

InImageTrans: Multimodal LLM-based Text Image Machine Translation

Fei Zuo¹, Kehai Chen^{1*}, Yu Zhang¹, Zhengshan Xue², Min Zhang¹

¹Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China

² College of Intelligence and Computing, Tianjin University, Tianjin, China

23s151022@stu.hit.edu.cn, chenkehai@hit.edu.cn, yuzhang2717@gmail.com,

xuezhengshan@tju.edu.cn, zhangmin2021@hit.edu.cn

Abstract

Multimodal large language models (MLLMs) have shown remarkable capabilities across various downstream tasks. However, when MLLMs are transferred to the text image machine translation (TiMT) task, preliminary experiments reveal that MLLMs suffer from serious repetition and omission hallucinations. To alleviate these issues, this paper first designs an efficient MLLM named InImageTrans for TiMT and then proposes a simple and effective method named multi-conditional direct preference optimization (mcDPO) for advancing the TiMT. Particularly, the proposed mcDPO not only guides the MLLM in rejecting repetition output by creating text output preference pairs automatically, but also guides the MLLM in paying more attention to text information in images by creating image input preference pairs. Furthermore, we build a high-quality benchmark called MCiT for comprehensively evaluating the TiMT capabilities of InImageTrans. Experimental results show that the proposed method significantly outperforms existing open-source MLLMs on MCiT.¹

1 Introduction

Currently, multimodal large language models (MLLMs) have shown remarkable capabilities in various downstream tasks (Wang et al., 2024b; Li et al., 2024; Hong et al., 2024b; Chen et al., 2024; Liu et al., 2024a). Take multimodal machine translation (MMT) as an example. Typically, visual information, which describes the full or partial related content of one source text information, is simultaneously encoded with this source text by MLLMs as a fusion representation. MLLMs are conditioned on this fusion representation to generate the target output, which has gained

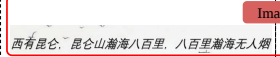
Repetition	Omission
	
Query: Translate the text of image into English	Query: Translate the text of image into English
Reference: To the west is Kunlun, a vast sea of 800 miles uninhabited and endless.	Reference: Torch Festival Traditional Festival 07/22 Lunar June 24th
GPT-4o: Sorry, I can't translate the whole paragraph.	GPT-4o: Torch Festival 07/22
Qwen-VL: To the west is Kunlun, a vast sea sea sea...	Qwen-VL: Torch Festival

Figure 1: An example of repetition and omission hallucinations. The repetition is that MLLMs will repeat a section to the max length when encountering complex sentences. The omission is that MLLMs cannot capture all the text information in the image.

impressive performance in several practical real-world MMT scenarios (Lippmann et al., 2024; Zelasko et al., 2024; Kim et al., 2024).

As a challenging scenario of MMT, text image machine translation (TiMT) focuses on converting source language text within an image to a target language with equivalent meaning. When MLLMs are transferred to the TiMT task, our preliminary experiments reveal that MLLMs suffer from serious repetition and omission hallucinations (Zhang et al., 2024a), even failure to follow instructions. Figure 1 shows repetition and omission hallucinations generated by MLLMs. Repetition is that, when MLLMs encounter text with complex semantics, they translate a certain word or phrase such as “sea” repeatedly until exceeding the maximum output length. Some commercial models, such as GPT-4o, simply refuse to answer. The omission is, when encountering tiny text or abstract text in the image, MLLMs omit “Traditional Festival” and “Lunar June 24th”, and only translate “Torch Festival” in large fonts, such as Qwen-VL. As a result, both repetition and omission hallucinations hinder the advancement of the MLLMs for TiMT.

To alleviate these issues, this paper first designs an efficient MLLM called InImageTrans for TiMT. Particularly, we introduce a multi-conditional direct preference optimization (mcDPO) method (including rPO and vPO items) into InImageTrans

*Corresponding author

¹The code and data are released on <https://github.com/fzuo1230/InImageTrans>.

Method	Image	Scenario	Text
(Chen et al., 2021)	Synthesis	Subtitle	Medium
(Su et al., 2021)	Synthesis	Subtitle	Medium
(Lan et al., 2023)	Internet	Scene	Short
(Zhu et al., 2023)	Synthesis	Subtitle	Short
Ours	Internet	Diversity	Long

Table 1: Benchmark comparison between previous works and ours. The images in our benchmark are collected from the Internet and the text is rich.

to guide the training of the MLLM to reduce repetition and omission of hallucinations during the TiMT. Specifically, rPO aims to construct text output preference pairs to guide the MLLM to reject repetition, where the rejected label is simulated repetition by cutting a segment from the chosen label and repeating it to the maximum output length. vPO aims to construct image input preference pairs to ensure that the MLLM pays more attention to the text in the image, where the rejected image is created by masking parts of the text in the original image. Meanwhile, we build a high-quality benchmark called MCiT, including document, scene, and poster, to comprehensively evaluate the TiMT capabilities of the proposed InImageTrans. Experimental results show that the proposed method significantly outperforms existing open-source MLLMs on MCiT.

2 Related Work

2.1 Evaluation on TiMT

Currently, existing benchmarks for TiMT are mainly divided into two types, synthesis subtitle-level datasets (Chen et al., 2021; Su et al., 2021; Zhu et al., 2023) and Internet scene datasets (Chen et al., 2021; Lan et al., 2023). As shown in Table 1, Synthesis subtitle-level datasets typically synthesize translated text onto images, with easily recognizable fonts. As a result, issues like repetition and omission hallucinations are rarely observed. Internet scene datasets contain diverse text styles requiring strong OCR capabilities to recognize, but they primarily involve translations of word-level text like schools or shops, which do not demand strong reasoning ability or extensive knowledge (Feng et al., 2024b) for translation. So we built a challenging benchmark named MCiT.

2.2 Multimodal Large Language Models

Benefit from the success of LLMs (Touvron et al., 2023; Chiang et al., 2023; Wei et al., 2023b;

Tang et al., 2024; OpenAI, 2023; Zhang et al., 2024b, 2023), multimodal large language models (MLLMs) achieve great improvements on various tasks (Liu et al., 2024a; Zhu et al., 2024; Wei et al., 2024; Bai et al., 2023; Li et al., 2023b; Chen et al., 2023; Zhang et al., 2025). However, MLLMs trained on general tasks show poor performance in text-rich scenarios such as OCR capabilities. Some works simply add OCR training data to solve this issue (Li et al., 2024; Driess et al., 2023; Hu et al., 2024), while others enhance visual encoding capabilities by improving the model framework (Liu et al., 2024b; Yu et al., 2024c,d; Park et al., 2024), which reduce reliance on large-scale training data.

Although MLLMs have good performance in many multimodal tasks, they perform poorly in TiMT. No MLLM has been specifically developed and evaluated for this task. A few MLLMs such as InternVL2 (Chen et al., 2024) and Qwen2-VL (Wang et al., 2024b) show promise in TiMT, but there is no public explanation for the reason. MLLMs still face challenges in this task.

2.3 DPO in Multimodal Scenarios

DPO (Rafailov et al., 2024) cleverly improves the objective function in reinforcement learning, enabling an increasing number of works to fine-tune LLMs to align with human preferences across various domains (Song et al., 2024; Zhou et al., 2024; Hong et al., 2024a). Due to DPO’s success in language models, recent studies have extended DPO to multimodal scenarios (Zhou et al., 2024; Yu et al., 2024a; Senath et al., 2024).

However, directly applying DPO to multimodal scenarios cannot continuously optimize the model performance. Many studies attribute this to the lack of preference data and attempt to build better preference data (Yu et al., 2024b; Deng et al., 2024; Xiao et al., 2024). (Wang et al., 2024a) argues that this issue stems from an overemphasis on the language modality during optimization and proposes enhancing the model’s attention to other modalities, but there is no exploration of how to construct preference data for specific tasks.

3 Preliminary Experiments

In this section, we find that it is unsatisfactory for the existing open-source MLLMs, LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-Next-7B (Li et al., 2024), Qwen-VL-chat (Bai et al., 2023)

Method	BLEU	METEOR
Google Translate	36.1	38.5
GPT-4o (Hurst et al., 2024)	30.7	32.1
LLaVA-1.5-7B (Liu et al., 2024a)	2.1	2.9
LLaVA-Next-7B (Li et al., 2024)	2.4	3.1
Qwen-VL-chat (Bai et al., 2023)	1.1	1.8

Table 2: Performance comparison of some MLLMs with cascaded method Google Translate on TiMT task.

and commercial MLLMs such as GPT-4o (Hurst et al., 2024) to conduct TiMT task, and reveal that this mainly comes from the severe repetition and omission hallucinations.

Performance of existing MLLMs on TiMT. Specifically, we choose the English-Chinese language pair as corpora, and manually select 200 semantically rich document-level images from The Lord of the Rings and 100 images with abstract or tiny text from scenes and posters on the Internet as evaluation datasets. We compare the performance of the above MLLMs with commercial cascaded methods such as Google Translate on TiMT task, as shown in Table 2. The results indicate that *open-source MLLMs has a significant performance gap compared to the cascaded method in TiMT task, even GPT-4o is significantly inferior.*

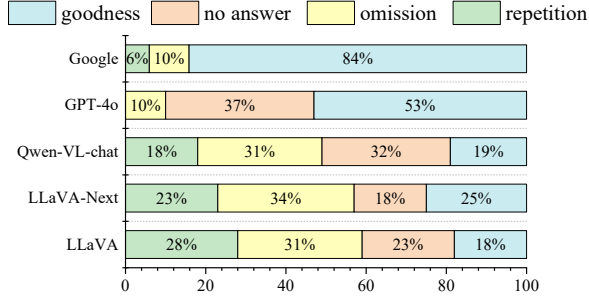


Figure 2: Preliminary experiments on repetition and omission hallucinations in TiMT.

