InImageTrans: Multimodal LLM-based Text Image Machine Translation

Fei Zuo[1], **Kehai Chen**[1*], **Yu Zhang**[1], **Zhengshan Xue**[2], **Min Zhang**[1]
[1]Institute 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 impressive performance in several practical real-world MMT scenarios (Lippmann et al., 2024; Żelasko et al., 2024; Kim et al., 2024).

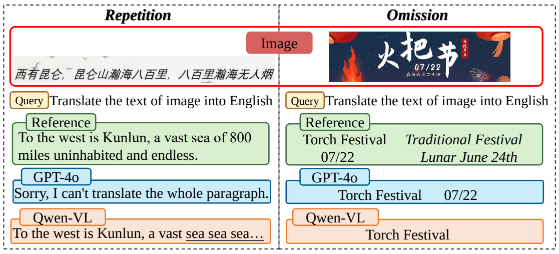

Repetition
Omission
Image
西有昆仑，昆仑山瀚海八百里，八百里瀚海无人烟
火把节
07/22
Query
Translate the text of image into English
Reference
To the west is Kunlun, a vast sea of 800 miles uninhabited and endless.
GPT-4o
Sorry, I can't translate the whole paragraph.
Qwen-VL
To the west is Kunlun, a vast sea sea sea…
Query
Translate the text of image into English
Reference
Torch Festival
07/22
Traditional Festival
Lunar June 24th
GPT-4o
Torch Festival 07/22
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.

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

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

Findings of the Association for Computational Linguistics: ACL 2025, pages 20256–20277
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

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*.

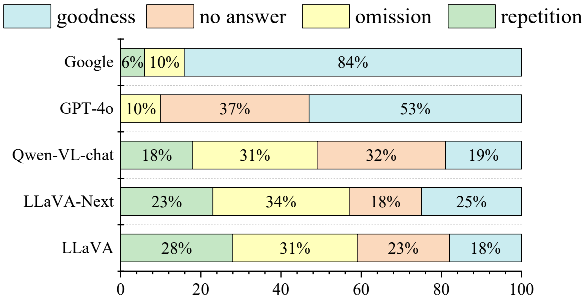


Figure 2: Preliminary experiments on repetition and omission hallucinations in TiMT.

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 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*.

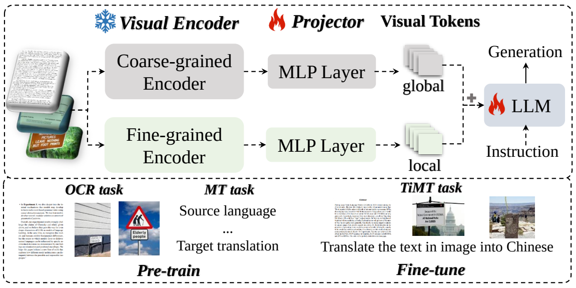


Figure 3: An overview of InImageTrans. We use a coarse-grained encoder and a fine-grained encoder to capture detailed visual information and feed them into LLM with instructions.

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:

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

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

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

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:

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

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:

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

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:

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

where θ represents the parameters involved in the training model, π_{ref} represents the reference model, σ is activation function, and β is a hyper-parameter that controls the degree of deviation.

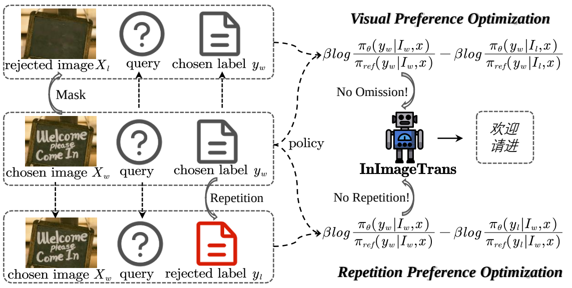

Visual Preference Optimization
rejected image X_l
query
chosen label y_w
$\beta log \frac{\pi_\theta(y_w|I_w,x)}{\pi_{ref}(y_w|I_w,x)} - \beta log \frac{\pi_\theta(y_w|I_l,x)}{\pi_{ref}(y_w|I_l,x)}$
Mask
No Omission!
chosen image X_w
query
chosen label y_w
policy
InImageTrans
欢迎
请进
Repetition
No Repetition!
chosen image X_w
query
rejected label y_l
$\beta log \frac{\pi_\theta(y_w|I_w,x)}{\pi_{ref}(y_w|I_w,x)} - \beta log \frac{\pi_\theta(y_l|I_w,x)}{\pi_{ref}(y_l|I_w,x)}$
Repetition Preference Optimization

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.

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-OCR[2] (Li et al., 2022) and DeepEraser[3] (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:

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

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↑
		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	36.4 / 38.1	27.7 / 37.5	30.5 / 43.1	**25.7 / 33.2**	30.1 / 46.5	36.5 / 46.2
InternVL2	26B	45.9 / 58.9	36.5 / 48.6	24.1 / 35.7	24.7 / 31.2	28.0 / 34.8	26.8 / 40.1	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	28.4 / 36.1	28.7 / 41.2	19.7 / 29.7	23.4 / 39.8	31.3 / 41.0
Ours										
InImageTrans	8B	59.3 / 70.8	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 / 70.8	**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↑	Repetition↓	Omission↓	BLEU↑	Repetition↓	Omission↓	BLEU↑	Repetition↓	Omission↓
InternVL-2-8B	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.

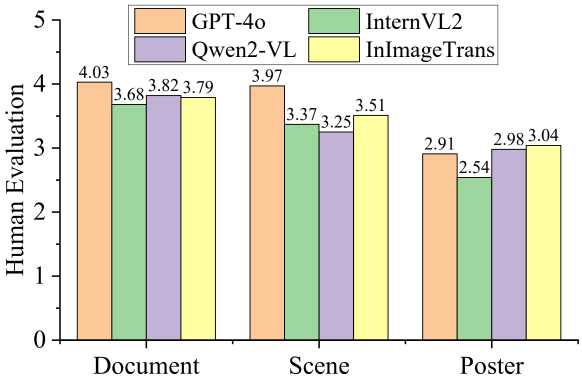


Figure 5: Overall human evaluation results of translation performance for different methods.

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 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.

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.

7 Discussion

7.1 Why rPO Can Relieve Repetition?

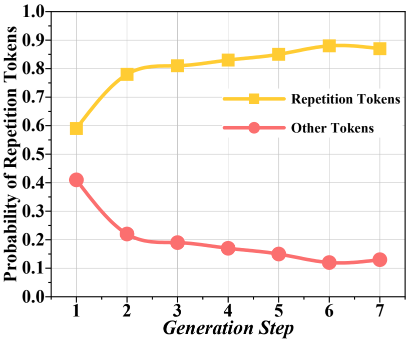


Figure 6: The probability of repetition tokens and other tokens with the generation step increases.

