Improving MLLM's Document Image Machine Translation via Synchronously Self-reviewing Its OCR Proficiency

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

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China

{liangyupu2021, zhangzhiyang2020, chenzhiyuan2023}@ia.ac.cn, {yaping.zhang, yang.zhao, lu.xiang, cqzong, yzhou}@nlpr.ia.ac.cn

Abstract

Multimodal Large Language Models (MLLMs) have shown strong performance in document image tasks, especially Optical Character Recognition (OCR). However, they struggle with Document Image Machine Translation (DIMT), which requires handling both cross-modal and cross-lingual challenges. Previous efforts to enhance DIMT capability through Supervised Fine-Tuning (SFT) on the DIMT dataset often result in the forgetting of the model's existing monolingual abilities, such as OCR. To address these challenges, we introduce a novel fine-tuning paradigm, named **S**ynchronously **S**elf-**R**eviewing (**SSR**) its OCR proficiency, inspired by the concept "*Bilingual Cognitive Advantage*". Specifically, SSR prompts the model to generate OCR text before producing translation text, which allows the model to leverage its strong monolingual OCR ability while learning to translate text across languages. Comprehensive experiments demonstrate the proposed SSR learning helps mitigate catastrophic forgetting, improving the generalization ability of MLLMs on both OCR and DIMT tasks.[1]

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved significant advancements in various document image understanding tasks, particularly in Optical Character Recognition (OCR), which plays a crucial role in extracting text from scanned documents or images. These improvements have led to notable progress in tasks, such as Visual Question Answering (VQA), and Information Extraction (IE) (Wei et al., 2024b; Liu et al., 2024; Wang et al., 2024; Wei et al., 2024a). However, MLLMs still face challenges towards Document Image Machine Translation (DIMT)—the task of translating text in document images from one language to another. (Zhang et al., 2023c,b; Liang et al., 2024).

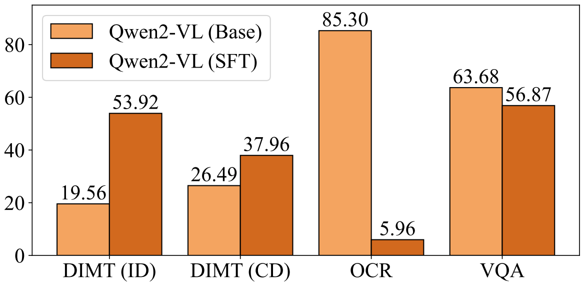


Figure 1: Performance of Qwen2-VL across various benchmarks. **Base** refers to the performance of the original MLLM, while **SFT** denotes the MLLM after fine-tuning on the DIMT dataset. **DIMT (ID)** and **DIMT (CD)** denote in-domain and cross-domain test separately. The evaluation metrics for DIMT, OCR, and VQA are BLEU, Character Accuracy (CA), and Average Normalized Levenshtein Similarity (ANLS), respectively.

An intuitive approach to enhancing MLLM's DIMT ability is to apply Supervised Fine-Tuning (SFT) (Ouyang et al., 2022) on annotated DIMT datasets. However, a major challenge with SFT is catastrophic forgetting, where fine-tuning MLLM on translation tasks often causes a loss of the model's original OCR capability. As shown in Figure 1, while the fine-tuned MLLM performs well on translation tasks, achieving a BLEU score of 53.92 on the in-domain DIMT task, it struggles to accurately extract text from images, with an accuracy of only 5.96 on the OCR task. This significant drop in OCR performance indicates a near-complete loss of the MLLM's OCR proficiency.

To address the challenges associated with SFT, existing continual learning methods have been proposed (Yin et al., 2022; Mok et al., 2023; Yang et al., 2024b; Shi et al., 2024; Wu et al., 2024). These methods aim to mitigate catastrophic forgetting and enhance domain generalization through

[*] Corresponding author.

[1] Our code is available at: https://github.com/liangyupu/SSR

Findings of the Association for Computational Linguistics: ACL 2025, pages 23659–23678
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

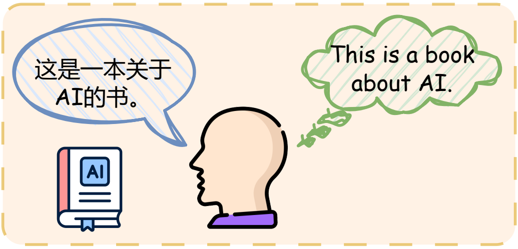


Figure 2: Bilingual individuals exhibit greater linguistic proficiency.

various strategies, such as replay-based methods and regularization-based methods. However, challenges persist in effectively balancing the retention of prior knowledge with the acquisition of new skills, especially in complex tasks like DIMT.

Inspired by the concept of "*Bilingual Cognitive Advantage*" (Bialystok, 1991; Hamers, 1998; Bialystok, 2001; Bialystok and Craik, 2010; Zhang et al., 2023a, 2024, 2025a), as shown in Figure 2, a learning paradigm that focuses on retaining and leveraging human's existing monolingual strengths while learning new languages, we introduce a simple yet effective fine-tuning paradigm called **S**ynchronized **S**elf-**R**eviewing (**SSR**), where the MLLM generates the OCR text in the source language first, followed by the translation text in the target language. By synchronous learning, SSR enables the MLLM to leverage its strong monolingual OCR proficiency while extending its capabilities to new languages, thereby improving its cross-lingual performance on the DIMT task. Additionally, SSR enhances the MLLM's generalization ability, making it more robust across various domains and tasks. Furthermore, the method benefits from the use of large amounts of unsupervised data, reducing the need for extensive parallel datasets, which are often scarce in the DIMT task.

In summary, this paper presents a novel method to improve DIMT performance by using synchronously self-reviewing to preserve monolingual OCR proficiency while enabling cross-lingual DIMT. We demonstrate, through extensive experiments, that SSR significantly enhances the MLLM's generalization across both OCR and DIMT tasks, addressing challenges such as catastrophic forgetting and poor domain generalization.

Our contributions are summarized as follows:

- We propose a novel fine-tuning paradigm, SSR, which leverages the strong monolingual capabilities of MLLMs to enhance their cross-lingual performance.
- We introduce synchronous self-reviewing to utilize the MLLM's OCR proficiency and preserve its monolingual capability.
- Extensive experiments validate the effectiveness of the proposed method in improving the generalization ability of MLLMs on the DIMT task while maintaining their monolingual competence.

2 Related Work

Different from text machine translation (Yang et al., 2023, 2024a, 2025), document image machine translation aims to translate text within document images from one language to another while preserving the logical layout (Liang et al., 2024). Recent advancements in DIMT can be categorized into two primary approaches: (1) **Cascade systems** (Hinami et al., 2021; Sable et al., 2023; Zhang et al., 2023c; Yao, 2023), which employ multiple models sequentially and encounter issues such as structural redundancy, error propagation, and high latency. (2) **End-to-end models** (Ma et al., 2022; Zhu et al., 2023; Zhang et al., 2023b; Liang et al., 2024; Ma et al., 2024; Zhang et al., 2025c,b; Guan et al., 2025), which streamline the process by optimizing a unified training objective, thereby improving structural efficiency. These end-to-end methods are increasingly attracting researchers' attention. Zhu et al. (2023) introduces an end-to-end TIMT framework that bridges the modality gap with pre-trained models. Liang et al. (2024) assembles multiple pre-trained models to complete the end-to-end DIMT task. Zhang et al. (2025b) proposes a framework to unify the geometric layout and logical layout of document images. While these end-to-end methods have demonstrated satisfactory performance, their effectiveness is restricted to respective training domains, with limited cross-domain generalization.

Recent advancements in MLLMs have significantly improved the processing and understanding of text-rich document images (Hu et al., 2024a,b; Wei et al., 2024b,a; Liu et al., 2024; Yu et al., 2024; Wang et al., 2024; Jian et al., 2024; Ren et al., 2025). Wei et al. (2024a) explores adding fine-grained vision perception for document images to the MLLM without affecting its existing natural image understanding capabilities. Liu et al. (2024) proposes shifted window attention to achieve cross-window connectivity at higher input resolutions and token

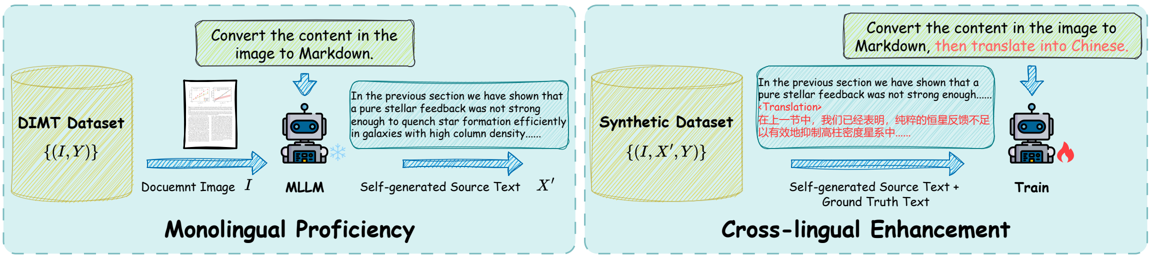

Convert the content in the image to Markdown.
DIMT Dataset
$\{(I, Y)\}$
Docuemnt Image I
MLLM
In the previous section we have shown that a pure stellar feedback was not strong enough to quench star formation efficiently in galaxies with high column density......
Self-generated Source Text X'
Monolingual Proficiency
Convert the content in the image to Markdown, then translate into Chinese.
Synthetic Dataset
$\{(I, X', Y)\}$
In the previous section we have shown that a pure stellar feedback was not strong enough......
<Translation>
在上一节中，我们已经表明，纯粹的恒星反馈不足以有效地抑制高柱密度星系中......
Self-generated Source Text + Ground Truth Text
Train
Cross-lingual Enhancement

Figure 3: Overview of our proposed fine-tuning paradigm SSR. It contains two steps: (1) **Monolingual proficiency**: Given a document image and the original OCR prompt, the MLLM generates the source text (OCR result). (2) **Cross-lingual enhancement**: Use the self-generated source text and the ground truth target text to fine-tune the MLLM, enabling it to learn the relationship between the image, source text, and target text, while also smoothing the training process.

resampler to filter out significant tokens. Wang et al. (2024) introduces dynamic resolution mechanism and multimodal rotary position embeddings to facilitating the effective fusion of text, images, and videos. Although MLLMs have demonstrated strong performance across various document image understanding tasks, their effectiveness diminishes for emerging tasks like DIMT.