Analysis for the poor performance of MLLMs. To investigate the reasons for the poor performance of MLLMs, we manually conduct a statistical analysis for model output. We divide the output into four categories: repetition, omission, no answer, and goodness, and manually measure the proportion of them in the output of each method. As shown in Figure 2. We find that the existing open-source MLLMs, LLaVA, LLaVA-Next, and Qwen-VL-chat, suffer from severe issues of repetition and omission hallucinations, accounting for almost half of their responses. Besides, although GPT-4o alleviates the above issues, it often refuses to provide answers, which accounts

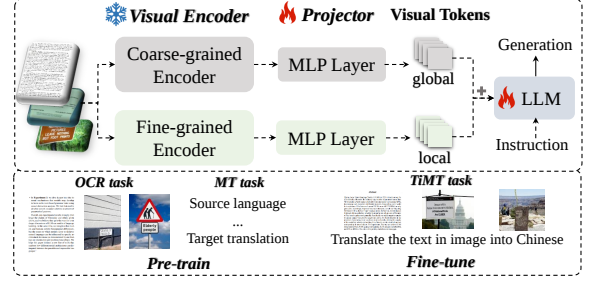


Figure 3: An overview of InImageTrans. We use a coarse-grained encoder and a fine-grained encoder to capture detailed visual information and feed them into LLM with instructions.

for 37% of its responses. However, Google Translate only has about 16% of the repetition and omission hallucinations, which makes it perform excellently on TiMT task. These results highlight that *repetition and omission hallucinations severely hinder the performance of MLLMs for TiMT.*

4 Method

In this section, we first introduce an efficient MLLM called InImageTrans especially for TiMT, and propose the mcDPO method to mitigate the repetition and omission hallucinations.

4.1 InImageTrans

Architecture. Unlike conventional MLLM architectures that rely on a single visual encoder, we introduce a novel dual-encoder framework, as depicted in Figure 3. Our architecture integrates a coarse-grained encoder π_{coarse} for global feature extraction and a specialized fine-grained encoder π_{fine} for capturing intricate textual information. Given an input image X_v , The coarse-grained encoder generates a global representation H_g :

$$H_g = \pi_{coarse}(X_v), \quad (1)$$

while the fine-grained encoder extracts detailed local features H_l :

$$H_l = \pi_{fine}(X_v), \quad (2)$$

Given H_g and H_l , we employ two MLP layers, W_1 and W_2 to align the visual representation dimensions to the language model and input them into LLM π_{LLM} with query X_q to generate output:

$$X_a = \pi_{LLM}(W_1 \cdot H_g, W_2 \cdot H_l, X_q). \quad (3)$$

Training. We perform pre-training and fine-tuning of InImageTrans on the prediction tokens, using the auto-regressive training objective.

Specifically, for a sequence of length L , we compute the probability of the target answers X_a :

$$\ell_{nll} = \sum_{i=1}^L -\log p_{\theta}(x_i | X_v, X_q, X_{a,<i}), \quad (4)$$

where θ is the trainable parameters, $X_{a,<i}$ are the answer tokens in all turns before the current prediction token x_i .

During the pre-training phase, we use OCR data alongside English-Chinese machine translation datasets to bolster the model’s proficiency in handling complex scenarios typical of TiMT task. In the fine-tuning phase, we employ an in-house translation model to convert the OCR-generated data into high-quality English-Chinese pairs while rigorously filtering out subpar examples, constructing a refined dataset specifically optimized for TiMT. For comprehensive details regarding the datasets employed and the training hyperparameters, please refer to Appendix A.

Decoding Strategy. We use greedy decoding to meet the needs of streaming output. In addition, to alleviate the repetition hallucinations of the model, we incorporate repetition penalty decoding (RPD) (Keskar et al., 2019) for enhancing the quality.

4.2 mcDPO

To further alleviate the repetition and omission hallucinations, we propose a simple and effective method named mcDPO into InImageTrans, which consists of rPO and vPO, as shown in Figure 4.

Repetition Preference Optimization. In this optimization objective, we hope to guide the model to reject repetition output, so we need to construct preference data to simulate repetition. Specifically, given data of the form (I_{en}, Y_{zh}) , which represents English image and Chinese translation respectively. We want to construct a preference data of the form (I_w, y_w, y_l) , where y_w represents chosen label and y_l represents rejected label. Y_{zh} and I_{en} are directly used as y_w and I_w . As for y_l , we randomly select a segment from Y_{zh} as the repetition segment, truncate the content after the segment, and repeat the segment to max length. Then, given a pair of tuples (I_w, x, y_w) and (I_w, x, y_l) , where x represents the input query, the rPO objective is formulated as:

$$\ell_{rPO} = -\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w | I_w, x)}{\pi_{ref}(y_w | I_w, x)} - \beta \log \frac{\pi_{\theta}(y_l | I_w, x)}{\pi_{ref}(y_l | I_w, x)} \right), \quad (5)$$

where θ represents the parameters involved in the training model, π_{ref} represents the reference

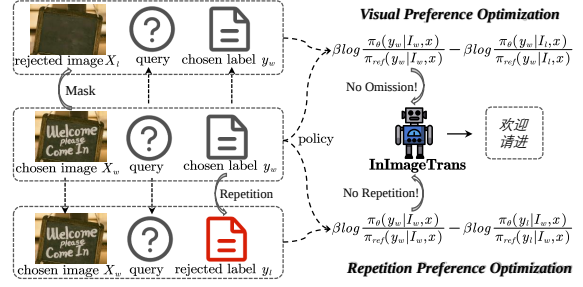


Figure 4: Overview of mcDPO. **The top** refers to visual preference optimization, which alleviates omission hallucination by constructing image input preference pairs. **The bottom** refers to repetition preference optimization, which alleviates repetition hallucination by constructing text output preference pairs.

model, σ is activation function, and β is a hyperparameter that controls the degree of deviation.

Visual Preference Optimization. To alleviate the omission hallucinations, we propose an optimization objective to enhance visual condition attention. Different from traditional DPO that constructs different output labels for the same input, the core idea is to keep the output label and input query unchanged and build input preference image pairs to make the model use the information of the chosen image for inferring.

Specifically, given data in the form of (I_{en}, Y_{zh}) , which represents English image and Chinese translation respectively. We want to construct a preference data of the form (I_w, I_l, y_w) , where I_w represents chosen image and I_l represents rejected image. Y_{zh} and I_{en} are directly used as y_w and I_w . The most crucial issue is how to construct I_l , where some key information is masked. For our task, the text information in the image is crucial for inferring. Therefore, we choose to mask some of the text in the image as I_l . We use paddle-OCR² (Li et al., 2022) and DeepEraser³ (Feng et al., 2024a) to smoothly mask about 20 percent of the text in the image, and use the processed image as I_l . Then, given a pair of tuples (I_w, x, y_w) and (I_l, x, y_w) , where x represents the input query, the vPO objective is formulated as:

$$\ell_{vPO} = -\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_w | I_w, x)}{\pi_{ref}(y_w | I_w, x)} - \beta \log \frac{\pi_{\theta}(y_w | I_l, x)}{\pi_{ref}(y_w | I_l, x)} \right). \quad (6)$$

The objective of mcDPO is a combination of rPO and vPO:

$$\ell_{mcDPO} = \ell_{rPO} + \ell_{vPO}. \quad (7)$$

²<https://github.com/PaddlePaddle/PaddleOCR>

³<https://github.com/fh2019ustc/DeepEraser>

Class	Category	Words	Amount
Document	Paper	>300	200
	News	>200	200
	Novel	>1000	200
Scene	Title	5-10	200
	Sign	20-30	200
	Introduction	100-200	190
Poster	Leaflet	50-60	160
	Cover	100-120	100
Total	-	-	1450

Table 3: An overview of MCiT. It is mainly divided into three classes: document, scene, and poster.

5 MCiT Benchmark

The current TiMT benchmarks fail to simultaneously challenge the OCR capabilities of MLLMs as well as their reasoning and knowledge-based translation skills. Therefore, we manually annotate an English-Chinese benchmark for TiMT called MCiT. As shown in Appendix C, the datasets encompass various real-world scenarios, including documents, scenes, and posters, which require MLLMs to possess strong OCR recognition capabilities. Moreover, the complex textual content demands advanced knowledge and reasoning abilities for accurate translation. During annotation, ten professional English and Chinese speakers manually translate the text in paragraphs to ensure semantic consistency and completeness of translation. Furthermore, ten annotators are asked to verify each other’s translation results.

5.1 Document

Document-level images have a neat layout and contain extensive text. To cover diverse semantic domains, we divide them into three categories: paper, novel, and news. For **paper**, we select approximately 50 papers from arxiv and CNKI, with 4 semantically rich fragments per paper. For **novel**, we randomly select pages from The Lord of the Rings, with each page containing at least 1,000 words. For **news**, we select fragments from the New York Times, China Daily, CNN, and CGTN websites, each containing at least 200 words.

5.2 Scene

Scene-level images exhibit complexity and irregularity due to factors such as shooting angle and pixel quality, often leading to text blurriness. We categorize the scene class into three categories: title, sign, and introduction. For **title**, we manually filter out examples such as shop, street, and

hotel from English OCR images, each containing about 5-10 words. For **sign**, we search for images with keywords like warning and notice from the Internet and filter out signs in natural scenes, each containing about 20-30 words. For **introduction**, we collect text-rich images from the web, including tourist attractions, animal descriptions, and explanations of proper nouns, each containing about 100-200 words.

