepSpeed Stage 3
 - Hardware: $4 \times$ NVIDIA A6000 GPUs

During inference, we set the temperature parameter to zero and employ greedy decoding. To prevent premature truncation of generated sequences, we specify the maximum number of new tokens as the greater of the ground-truth sequence or length 4096.

C Complete Results of Main Experiments

This section presents the complete results of our main experiments. Table 10 reports the results for the region-specific translation task, while Table 11 provides detailed results for the full-image translation with grounding task. From these tables, we observe that Qwen2.5-VL-3B achieves the best performance across most metrics after fine-tuning. Additionally, models such as Aquila-VL-2B and DeepseekVL2-Tiny demonstrate strong

performance despite their relatively limited foundational capabilities.

Moreover, our evaluation reveals distinct performance patterns across different domains:

- **Easy Domains.** Most LVLMs achieve high performance in easy domains such as ads&books&posters, and natural scenes&street view. The improvement is limited in these domains because the number of text regions is usually small and often dominates the image, making them easier to recognize. In contrast, domains like charts&tables and hand-written text show significant improvement. Charts&tables contain small characters, while hand-written text contains characters that are harder to recognize.
- **Hard Domains.** Performance in document and infographic domains is similar across all LVLMs. Both domains contain long paragraphs and small characters. The primary difference lies in layout: documents typically have a structured layout, while infographics have a more random layout. However, experiments show that this difference does not significantly impact performance. We attribute this to the models’ ability to accurately locate texts after fine-tuning across multiple domains.
- **However, we observe the opposite pattern in ZH-EN full-image translation with grounding, where models perform better on hard domains than on easy domains.** We attribute this to the fact that difficult scenarios typically contain longer, semantically coherent text, which provides richer contextual information to guide translation. In contrast, simple scenarios often feature short, context-deficient phrases, such as advertising slogans or highly localized Chinese expressions, that most models are poorly trained to handle (e.g. 新华书店(Xinhua Bookstore) is translated to New China Bookstore). Even when these phrases are accurately recognized, they show suboptimal translation quality for such culturally embedded phrases.

Model	ads&book&poster		chart&table		document		hand-written		infographics		natural&street	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
<i>Proprietary LVLs</i>												
Qwen2.5-VL-72B	47.2	78.3	50.5	78.5	41.1	76.0	37.6	74.0	36.6	71.1	56.8	81.5
GPT-4o	36.6	72.8	16.2	54.5	8.1	45.3	28.8	68.3	11.7	51.6	35.4	70.9
<i>Compact LVLs</i>												
Aquila-VL-2B	2.7	50.1	2.7	43.1	2.1	40.5	3.4	41.7	2.4	44.1	5.3	55.6
Aquila-VL-2B*	39.6	80.9	47.0	81.4	38.2	79.4	36.1	81.1	34.2	76.5	46.9	79.0
InternVL2.5-2B	23.3	68.4	13.3	54.3	8.6	48.7	20.2	64.8	10.6	53.3	29.7	69.0
InternVL2.5-2B*	49.8	81.8	47.5	83.2	39.4	79.3	36.8	76.5	38.7	77.8	52.0	80.0
DeepseekVL2-Tiny	2.4	49.9	2.4	41.8	2.0	41.7	4.4	44.0	3.8	44.8	4.4	54.5
DeepseekVL2-Tiny*	46.7	86.2	54.5	86.7	41.2	83.9	35.8	85.1	33.8	88.1	51.0	82.6
SmolVLM2-2.2B	3.9	47.9	1.2	38.2	1.2	34.6	0.5	37.8	1.6	37.5	4.2	51.2
SmolVLM2-2.2B*	26.3	72.3	22.9	66.6	21.1	66.2	15.8	62.0	17.9	61.8	27.8	69.3
PaliGemma2-3B	0.3	43.5	0.1	34.6	0.1	28.3	0.0	25.0	0.1	34.6	0.0	42.8
PaliGemma2-3B*	36.6	76.3	23.4	65.4	15.0	59.0	12.9	56.2	20.9	62.1	39.9	73.4
Qwen2.5-VL-3B	31.3	68.2	20.9	62.2	8.6	56.0	6.0	61.9	12.2	53.7	38.0	75.7
Qwen2.5-VL-3B*	52.6	86.5	60.3	90.2	51.3	86.8	52.7	89.7	47.0	86.4	57.8	86.7
<i>Cascade Pipelines</i>												
EasyOCR + LLM	20.2	59.0	23.6	63.3	31.6	69.3	3.8	32.8	34.2	69.2	14.5	54.6
GOT-OCR + LLM	40.9	77.4	41.3	76.2	33.1	73.1	43.6	80.1	32.8	73.3	37.5	73.1

(a) Detailed results for region-specific translation task (EN-ZH).