3 Method

In this section, we will introduce SSR, a novel fine-tuning paradigm that leverage the MLLM's monolingual (OCR) proficiency to enhance its cross-lingual (DIMT) ability. The overview of our approach is shown in Figure 3. The key idea is to train the model to first generate the source text (OCR result) before producing the target text (translation text). This approach enables the model to incorporate both image and source text information when generating the target text. Although the self-generated source text may contain misrecognized or repeated text, since it is sampled from the model's original distribution, it contributes to a smoother convergence of the model's loss curve during training, which will be discussed in Section 5.1. Furthermore, this self-review process helps in the retention of the model's original monolingual capabilities.

3.1 Monolingual Proficiency

This process involves prompting the MLLM with its original OCR instruction to perform OCR on the document image. Since the generated text is sampled from the MLLM's original distribution, it is better suited for maintaining its inherent monolingual capabilities.

Given a DIMT dataset $\mathcal{D} = \{(\boldsymbol{I}, \boldsymbol{Y})\}$, where $\boldsymbol{I}$ and $\boldsymbol{Y}$ denote the document image and corresponding ground truth target text, we prompt the MLLM to generate the OCR result $\boldsymbol{X}'$ for each document image $\boldsymbol{I}$ based on its original OCR instruction.

$$\boldsymbol{X}' \sim \mathrm{MLLM}(\boldsymbol{P}_{\mathrm{ocr}}, \boldsymbol{I}) \tag{1}$$

where $\boldsymbol{P}_{\mathrm{ocr}}$ is the MLLM's original OCR instruction.

This process is similar to some replay methods (Shi et al., 2024) in continual learning; however, the key difference is that we allow the MLLM to generate its own replay data.

3.2 Cross-lingual Enhancement

This process concatenates the self-generated source text and ground truth target text to fine-tune the MLLM. This approach enables the model to learn the relationship between different modalities while leveraging its monolingual capabilities to enhance cross-lingual performance, simultaneously facilitating self-review of its monolingual proficiency.

SSR constructs a prompt template based on the original OCR template. Take Qwen2-VL (Wang et al., 2024) as an example, the prompt construction is as follows:

SSR-constrained Prompt Template

Instruction:
Convert the content in the image to Markdown (original OCR instruction of the MLLM), then translate into Chinese.

Response:
$\boldsymbol{X}'$ (self-generated source text)
<Translation> (special token)
$\boldsymbol{Y}$ (ground truth target text)

The constructed instruction-response pair is subsequently used to train the MLLM using the standard negative log-likelihood loss, which can be formulated as follows:

$$\mathcal{L} = -\sum_{t=1}^{r} \log p(\boldsymbol{R}_t | \boldsymbol{R}_{<t}, \boldsymbol{P}, \boldsymbol{I}; \boldsymbol{\theta}) \quad (2)$$

$$\boldsymbol{R} = \text{CONCAT}(\boldsymbol{X}', < \text{Translation} >, \boldsymbol{Y}) \quad (3)$$

where $\boldsymbol{R}_t$ denotes the t-th token of the response, $\boldsymbol{P}$ represents the instruction, $\boldsymbol{\theta}$ refers to the trainable parameters, and r denotes the length of $\boldsymbol{R}$.

This approach trains the MLLM to gradually learn to generate target text, using the generated source text as a reference to guide target text generation. This aligns more closely with the MLLM's original output distribution, resulting in a smoother training curve for the MLLM.

4 Experiment

4.1 Dataset & Metrics

We randomly select 10K samples from the DoTA dataset (Liang et al., 2024) and comprehensively evaluate the model on the DoTA dataset for in-domain test and DITrans dataset (Zhang et al., 2023b) for cross-domain test. Detailed settings can be seen in Appendix A.1.

We thoroughly evaluate the models' capabilities in three aspects: (1) **Full-text translation**, which means the translation quality of all the text in the image - BLEU. (2) **Plain-text translation**, which means the translation quality of the text after removing formulas and tables - BLEU-PT. (3) **Structure preserving**, which means the model's ability to restore the layout structure of the document images - STEDS (Structure Tree-Edit-Distance-based Similarity). All metric calculations follow the same procedure as described by Liang et al. (2024).

4.2 Settings

We select four MLLMs with different numbers of parameters: Vary-toy (Wei et al., 2024b), Vary-base (Wei et al., 2024a), Textmonkey (Liu et al., 2024) and Qwen2-VL (Wang et al., 2024). Given the constraints of our computational resources, the Low-Rank Adaptation (LoRA) technique (Hu et al., 2022) is utilized in our experiments. Specifically, a LoRA adapter with a rank of 16 is integrated into all the linear layers of the LLM part in the MLLM and exclusively trains the adapter. The MLLMs are fine-tuned for 3 epochs on the train set. We use the Adam optimizer and employ a linear decay learning rate schedule with a learning rate of 1e-4. The batch is set to 32 for stable training. The greedy search is used for inference. More detailed settings are in Appendix A.2.

4.3 Baselines

We evaluate our method against diverse baselines, including small models, MLLMs with Chain of Thought (CoT), Supervised Fine-tuning (SFT), and replay method, to comprehensively assess its performance and validate its effectiveness.

- **Small Model Baselines**

LARDIT (Zhang et al., 2023c) This cascade system employs a layout analysis model (Yao, 2023), the OCR tool, and a text-only machine translation model trained on the DoTA dataset, sequentially.

Nougat-trans (Blecher et al., 2024) We utilize the Nougat model for combined layout analysis and OCR and the text-only machine translation model is employed for translation.

DIMTDA (Liang et al., 2024) This end-to-end DIMT model uses a model assembler to integrate multiple pre-trained models to enhance the understanding of layout and translation capabilities.

UMTIT (Niu et al., 2024) This model consists of two image-text modality conversion steps. We only use the result of the first step for evaluation, which converts images to text to recognize the source text and generate translations.

MTKD (Ma et al., 2023) This method can effectively distillate knowledge from the pipeline model and utilizes three teacher models to improve the performance of the end-to-end TIMT model.

AnyTrans (Qian et al., 2024) This paper presents a framework entirely using open-source models, such as LLMs and text-guided diffusion models, to complete in-image machine translation. We only use the result of the translated text for evaluation.

The following lists the baselines based on MLLMs. The detailed prompts for each method can be seen in Appendix A.3.

Base We directly prompt the original MLLM to perform the DIMT task.

- **CoT Baselines**

CoT (Direct) (Wei et al., 2022) We directly prompt the original MLLM to perform "OCR than translation" on the document image.

CoT (Cascade) (Wei et al., 2022) We first prompt the original MLLM to perform OCR, and then prompt it to generate the translation based on both the image and the OCR result.

- **SFT Baselines**

SFT (MT) (Ouyang et al., 2022) The MLLM is first fine-tuned on the English-Chinese parallel corpus from the training set, and then CoT (Cascade)

	Academic Article (ID)			Political Report (CD)			Ads & News (CD)			Time
	BLEU	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**	**s/page (↓)**
Baselines										
LARDIT	35.58	41.75	75.83	14.66	16.58	57.77	1.64	1.71	41.63	12.46
Nougat-trans	43.37	50.79	88.16	18.39	19.21	52.12	2.71	2.83	40.53	17.03
DIMTDA	38.68	42.34	84.44	12.64	15.03	60.86	2.06	2.17	40.75	9.82
UMTIT	37.40	40.02	82.37	10.06	10.67	51.90	2.77	2.08	40.87	14.76
MTKD	37.32	39.96	82.28	13.24	15.33	59.58	2.42	2.39	40.89	9.24
AnyTrans	32.98	34.94	75.83	31.05	31.05	57.77	16.47	17.89	41.63	14.81
Vary-base (8.1B)										
Base	13.45	5.79	76.26	2.84	2.79	56.21	1.06	1.06	44.17	**47.62**
CoT (Direct)	11.41	4.60	79.89	2.37	2.31	57.11	0.95	0.96	**51.05**	52.32
CoT (Cascade)	3.42	1.81	42.11	2.90	2.73	41.17	0.87	0.87	37.14	120.54
SFT (MT)	3.94	2.48	48.00	3.29	3.16	**57.89**	1.18	1.18	49.91	233.08
SFT (DIMT)	19.84	18.60	75.71	4.46	4.49	46.9	0.94	0.94	36.70	92.25
SDFT	11.56	11.51	67.30	2.99	3.02	42.13	0.79	0.82	33.96	137.93
SSR	**33.86**	**34.50**	**81.72**	**21.47**	**22.03**	50.92	**6.68**	**6.69**	49.07	150.44
Textmonkey (9.7B)										
Base	0.12	0.21	29.37	0.36	0.62	31.90	0.32	0.67	26.65	**64.98**
CoT (Direct)	0.34	0.33	33.65	0.99	0.94	37.85	0.88	0.48	33.75	71.88
CoT (Cascade)	0.47	0.61	29.43	0.52	0.74	31.90	0.34	0.70	26.69	123.21
SFT (MT)	16.69	18.93	69.42	12.26	12.26	**61.06**	5.26	5.26	52.21	259.22
SFT (DIMT)	21.10	24.50	73.07	15.98	16.07	60.46	6.07	6.07	54.25	97.99
SDFT	20.50	24.04	71.80	26.62	27.31	58.51	9.26	9.28	**55.68**	137.74
SSR	**26.45**	**28.55**	**75.97**	**32.66**	**33.57**	59.37	**12.40**	**12.40**	54.31	147.83
Qwen2-VL (8.3B)										
Base	19.56	15.38	57.29	26.49	26.51	58.10	11.19	11.19	58.81	**33.58**
CoT (Direct)	12.71	8.01	57.94	22.16	22.30	61.34	6.12	6.12	57.89	40.71
CoT (Cascade)	29.44	27.07	57.75	36.37	36.31	63.50	28.92	28.92	68.69	58.69
SFT (MT)	33.07	35.30	63.91	35.79	35.78	64.17	18.68	18.68	50.67	113.72
SFT (DIMT)	53.92	53.20	87.27	37.96	37.93	63.08	23.48	23.49	69.72	51.64
SDFT	53.55	55.11	87.17	39.01	38.97	63.25	27.65	27.65	67.78	54.26
SSR	**57.23**	**58.88**	**89.65**	**41.91**	**41.80**	**67.28**	**33.61**	**33.59**	**71.98**	95.48