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

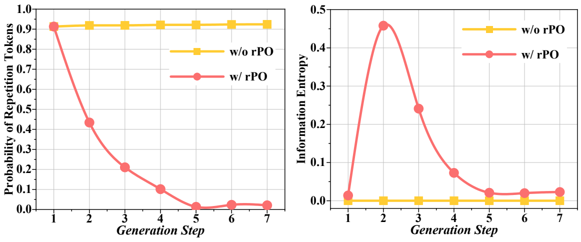


Figure 7: **The left** represents probabilities of repetition tokens and **the right** represents information entropy of the model output across generation steps. We report the average scores among 20 selected examples whose repetition hallucinations are resolved by rPO.

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, moext 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, Weihan 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, Hyeongwoo 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. Silkie: 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, Prajjwal 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*.

Weihan 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, Zhiquan 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*.

Wenyi 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, Xiaojiang 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. Rlaif-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, Quanjiang 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 synthdog[4]. 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

schudule. 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×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×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×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.

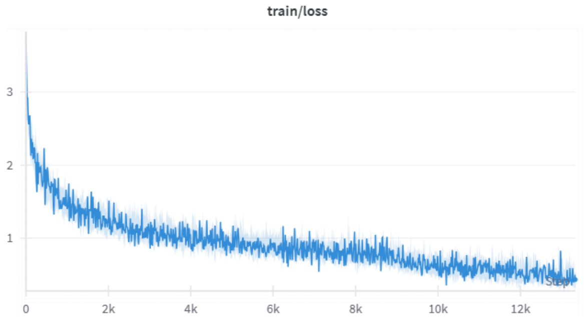


Figure 8: Training loss of mcDPO training process.

A.5 Evaluation Details

Translation Quality We use sacreBLEU[5] (Post, 2018) to calculate BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) scores for evaluating translation quality. Furthermore, we use evaluate[6] to calculate METEOR (Banerjee and Lavie, 2005) scores and use Unbabel-comet[7] 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:

$$Repetition\_4 = 1.0 - \frac{|unique(4\_gram)|}{|4\_gram|}. \quad (8)$$

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↑
		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**	66.7	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	85.0	80.9	**76.5**	70.6	**66.9**	**71.7**	62.3	70.8	**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↓
		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	105.1	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	114.5	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	111.7	193.5	107.4	113.7	131.5	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	106.4	**111.3**	**173.6**	**104.2**	**107.5**	**126.1**	115.7	**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 Translate[8] and Baidu Translate[9].

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, 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.

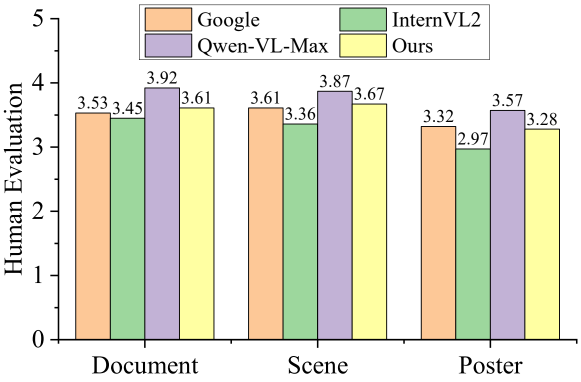


Figure 9: Overall human evaluation results of paragraph merging performance for different methods.

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 scenrios, 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

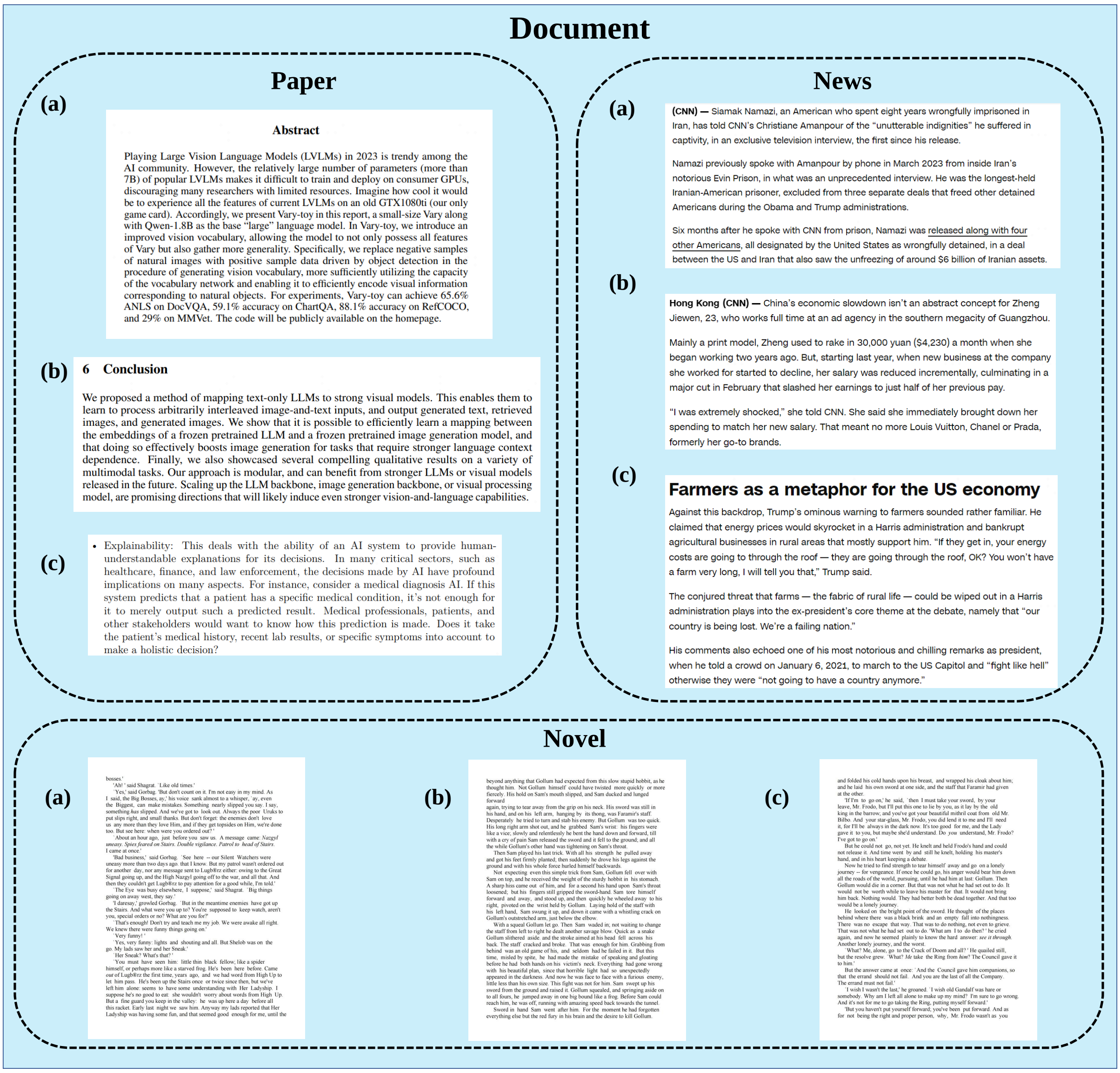
In order to more intuitively demonstrate the difference between MCiT and the benchmarks in previous works, we list here various scenario and types of image examples in MCiT, as shown in Figure 10, 11, and 12. The document class in Figure 10 has a large amount of text, the scene class in Figure 11 has 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