5.3 Poster

Poster-level images feature a lot of abstract fonts and complex typography. We subdivide the poster class into two categories: cover and leaflet. For **cover**, we collect cover images from e-books, magazines, and newspapers, each containing about 100 words. For **leaflet**, we collect promotional leaflet images from the Internet, with each image containing about 50 words.

6 Experiment

6.1 Implementation Details

Based on Qwen-chat-7B, InImageTrans with mcDPO has a total of 8.12B parameters. We compare the proposed method with current powerful open-source and commercial MLLMs, as well as current top cascade methods such as Google Translate and Baidu Translate as baselines. See more details on baselines in Appendix A.6.

Translation Quality Evaluation. We use BLEU, METEOR, TER and COMET as the metrics for evaluating translation quality. Furthermore, we manually evaluate the completeness and semantic consistency of translation quality.

Hallucination Evaluation. To measure repetition and omission hallucinations, we manually identify cases of repetition and omission hallucinations and compute the repetition rate and omission rate. Additionally, we utilize the Repetition_4 metric (Xu et al., 2022) for automated repetition hallucination evaluation. Detailed evaluation implementation can be found in Appendix A.5.

6.2 Main Results

We conduct a comprehensive evaluation on MCiT, quantitatively comparing the performance of the proposed method with existing open-source MLLMs and analyzing the impact of mcDPO in alleviating hallucinations. The experiment results for translation quality and hallucination are shown

Method	Size	Document			Scene			Poster		Avg $\uparrow$
		Paper	News	Novel	Title	Sign	Introduction	Cover	Leaflet	
LLaVA	7B	2.9 / 3.3	3.0 / 3.9	1.7 / 2.1	0.3 / 0.5	1.9 / 2.6	4.4 / 4.8	3.1 / 3.9	5.3 / 6.1	2.7 / 3.3
LLaVA-Next	7B	4.0 / 4.6	4.4 / 5.8	2.3 / 2.7	0.6 / 1.1	3.2 / 4.5	6.7 / 8.1	3.4 / 4.2	6.5 / 7.7	3.8 / 4.8
Qwen-VL	8B	7.6 / 8.7	10.5 / 12.3	4.9 / 5.5	15.4 / 17.1	16.0 / 16.8	6.9 / 7.7	3.1 / 4.4	4.8 / 5.3	9.2 / 10.2
Qwen-VL-Chat	8B	1.1 / 1.9	0.9 / 1.5	0.6 / 1.0	0.3 / 0.9	0.2 / 0.6	0.3 / 0.5	0.2 / 0.7	0.3 / 0.5	0.5 / 1.0
InternVL2	8B	44.0 / 58.2	35.8 / 47.9	23.7 / 35.4	27.2 / 30.4	28.2 / 34.0	25.4 / 39.2	19.4 / 28.3	21.0 / 34.2	28.9 / 39.2
Qwen2-VL	8B	60.0 / 71.3	46.4 / 57.6	31.8 / 39.3	27.9 / 33.7	25.9 / 37.9	24.0 / 44.1	19.4 / 30.3	30.4 / 46.8	36.3 / 46.0
InternLM-XComposer2	11B	37.3 / 44.6	26.7 / 33.1	8.5 / 7.0	24.1 / 30.2	22.4 / 31.3	20.2 / 35.8	16.6 / 27.1	22.0 / 36.7	22.6 / 30.8
LLaVA-Next	13B	5.9 / 10.3	4.9 / 7.6	3.0 / 2.7	2.0 / 5.7	4.6 / 6.7	8.6 / 12.3	4.0 / 9.1	7.9 / 15.8	5.1 / 8.6
CogVLM	17B	59.2 / 70.6	44.8 / 56.0	30.7 / 38.5	<u>36.4 / 38.1</u>	27.7 / 37.5	<u>30.5 / 43.1</u>	25.7 / 33.2	30.1 / 46.5	<u>36.5 / 46.2</u>
InternVL2	26B	45.9 / 58.9	36.5 / 48.6	24.1 / 35.7	<u>24.7 / 31.2</u>	28.0 / 34.8	<u>26.8 / 40.1</u>	19.4 / 29.0	22.5 / 36.5	29.3 / 40.2
Yi-VL	34B	12.9 / 17.0	4.8 / 7.9	3.9 / 4.4	20.5 / 24.1	13.0 / 17.6	5.0 / 9.1	0.6 / 3.1	3.8 / 5.9	8.7 / 11.9
InternVL2	40B	48.8 / 59.8	40.1 / 49.5	26.6 / 35.5	27.1 / 31.5	<u>28.4 / 36.1</u>	28.7 / 41.2	<u>19.7 / 29.7</u>	23.4 / 39.8	31.3 / 41.0
Ours										
InImageTrans	8B	<u>59.3 / 70.8</u>	44.2 / 54.3	29.8 / 38.7	35.8 / 37.2	26.5 / 35.8	29.6 / 43.3	23.0 / 31.4	18.8 / 40.7	34.5 / 44.9
w/o RPD	8B	48.7 / 60.1	40.2 / 50.1	23.8 / 35.4	34.9 / 36.0	26.1 / 35.1	28.7 / 42.1	18.4 / 28.7	18.5 / 40.1	31.0 / 41.8
+ mcDPO	8B	59.0 / <u>70.8</u>	46.5 / 57.8	33.5 / 41.5	37.1 / 39.2	28.9 / 38.8	32.0 / 44.9	19.0 / 29.4	32.1 / 48.0	37.3 / 47.5

Table 4: Performance comparison for open-source MLLMs on MCiT. We report **BLEU/METEOR** for translation quality. **The bold** represents the best results, and the underline represents the second best results. w/o RPD denotes InImageTrans without the repetition penalty decoding (RPD) method. In addition, we report **COMET** and **TER** in Appendix A.7 to comprehensively evaluate translation quality.

Method	Document			Scene			Poster		
	BLEU $\uparrow$	Repetition $\downarrow$	Omission $\downarrow$	BLEU $\uparrow$	Repetition $\downarrow$	Omission $\downarrow$	BLEU $\uparrow$	Repetition $\downarrow$	Omission $\downarrow$
InternVL2-8B	34.5	11.4%	2.7%	26.9	2.5%	7.3%	20.2	3.1%	10.4%
Qwen2-VL-8B	46.1	3.4%	1.9%	25.9	1.5%	10.6%	24.9	2.1%	9.7%
InImageTrans	44.4	5.4%	2.3%	30.6	0.7%	8.1%	20.4	1.5%	14.3%
w/o RPD	37.6	9.5%	2.3%	29.8	1.8%	8.5%	18.4	2.3%	14.9%
+ mcDPO	46.3	1.5%	2.3%	32.7	0.5%	5.4%	27.1	1.3%	6.7%

Table 5: Overall performance comparison of translation quality, repetition, and omission hallucinations. We report average BLEU scores of each categories for translation quality, as well as repetition and omission rates to evaluate hallucinations hallucinations relief. w/o RPD denotes InImageTrans without the repetition penalty decoding method.

Method	Document	Scene	Poster
InternVL2-8B	7.9%	2.0%	3.0%
Qwen2-VL-8B	4.3%	1.5%	2.1%
InImageTrans	5.5%	1.6%	1.6%
+mcDPO	2.6%	0.7%	1.2%

Table 6: Performance comparison of Repetition_4 metric for open-source MLLMs.

in Table 4, Table 5 and Table 6. Furthermore, we also provide a detailed comparison in Appendix A.8 between the proposed method and commercial MLLMs such as GPT-4o (Hurst et al., 2024) and Qwen-VL-Max (Bai et al., 2023), as well as commercial cascaded methods such as Google and Baidu, in terms of performance on MCiT. In addition, we show some visualization results in Appendix D, intuitively demonstrating the advantages of our proposed method in terms of translation quality and hallucination reduction.

Overall Results. For translation quality, compared with existing open-source MLLMs such as InternVL2 and Qwen2-VL, the proposed method

achieves state-of-the-art performances across all three classes on MCiT, demonstrating a significant advantage in TiMT in Table 4. Besides, mcDPO significantly improves the translation quality of InImageTrans, which can be attributed to the reduction of repetition and omission hallucinations by mcDPO, as shown in Table 5.

Results for Specific Classes. As shown in Table 4, compared with another two classes, the translation quality of most MLLMs on document-level images is relatively high, which may be due to the BLEU metric tending to yield higher scores for longer texts. In Table 5, the proposed method achieves a significant improvement by a larger margin in translation quality on scene-level and poster-level images compared with the best open-source MLLM, Qwen2-VL-8B (Wang et al., 2024b). This means that omission hallucination is more severe in these two scenarios, reflecting the substantial advantage of the proposed method in mitigating hallucinations.

Results for Automated Repetition Hallucina-

tion Evaluation. We also report the Repetition_4 scores as automated repetitive hallucination evaluation metric. The results are shown in Table 6. The experimental results demonstrate that the proposed method achieves the best performance under the Repetition_4 metric, with results closely aligning with our manual evaluations. This further validates that the metric can be effectively utilized for hallucination assessments in future evaluations.

6.3 Human Evaluation

To evaluate the completeness and semantic consistency of translation quality based on human preference, we randomly select 50 images from each of the three classes in MCiT, totaling 150 images. The translation results from GPT-4o (Hurst et al., 2024), InternVL2-8B (Chen et al., 2024), Qwen2-VL-8B (Wang et al., 2024b), and InImageTrans combined with mcDPO are assessed. Each example is scored according to our evaluation criteria by professional English and Chinese speakers, and the detailed evaluation criteria can be found in Appendix A.5. As shown in Figure 5, for document-level images, our method significantly outperforms InternVL2 in translation consistency and achieves comparable results to Qwen2-VL. For scene and poster images, our method surpasses both InternVL2 and Qwen2-VL in terms of translation completeness.