Table 1: Results on DoTA and DITrans dataset. All MLLMs are fine-tuned on the DoTA dataset and tested on both the DoTA dataset, which contains images from the **Academic Article** domain, serving as the **in-domain (ID)** test and the DITrans dataset, which includes images from the **Political Report**, **Ads & News** domains, serving as the zero-shot **cross-domain (CD)** test. The number of parameters for each MLLM is provided alongside its respective model. The **Time** refers to the average inference time on a single NVIDIA A100 GPU. (↓) indicates that for this metric, lower values are better. The **bold numbers** indicate the best performance achieved by each MLLM.

method above is applied to generate translations.

SFT (DIMT) (Ouyang et al., 2022) The MLLM is directly fine-tuned on the train set.

- **Replay Baseline**

SDFT (Yang et al., 2024b) This method fine-tunes the model with a distilled dataset generated by the model itself to match its original distribution.

5 Results & Analysis

5.1 Main Results

Table 1 reports the performance of all methods. It can be observed that our method outperforms the baselines in terms of translation quality across all MLLMs with varying sizes and structures. The results of Vary-toy experiment can be seen in the Appendix B.1.

- **MLLM with Limited Instruction-following Ability** In the Vary-base and Textmonkey experiments, the performance of SSR significantly surpasses all other methods. Take the Vary-base experiment as an example, the improvements are 14.02 BLEU in the in-domain test, and 17.01 BLEU and 5.74 BLEU in two zero-shot cross-domain tests, compared to SFT (DIMT). These results show that our approach can be applied to larger MLLMs, thereby validating its effectiveness in enhancing translation quality and generalization.

- **MLLM with Strong Instruction-following Ability** In the Qwen2-VL experiment, the original MLLM achieves high translation quality (53.92 BLEU in the in-domain test) despite requiring only minimal training data, our proposed method still outperforms SFT (DIMT) by margins of 3.31 BLEU and 5.68 BLEU-PT. Furthermore, in zero-shot cross-domain evaluations on ads & news domains, SSR surpasses SFT (DIMT) by substantial

		OCR (Document)		OCR (Scene)		DocVQA	InfoVQA	ChartQA
		CA	WA	CA	WA	ANLS	ANLS	ANLS
Vary-toy	Base	**68.46**	**65.17**	45.74	41.73	**47.76**	5.13	**7.87**
	SFT (MT)	16.95	13.10	49.09	46.54	27.97	1.16	2.07
	SFT (DIMT)	14.53	8.81	31.61	29.56	40.10	2.73	5.25
	SDFT	64.80	61.50	43.02	39.67	42.59	**5.45**	5.98
	SSR	61.10	56.86	**51.8**	**46.66**	43.61	4.02	4.92
Vary-base	Base	**68.48**	**64.91**	**81.41**	**76.83**	**66.38**	**12.75**	**12.51**
	SFT (MT)	50.39	46.15	80.61	47.78	56.25	6.67	6.55
	SFT (DIMT)	47.65	42.46	70.01	33.33	60.82	11.02	11.7
	SDFT	67.35	63.59	54.91	49.41	62.68	11.49	12.46
	SSR	66.14	62.45	81.26	75.92	62.19	10.15	9.78
Textmonkey	Base	**75.52**	**70.56**	**84.74**	**79.40**	**58.39**	22.21	**8.69**
	SFT (MT)	9.23	7.49	78.46	74.53	39.17	12.15	6.10
	SFT (DIMT)	8.98	5.69	72.86	68.82	52.52	22.58	7.47
	SDFT	73.51	69.77	57.96	52.67	39.78	20.45	7.38
	SSR	72.76	67.57	83.11	78.59	55.02	**22.78**	8.11
Qwen2-VL	Base	85.30	78.20	70.29	64.75	**93.55**	**63.07**	**63.68**
	SFT (MT)	5.83	3.08	19.83	17.45	84.90	54.33	46.57
	SFT (DIMT)	5.96	2.17	33.47	31.06	88.98	57.57	56.87
	SDFT	**86.72**	**80.12**	71.98	67.84	90.55	59.72	60.51
	SSR	85.18	78.12	**82.03**	**77.48**	92.47	60.56	61.37

Table 2: Results of MLLMs' monolingual ability preserving after fine-tuning with different methods. The **bold numbers** indicate the best performance achieved by each MLLM, and the underline numbers are the second best.

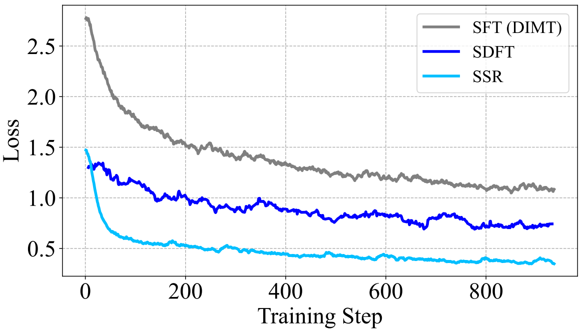


Figure 4: Training loss curves of different methods in the Vary-base experiment.

increments of 10.13 BLEU and 10.10 BLEU-PT. These findings demonstrate that our approach remains applicable to more advanced MLLMs exhibiting superior instruction-following capabilities, aligning with the ongoing research direction in MLLM development. The output samples for the DIMT test can be seen in Appendix D.

We also shows the training loss in the Vary-base experiment in Figure 4. As shown in the figure, the training loss curve of SSR is smoother and achieves the lowest loss value. This is due to the fact that the constructed data in the monolingual demonstration is sampled from the original distribution of Vary-base, making it more suitable for training.

5.2 Monolingual Ability Preserving

To assess the preservation of monolingual capabilities in base MLLMs across different methods, we perform a comprehensive evaluation using various benchmarks. For OCR performance evaluation, we employ the DITrans dataset (Zhang et al., 2023c) for document image testing and the FST dataset (Karatzas et al., 2015) for scene text image testing, with Character Accuracy (CA) and Word Accuracy (WA) as quantitative metrics. Visual Question Answering (VQA) capabilities are examined through the DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), and ChartQA (Masry et al., 2022) benchmarks, assessed via the Average Normalized Levenshtein Similarity (ANLS) metric. Notably, all evaluations are conducted in a zero-shot manner without additional fine-tuning on downstream task-specific datasets. The results are listed in Table 2.

For OCR performance, both SFT-based methods result in a significant decline in OCR effectiveness across both scenarios, illustrating a classic case of catastrophic forgetting. In contrast, SSR exhibits remarkable proficiency in maintaining the OCR capabilities of the base MLLMs. Taking the Qwen2-VL experiment as an example, SSR causes only a 0.12 decrease in CA and a 0.08 decrease

		DocVQA	InfoVQA	ChartQA
Vary-toy	Base	6.65	0.07	0.01
	SSR	7.57	0.21	0.38
Vary-base	Base	8.64	1.10	0.90
	SSR	9.13	1.93	1.64
Textmonkey	Base	19.20	9.73	13.71
	SSR	21.01	10.80	8.31
Qwen2-VL	Base	46.99	38.32	50.27
	SSR	55.16	40.32	50.37

Table 3: Results of MLLMs' cross-lingual ability generalization after fine-tuning with SSR. The text in the input image is in English, while the questions and answers are in Chinese. The ANLS scores are reported.

in WA in document image scenarios. In the scene text image scenarios, SSR even surpasses the base MLLM, achieving a increase of 11.74 in CA and 12.73 in WA. These results underscore the effectiveness of monolingual demonstrations in preserving the OCR capabilities of the base MLLMs. In document image scenarios, SDFT achieves the best performance, as it is fine-tuned with document image OCR task data. However, SSR still delivers comparable performance and surpasses SDFT in scene text image scenarios, highlighting its superior generalization capability.

In terms of VQA performance, our method also exhibits impressive preservation of monolingual abilities. In the Qwen2-VL experiment, the MLLM experiences only a 1.08 drop in ANLS on the DocVQA dataset, a negligible cost compared to the 4.57 ANLS drop seen with SFT (DIMT). This highlights the effectiveness of our method in preserving unseen general monolingual capabilities. The output samples for the OCR and VQA test can be seen in Appendix D.

5.3 Cross-lingual Ability Generalization

In our preliminary experiments, we observed that MLLMs, after fine-tuning with SSR, generalize to cross-lingual document image understanding abilities. Therefore, we conduct further comprehensive experiments to evaluate their cross-lingual capabilities. We translate both the questions and answers in several VQA benchmarks into Chinese using Google Translate and perform evaluation in a zero-shot manner. The results are shown in Table 3.