Document
Paper
News
Novel
(a)
(b)
(c)
(a)
(b)
(c)
(a)
(b)
(c)

Abstract

Playing Large Vision Language Models (LVLMs) in 2023 is trendy among the AI community. However, the relatively large number of parameters (more than 7B) of popular LVLMs makes it difficult to train and deploy on consumer GPUs, discouraging many researchers with limited resources. Imagine how cool it would be to experience all the features of current LVLMs on an old GTX1080ti (our only game card). Accordingly, we present Vary-toy in this report, a small-size Vary along with Qwen-1.8B as the base "large" language model. In Vary-toy, we introduce an improved vision vocabulary, allowing the model to not only possess all features of Vary but also gather more generality. Specifically, we replace negative samples of natural images with positive sample data driven by object detection in the procedure of generating vision vocabulary, more sufficiently utilizing the capacity of the vocabulary network and enabling it to efficiently encode visual information corresponding to natural objects. For experiments, Vary-toy can achieve 65.6% ANLS on DocVQA, 59.1% accuracy on ChartQA, 88.1% accuracy on RefCOCO, and 29% on MMVet. The code will be publicly available on the homepage.

6 Conclusion

We proposed a method of mapping text-only LLMs to strong visual models. This enables them to learn to process arbitrarily interleaved image-and-text inputs, and output generated text, retrieved images, and generated images. We show that it is possible to efficiently learn a mapping between the embeddings of a frozen pretrained LLM and a frozen pretrained image generation model, and that doing so effectively boosts image generation for tasks that require stronger language context dependence. Finally, we also showcased several compelling qualitative results on a variety of multimodal tasks. Our approach is modular, and can benefit from stronger LLMs or visual models released in the future. Scaling up the LLM backbone, image generation backbone, or visual processing model, are promising directions that will likely induce even stronger vision-and-language capabilities.

- Explainability: This deals with the ability of an AI system to provide human-understandable explanations for its decisions. In many critical sectors, such as healthcare, finance, and law enforcement, the decisions made by AI have profound implications on many aspects. For instance, consider a medical diagnosis AI. If this system predicts that a patient has a specific medical condition, it's not enough for it to merely output such a predicted result. Medical professionals, patients, and other stakeholders would want to know how this prediction is made. Does it take the patient's medical history, recent lab results, or specific symptoms into account to make a holistic decision?

(CNN) — Siamak Namazi, an American who spent eight years wrongfully imprisoned in Iran, has told CNN's Christiane Amanpour of the "unutterable indignities" he suffered in captivity, in an exclusive television interview, the first since his release.

Namazi previously spoke with Amanpour by phone in March 2023 from inside Iran's notorious Evin Prison, in what was an unprecedented interview. He was the longest-held Iranian-American prisoner, excluded from three separate deals that freed other detained Americans during the Obama and Trump administrations.

Six months after he spoke with CNN from prison, Namazi was released along with four other Americans, all designated by the United States as wrongfully detained, in a deal between the US and Iran that also saw the unfreezing of around $6 billion of Iranian assets.

Hong Kong (CNN) — China's economic slowdown isn't an abstract concept for Zheng Jiewen, 23, who works full time at an ad agency in the southern megacity of Guangzhou.

Mainly a print model, Zheng used to rake in 30,000 yuan ($4,230) a month when she began working two years ago. But, starting last year, when new business at the company she worked for started to decline, her salary was reduced incrementally, culminating in a major cut in February that slashed her earnings to just half of her previous pay.

"I was extremely shocked," she told CNN. She said she immediately brought down her spending to match her new salary. That meant no more Louis Vuitton, Chanel or Prada, formerly her go-to brands.

Farmers as a metaphor for the US economy

Against this backdrop, Trump's ominous warning to farmers sounded rather familiar. He claimed that energy prices would skyrocket in a Harris administration and bankrupt agricultural businesses in rural areas that mostly support him. "If they get in, your energy costs are going to through the roof — they are going through the roof, OK? You won't have a farm very long, I will tell you that," Trump said.

The conjured threat that farms — the fabric of rural life — could be wiped out in a Harris administration plays into the ex-president's core theme at the debate, namely that "our country is being lost. We're a failing nation."

His comments also echoed one of his most notorious and chilling remarks as president, when he told a crowd on January 6, 2021, to march to the US Capitol and "fight like hell" otherwise they were "not going to have a country anymore."

bosses.'

'Ah! ' said Shagrat. `Like old times.'

`Yes,' said Gorbag. 'But don't count on it. I'm not easy in my mind. As I said, the Big Bosses, ay,' his voice sank almost to a whisper, `ay, even the Biggest, can make mistakes. Something nearly slipped you say. I say, something *has* slipped. And we've got to look out. Always the poor Uruks to put slips right, and small thanks. But don't forget: the enemies don't love us any more than they love Him, and if they get topsides on Him, we're done too. But see here: when were you ordered out? '

`About an hour ago, just before you saw us. A message came: *Nazgyl uneasy. Spies feared on Stairs. Double vigilance. Patrol to head of Stairs.* I came at once.'

'Bad business,' said Gorbag. `See here -- our Silent Watchers were uneasy more than two days ago. that I know. But my patrol wasn't ordered out for another day, nor any message sent to Lugb®rz either: owing to the Great Signal going up, and the High Nazgyl going off to the war, and all that. And then they couldn't get Lugb®rz to pay attention for a good while, I'm told.'

`The Eye was busy elsewhere, I suppose,' said Shagrat. `Big things going on away west, they say.'

'I daresay,' growled Gorbag. `But in the meantime enemies have got up the Stairs. And what were you up to? You're supposed to keep watch, aren't you, special orders or no? What are you for?'

`That's enough! Don't try and teach me my job. We were awake all right. We knew there were funny things going on.'

`Very funny! '

`Yes, very funny: lights and shouting and all. But Shelob was on the go. My lads saw her and her Sneak.'

`Her Sneak? What's that? '

`You must have seen him: little thin black fellow; like a spider himself, or perhaps more like a starved frog. He's been here before. Came *out* of Lugb®rz the first time, years ago, and we had word from High Up to let him pass. He's been up the Stairs once or twice since then, but we've left him alone: seems to have some understanding with Her Ladyship. I suppose he's no good to eat: she wouldn't worry about words from High Up. But a fine guard you keep in the valley: he was up here a day before all this racket. Early last night we saw him. Anyway my lads reported that Her Ladyship was having some fun, and that seemed good enough for me, until the

beyond anything that Gollum had expected from this slow stupid hobbit, as he thought him. Not Gollum himself could have twisted more quickly or more fiercely. His hold on Sam's mouth slipped, and Sam ducked and lunged forward
again, trying to tear away from the grip on his neck. His sword was still in his hand, and on his left arm, hanging by its thong, was Faramir's staff. Desperately he tried to turn and stab his enemy. But Gollum was too quick. His long right arm shot out, and he grabbed Sam's wrist: his fingers were like a vice; slowly and relentlessly he bent the hand down and forward, till with a cry of pain Sam released the sword and it fell to the ground; and all the while Gollum's other hand was tightening on Sam's throat.