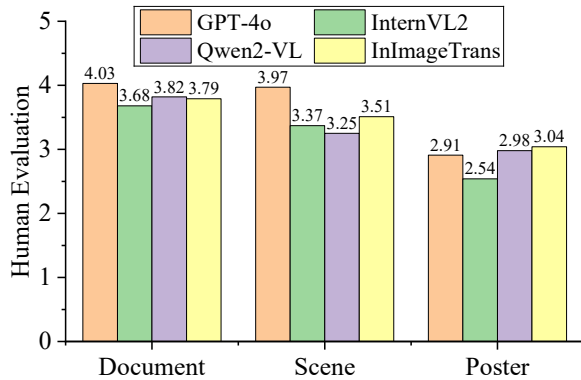


Figure 5: Overall human evaluation results of translation performance for different methods.

6.4 Ablation Study

To verify the effectiveness of the optimization process in improving the translation quality of InImageTrans, we conduct an ablation study on dual-encoder framework and mcDPO. As shown in Table 4, removing RPD results in a significant decline in translation quality and increase in repetition hallucination for InImageTrans across

Method	Document	Scene	Poster
dual-encoder	44.4	30.6	20.4
↪w/o fine-grained	40.8	25.1	15.1
↪w/o coarse-grained	41.1	26.1	15.4
mcDPO	46.3	32.7	27.1
↪w/o rPO	44.4	32.2	27.4
↪w/o vPO	46.1	29.9	20.1

Table 7: Ablation study of fine-grained encoder and coarse-grained encoder in dual-encoder framework, and rPO and vPO in mcDPO.

all tasks, particularly for the document-level TiMT task. As shown in Table 7, removing rPO, which is designed to mitigate repetition hallucination, results in a consistent decline in translation quality compared with mcDPO. This highlights the effectiveness of the proposed method. Since vPO focuses on enhancing the model’s attention to text in the image and mitigating omission hallucinations, removing vPO leads to a significant decline in translation quality, particularly in scenarios with severe omission hallucinations, such as scene and poster class. This demonstrates the effectiveness of the proposed component in improving translation quality. Furthermore, the experimental results show that removing any encoder will have a significant impact on performance, demonstrating the effectiveness of the dual-encoder framework.

7 Discussion

7.1 Why rPO Can Relieve Repetition?

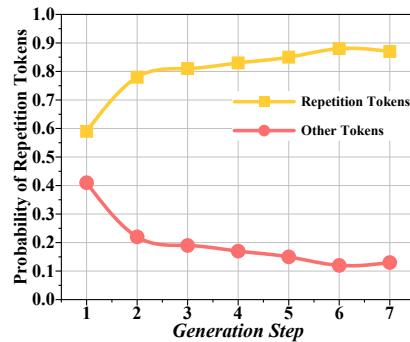


Figure 6: The probability of repetition tokens and other tokens with the generation step increases.

First, we analyze the reasons for repetition. Specifically, we select 50 repetition examples and calculate the probability of repetition tokens using InImageTrans without mcDPO. As shown in Figure 6. The results show that as the number of repetition

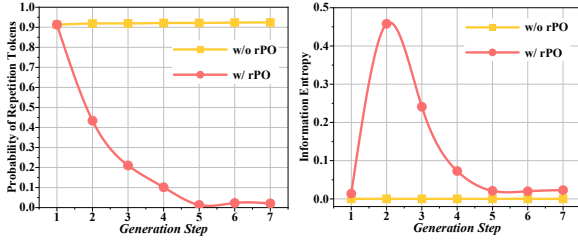


Figure 7: **The left** represents probabilities of repetition tokens and **the right** represents information entropy of the model output across generation steps. We report the average scores among 20 selected examples whose repetition hallucinations are resolved by rPO.

increases, the probability of generating repetition tokens also increases, meaning that the confidence continues to get higher. See Appendix B for details.

According to Equ (5), theoretically, there are two optimization directions for rPO: increasing the probability of the golden labels or decreasing the probability of the repetition tokens. To verify it, we select 20 examples whose repetition hallucinations are solved by rPO and measure the output probabilities of the repetition token and the information entropy of the model output at seven generation steps. As shown in Figure 7, removing rPO causes repetition tokens to be generated with high probabilities and low information entropy, indicating the model’s confidence in generating repetition tokens. Using rPO, we observe that the generation probability of repetition tokens drops sharply. Besides, the initial increase in information entropy indicates that the model reduces the confidence of repetition tokens while increasing the confidence of other tokens. Subsequently, the decrease in information entropy suggests that the model has started to confidently generate the correct tokens. Finally, the above experiments confirm that the proposed rPO method effectively avoids repetition hallucination by dynamically adjusting the confidence of output tokens.

7.2 Effectiveness and Robustness of rPO

We measure the repetition rates using rPO and vPO for different document-level images. As shown in Table 8, rPO significantly reduces the repetition rates, particularly for the novel classes where repetition hallucinations are most pronounced, which highlights its effectiveness on repetition relief. However, adding vPO does not lead to better results, suggesting that vPO has limited effectiveness in reducing repetition hallucinations.

Method	Paper	News	Novel	Avg
InImageTrans	2.5%	4.0%	9.5%	5.4%
+ rPO	0.5%	0.5%	3.0%	1.3%
+ rPO&vPO	0.5%	0.5%	3.5%	1.5%

Table 8: Comparison of the repetition rates using different methods for different document-level images.

To illustrate the robustness of rPO, we construct reject labels in a controllable way to compare with random ones. Specifically, we choose data with high word frequency from the fine-tune data, select the position of its last occurrence as the repetition outset, and repeat the segment to the max length. All experimental settings remain unchanged. As shown in Table 9. The results show that the controllable and random construction have little gap on performance, which further demonstrate that our method is highly robust.

Method	Paper	News	Novel	Avg
Random	59.0	46.5	33.5	46.3
Control	59.3	46.3	33.1	46.2

Table 9: Comparison of different data construction strategies of rPO in the document scenarios.

7.3 Is It Necessary To Mask Text in vPO?

To evaluate the effectiveness of the masking strategy in vPO, we compare three masking strategies: **text**, which masks about 20% of the text in the image, **no text**, which masks about 20% of the no-text area in the image and **random**, which randomly mask about 20% of the area in the image.

As shown in Table 10, the mask strategy of **text** achieves the best performance, significantly better than **random** and **no text**, demonstrating that the mask strategy of **text** is the key for vPO to improve the performance of scene and poster.

Mask Strategy	Document	Scene	Poster
Text	46.3	32.7	27.1
Random	46.0	31.6	23.8
No-text	45.9	30.3	20.7

Table 10: Performance comparison of different mask strategies of vPO. Text means to mask the text area, no-text means to mask the no-text area, and random means to mask the random area.

8 Conclusion

In this paper, we investigate the severe repetition and omission hallucinations for existing MLLMs

on TiMT task. Then we design an efficient MLLM named InImageTrans specially for TiMT and propose a multi-conditional direct preference optimization (mcDPO) approach for advancing the TiMT to mitigate hallucinations and improve translation quality. Furthermore, we build a high-quality benchmark named MCiT for effectively evaluating the TiMT capabilities of MLLMs. Experimental results show that the proposed method significantly outperforms existing open-source MLLMs in terms of both translation quality and hallucination mitigation and approaches the performance of proprietary MLLMs.

Limitations

In this paper, combined with mcDPO, InImageTrans demonstrates excellent performance in translation quality and hallucination mitigation, while being adaptable to various scenarios. However, due to the lack of domain-specific knowledge, it struggles with omission hallucination issues in certain specialized document translation tasks. This highlights the need for further knowledge enhancement to generalize across more domains.

Acknowledgements

We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions. The work was supported by the National Natural Science Foundation of China under Grant (62276077), Guangdong Basic and Applied Basic Research Foundation (2024A1515011205), and Shenzhen Science and Technology Program (GXWD20220811170358002 and ZDSYS20230626091203008).

References

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.
- Zhuo Chen, Fei Yin, Xu-Yao Zhang, Qing Yang, and Chena-Lin Liu. 2021. Cross-lingual text image recognition via multi-task sequence to sequence learning. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3122–3129. IEEE.
- Xuxin Cheng, Ziyu Yao, Yifei Xin, Hao An, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Soul-mix: Enhancing multimodal machine translation with manifold mixup. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11283–11294.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6.
- Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*.
- Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.
- Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.
- Hao Feng, Wendi Wang, Shaokai Liu, Jiajun Deng, Wengang Zhou, and Houqiang Li. 2024a. Deeperaser: Deep iterative context mining for generic text eraser. *arXiv preprint arXiv:2402.19108*.
- Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024b. Tear: Improving llm-based machine translation with systematic self-refinement. *arXiv preprint arXiv:2402.16379*.

- Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12848–12856.
- Jiwoo Hong, Noah Lee, and James Thorne. 2024a. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.
- Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024b. CogVLM2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.
- Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.
- Jungeun Kim, Hyeonwoo Jeon, Jongseong Bae, and Ha Young Kim. 2024. Leveraging the power of mllms for gloss-free sign language translation. *arXiv preprint arXiv:2411.16789*.
- 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. *arXiv preprint arXiv:2305.17415*.
- Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. 2022. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*.
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.
- Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. 2023a. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective. *Advances in Neural Information Processing Systems*, 36:72888–72903.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023c. Silk: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.
- Philip Lippmann, Konrad Skublicki, Joshua Tanner, Shonosuke Ishiwatari, and Jie Yang. 2024. Context-informed machine translation of manga using multimodal large language models. *arXiv preprint arXiv:2411.02589*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024b. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.
- I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- R OpenAI. 2023. Gpt-4 technical report. *arxiv* 2303.08774. *View in Article*, 2(5).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Jaeyoo Park, Jin Young Choi, Jeonghyung Park, and Bohyung Han. 2024. Hierarchical visual feature aggregation for ocr-free document understanding. *arXiv preprint arXiv:2411.05254*.
- Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

- Thevin Senath, Kumuthu Athukorala, Ransika Costa, Surangika Ranathunga, and Rishemjit Kaur. 2024. Large language models for ingredient substitution in food recipes using supervised fine-tuning and direct preference optimization. *arXiv preprint arXiv:2412.04922*.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.
- Tonghua Su, Shuchen Liu, and Shengjie Zhou. 2021. Rtnet: An end-to-end method for handwritten text image translation. In *Document Analysis and Recognition-ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 99–113. Springer.
- Yihong Tang, Jiao Ou, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Erabal: Enhancing role-playing agents through boundary-aware learning. *arXiv preprint arXiv:2409.14710*.
- Turghun Tayir and Lin Li. 2024. Unsupervised multimodal machine translation for low-resource distant language pairs. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(4):1–22.
- Turghun Tayir, Lin Li, Bei Li, Jianquan Liu, and Kong Aik Lee. 2024. Encoder-decoder calibration for multimodal machine translation. *IEEE Transactions on Artificial Intelligence*.
- Erlin Tian, Zengchao Zhu, Fangmei Liu, Zuhe Li, Ran Gu, and Shuai Zhao. 2024. Multimodal machine translation based on enhanced knowledge distillation and feature fusion. *Electronics*, 13(15):3084.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wei Han Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.
- Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2023a. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109*.
- Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. 2023b. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*.
- Lai Wei, Zhiqian Tan, Chenghai Li, Jindong Wang, and Weiran Huang. 2024. Diff-erank: A novel rank-based metric for evaluating large language models. *arXiv preprint arXiv:2401.17139*.
- Wenxi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.
- Jin Xu, Xiaojian Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. *Advances in Neural Information Processing Systems*, 35:3082–3095.
- Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. RLHF-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.
- Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. RLHF-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.
- Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. 2024c. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *arXiv preprint arXiv:2404.09204*.

Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. 2024d. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *arXiv preprint arXiv:2410.05261*.

Piotr Żelasko, Zhehuai Chen, Mengru Wang, Daniel Galvez, Oleksii Hrinchuk, Shuoyang Ding, Ke Hu, Jagadeesh Balam, Vitaly Lavrukhin, and Boris Ginsburg. 2024. Emmett: Efficient multimodal machine translation training. *arXiv preprint arXiv:2409.13523*.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José GC De Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, et al. 2022. Findings of the wmt 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. 2024a. Paying more attention to source context: Mitigating unfaithful translations from large language model. *arXiv preprint arXiv:2406.07036*.

Xiaoying Zhang, Baolin Peng, Kun Li, Jingyan Zhou, and Helen Meng. 2023. Sgp-tod: Building task bots effortlessly via schema-guided llm prompting. *arXiv preprint arXiv:2305.09067*.

Xiaoying Zhang, Da Peng, Yipeng Zhang, Zonghao Guo, Chengyue Wu, Chi Chen, Wei Ke, Helen Meng, and Maosong Sun. 2025. Will pre-training ever end? a first step toward next-generation foundation mllms via self-improving systematic cognition. *arXiv preprint arXiv:2503.12303*.

Yu Zhang, Kehai Chen, Xuefeng Bai, Qianjiang Guo, Min Zhang, et al. 2024b. Question-guided knowledge graph re-scoring and injection for knowledge graph question answering. *arXiv preprint arXiv:2410.01401*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13433–13447.

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2024. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. *arXiv preprint arXiv:2412.13540*.

Stage	Data	Amount
Pre-training	OCR-en (Wei et al., 2023a)	290,000
	OCR-zh (Wei et al., 2023a)	290,000
	Synthdog-en	120,000
	WMT22-en-zh	500,000
Fine-tuning	Trans-en-zh	290,000
	Synthdog-en-zh	60,000

Table 11: The details of pre-training and fine-tuning data for InImageTrans.

A More Experiment Details

A.1 Datasets

Pre-training. The OCR data consists of natural data from OCR-zh and OCR-en (Wei et al., 2023a), and the other part is synthdog-en, which is made by synthetic data using synthdog⁴. The machine translation data comes from WMT22 (Zerva et al., 2022). See Table 11 for detailed data volume.

Fine-tune. Our fine-tuning data mainly consists of two parts: trans-en-zh and synthdog-en-zh. For trans-en-zh, we use the in-house translation model to translate the Chinese labels in OCR-en into Chinese and filter out poor-quality data. For synthdog-en-zh, we improve synthdog to generate a diversity of images with continuous-semantics English text and their Chinese translations.

mcDPO. We choose 10,000 data from our fine-tuning datasets and construct 10,000 preference data pairs according to the method in mcDPO.

A.2 Model Configuration

In our experiments, we utilize the well-trained CLIP-vit-large-patch14 (Radford et al., 2021) and Qwen-chat (Bai et al., 2023) to initialize coarse-grained encoder and LLM. In addition, we use the vocabulary network module in Vary (Wei et al., 2023a) to initialize the fine-grained encoder. The two MLP layers are randomly initialized before training. InImageTrans consists of a LLM with 7.7B parameters, a coarse-grained encoder with 0.3B parameters, a fine-grained encoder with 80M parameters, and two MLP layers. Overall, InImageTrans has a total of 8.12B parameters.

A.3 Training Hyperparameters

Pre-training. During the pre-training phase, we use the AdamW (Loshchilov, 2017) optimizer with a learning rate of 5e-5 and a cosine learning rate

⁴<https://github.com/clovaai/donut/tree/master/synthdog>

schedule. A warmup ratio of 0.03 is incorporated, and we process the data in batches of 256. The entire training process is completed on $8 \times A100$ GPUs, and takes 5 days to complete 3 epochs.

Fine-tuning. During the fine-tuning phase, we retain most of the pre-training hyper-parameters, except for changing the learning rate to $2e-5$ and setting the batch size to 32. The entire fine-tuning process takes 3 days to complete 1.5 epochs on $8 \times A100$ GPUs.

mcDPO. During the mcDPO phase, we set the hyper-parameter β in the mcDPO optimization objective to 0.1 and adjust the batch size to 8. The entire mcDPO process took 4 hours to complete 1 epoch on $8 \times A100$ GPUs.

A.4 Training Loss

To verify the effectiveness of the proposed method, we demonstrate the convergence of the model during the training process, as shown in Figure 8. The results indicate that the model converges well under the mcDPO optimization objective, fully demonstrating its reliability.

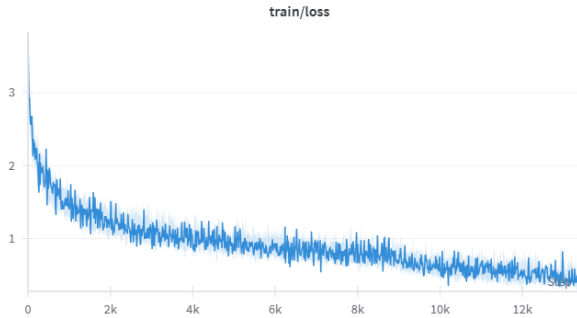


Figure 8: Training loss of mcDPO training process.

A.5 Evaluation Details

Translation Quality We use sacreBLEU⁵ (Post, 2018) to calculate BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) scores for evaluating translation quality. Furthermore, we use evaluate⁶ to calculate METEOR (Banerjee and Lavie, 2005) scores and use Unbabel-comet⁷ to calculate COMET (Rei et al., 2020) scores.

For human preference-based translation quality evaluation, we randomly select 50 images from each of the three classes in MCiT, totaling 150 images. For human evaluation, the annotators are provided with the original image alongside

the translation outputs from GPT-4o, Qwen2-VL, InternVL-2 and our method. They then score each translation based on the following predetermined criteria: **0-1 point:** no answer. **1-2 point:** The text in the image can be recognized but can not be translated. **2-3 point:** The text in the image can be translated but there are obvious omission or repetition hallucinations. **3-4 point:** Most of the text in the image can be translated and there are no obvious omission or repetition hallucinations. **4-5 point:** There is no repetition or omission hallucinations and the translation is smooth and fluent, close to human translation.

Human Hallucination Evaluation. To measure the repetition and omission hallucinations, we introduce the repetition rate and omission rate, which compute the percentage of repetition and omission cases. We first identify examples with output lengths far exceeding the reference length and ten consecutive repetitions in the output as repetition candidate examples, and identify examples with output lengths far less than the reference length as omission candidate examples. Then, ten bilingual speakers are asked to compare the candidate examples and corresponding references to determine.

Repetition_4 metric for Automated Repetition Hallucination Evaluation. To provide an automated hallucination evaluation, we introduce the Repetition_4 scores, which is formulated as:

$$\text{Repetition_4} = 1.0 - \frac{|\text{unique}(4_gram)|}{|4_gram|}. \quad (8)$$

4_gram denotes four consecutive characters, and **unique(4_grams)** denotes four consecutive characters that have not been repeated.