It is evident that the cross-lingual document image understanding ability of MLLMs is significantly enhanced after fine-tuning with SSR. Specifically, after fine-tuning, Qwen2-VL achieves improvements of 8.17 and 2.00 ANLS on the DocVQA and InfoVQA test sets, respectively. Moreover, by comparing the performance of Vary-base, Textmonkey and Qwen2-VL, MLLMs with stronger instruction-following capabilities demonstrate more substantial improvements. The output samples for the cross-lingual VQA test can be seen in Appendix D.

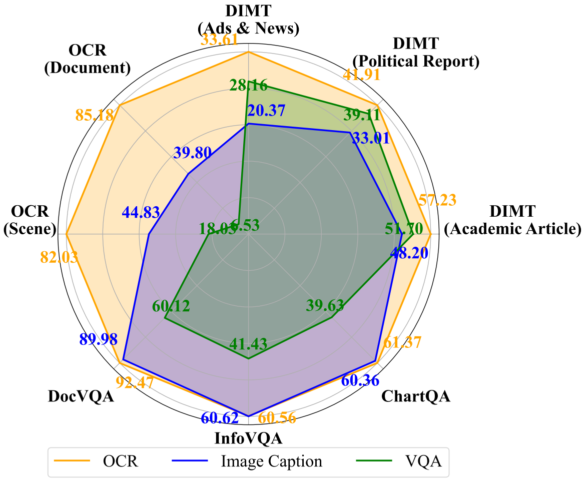


Figure 5: Results of Qwen2-VL through SSR fine-tuning using different monolingual tasks. Detailed data can be seen in Appendix C. It is better to zoom in for a clearer view.

5.4 Monolingual Task Selection

To investigate the impact of different monolingual tasks on our method, we select OCR, image caption, and VQA as the monolingual abilities to demonstrate, constructing synthetic data separately for fine-tuning Qwen2-VL. Detailed prompt templates can be found in Appendix A.3. All other settings remain consistent with the main experiment. We use BLEU, CA, and ANLS as metrics.

Figure 5 shows that using OCR as the demonstration task yields the best performance across all test sets, effectively enhancing cross-lingual ability while preserving monolingual proficiency. We believe this is because, to complete the OCR task, the MLLM needs to generate the longest text, thereby preserving the most information from the original MLLM's output distribution while also providing more context for generating target text.

5.5 Extension to Unsupervised Data

A principal advantage of our method lies in its capacity to harness the MLLM's OCR capability alongside extensive unsupervised data (only document images) to generate synthetic data, thereby augmenting the model's translation performance.

	Academic Articles (ID)			Political Report (CD)			Ads & News (CD)		
	BLEU	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**
Vary-base (8.1B)									
Base	13.45	5.79	76.26	2.84	2.79	56.21	1.06	1.06	44.17
SSR w Ground Truth Text	33.71	32.50	**83.14**	**26.05**	**26.90**	**56.82**	4.63	5.00	46.15
SSR w OCR Text	27.35	25.57	72.16	23.78	24.08	52.50	5.05	5.05	48.35
SSR w Self-generated Text	**33.86**	**34.50**	81.72	21.47	22.03	50.92	**6.68**	**6.69**	**49.07**
Qwen2-VL (8.3B)									
Base	19.56	15.38	57.29	26.49	26.51	58.1	11.19	11.19	58.81
SSR w Ground Truth Text	54.55	58.07	87.62	41.43	41.38	60.43	32.55	32.55	68.14
SSR w OCR Text	52.83	52.03	84.68	37.63	38.24	61.70	29.09	29.09	64.08
SSR w Self-generated Text	**57.23**	**58.88**	**89.65**	**41.91**	**41.80**	**67.28**	**33.61**	**33.59**	**71.98**

Table 4: Results of Vary-base and Qwen2-VL through SSR fine-tuning using heterogeneous source texts. **ID** and **CD** denote in-domain and cross-domain test, respectively. The **bold numbers** indicate the best performance.

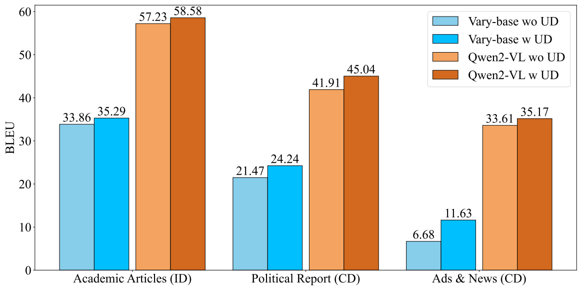


Figure 6: Results of Vary-base and Qwen2-VL through SSR fine-tuning using unsupervised data. **UD** denotes unsupervised data. Detailed data can be seen in Appendix C. It is better to zoom in for a clearer view.

To investigate the effectiveness of incorporating additional unsupervised data, we randomly select 10K document images from the DocVQA training set as the unsupervised data, obtain their OCR results using MLLMs, and translate them into Chinese using Google Translate. These synthetic data are then integrated with the original training set, and we conduct experiments with Vary-base and Qwen2-VL under the same settings as SSR in the main experiment. The results are shown in Figure 6.

As shown in the figure, introducing unsupervised data further enhances the DIMT performance of MLLMs in both in-domain and cross-domain settings compared to the main experiment. Taking Qwen2-VL as an example, although SSR has already achieved 57.23 BLEU in the academic article domain, our method, which leverages unsupervised data to generate synthetic data, leads to an improvement of 1.35 BLEU in the in-domain test and 3.13 BLEU in the cross-domain test. This demonstrates the significant potential of our approach for practical applications.

5.6 Extension to Heterogeneous Source Texts

Another advantage of our method lies in the extensibility to accommodate heterogeneous source texts. To validate this capability, we conduct evaluations comparing performance when utilizing ground truth source texts from the DoTA dataset and source texts generated by the OCR tool. Experiments are applied to both Vary-base and Qwen2-VL, following the same settings as the main experiment. The results are listed in Table 4.

Table 4 demonstrates that when the ground truth source text formatting aligns with the MLLM's OCR output format, as observed in Vary-base, SSR achieves performance parity using either ground truth text or self-generated text. In contrast, significant formatting discrepancies in Qwen2-VL lead to self-generated text consistently outperforming ground truth text in SSR across all evaluated domains. Notably, OCR text proves to be a suboptimal variant of ground truth text, with performance degradation attributed to inherent OCR noise artifacts. This disparity highlights the importance of format alignment between source texts and the MLLM OCR output for optimal SSR performance.

6 Conclusion

In this paper, we propose a novel fine-tuning paradigm, SSR, to enhance MLLMs' DIMT capabilities by leveraging their OCR proficiency, offering three key advantages. First, monolingual proficiency preserves the MLLM's original monolingual competence by maintaining the source text format. Second, cross-lingual enhancement enables the MLLM to establish relationships between different modalities, enriching target text generation with additional information. Finally, our approach can be extended to utilize large-scale unsupervised data to further enhance performance. Extensive experiments validate the effectiveness of SSR, demonstrating its superiority in strengthening cross-lingual capabilities while preserving monolingual proficiency.

Limitations

Although SSR achieves notable results on the DIMT task, its instruction-following ability and user interaction can be further improved. In the future, we plan to leverage MLLMs' text-grounding capabilities and explore the integration of user prompts to translate text within specific image regions, thereby enhancing translation alignment with user preferences.

Acknowledgements

We thank anonymous reviewers for helpful suggestions. This work is supported by the National Natural Science Foundation of China (No. 62476275 and No. 62476271).

References

Ellen Bialystok. 1991. *Language processing in bilingual children*. Cambridge University Press.

Ellen Bialystok. 2001. *Bilingualism in development: Language, literacy, and cognition*. Cambridge University Press.

Ellen Bialystok and Fergus IM Craik. 2010. Cognitive and linguistic processing in the bilingual mind. *Current directions in psychological science*, 19(1):19–23.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Boyu Guan, Yining Zhang, Yang Zhao, and Chengqing Zong. 2025. TriFine: A large-scale dataset of vision-audio-subtitle for tri-modal machine translation and benchmark with fine-grained annotated tags. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8215–8231, Abu Dhabi, UAE. Association for Computational Linguistics.

Josiane F Hamers. 1998. Cognitive and language development of bilingual children. *Cultural and language diversity and the deaf experience*, pages 51–75.

Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. Towards fully automated manga translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12998–13008. AAAI Press.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024a. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120, Miami, Florida, USA. Association for Computational Linguistics.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024b. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *CoRR*, abs/2409.03420.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and et al. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Pu Jian, Donglei Yu, and Jiajun Zhang. 2024. Large language models know what is key visual entity: An llm-assisted multimodal retrieval for vqa. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10939–10956.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman K. Ghosh, Andrew D. Bagdanov, Masakazu Iwamura, Jiri Matas, Lukás Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. 2015. ICDAR 2015 competition on robust reading. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 1156–1160. IEEE Computer Society.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7084–7095, Mexico City, Mexico. Association for Computational Linguistics.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024. Textmonkey: An ocr-free large multimodal model for understanding document. *CoRR*, abs/2403.04473.

Cong Ma, Yaping Zhang, Mei Tu, Xu Han, Linghui Wu, Yang Zhao, and Yu Zhou. 2022. Improving end-to-end text image translation from the auxiliary text translation task. In *26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*, pages 1664–1670. IEEE.

Cong Ma, Yaping Zhang, Mei Tu, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023. Multi-teacher knowledge distillation for end-to-end text image machine translation. In *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part I*, volume 14187 of *Lecture Notes in Computer Science*, pages 484–501. Springer.