Then Sam played his last trick. With all his strength he pulled away and got his feet firmly planted; then suddenly he drove his legs against the ground and with his whole force hurled himself backwards.

Not expecting even this simple trick from Sam, Gollum fell over with Sam on top, and he received the weight of the sturdy hobbit in his stomach. A sharp hiss came out of him, and for a second his hand upon Sam's throat loosened; but his fingers still gripped the sword-hand. Sam tore himself forward and away, and stood up, and then quickly he wheeled away to his right, pivoted on the wrist held by Gollum. Laying hold of the staff with his left hand, Sam swung it up, and down it came with a whistling crack on Gollum's outstretched arm, just below the elbow.

With a squeal Gollum let go. Then Sam waded in; not waiting to change the staff from left to right he dealt another savage blow. Quick as a snake Gollum slithered aside. and the stroke aimed at his head fell across his back. The staff cracked and broke. That was enough for him. Grabbing from behind was an old game of his, and seldom had he failed in it. But this time, misled by spite, he had made the mistake of speaking and gloating before he had both hands on his victim's neck. Everything had gone wrong with his beautiful plan, since that horrible light had so unexpectedly appeared in the darkness. And now he was face to face with a furious enemy, little less than his own size. This fight was not for him. Sam swept up his sword from the ground and raised it. Gollum squealed, and springing aside on to all fours, he jumped away in one big bound like a frog. Before Sam could reach him, he was off, running with amazing speed back towards the tunnel.

Sword in hand Sam went after him. For the moment he had forgotten everything else but the red fury in his brain and the desire to kill Gollum.

and folded his cold hands upon his breast, and wrapped his cloak about him; and he laid his own sword at one side, and the staff that Faramir had given at the other.

'If I'm to go on,' he said, `then I must take your sword, by your leave, Mr. Frodo, but I'll put this one to lie by you, as it lay by the old king in the barrow; and you've got your beautiful mithril coat from old Mr. Bilbo. And your star-glass, Mr. Frodo, you did lend it to me and I'll need it, for I'll be always in the dark now. It's too good for me, and the Lady gave it to you, but maybe she'd understand. Do *you* understand, Mr. Frodo? I've got to go on.'

But he could not go, not yet. He knelt and held Frodo's hand and could not release it. And time went by and still he knelt, holding his master's hand, and in his heart keeping a debate.

Now he tried to find strength to tear himself away and go on a lonely journey -- for vengeance. If once he could go, his anger would bear him down all the roads of the world, pursuing, until he had him at last: Gollum. Then Gollum would die in a corner. But that was not what he had set out to do. It would not be worth while to leave his master for that. It would not bring him back. Nothing would. They had better both be dead together. And that too would be a lonely journey.

He looked on the bright point of the sword. He thought of the places behind where there was a black brink and an empty fall into nothingness. There was no escape that way. That was to do nothing, not even to grieve. That was not what he had set out to do. 'What am I to do then? ' he cried again, and now he seemed plainly to know the hard answer: *see it through*. Another lonely journey, and the worst.

`What? Me, alone, go to the Crack of Doom and all? ' He quailed still, but the resolve grew. `What? *Me* take the Ring from *him*? The Council gave it to him.'

But the answer came at once: `And the Council gave him companions, so that the errand should not fail. And you are the last of all the Company. The errand must not fail.'

`I wish I wasn't the last,' he groaned. `I wish old Gandalf was hare or somebody. Why am I left all alone to make up my mind? I'm sure to go wrong. And it's not for me to go taking the Ring, putting myself forward.'

'But you haven't put yourself forward; you've been put forward. And as for not being the right and proper person, why, Mr. Frodo wasn't as you

Figure 10: Some examples of document images in MCiT. **The upper left** is examples of paper, **the upper right** is examples of news, and **the bottom** is examples of novel.

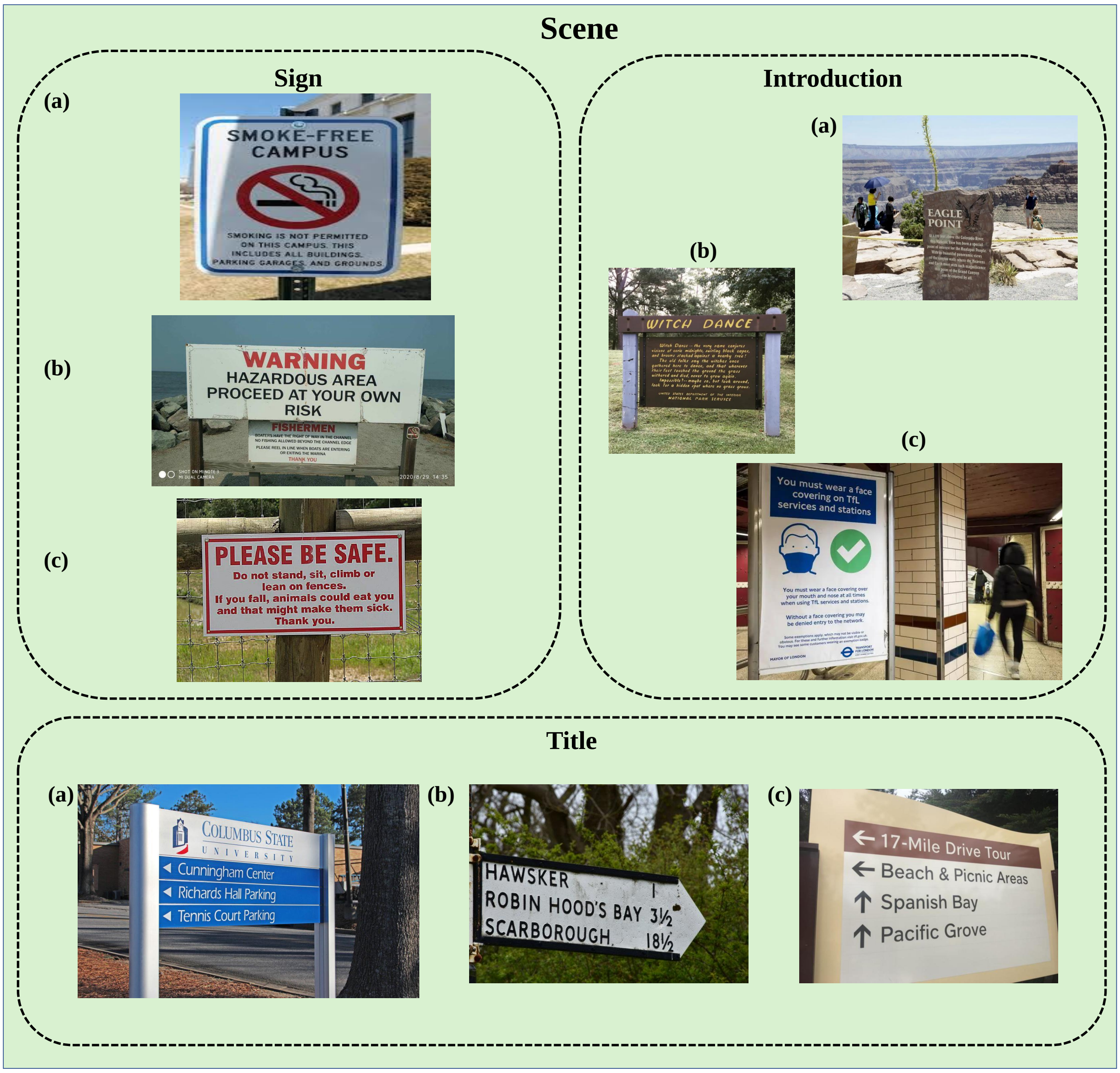


Figure 11: Some examples of scene images in MCiT. **The upper left** is examples of sign, **the upper right** is examples of introduction, and **the bottom** is examples of title.