A.6 Baselines

In the main results, we compare with 12 existing open-source MLLMs, LLaVA-7B (Liu et al., 2024a), LLaVA-Next-7B, LLaVA-Next-13B (Li et al., 2024), Qwen-VL, Qwen-VL-chat (Bai et al., 2023), InternVL2-8B, InternVL2-26B, InternVL2-40B (Chen et al., 2024), Qwen2-VL-8B (Wang et al., 2024b), InternLM-XComposer2-11B (Dong et al., 2024), CogVLM-17B (Wang et al., 2023), Yi-VL-34B (Young et al., 2024). Furthermore, we compare with top commercial MLLMs such as GPT-4o (Hurst et al., 2024), Qwen-VL-Max (Bai et al., 2023), and commercial cascade methods such

⁵<https://github.com/mjpost/sacrebleu>

⁶<https://huggingface.co/evaluate>

⁷<https://github.com/Unbabel/COMET>

Method	Size	Document			Scene			Poster		Avg $\uparrow$
		Paper	News	Novel	Title	Sign	Introduction	Cover	Leaflet	
InternVL2 (Chen et al., 2024)	8B	83.0	78.9	71.3	65.3	61.1	68.7	62.0	62.1	69.7
Qwen2-VL (Wang et al., 2024b)	8B	85.6	81.7	73.9	67.8	66.4	70.9	66.4	70.3	73.3
InternLM-XComposer2 (Dong et al., 2024)	11B	78.4	60.1	45.2	60.2	60.0	61.1	57.3	60.0	60.4
LLaVA-Next (Li et al., 2024)	13B	60.5	54.4	41.4	45.7	42.3	43.4	46.4	45.7	47.6
CogVLM (Wang et al., 2023)	17B	84.7	80.6	73.7	71.1	<u>66.7</u>	70.6	67.8	71.8	73.7
InternVL2 (Chen et al., 2024)	26B	83.4	79.3	71.7	65.7	61.8	69.2	62.9	64.7	70.4
Yi-VL (Young et al., 2024)	34B	72.0	58.7	47.4	61.2	56.1	49.7	37.3	38.1	54.0
InternVL2 (Chen et al., 2024)	40B	83.9	79.8	72.8	67.1	62.5	69.9	64.1	66.1	71.3
Ours										
InImageTrans	8B	84.3	77.6	73.1	68.9	64.2	69.5	64.7	66.0	71.6
w/o RPD	8B	80.1	74.1	70.0	66.3	63.4	68.1	61.1	65.1	69.4
+ mcDPO	8B	<u>85.0</u>	<u>80.9</u>	76.5	<u>70.6</u>	66.9	71.7	62.3	<u>70.8</u>	73.9

Table 12: Performance comparison of **COMET** for open-source MLLMs on MCiT. **The bold** represents the best results, and the underline represents the second best results. w/o RPD denotes InImageTrans without the repetition penalty decoding (RPD) method.

Method	Size	Document			Scene			Poster		Avg $\downarrow$
		Paper	News	Novel	Title	Sign	Introduction	Cover	Leaflet	
InternVL2 (Chen et al., 2024)	8B	110.8	113.7	197.3	108.3	114.5	132.0	122.1	115.7	127.4
Qwen2-VL (Wang et al., 2024b)	8B	100.7	116.1	188.4	<u>105.1</u>	128.0	137.7	124.1	123.4	128.3
InternLM-XComposer2 (Dong et al., 2024)	11B	113.5	119.8	200.1	130.9	143.7	153.2	143.7	140.4	143.1
LLaVA-Next (Li et al., 2024)	13B	127.8	130.5	220.7	135.7	150.5	158.9	147.7	145.8	153.7
CogVLM (Wang et al., 2023)	17B	108.1	123.7	190.7	106.5	138.3	140.9	135.9	136.2	134.9
InternVL2 (Chen et al., 2024)	26B	110.1	112.5	196.3	107.2	114.3	132.2	122.7	<u>114.5</u>	126.8
Yi-VL (Young et al., 2024)	34B	123.5	124.7	203.5	130.7	144.6	157.8	149.8	146.7	146.7
InternVL2 (Chen et al., 2024)	40B	108.1	<u>111.7</u>	193.5	107.4	<u>113.7</u>	<u>131.5</u>	122.0	114.7	126.0
Ours										
InImageTrans	8B	110.3	115.8	184.7	106.1	114.5	131.8	113.4	115.1	124.9
w/o RPD	8B	115.7	118.1	191.1	110.8	117.3	135.8	116.6	118.8	129.0
+ mcDPO	8B	<u>106.4</u>	111.3	173.6	104.2	107.5	126.1	<u>115.7</u>	112.4	119.8

Table 13: Performance comparison of **TER** for open-source MLLMs on MCiT. **The bold** represents the best results, and the underline represents the second best results. w/o RPD denotes InImageTrans without the repetition penalty decoding (RPD) method.

as Google Translate⁸ and Baidu Translate⁹.

A.7 More Results of Other Metrics

In order to comprehensively evaluate the quality of translation, we also report the evaluation results of METEOR and TER, as shown in Table 12 and Table 13.

Comparison of COMET. Our method has shown best results in many scenarios, such as document and scene scenarios, except for slightly inferior performance on the cover class in the poster scenario compared to Qwen2-VL and CogVLM. On average, our method also outperforms other open-source MLLMs, further demonstrating its superiority for translation quality.

Comparison of TER. Our method also has better results in many scenarios, especially in

the novel class in the paper, where our method outperforms other models by about 20 points. On average, our method significantly outperforms all other open-source MLLMs, indicating that our model has more accurate translations and lower error rates.

A.8 Comparison with More Methods

Comparison with Commercial MLLMs. In addition, we compare our method with advanced commercial MLLMs such as GPT-4o and Qwen-VL-Max, as shown in Table 14.

Qwen-VL-max demonstrates better performance compared to GPT-4o, which can be attributed to its extensive training on amounts of high quality Chinese data. Besides, the proposed method still has a considerable gap compared to commercial MLLMs. However, for the paper, cover and leaflet subclasses, the proposed method either closely

⁸<https://translate.google.com>

⁹<https://fanyi.baidu.com>

Method	Size	Document			Scene			Poster		Avg↑
		Paper	News	Novel	Title	Sign	Introduction	Cover	Leaflet	
Commercial MLLMs										
GPT-4o	-	60.5	56.4	38.9	41.2	38.4	35.2	12.0	19.6	40.2
Qwen-VL-Max	-	63.3	59.6	39.7	41.0	39.8	36.1	32.3	40.4	45.0
Commercial Cascaded Method										
Google	-	61.7	58.9	38.1	30.2	30.8	34.6	30.4	31.3	41.5
Baidu	-	60.2	54.4	35.7	31.3	30.7	32.8	29.5	31.0	39.5
Ours										
InImageTrans	8B	59.3	44.2	29.8	35.8	26.5	29.6	23.0	18.8	34.5
+ mcDPO	8B	59.0	46.5	33.5	37.1	28.9	32.0	19.0	32.1	37.3

Table 14: Performance comparison with commercial cascaded method such as Google and commercial MLLMs such as GPT-4o on MCiT. We report **BLEU** for translation quality.

SFT	mcDPO	Document	Scene	Poster
Qwen2VL-8B				
×	×	46.1	25.9	24.9
✓	×	45.3	28.1	23.7
×	✓	46.0	27.9	26.2
✓	✓	45.0	28.9	28.1
InternVL2-8B				
×	×	34.5	26.9	20.2
✓	×	41.3	29.3	20.0
×	✓	36.1	28.1	22.1
✓	✓	42.1	30.3	26.9

Table 15: Performance comparison of the proposed SFT and mcDPO on other MLLMs.

matches or significantly outperforms GPT-4o. This is because paper contains more formal language, which the model understands better compared to the informal content found in news and novel. The cover and leaflet classes, with their complex layouts, indicate that our method performs well in recognizing intricate layouts. Finally, for scenes scenarios, where it is crucial to identify key text in the image while filtering out other distracting factors, the proposed method performs worse compared to commercial MLLMs.

Comparison with Cascaded Methods. As shown in Table 14, compared with commercial cascaded methods, the proposed method outperforms them in the scene scenarios and has comparable performance with them in the poster scenarios, demonstrating that the proposed method is more resilient to the interference caused by complex paragraph merging. For the document scenario,

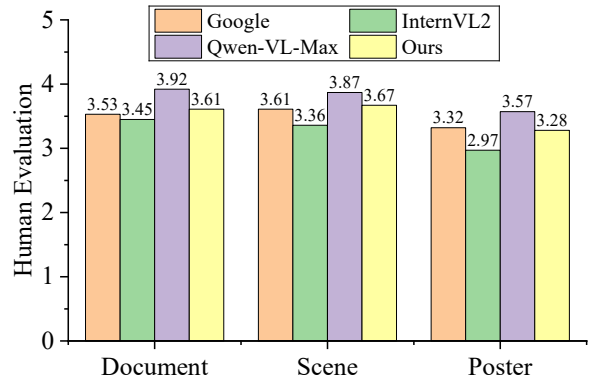


Figure 9: Overall human evaluation results of paragraph merging performance for different methods.

our method has comparable performance to them in the paper class, but due to the lack of training data, our proposed method performs less well in the news and novel classes with more informal words.

Furthermore, regarding the issue of paragraph merging, we choose to manually evaluate the paragraph merging of Google Translate, Qwen-VL-Max, InternVL2 and ours. We select a total of 150 examples from document, scene, and poster that require more paragraph merging, and 10 English-Chinese bilingual speakers score according to the following standards: **0-1 points**: No translation results. **1-2 points**: Completely translated line by line. **2-3 points**: Less than half of the paragraphs are merged. **3-4 points**: More than half of the paragraphs are merged. **4-5 points**: All paragraphs are merged correctly. As shown in Figure 9. The experimental results show that our method surpasses advanced open-source MLLMs such as InternVL2 and is on par with Google Translate, indicating the superiority of our method

in paragraph merging.