Cong Ma, Yaping Zhang, Zhiyang Zhang, Yupu Liang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2024. Born a BabyNet with hierarchical parental supervision for end-to-end text image machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2468–2479, Torino, Italia. ELRA and ICCL.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Jisoo Mok, Jaeyoung Do, Sungjin Lee, Tara Taghavi, Seunghak Yu, and Sungroh Yoon. 2023. Large-scale lifelong learning of in-context instructions and how to tackle it. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12573–12589, Toronto, Canada. Association for Computational Linguistics.

Liqiang Niu, Fandong Meng, and Jie Zhou. 2024. UMTIT: unifying recognition, translation, and generation for multimodal text image translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16953–16972. ELRA and ICCL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhipeng Qian, Pei Zhang, Baosong Yang, Kai Fan, Yiwei Ma, Derek F. Wong, Xiaoshuai Sun, and Rongrong Ji. 2024. AnyTrans: Translate AnyText in the image with large scale models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2432–2444, Miami, Florida, USA. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. 2025. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*.

Nilesh P Sable, Priya Shelke, Ninad Deogaonkar, Nachiket Joshi, Rudra Kabadi, and Tushar Joshi. 2023. Doc-handler: Document scanner, manipulator, and translator based on image and natural language processing. In *2023 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 1–6. IEEE.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *CoRR*, abs/2404.16789.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Vary: Scaling up the vision vocabulary for large vision-language model. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IV*, volume 15062 of *Lecture Notes in Computer Science*, pages 408–424. Springer.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, En Yu, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024b. Small language model meets with reinforced vision vocabulary. *CoRR*, abs/2401.12503.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Junhong Wu, Yang Zhao, Yangyifan Xu, Bing Liu, and Chengqing Zong. 2024. Boosting LLM translation skills without general ability loss via rationale distillation. *CoRR*, abs/2410.13944.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2024a. Language imbalance driven rewarding for multilingual self-improving. *arXiv preprint arXiv:2410.08964*.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025. Implicit cross-lingual rewarding for efficient multilingual preference alignment. *arXiv preprint arXiv:2503.04647*.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024b. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.

Cong Yao. 2023. Docxchain: A powerful open-source toolchain for document parsing and beyond. *CoRR*, abs/2310.12430.

Wenpeng Yin, Jia Li, and Caiming Xiong. 2022. ConTinTin: Continual learning from task instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3062–3072, Dublin, Ireland. Association for Computational Linguistics.

Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. 2024. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *CoRR*, abs/2404.09204.

Yunhao Zhang, Shaonan Wang, Xinyi Dong, Jiajun Yu, and Chengqing Zong. 2023a. Navigating brain language representations: A comparative analysis of neural language models and psychologically plausible models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Yunhao Zhang, Shaonan Wang, Nan Lin, Lingzhong Fan, and Chengqing Zong. 2025a. A simple clustering approach to map the human brain's cortical semantic network organization during task. *NeuroImage*, 309:121096.

Yunhao Zhang, Xiaohan Zhang, Chong Li, Shaonan Wang, and Chengqing Zong. 2024. Mulcogbench: A multi-modal cognitive benchmark dataset for evaluating chinese and english computational language models. *arXiv preprint arXiv:2403.01116*.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Cong Ma, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2025b. Understand layout and translate text: Unified feature-conductive end-to-end document image translation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3358–3376.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023b. LayoutDIT: Layout-aware end-to-end document image translation with multi-step conductive decoder. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10043–10053, Singapore. Association for Computational Linguistics.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2025c. From chaotic OCR words to coherent document: A fine-to-coarse zoom-out network for complex-layout document image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10877–10890, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhiyang Zhang, Yaping Zhang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023c. A novel dataset and benchmark analysis on document image translation. In *China Conference on Machine Translation*, pages 103–115. Springer.

Shaolin Zhu, Shangjie Li, Yikun Lei, and Deyi Xiong. 2023. PEIT: Bridging the modality gap with pretrained 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, Toronto, Canada. Association for Computational Linguistics.

Appendix

A Setting Details

A.1 Dataset Settings

We randomly select 10K samples from the DoTA dataset to form the train set, and use the original valid and test sets. In the DITrans dataset, the sample sizes for the advertisement, news, and political report subdomains are 485, 610, and 1397, respectively. Due to the small number of images in the advertisement and news domains and their similar layout structures as scanned document images, we merge these two domains. We then randomly select 100 images as the test set. For the political report domain, we also randomly select 100 images as the test set.

A.2 Main Experiment Settings

We select four MLLMs with different numbers of parameters: Vary-toy (Wei et al., 2024b), Vary-base (Wei et al., 2024a), Textmonkey (Liu et al., 2024) and Qwen2-VL (Wang et al., 2024). We use the LoRA fine-tuning in our experiments. The LoRA adapter is added to all the linear layers of the LLM part in the MLLM. The LoRA rank and alpha are both equal to 16. We only fine-tune the adapter for 3 epochs with a batch size of 32. A linear decay learning rate schedule with a learning rate of 1e-4 and a warmup ratio of 0.1 is used. We use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$ for both training stages. We used two NVIDIA A100 GPUs and spent 16 hours to complete all the training task of SSR in the main experiment. The greedy search is used for inference.

A.3 Detailed Prompts

The OCR instructions used in the main experiment are listed as follows.

OCR Instruction for Vary-toy/base

Convert the document to markdown format.

OCR Instruction for Textmonkey

Read all the text in the image.

OCR Instruction for Qwen2-VL

Convert the content in the image to Markdown.

The instructions of baselines in the main experiment are listed as follows.

Instruction for CoT (Direct)

Convert the content in the image to Markdown (original OCR instruction of the MLLM), then translate into Chinese.

Prompt Template for CoT (Cascade)

<Round 1>
Instruction:
Convert the content in the image to Markdown. (original image caption instruction of the MLLM)

Response:
$\boldsymbol{X}$ (self-generated source text)

<Round 2>
Instruction:
Translate these text into Chinese.

Response:
$\boldsymbol{Y}$ (generated target text)

Instruction for SFT (DIMT)

Translate all the text in the image into Chinese and output in Markdown format.

The prompt templates used in the monolingual task selection experiment are listed as follows.

Prompt Template for Image Caption

Instruction:
Describe this image (original image caption instruction of the MLLM), then translate into Chinese.

Response:
$\boldsymbol{X}$ (self-generated image caption text)
<Translation> (special token)
$\boldsymbol{Y}$ (ground truth target text)

Prompt Template for VQA

Instruction:
Convert the content in the image to Markdown (original OCR instruction of the MLLM), then answer the following question:
$\boldsymbol{Q}$ (question from DocVQA, translated into Chinese)

Response:
$\boldsymbol{X}$ (self-generated source text)
<Answer> (special token)
$\boldsymbol{A}$ (answer from DocVQA, translated into Chinese)

B Detailed Analysis

B.1 Small MLLM Results in the Main Experiment

The results are shown in Table 5. In the Vary-toy experiment, SSR surpasses SFT (DIMT) by 4.64

	Academic Article (ID)			Political Report (CD)			Ads & News (CD)			Time
	BLEU	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**	**s/page (↓)**
Vary-toy (2.2B)										
Base	10.64	4.92	66.23	2.07	2.10	45.12	0.70	0.70	29.60	**43.53**
CoT (Direct)	9.17	3.87	**73.45**	2.40	2.42	**59.58**	0.68	0.68	**57.91**	46.88
CoT (Cascade)	3.99	1.68	38.13	1.09	0.99	36.06	0.23	0.27	38.39	62.64
SFT (MT)	1.99	1.38	32.14	1.30	1.33	41.04	0.47	0.47	40.02	185.38
SFT (DIMT)	9.31	8.37	62.73	1.49	1.47	38.39	0.42	0.46	41.06	86.79
SDFT	7.35	7.44	57.86	1.54	1.56	37.00	0.54	0.55	50.79	98.09
SSR	**13.95**	**14.21**	65.49	**8.15**	**8.22**	49.25	**1.26**	**1.34**	42.84	142.29

Table 5: Results of different settings of Vary-toy on DoTA and DITrans dataset.

	Academic Articles (ID)			Political Report (CD)			Ads & News (CD)		
	BLEU	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**
GPT-4o	29.70	31.95	59.45	38.66	38.66	60.54	21.75	21.75	59.48
Gemini	30.31	31.69	63.32	40.11	40.11	69.58	26.83	26.83	65.31
Qwen2-VL (8.3B)									
Base	19.56	15.38	57.29	26.49	26.51	58.10	11.19	11.19	58.81
SFT (DIMT)	53.92	53.20	87.27	37.96	37.93	63.08	23.48	23.49	69.72
SSR	**57.23**	**58.88**	**89.65**	**41.91**	**41.80**	**67.28**	**33.61**	**33.59**	**71.98**

Table 6: Results on comparison with commercial MLLMs. The **bold numbers** indicate the best performance of all models, including the commercial MLLMs.

BLEU on the in-domain test, and also achieves 8.15 BLEU in the political report zero-shot cross-domain test. These results demonstrate the effectiveness of our method in enhancing both translation quality and generalization in small MLLMs. Although our method increases inference time, the performance improvement makes this trade-off acceptable.

B.2 Comparison with Commercial MLLMs

With the rapid development of MLLMs, some commercial MLLMs (Hurst et al., 2024; Reid et al., 2024) demonstrate the capability of understanding text-rich document images. To assess their ability to accomplish the DIMT task, we randomly choose 200 samples from the test set of the DoTA dataset and the original DITrans test sets in the main experiments, then prompt GPT-4o and Gemini with three different prompts to complete the document image machine translation task. The prompts we used are as follows.