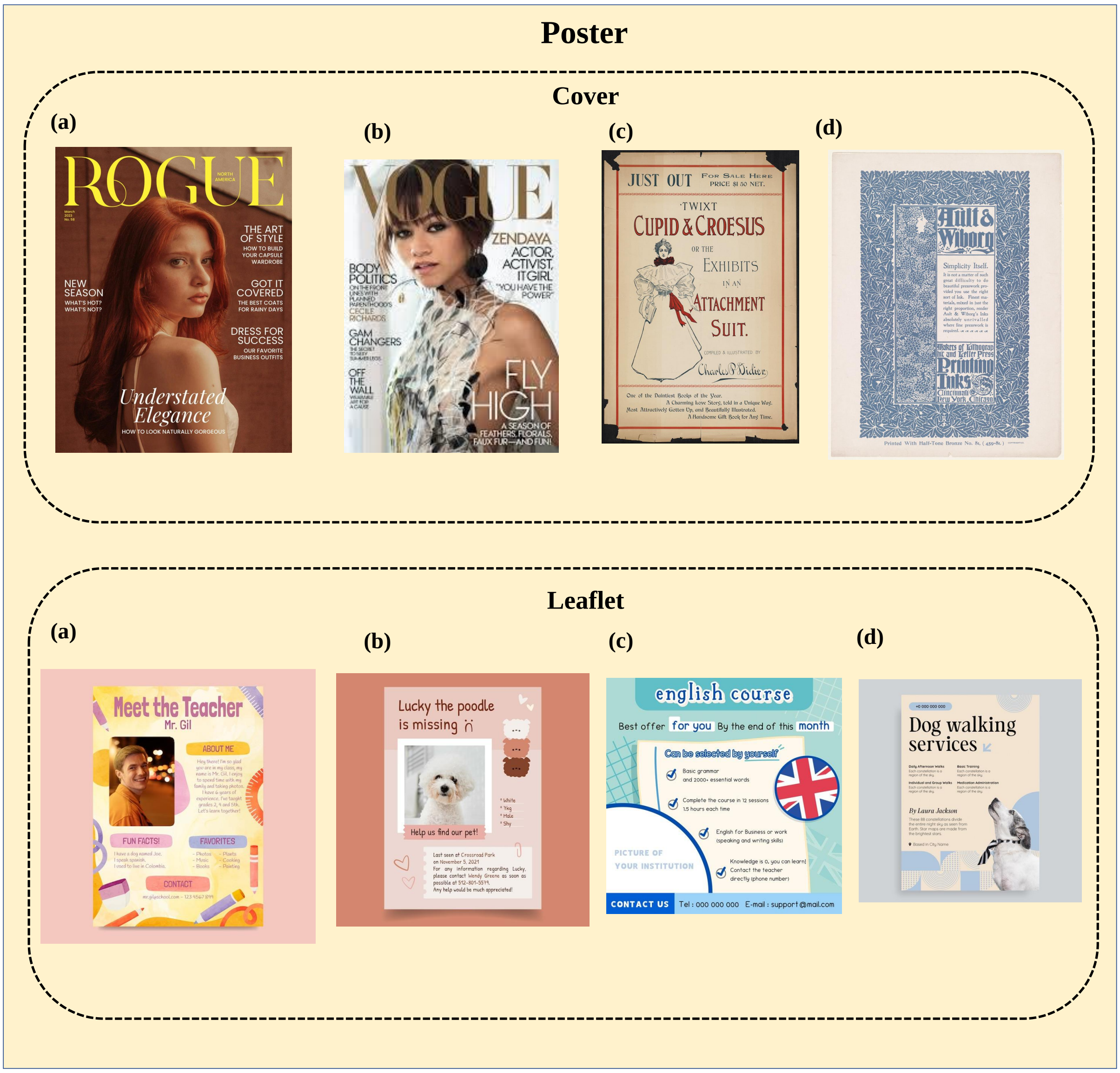
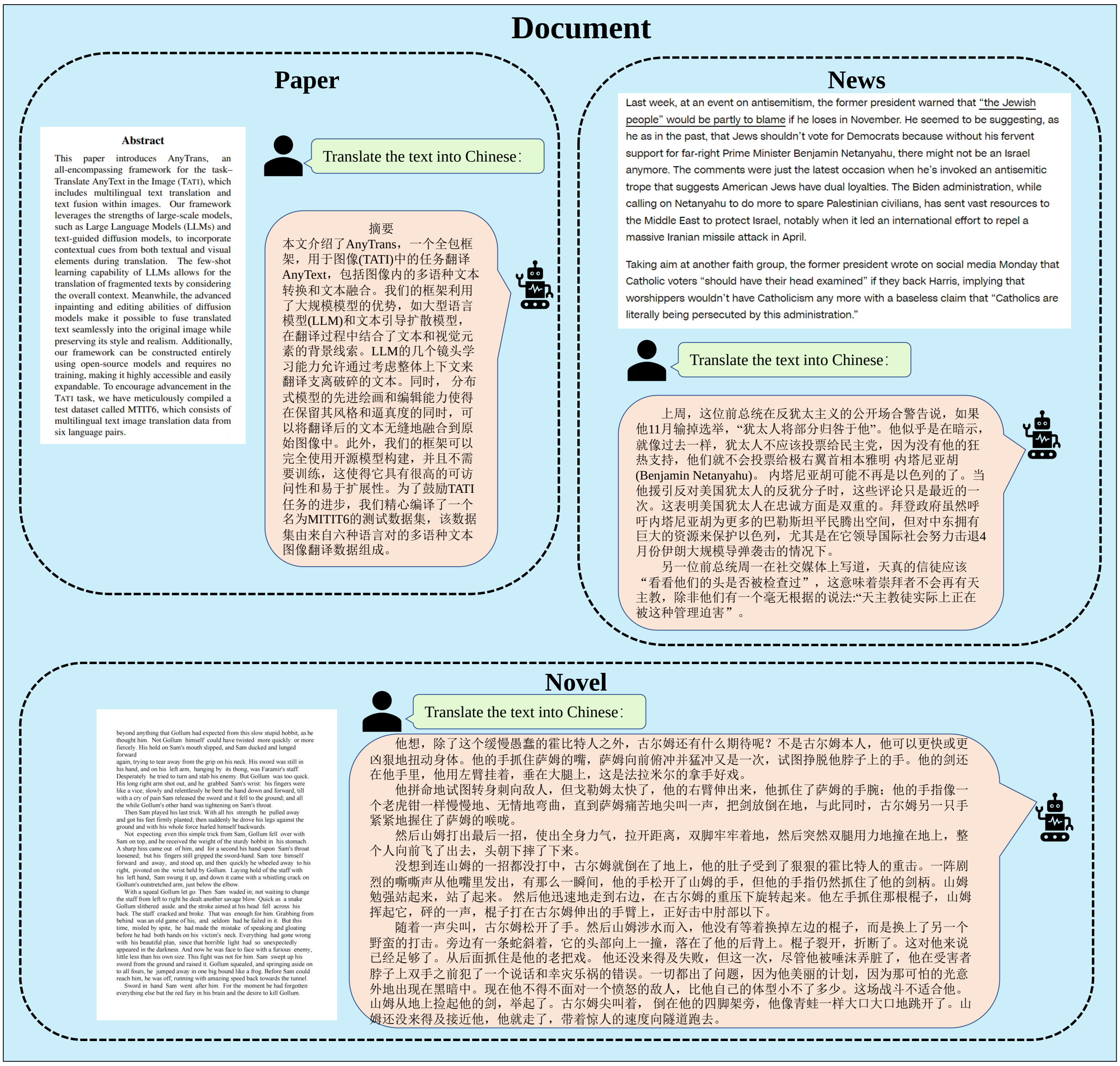