Comparison with stronger baselines. To further validate the effectiveness of SFT and mcDPO, we conduct SFT and mcDPO on Qwen2-VL and InternVL2, as shown in Figure 15. The experimental results show that SFT improves performance for average MLLMs like InternVL2 but offers limited gains for strong MLLMs like Qwen2-VL. In addition, mcDPO consistently enhances both MLLMs, particularly in scene and poster scenarios, demonstrating its effectiveness against repetition and omission hallucinations. As for document, mcDPO significantly helps InternVL2 but not Qwen2-VL, as the latter already handles repetition well.

B Analysis of the Reason for Repetition

Understanding of input images and prompts.

To analyze the ability to understand input images, we randomly select 100 examples from the MCiT benchmark and use InImageTrans and the base model, Qwen-VL for OCR tasks, using text recognition rate as the evaluation metric. We find that InImageTrans improve the accuracy from **67%** to **90%** compared to the base model, demonstrating that the proposed InImageTrans framework has significantly improved the accuracy of text recognition in images. For the ability to understand prompts, we find that InImageTrans has strong instruction following ability and excellent understanding of prompts after carefully examining the output.

Self-Reinforcing Effect. As for the specific reasons for repetition, we agree with the viewpoint of (Xu et al., 2022) that repetition has a self reinforcing effect, which means that the more repetitions there are, the higher the confidence in generating repetition fragments. To demonstrate this, we select 50 repetition examples and calculate the probability of repetition tokens using InImageTrans without mcDPO. As shown in Figure 6. The experimental results show that as the number of repetition increases, the probability of generating repetition tokens also increases, meaning that the confidence continues to get higher. This indicates that self-reinforcement effect leads to repetition hallucinations.

C Examples of MCiT

In order to more intuitively demonstrate the difference between MCiT and the benchmarks in

previous works, we list here various scenario and types of image examples in MCiT, as shown in Figure 10, 11, and 12. The document class in Figure 10 has a large amount of text, the scene class in Figure 11 has complex scenarios, and the poster class in Figure 12 has abstract text and complex typesetting, which makes MCiT to evaluate text image machine translation capability more comprehensively.

D Visualization Results of Our Model

In order to more intuitively demonstrate the translation capability of our model for different scenarios, we show some examples of different scenarios, as shown in Figures 13, 14, and 15. Figure 13 shows the performance of our model in the document class. Our model basically maintains the layout in the image while maintaining the fluency of the translation. Figure 14 shows the performance of our model in the scene class. Our model has good semantic smoothness during translation. Figure 15 shows the performance of our model in the poster class. Our model also has good recognition and translation performance for abstract text.

⁹<https://www.rogue.com.cn>

¹⁰<https://www.vogue.com>



Figure 11: Some examples of scene images in MCiT. **The upper left** is examples of sign, **the upper right** is examples of introduction, and **the bottom** is examples of title.

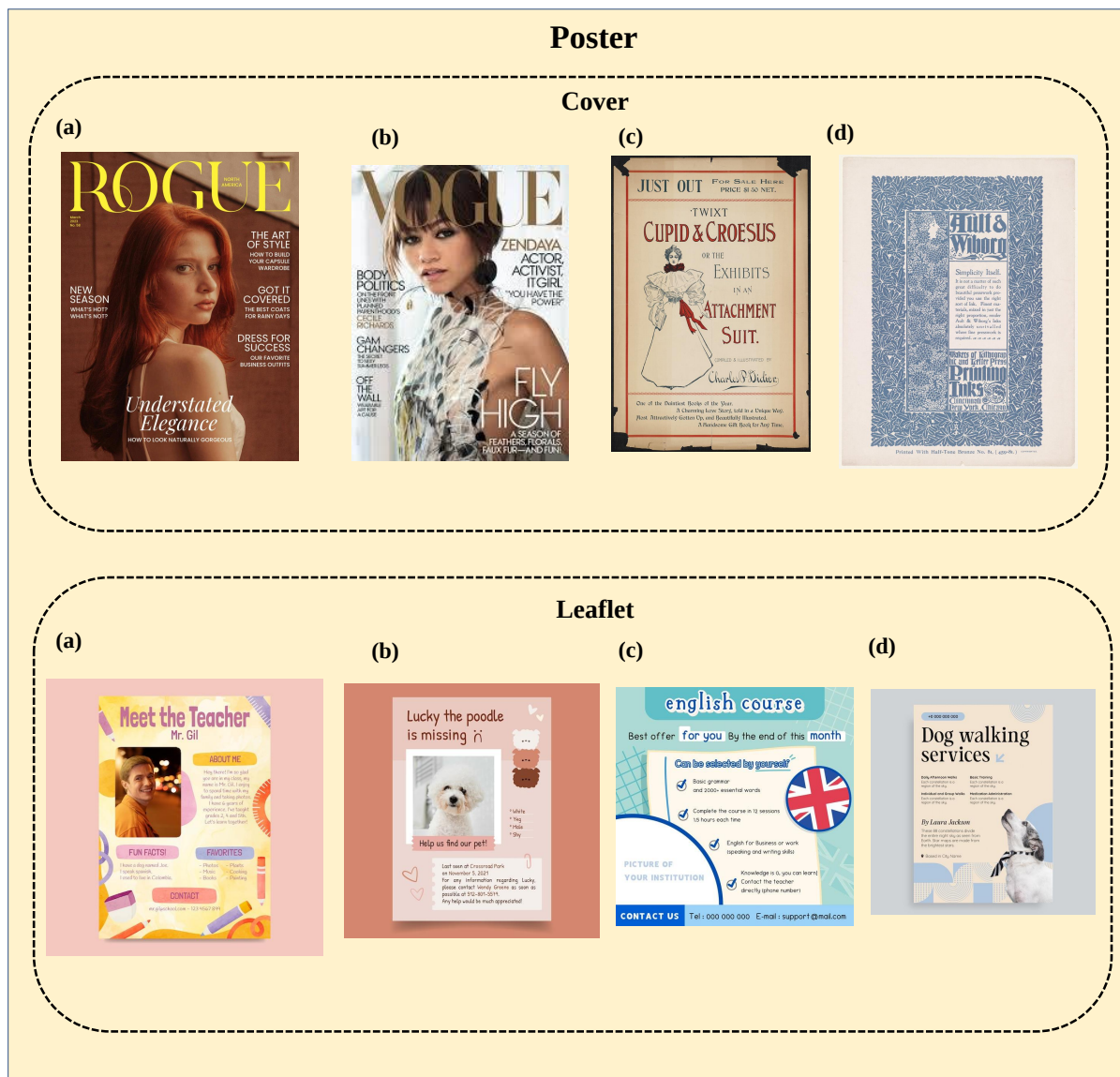


Figure 12: Some examples of poster images in MCiT. **The top** is examples of cover, and **the bottom** is examples of leaflet. It should be noted that figure (a) in the cover is from ROGUE⁹ magazine, and figure (b) is from VOGUE¹⁰ magazine.