Prompts for GPT-4o and Gemini to complete DIMT task

`<Prompt 1>`
Output the Chinese translations of this image in markdown format.

`<Prompt 2>`
Please extract and provide the Chinese translations of the text contained within this image, ensuring that the translations are accurately represented, and format them using markdown for clear presentation.

`<Prompt 3>`
Please translate the all texts in this image into English and adhere to the following translation standards:
Accuracy: Ensure that the translation fully captures the meaning of all the texts in the image without adding or omitting any information.
Fluency: The translation should read naturally and smoothly, reflecting the conventions of the target language and the translation should follow the reading order of the image.
Format: The translation should be presented in markdown format.

We average the metric values of the translation results obtained from different prompts to determine the final results. As the output format of MLLMs may be unstable, we filter the English parts of the output text and only keep the Chinese parts.

Table 6 demonstrates that while GPT-4o and Gemini exhibit inherent capability to execute the DIMT task, surpassing the baseline Qwen2-VL model, they exhibit inferior performance compared to SSR-fine-tuned Qwen2-VL. This discrepancy stems from commercial MLLMs' lack of training on the DoTA dataset and their divergent output formats relative to the reference standards, resulting in

	Academic Articles (ID)			Political Report (CD)			Ads & News (CD)		
	BLEU	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**	**BLEU**	**BLEU-PT**	**STEDS**
En-Fr Vary-base (8.1B)									
Base	18.27	11.88	**81.87**	5.46	5.35	**56.61**	3.85	3.85	**49.41**
SFT (DIMT)	30.97	29.23	79.16	10.50	10.50	48.17	3.42	3.41	41.03
SSR	**44.9**	**45.37**	81.71	**35.99**	**36.68**	56.21	**10.19**	**10.19**	48.79
En-Fr Qwen2-VL (8.3B)									
Base	30.42	27.76	60.22	40.58	40.57	60.74	21.16	21.16	66.56
SFT (DIMT)	61.19	62.81	87.13	41.95	41.68	62.21	27.82	27.82	57.99
SSR	**65.25**	**68.18**	**89.85**	**51.39**	**51.22**	**63.34**	**38.96**	**38.97**	**66.58**
En-De Vary-base (8.1B)									
Base	18.95	12.62	**82.23**	6.10	5.97	56.10	4.47	4.47	**50.03**
SFT (DIMT)	29.11	27.76	78.35	6.66	6.72	48.92	2.73	2.87	41.84
SSR	**38.48**	**37.66**	79.57	**22.22**	**22.76**	**56.47**	**7.93**	**8.41**	49.25
En-De Qwen2-VL (8.3B)									
Base	25.38	22.10	60.52	27.23	27.28	57.29	19.09	19.09	64.78
SFT (DIMT)	56.15	55.47	86.40	35.01	34.84	61.05	27.18	27.18	63.52
SSR	**58.60**	**60.32**	**90.09**	**43.43**	**43.18**	**65.61**	**27.99**	**27.99**	**65.31**

Table 7: Results on English-French and English-German DIMT test. The **bold numbers** indicate the best performance of all methods.

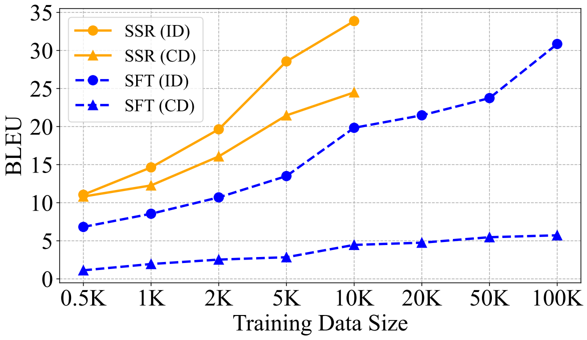


Figure 7: Results of Vary-base through SSR fine-tuning under low-resource scenarios. Detailed data can be seen in Appendix C.

substantially poorer performance on metrics including BLEU and STEDS, compared to Qwen2-VL after fine-tuning. Notably, Qwen2-VL, after fine-tuning with SSR, maintains superior performance over commercial MLLMs in the political report and ads & news domains, which are absent from its original training data. In contrast, Qwen2-VL fine-tuned with SFT does not exhibit comparable performance. This comparative analysis substantiates SSR's efficacy in enhancing MLLMs' generalization capabilities for DIMT tasks.

B.3 Low-resource Scenarios

To investigate the performance of our method in low-resource scenarios, we fine-tune Vary-base with SSR using different sizes of training data. The results are presented in Figure 7.

It can be observed that as the training data size increases, the performance of both methods improves. However, SSR consistently outperforms SFT across all data sizes and in both testing scenarios. With only 10K training samples, SSR surpasses SFT, which utilizes 100K training samples, by 3.01 BLEU on the in-domain test and 18.77 BLEU on the cross-domain test. Even with just 500 training samples, SSR still outperforms SFT (100K) by 5.09 BLEU on the cross-domain test, highlighting the exceptional potential of our approach in low-resource scenarios.

B.4 Evaluation on Other Languages

To verify our method's effectiveness in other languages, we conduct English-French and English-German DIMT experiments. We randomly choose 10K samples from the En-Fr and En-De subsets of the DoTA dataset to fine-tune MLLMs. The rest of the settings remain the same as the main experiment. The results are shown in Table 7.

Taking Qwen2-VL as an example, in the English-French DIMT test, SSR outperforms SFT (DIMT) across all test scenarios, achieving a BLEU score of 65.25 in the in-domain test. Similarly, in the English-German DIMT test, SSR surpasses SFT (DIMT) in all test scenarios, reaching a BLEU

	DIMT	DIMT	DIMT	OCR	OCR	DocVQA	InfoVQA	ChartQA
	(Academic Article)	(Political Report)	(Ads & News)	(Document)	(Scene)			
	BLEU	BELU	BLEU	CA	CA	ANLS	ANLS	ANLS
OCR	57.23	41.91	33.61	85.18	82.03	92.47	60.56	61.37
Image Caption	48.20	33.01	20.37	39.80	44.83	89.98	60.62	60.36
VQA	51.70	39.11	28.16	6.53	18.03	60.12	41.43	39.63

Table 8: Detailed data of Figure 5.

	Academic Articles (ID)			Political Report (CD)			Ads & News (CD)		
	BLEU	BLEU-PT	STEDS	BLEU	BLEU-PT	STEDS	BLEU	BLEU-PT	STEDS
Vary-base (8.1B)									
w/o UD	33.86	34.50	81.72	21.47	22.03	50.92	6.68	6.69	49.07
w UD	35.29	37.07	84.61	24.24	24.74	52.65	11.63	11.56	51.54
Qwen2-VL (8.3B)									
w/o UD	57.23	58.88	89.65	41.91	41.80	67.28	33.61	33.59	71.98
w UD	58.58	60.14	89.94	45.04	45.04	63.49	35.17	35.17	73.12

Table 9: Detailed data of Figure 6. **UD** denotes unsupervised data.

	SFT (ID)	SFT (CD)	SSR (ID)	SSR (CD)
0.5K	6.82	1.12	11.05	10.81
1K	8.55	1.95	14.64	12.27
2K	10.70	2.53	19.64	16.08
5K	13.50	2.84	28.56	21.47
10K	19.84	4.46	33.86	24.49
20K	21.48	4.75		
50K	23.74	5.47		
100K	30.85	5.72		

Table 10: Detailed data of Figure 7.

score of 58.60 in the in-domain test. These results demonstrate the effectiveness of SSR in enhancing the DIMT capability of MLLMs and improving their generalization in DIMT tasks across different languages.

C Detailed Data

Table 8 presents the detailed data corresponding to the results of Qwen2-VL through SSR fine-tuning using different monolingual tasks, as shown in Figure 5. Table 9 provides the detailed data corresponding to the results of Vary-base and Qwen2-VL through SSR fine-tuning using unsupervised data, as shown in Figure 6. Table 10 lists the detailed data corresponding to the results of Vary-base through SSR fine-tuning using different training data sizes, as shown in Figure 7.

D Output Samples

We provide the output samples of Qwen2-VL (after fine-tuning with SSR in the main experiment) on the DIMT test in Figure 8, Figure 9, and Figure 10. It is evident that the MLLM fine-tuned with SSR on the DoTA dataset can understand complex layout relationships and generate translation texts in markdown format following human reading order. Moreover, it can transfer this capability across domains to political report and ads & news domains.

Figure 11 and Figure 12 show the output samples of Qwen2-VL (after fine-tuning with SSR in the main experiment) on the OCR and VQA test. As shown in the figure, the MLLM retains strong OCR and VQA capabilities even after being fine-tuned with SSR. Furthermore, during SSR fine-tuning, the MLLM learns the relationships between English and Chinese, enabling it to generalize cross-lingual VQA capability—allowing it to answer in Chinese when given an English image and a Chinese question.