Figure 12: Some examples of poster images in MCiT. **The top** is examples of cover, and **the bottom** is examples of leaflet. It should be noted that figure **(a)** in the cover is from ROGUE [9] magazine, and figure (**b**) is from VOGUE [10] magazine.



Abstract

This paper introduces AnyTrans, an all-encompassing framework for the task–Translate AnyText in the Image (TATI), which includes multilingual text translation and text fusion within images. Our framework leverages the strengths of large-scale models, such as Large Language Models (LLMs) and text-guided diffusion models, to incorporate contextual cues from both textual and visual elements during translation. The few-shot learning capability of LLMs allows for the translation of fragmented texts by considering the overall context. Meanwhile, the advanced inpainting and editing abilities of diffusion models make it possible to fuse translated text seamlessly into the original image while preserving its style and realism. Additionally, our framework can be constructed entirely using open-source models and requires no training, making it highly accessible and easily expandable. To encourage advancement in the TATI task, we have meticulously compiled a test dataset called MTIT6, which consists of multilingual text image translation data from six language pairs.

Translate the text into Chinese：

摘要

本文介绍了AnyTrans，一个全包框架，用于图像(TATI)中的任务翻译AnyText，包括图像内的多语种文本转换和文本融合。我们的框架利用了大规模模型的优势，如大型语言模型(LLM)和文本引导扩散模型，在翻译过程中结合了文本和视觉元素的背景线索。LLM的几个镜头学习能力允许通过考虑整体上下文来翻译支离破碎的文本。同时， 分布式模型的先进绘画和编辑能力使得在保留其风格和逼真度的同时，可以将翻译后的文本无缝地融合到原始图像中。此外，我们的框架可以完全使用开源模型构建，并且不需要训练，这使得它具有很高的可访问性和易于扩展性。为了鼓励TATI任务的进步，我们精心编译了一个名为MITIT6的测试数据集，该数据集由来自六种语言对的多语种文本图像翻译数据组成。

Last week, at an event on antisemitism, the former president warned that "the Jewish people" would be partly to blame if he loses in November. He seemed to be suggesting, as he as in the past, that Jews shouldn't vote for Democrats because without his fervent support for far-right Prime Minister Benjamin Netanyahu, there might not be an Israel anymore. The comments were just the latest occasion when he's invoked an antisemitic trope that suggests American Jews have dual loyalties. The Biden administration, while calling on Netanyahu to do more to spare Palestinian civilians, has sent vast resources to the Middle East to protect Israel, notably when it led an international effort to repel a massive Iranian missile attack in April.

Taking aim at another faith group, the former president wrote on social media Monday that Catholic voters "should have their head examined" if they back Harris, implying that worshippers wouldn't have Catholicism any more with a baseless claim that "Catholics are literally being persecuted by this administration."

Translate the text into Chinese：

上周，这位前总统在反犹太主义的公开场合警告说，如果他11月输掉选举，“犹太人将部分归咎于他”。他似乎是在暗示，就像过去一样，犹太人不应该投票给民主党，因为没有他的狂热支持，他们就不会投票给极右翼首相本雅明 ·内塔尼亚胡(Benjamin Netanyahu)。 内塔尼亚胡可能不再是以色列的了。当他援引反对美国犹太人的反犹分子时，这些评论只是最近的一次。这表明美国犹太人在忠诚方面是双重的。拜登政府虽然呼吁内塔尼亚胡为更多的巴勒斯坦平民腾出空间，但对中东拥有巨大的资源来保护以色列，尤其是在它领导国际社会努力击退4月份伊朗大规模导弹袭击的情况下。

另一位前总统周一在社交媒体上写道，天真的信徒应该“看看他们的头是否被检查过”，这意味着崇拜者不会再有天主教，除非他们有一个毫无根据的说法:“天主教徒实际上正在被这种管理迫害”。

beyond anything that Gollum had expected from this slow stupid hobbit, as he thought him. Not Gollum himself could have twisted more quickly or more fiercely. His hold on Sam's mouth slipped, and Sam ducked and lunged forward

again, trying to tear away from the grip on his neck. His sword was still in his hand, and on his left arm, hanging by its thong, was Faramir's staff. Desperately he tried to turn and stab his enemy. But Gollum was too quick. His long right arm shot out, and he grabbed Sam's wrist: his fingers were like a vice; slowly and relentlessly he bent the hand down and forward, till with a cry of pain Sam released the sword and it fell to the ground; and all the while Gollum's other hand was tightening on Sam's throat.

Then Sam played his last trick. With all his strength he pulled away and got his feet firmly planted; then suddenly he drove his legs against the ground and with his whole force hurled himself backwards.

Not expecting even this simple trick from Sam, Gollum fell over with Sam on top, and he received the weight of the sturdy hobbit in his stomach. A sharp hiss came out of him, and for a second his hand upon Sam's throat loosened; but his fingers still gripped the sword-hand. Sam tore himself forward and away, and stood up, and then quickly he wheeled away to his right, pivoted on the wrist held by Gollum. Laying hold of the staff with his left hand, Sam swung it up, and down it came with a whistling crack on Gollum's outstretched arm, just below the elbow.

With a squeal Gollum let go. Then Sam waded in; not waiting to change the staff from left to right he dealt another savage blow. Quick as a snake Gollum slithered aside. and the stroke aimed at his head fell across his back. The staff cracked and broke. That was enough for him. Grabbing from behind was an old game of his, and seldom had he failed in it. But this time, misled by spite, he had made the mistake of speaking and gloating before he had both hands on his victim's neck. Everything had gone wrong with his beautiful plan, since that horrible light had so unexpectedly appeared in the darkness. And now he was face to face with a furious enemy, little less than his own size. This fight was not for him. Sam swept up his sword from the ground and raised it. Gollum squealed, and springing aside on to all fours, he jumped away in one big bound like a frog. Before Sam could reach him, he was off, running with amazing speed back towards the tunnel.