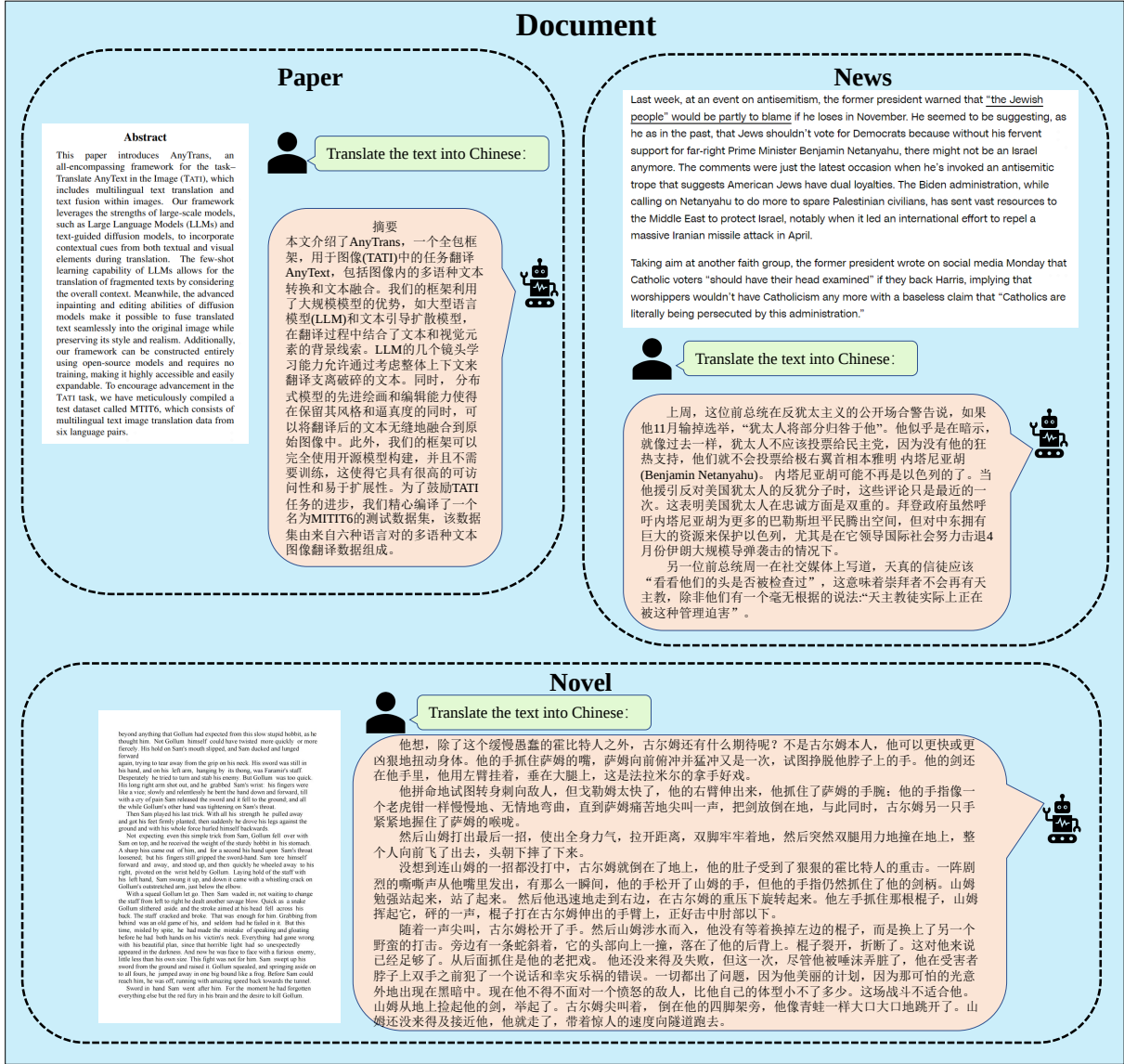
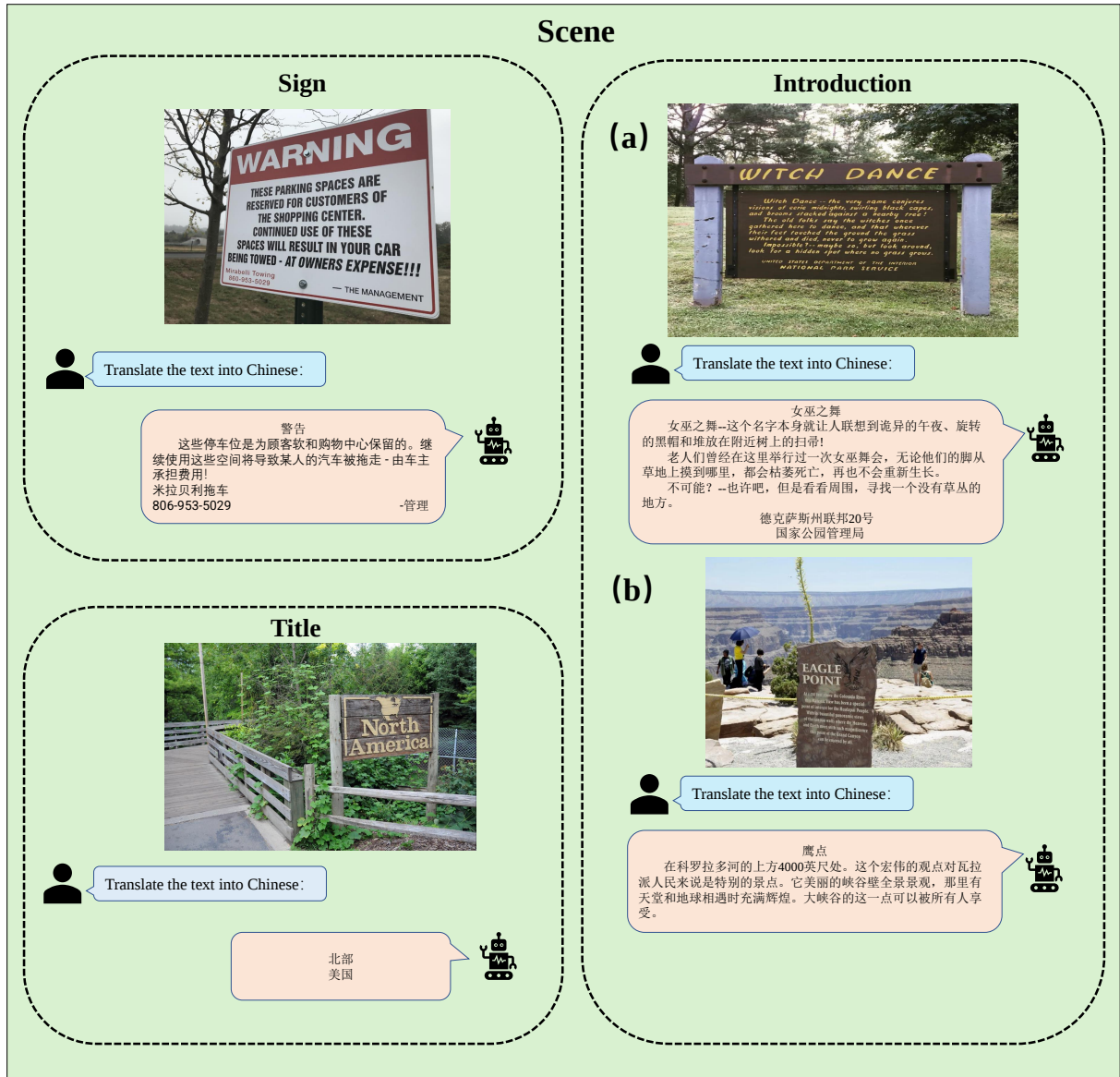


Figure 13: Some visualization results of our model on document images. **The top left** is the result for paper, **the top right** is the result for news, and **the bottom** is the result for novel.



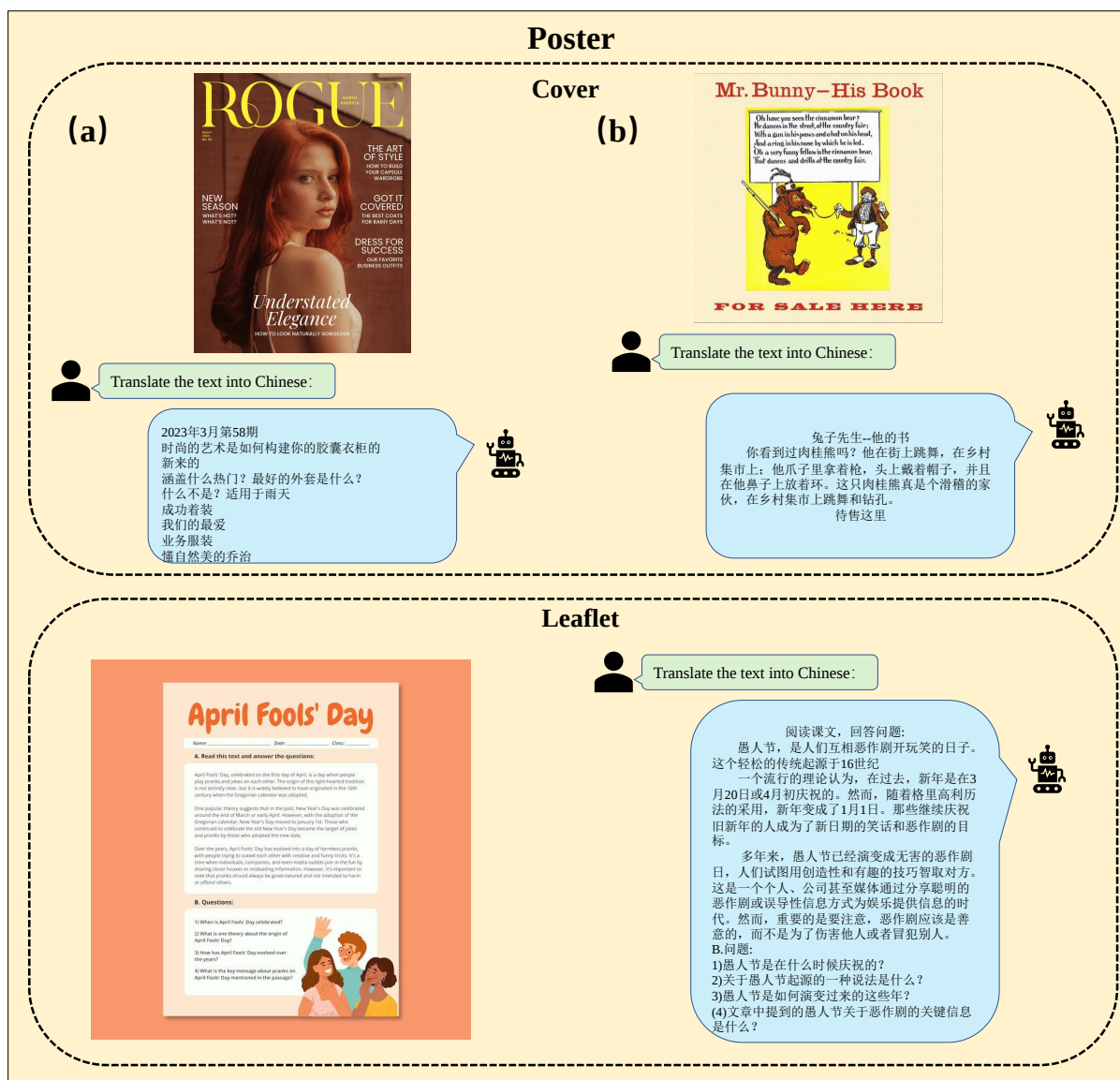


Figure 15: Some visualizations of our model on the poster images. **The top** is the cover result, **the bottom** is the leaflet result. Figure (a) in the cover is from ROGUE⁹ magazine.