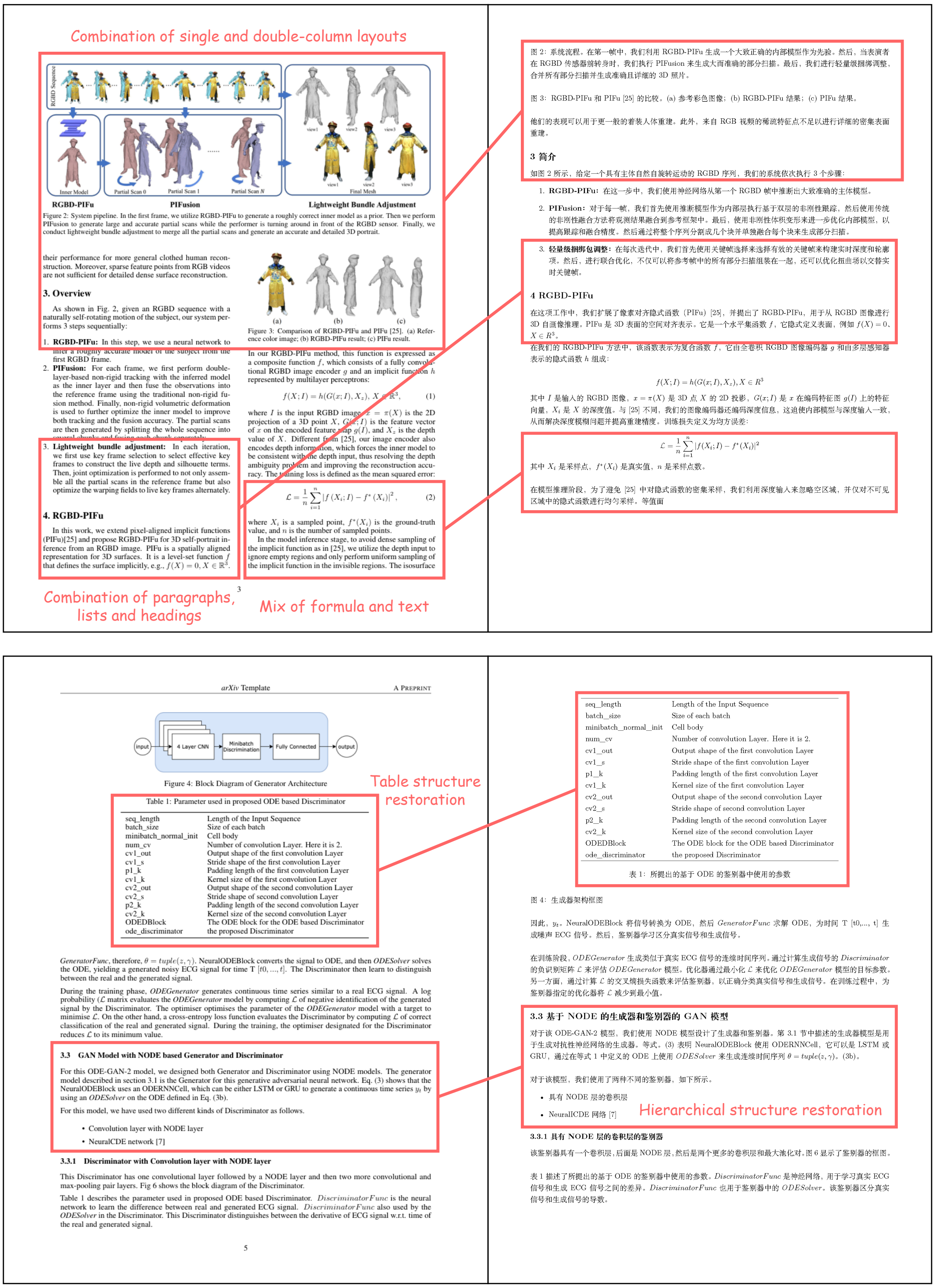


Figure 2: System pipeline. In the first frame, we utilize RGBD-PIFu to generate a roughly correct inner model as a prior. Then we perform PIFusion to generate large and accurate partial scans while the performer is turning around in front of the RGBD sensor. Finally, we conduct lightweight bundle adjustment to merge all the partial scans and generate an accurate and detailed 3D portrait.

their performance for more general clothed human reconstruction. Moreover, sparse feature points from RGB videos are not sufficient for detailed dense surface reconstruction.

3. Overview

As shown in Fig. 2, given an RGBD sequence with a naturally self-rotating motion of the subject, our system performs 3 steps sequentially:

1. **RGBD-PIFu:** In this step, we use a neural network to infer a roughly accurate model of the subject from the first RGBD frame.
2. **PIFusion:** For each frame, we first perform double-layer-based non-rigid tracking with the inferred model as the inner layer and then fuse the observations into the reference frame using the traditional non-rigid fusion method. Finally, non-rigid volumetric deformation is used to further optimize the inner model to improve both tracking and the fusion accuracy. The partial scans are then generated by splitting the whole sequence into several chunks and fusing each chunk separately.
3. **Lightweight bundle adjustment:** In each iteration, we first use key frame selection to select effective key frames to construct the live depth and silhouette terms. Then, joint optimization is performed to not only assemble all the partial scans in the reference frame but also optimize the warping fields to live key frames alternately.

4. RGBD-PIFu

In this work, we extend pixel-aligned implicit functions (PIFu)[25] and propose RGBD-PIFu for 3D self-portrait inference from an RGBD image. PIFu is a spatially aligned representation for 3D surfaces. It is a level-set function f that defines the surface implicitly, e.g., $f(X) = 0, X \in \mathbb{R}^3$.

Figure 3: Comparison of RGBD-PIFu and PIFu [25]. (a) Reference color image; (b) RGBD-PIFu result; (c) PIFu result.

In our RGBD-PIFu method, this function is expressed as a composite function f, which consists of a fully convolutional RGBD image encoder g and an implicit function h represented by multilayer perceptrons:

$$f(X;I) = h(G(x;I), X_z),\ X \in \mathbb{R}^3, \tag{1}$$

where I is the input RGBD image, $x = \pi(X)$ is the 2D projection of a 3D point X, $G(x;I)$ is the feature vector of x on the encoded feature map $g(I)$, and X_z is the depth value of X. Different from [25], our image encoder also encodes depth information, which forces the inner model to be consistent with the depth input, thus resolving the depth ambiguity problem and improving the reconstruction accuracy. The training loss is defined as the mean squared error:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\left|f\left(X_i;I\right) - f^*\left(X_i\right)\right|^2, \tag{2}$$

where X_i is a sampled point, $f^*(X_i)$ is the ground-truth value, and n is the number of sampled points.

In the model inference stage, to avoid dense sampling of the implicit function as in [25], we utilize the depth input to ignore empty regions and only perform uniform sampling of the implicit function in the invisible regions. The isosurface

3

图 2：系统流程。在第一帧中，我们利用 RGBD-PIFu 生成一个大致正确的内部模型作为先验。然后，当表演者在 RGBD 传感器前转身时，我们执行 PIFusion 来生成大而准确的部分扫描。最后，我们进行轻量级捆绑调整，合并所有部分扫描并生成准确且详细的 3D 照片。

图 3：RGBD-PIFu 和 PIFu [25] 的比较。(a) 参考彩色图像；(b) RGBD-PIFu 结果；(c) PIFu 结果。

他们的表现可以用于更一般的着装人体重建。此外，来自 RGB 视频的稀疏特征点不足以进行详细的密集表面重建。

3 简介

如图 2 所示，给定一个具有主体自然自旋转运动的 RGBD 序列，我们的系统依次执行 3 个步骤：

1. **RGBD-PIFu：**在这一步中，我们使用神经网络从第一个 RGBD 帧中推断出大致准确的主体模型。
2. **PIFusion：**对于每一帧，我们首先使用推断模型作为内部层执行基于双层的非刚性跟踪，然后使用传统的非刚性融合方法将观测结果融合到参考框架中。最后，使用非刚性体积变形来进一步优化内部模型，以提高跟踪和融合精度。然后通过将整个序列分割成几个块并单独融合每个块来生成部分扫描。
3. **轻量级捆绑包调整：**在每次迭代中，我们首先使用关键帧选择来选择有效的关键帧来构建实时深度和轮廓项。然后，进行联合优化，不仅可以将参考帧中的所有部分扫描组装在一起，还可以优化扭曲场以交替实时关键帧。

4 RGBD-PIFu

在这项工作中，我们扩展了像素对齐隐式函数（PIFu）[25]，并提出了 RGBD-PIFu，用于从 RGBD 图像进行 3D 自画像推理。PIFu 是 3D 表面的空间对齐表示。它是一个水平集函数 f，它隐式定义表面，例如 $f(X) = 0$、$X \in R^3$。

在我们的 RGBD-PIFu 方法中，该函数表示为复合函数 f，它由全卷积 RGBD 图像编码器 g 和由多层感知器表示的隐式函数 h 组成：

$$f(X;I) = h(G(x;I), X_z), X \in R^3$$

其中 I 是输入的 RGBD 图像，$x = \pi(X)$ 是 3D 点 X 的 2D 投影，$G(x;I)$ 是 x 在编码特征图 $g(I)$ 上的特征向量，X_i 是 X 的深度值。与 [25] 不同，我们的图像编码器还编码深度信息，这迫使内部模型与深度输入一致，从而解决深度模糊问题并提高重建精度。训练损失定义为均方误差：

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}|f(X_i;I) - f^*(X_i)|^2$$

其中 X_i 是采样点，$f^*(X_i)$ 是真实值，n 是采样点数。

在模型推理阶段，为了避免 [25] 中对隐式函数的密集采样，我们利用深度输入来忽略空区域，并仅对不可见区域中的隐式函数进行均匀采样。等值面

arXiv Template　　A PREPRINT

Figure 4: Block Diagram of Generator Architecture

Table 1: Parameter used in proposed ODE based Discriminator

seq_length	Length of the Input Sequence
batch_size	Size of each batch
minibatch_normal_init	Cell body
num_cv	Number of convolution Layer. Here it is 2.
cv1_out	Output shape of the first convolution Layer
cv1_s	Stride shape of the first convolution Layer
p1_k	Padding length of the first convolution Layer
cv1_k	Kernel size of the first convolution Layer
cv2_out	Output shape of the second convolution Layer
cv2_s	Stride shape of second convolution Layer
p2_k	Padding length of the second convolution Layer
cv2_k	Kernel size of the second convolution Layer
ODEDBlock	The ODE block for the ODE based Discriminator
ode_discriminator	the proposed Discriminator

GeneratorFunc, therefore, $\theta = tuple(z, \gamma)$. NeuralODEBlock converts the signal to ODE, and then *ODESolver* solves the ODE, yielding a generated noisy ECG signal for time T $[t0, ..., t]$. The Discriminator then learn to distinguish between the real and the generated signal.