Model	ads&book&poster		chart&table		document		hand-written		infographics		natural&street	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
<i>Proprietary LVLs</i>												
Qwen2.5-VL-72B	30.4	73.5	48.7	83.6	40.3	76.2	28.0	67.3	46.6	81.4	29.6	69.3
GPT-4o	26.1	65.1	19.5	64.9	8.6	53.1	17.0	57.8	9.7	51.9	16.9	59.6
<i>Compact LVLs</i>												
Aquila-VL-2B	3.2	44.6	2.6	46.7	0.4	43.1	1.9	41.5	2.3	44.1	2.9	44.2
Aquila-VL-2B*	19.6	66.4	27.2	71.7	9.5	65.1	8.9	52.0	21.4	64.9	17.3	60.6
InternVL2.5-2B	14.0	60.4	10.9	59.7	5.7	55.5	14.4	61.0	15.0	56.3	13.9	55.4
InternVL2.5-2B*	25.5	74.3	34.2	77.9	35.7	80.5	21.4	67.5	33.9	74.4	24.4	69.5
DeepseekVL2-Tiny	3.2	49.9	3.4	51.2	1.3	54.2	3.8	51.7	1.9	49.3	5.3	49.0
DeepseekVL2-Tiny*	24.2	75.4	40.6	84.5	33.6	83.4	20.8	71.7	30.0	76.5	29.7	75.0
SmolVLM2-2.2B	1.9	39.2	1.0	36.6	0.3	36.1	0.9	37.3	0.6	36.1	1.8	38.3
SmolVLM2-2.2B*	11.2	53.2	10.8	52.7	9.1	55.0	8.9	51.1	13.9	48.0	15.7	58.6
PaliGemma2-3B	1.5	39.3	1.3	42.6	0.1	29.4	1.4	39.3	1.4	38.7	1.0	36.5
PaliGemma2-3B*	18.0	63.1	13.5	54.0	8.7	51.3	13.3	57.1	12.3	49.7	14.7	57.4
Qwen2.5-VL-3B	9.2	55.8	11.7	62.8	7.4	58.4	14.1	63.8	11.0	55.2	9.7	53.9
Qwen2.5-VL-3B*	29.3	77.8	46.5	86.2	42.7	83.0	28.6	73.6	42.5	86.5	31.1	76.1
<i>Cascade Pipelines</i>												
EasyOCR + LLM	7.6	53.0	26.7	71.4	37.2	81.0	5.5	49.0	28.6	72.9	9.0	50.5
GOT-OCR + LLM	23.2	71.9	33.1	77.5	39.3	81.4	13.8	59.9	32.8	76.2	21.5	63.0

(b) Detailed results for region-specific translation task (ZH-EN).

Table 10: Detailed evaluation results for region-specific translation task. Models marked with * indicate fine-tuning on our PATIMT train set. Best results are masked in **bold**.

Model	ads&book&poster			chart&table			document			hand-written			infographics			natural&street		
	BLEU	COMET	IoU	BLEU	COMET	IoU	BLEU	COMET	IoU	BLEU	COMET	IoU	BLEU	COMET	IoU	BLEU	COMET	IoU
<i>Proprietary LVLMS</i>																		
Qwen2.5-VL-72B	19.5	65.9	0.274	9.8	46.6	0.110	3.0	37.7	0.056	12.7	37.8	0.153	1.7	35.7	0.036	32.8	68.4	0.480
GPT-4o	12.1	56.6	0.138	6.3	48.8	0.059	2.1	40.3	0.057	2.3	39.4	0.050	3.1	43.9	0.058	15.7	55.2	0.045
<i>Compact LVLMS</i>																		
Aquila-VL-2B	1.0	20.9	0.062	0.0	10.8	0.001	0.0	5.0	0.004	0.0	11.0	0.058	0.0	17.8	0.005	5.1	28.5	0.090
Aquila-VL-2B*	19.8	67.2	0.422	9.7	55.7	0.139	15.2	61.2	0.355	29.6	77.4	0.679	14.4	59.6	0.210	28.9	68.9	0.347
InternVL2.5-2B	8.1	53.7	0.056	2.2	43.7	0.015	1.6	40.8	0.035	9.3	51.1	0.185	1.6	41.8	0.030	16.6	53.8	0.022
InternVL2.5-2B*	28.8	72.7	0.512	10.6	47.4	0.212	14.1	52.8	0.428	25.3	68.6	0.641	9.2	45.1	0.246	32.2	70.9	0.428
DeepseekVL2-Tiny	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000
DeepseekVL2-Tiny*	21.9	66.3	0.364	6.8	51.9	0.076	4.1	49.4	0.085	21.5	70.8	0.109	12.1	58.5	0.182	31.3	69.2	0.376
SmolVLM2-2.2B	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000
SmolVLM2-2.2B*	17.7	67.3	0.446	6.0	51.4	0.077	9.9	53.5	0.191	12.3	57.1	0.471	6.6	51.3	0.098	21.2	64.5	0.256
PaliGemma2-3B	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000
PaliGemma2-3B*	22.0	57.5	0.143	9.2	51.0	0.041	13.9	52.0	0.206	9.5	48.9	0.141	11.0	53.1	0.063	22.8	64.8	0.043
Qwen2.5-VL-3B	5.3	19.0	0.093	9.8	46.6	0.110	1.2	15.9	0.071	0.0	18.1	0.084	3.5	16.2	0.059	0.1	1.7	0.023
Qwen2.5-VL-3B*	26.4	69.8	0.439	15.0	57.0	0.279	23.3	64.9	0.539	36.4	79.9	0.669	20.9	60.4	0.366	36.2	69.8	0.448
<i>Cascade Pipelines</i>																		
EasyOCR + LLM	11.0	54.9	0.312	5.7	52.7	0.251	1.2	42.3	0.094	0.3	29.0	0.087	3.1	49.5	0.222	11.9	53.7	0.372
GOT-OCR + LLM	6.9	41.3	-	2.3	35.7	-	14.6	48.2	-	38.6	75.6	-	2.9	34.3	-	4.1	40.3	-