Sword in hand Sam went after him. For the moment he had forgotten everything else but the red fury in his brain and the desire to kill Gollum.

Translate the text into Chinese：

他想，除了这个缓慢愚蠢的霍比特人之外，古尔姆还有什么期待呢？不是古尔姆本人，他可以更快或更凶狠地扭动身体。他的手抓住萨姆的嘴，萨姆向前俯冲并猛冲又是一次，试图挣脱他脖子上的手。他的剑还在他手里，他用左臂挂着，垂在大腿上，这是法拉米尔的拿手好戏。

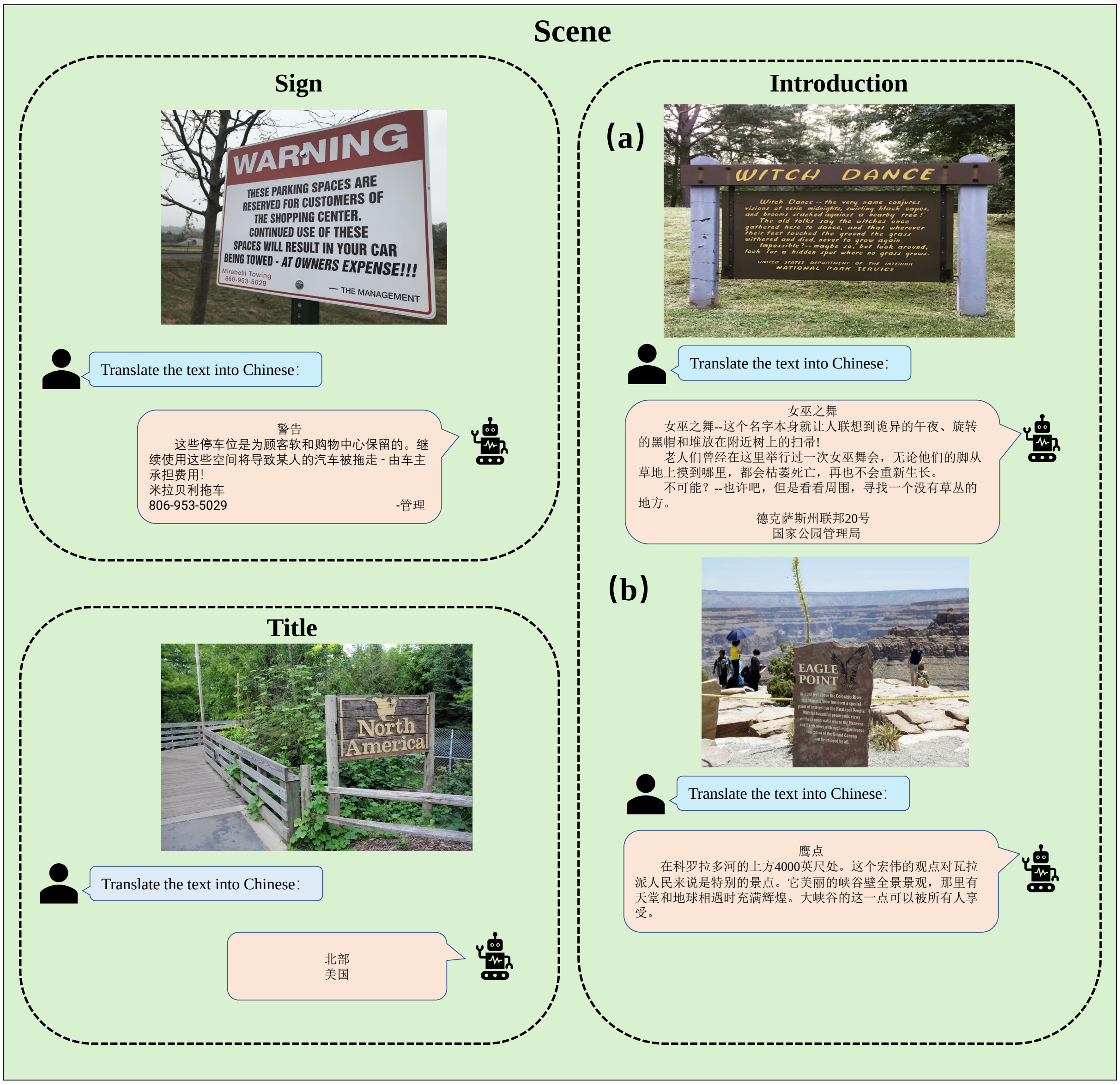
他拼命地试图转身刺向敌人，但戈勒姆太快了，他的右臂伸出来，他抓住了萨姆的手腕；他的手指像一个老虎钳一样慢慢地、无情地弯曲，直到萨姆痛苦地尖叫一声，把剑放倒在地，与此同时，古尔姆另一只手紧紧地握住了萨姆的喉咙。

然后山姆打出最后一招，使出全身力气，拉开距离，双脚牢牢着地，然后突然双腿用力地撞在地上，整个人向前飞了出去，头朝下摔了下来。

没想到连山姆的一招都没打中，古尔姆就倒在了地上，他的肚子受到了狠狠的霍比特人的重击。一阵剧烈的嘶嘶声从他嘴里发出，有那么一瞬间，他的手松开了山姆的手，但他的手指仍然抓住了他的剑柄。山姆勉强站起来，站了起来。 然后他迅速地走到右边，在古尔姆的重压下旋转起来。他左手抓住那根棍子，山姆挥起它，砰的一声，棍子打在古尔姆伸出的手臂上，正好击中肘部以下。

随着一声尖叫，古尔姆松开了手。然后山姆涉水而入，他没有等着换掉左边的棍子，而是换上了另一个野蛮的打击。旁边有一条蛇斜着，它的头部向上一撞，落在了他的后背上。棍子裂开，折断了。这对他来说已经足够了。从后面抓住是他的老把戏。 他还没来得及失败，但这一次，尽管他被唾沫弄脏了，他在受害者脖子上双手之前犯了一个说话和幸灾乐祸的错误。一切都出了问题，因为他美丽的计划，因为那可怕的光意外地出现在黑暗中。现在他不得不面对一个愤怒的敌人，比他自己的体型小不了多少。这场战斗不适合他。山姆从地上捡起他的剑，举起了。古尔姆尖叫着， 倒在他的四脚架旁，他像青蛙一样大口大口地跳开了。山姆还没来得及接近他，他就走了，带着惊人的速度向隧道跑去。

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.