(a) Detailed results for full-image translation with grounding task (EN-ZH).

Model	ads&book&poster			chart&table			document			hand-written			infographics			natural&street		
	BLEU	COMET	IoU	BLEU	COMET	IoU	BLEU	COMET	IoU	BLEU	COMET	IoU	BLEU	COMET	IoU	BLEU	COMET	IoU
<i>Proprietary LVLMS</i>																		
Qwen2.5-VL-72B	19.5	66.1	0.51	9.0	47.6	0.053	4.2	43.1	0.034	16.0	59.5	0.409	2.0	39.9	0.058	16.8	62.0	0.439
GPT-4o	16.2	54.7	0.149	8.7	53.5	0.059	3.0	47.2	0.078	8.1	42.3	0.105	3.0	47.1	0.086	10.1	46.0	0.089
<i>Compact LVLMS</i>																		
Aquila-VL-2B	0.6	23.0	0.069	0.3	21.4	0.007	0.0	12.9	0.014	0.2	24.5	0.144	0.1	17.9	0.008	0.4	25.2	0.094
Aquila-VL-2B*	10.6	56.6	0.419	7.5	54.3	0.124	5.5	54.7	0.339	7.8	48.9	0.482	4.8	51.0	0.247	8.4	53.0	0.379
InternVL2.5-2B	7.6	51.1	0.085	2.9	46.4	0.021	3.2	49.5	0.037	9.0	48.0	0.064	1.7	43.0	0.042	7.2	45.4	0.032
InternVL2.5-2B*	13.1	63.3	0.487	7.6	51.2	0.21	4.1	42.0	0.315	16.0	62.6	0.669	8.5	46.0	0.346	15.9	62.4	0.53
DeepseekVL2-Tiny	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000
DeepseekVL2-Tiny*	13.5	59.6	0.393	9.4	56.8	0.139	11.7	56.8	0.197	13.8	60.0	0.583	5.1	51.0	0.153	13.8	59.4	0.414
SmolVLM2-2.2B	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000
SmolVLM2-2.2B*	9.2	50.3	0.274	9.2	49.8	0.143	8.4	48.9	0.080	7.3	47.2	0.513	10.5	52.7	0.107	16.8	58.5	0.292
PaliGemma2-3B	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000	0.0	0.0	0.000
PaliGemma2-3B*	9.1	52.3	0.145	9.7	48.7	0.106	15.7	61.3	0.088	8.1	47.5	0.208	9.0	49.9	0.095	13.7	50.7	0.301
Qwen2.5-VL-3B	3.0	18.5	0.117	2.9	28.9	0.023	1.0	15.4	0.022	3.6	22.6	0.152	0.9	11.5	0.035	2.0	9.8	0.061
Qwen2.5-VL-3B*	15.9	64.5	0.481	15.8	58.5	0.236	21.3	56.7	0.381	22.6	63.1	0.62	15.7	54.4	0.389	14.0	59.0	0.452
<i>Cascade Pipelines</i>																		
EasyOCR + LLM	7.3	52.5	0.44	11.3	60.9	0.332	2.9	50.4	0.125	5.0	40.6	0.382	2.2	46.6	0.195	3.9	43.1	0.357
GOT-OCR + LLM	3.6	47.0	-	3.8	40.5	-	16.7	61.5	-	5.2	48.4	-	5.6	42.7	-	3.1	44.1	-

(b) Detailed results for full-image translation with grounding task (ZH-EN).

Table 11: Detailed evaluation results for full-image translation with grounding task. Models marked with * indicate fine-tuning on our PATIMT train set. Best results are masked in **bold**.