Scene
Sign
WARNING
THESE PARKING SPACES ARE RESERVED FOR CUSTOMERS OF THE SHOPPING CENTER. CONTINUED USE OF THESE SPACES WILL RESULT IN YOUR CAR BEING TOWED - AT OWNERS EXPENSE!!!
Mirabelli Towing 860-953-5029
— THE MANAGEMENT
Title
North America
Introduction
(a)
WITCH DANCE
Witch Dance -- the very name conjures visions of eerie midnights, swirling black capes, and brooms stacked against a nearby tree! The old folks say the witches once gathered here to dance, and that wherever their feet touched the ground the grass withered and died, never to grow again. Impossible?--maybe so, but look around, look for a hidden spot where no grass grows.
UNITED STATES DEPARTMENT OF THE INTERIOR NATIONAL PARK SERVICE
(b)
EAGLE POINT
this point of the Grand Canyon can be enjoyed by all.

Translate the text into Chinese:

警告
这些停车位是为顾客软和购物中心保留的。继续使用这些空间将导致某人的汽车被拖走 - 由车主承担费用!
米拉贝利拖车
806-953-5029 -管理

Translate the text into Chinese:

北部
美国

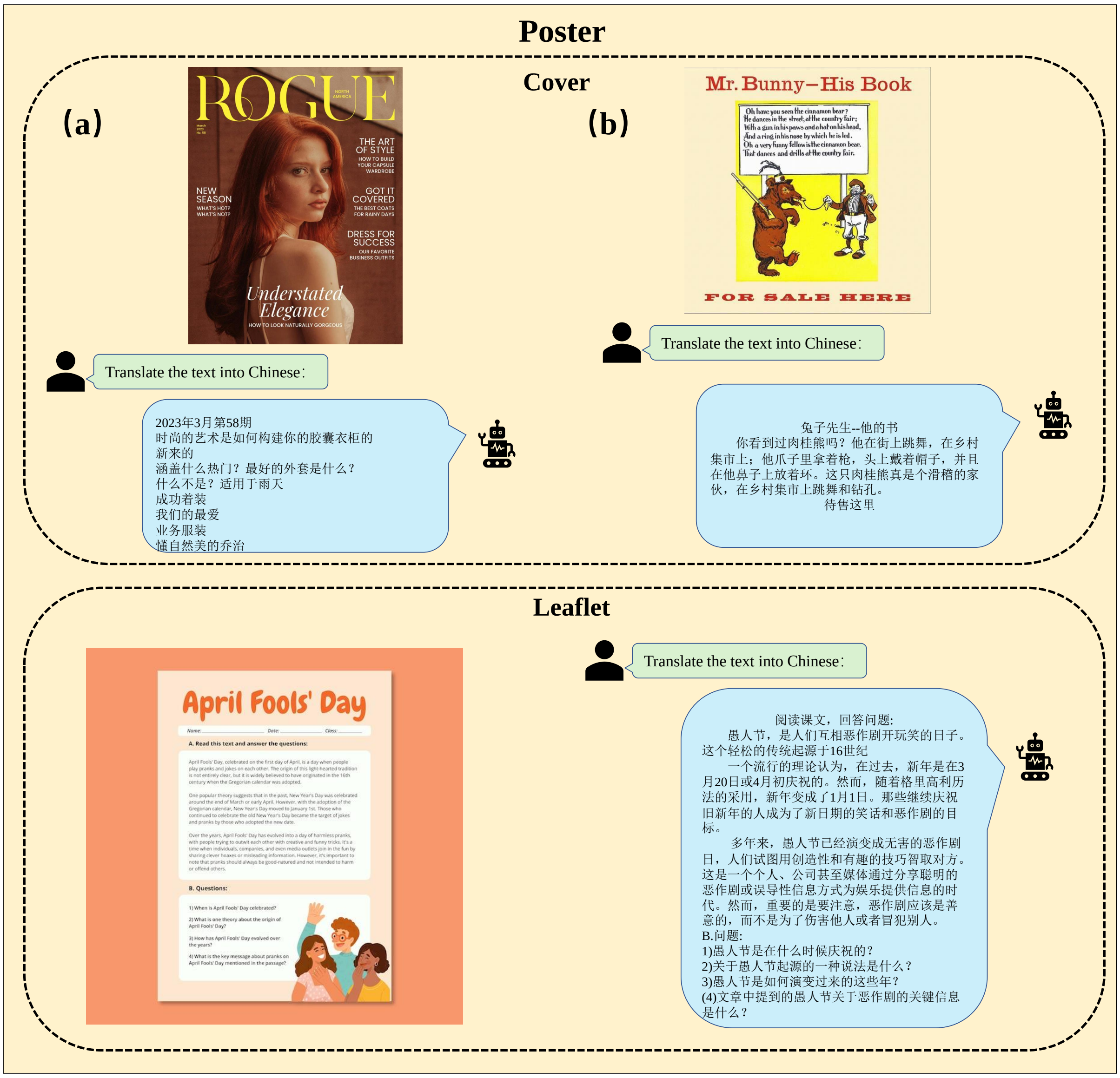
Translate the text into Chinese:

女巫之舞
女巫之舞--这个名字本身就让人联想到诡异的午夜、旋转的黑帽和堆放在附近树上的扫帚!
老人们曾经在这里举行过一次女巫舞会，无论他们的脚从草地上摸到哪里，都会枯萎死亡，再也不会重新生长。
不可能？--也许吧，但是看看周围，寻找一个没有草丛的地方。
德克萨斯州联邦20号
国家公园管理局

Translate the text into Chinese:

鹰点
在科罗拉多河的上方4000英尺处。这个宏伟的观点对瓦拉派人民来说是特别的景点。它美丽的峡谷壁全景景观，那里有天堂和地球相遇时充满辉煌。大峡谷的这一点可以被所有人享受。

Figure 14: Some visualization results of our model on scene images. **The upper left** is the result of sign, **the lower left** is the result of title, and **the right** is the result of introduction.



Translate the text into Chinese：

2023年3月第58期
时尚的艺术是如何构建你的胶囊衣柜的
新来的
涵盖什么热门？最好的外套是什么？
什么不是？适用于雨天
成功着装
我们的最爱
业务服装
懂自然美的乔治

Translate the text into Chinese：

兔子先生--他的书

你看到过肉桂熊吗？他在街上跳舞，在乡村集市上；他爪子里拿着枪，头上戴着帽子，并且在他鼻子上放着环。这只肉桂熊真是个滑稽的家伙，在乡村集市上跳舞和钻孔。

待售这里

Translate the text into Chinese：

阅读课文，回答问题:

愚人节，是人们互相恶作剧开玩笑的日子。这个轻松的传统起源于16世纪

一个流行的理论认为，在过去，新年是在3月20日或4月初庆祝的。然而，随着格里高利历法的采用，新年变成了1月1日。那些继续庆祝旧新年的人成为了新日期的笑话和恶作剧的目标。

多年来，愚人节已经演变成无害的恶作剧日，人们试图用创造性和有趣的技巧智取对方。这是一个个人、公司甚至媒体通过分享聪明的恶作剧或误导性信息方式为娱乐提供信息的时代。然而，重要的是要注意，恶作剧应该是善意的，而不是为了伤害他人或者冒犯别人。

B.问题:
1)愚人节是在什么时候庆祝的？
2)关于愚人节起源的一种说法是什么？
3)愚人节是如何演变过来的这些年？
(4)文章中提到的愚人节关于恶作剧的关键信息是什么？

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 ROGUE [9] magazine.