During the training phase, *ODEGenerator* generates continuous time series similar to a real ECG signal. A log probability ($\mathcal{L}$ matrix evaluates the *ODEGenerator* model by computing $\mathcal{L}$ of negative identification of the generated signal by the Discriminator. The optimiser optimises the parameter of the *ODEGenerator* model with a target to minimise $\mathcal{L}$. On the other hand, a cross-entropy loss function evaluates the Discriminator by computing $\mathcal{L}$ of correct classification of the real and generated signal. During the training, the optimiser designated for the Discriminator reduces $\mathcal{L}$ to its minimum value.

3.3 GAN Model with NODE based Generator and Discriminator

For this ODE-GAN-2 model, we designed both Generator and Discriminator using NODE models. The generator model described in section 3.1 is the Generator for this generative adversarial neural network. Eq. (3) shows that the NeuralODEBlock uses an ODERNNCell, which can be either LSTM or GRU to generate a continuous time series y_t by using an *ODESolver* on the ODE defined in Eq. (3b).

For this model, we have used two different kinds of Discriminator as follows.

- Convolution layer with NODE layer
- NeuralCDE network [7]

3.3.1 Discriminator with Convolution layer with NODE layer

This Discriminator has one convolutional layer followed by a NODE layer and then two more convolutional and max-pooling pair layers. Fig 6 shows the block diagram of the Discriminator.

Table 1 describes the parameter used in proposed ODE based Discriminator. $DiscriminatorFunc$ is the neural network to learn the difference between real and generated ECG signal. $DiscriminatorFunc$ also used by the *ODESolver* in the Discriminator. This Discriminator distinguishes between the derivative of ECG signal w.r.t. time of the real and generated signal.

5

seq_length	Length of the Input Sequence
batch_size	Size of each batch
minibatch_normal_init	Cell body
num_cv	Number of convolution Layer. Here it is 2.
cv1_out	Output shape of the first convolution Layer
cv1_s	Stride shape of the first convolution Layer
p1_k	Padding length of the first convolution Layer
cv1_k	Kernel size of the first convolution Layer
cv2_out	Output shape of the second convolution Layer
cv2_s	Stride shape of second convolution Layer
p2_k	Padding length of the second convolution Layer
cv2_k	Kernel size of the second convolution Layer
ODEDBlock	The ODE block for the ODE based Discriminator
ode_discriminator	the proposed Discriminator

表 1：所提出的基于 ODE 的鉴别器中使用的参数

图 4：生成器架构框图

因此，y_t。NeuralODEBlock 将信号转换为 ODE，然后 $GeneratorFunc$ 求解 ODE，为时间 T [t0,..., t] 生成噪声 ECG 信号。然后，鉴别器学习区分真实信号和生成信号。

在训练阶段，$ODEGenerator$ 生成类似于真实 ECG 信号的连续时间序列。通过计算生成信号的 $Discriminator$ 的负识别矩阵 $\mathcal{L}$ 来评估 $ODEGenerator$ 模型。优化器通过最小化 $\mathcal{L}$ 来优化 $ODEGenerator$ 模型的目标参数。另一方面，通过计算 $\mathcal{L}$ 的交叉熵损失函数来评估鉴别器，以正确分类真实信号和生成信号。在训练过程中，为鉴别器指定的优化器将 $\mathcal{L}$ 减少到最小值。

3.3 基于 NODE 的生成器和鉴别器的 GAN 模型

对于该 ODE-GAN-2 模型，我们使用 NODE 模型设计了生成器和鉴别器。第 3.1 节中描述的生成器模型是用于生成对抗性神经网络的生成器。等式。(3) 表明 NeuralODEBlock 使用 ODERNNCell，它可以是 LSTM 或 GRU，通过在等式 1 中定义的 ODE 上使用 $ODESolver$ 来生成连续时间序列 $\theta = tuple(z, \gamma)$。(3b)。

对于该模型，我们使用了两种不同的鉴别器，如下所示。

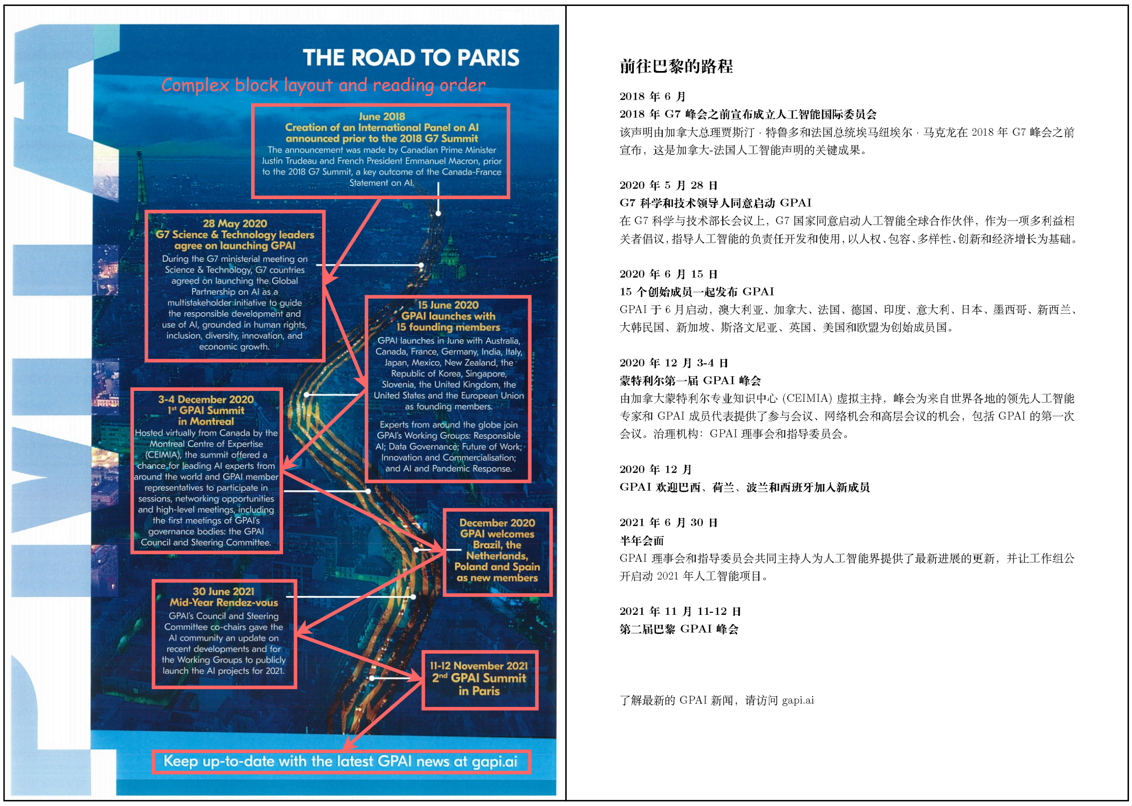
- 具有 NODE 层的卷积层
- NeuralICDE 网络 [7]

3.3.1 具有 NODE 层的卷积层的鉴别器

该鉴别器具有一个卷积层，后面是 NODE 层，然后是两个更多的卷积层和最大池化对。图 6 显示了鉴别器的框图。

表 1 描述了所提出的基于 ODE 的鉴别器中使用的参数。$DiscriminatorFunc$ 是神经网络，用于学习真实 ECG 信号和生成 ECG 信号之间的差异。$DiscriminatorFunc$ 也用于鉴别器中的 $ODESolver$。该鉴别器区分真实信号和生成信号的导数。

Figure 8: The output samples of Qwen2-VL (after fine-tuning with SSR in the main experiment) on the DoTA test set (Academic Articles). For each image pair, the left is the input document image, and the right is the output translations in markdown format after rendering. It is better to zoom in for a clearer view.



THE ROAD TO PARIS

June 2018
Creation of an International Panel on AI announced prior to the 2018 G7 Summit
The announcement was made by Canadian Prime Minister Justin Trudeau and French President Emmanuel Macron, prior to the 2018 G7 Summit, a key outcome of the Canada-France Statement on AI.

28 May 2020
G7 Science & Technology leaders agree on launching GPAI
During the G7 ministerial meeting on Science & Technology, G7 countries agreed on launching the Global Partnership on AI as a multistakeholder initiative to guide the responsible development and use of AI, grounded in human rights, inclusion, diversity, innovation, and economic growth.

15 June 2020
GPAI launches with 15 founding members
GPAI launches in June with Australia, Canada, France, Germany, India, Italy, Japan, Mexico, New Zealand, the Republic of Korea, Singapore, Slovenia, the United Kingdom, the United States and the European Union as founding members.

Experts from around the globe join GPAI's Working Groups: Responsible AI; Data Governance; Future of Work; Innovation and Commercialisation; and AI and Pandemic Response.

3-4 December 2020
1st GPAI Summit in Montreal
Hosted virtually from Canada by the Montreal Centre of Expertise (CEIMIA), the summit offered a chance for leading AI experts from around the world and GPAI member representatives to participate in sessions, networking opportunities and high-level meetings, including the first meetings of GPAI's governance bodies: the GPAI Council and Steering Committee.

December 2020
GPAI welcomes Brazil, the Netherlands, Poland and Spain as new members

30 June 2021
Mid-Year Rendez-vous
GPAI's Council and Steering Committee co-chairs gave the AI community an update on recent developments and for the Working Groups to publicly launch the AI projects for 2021.

11-12 November 2021
2nd GPAI Summit in Paris

Keep up-to-date with the latest GPAI news at gapi.ai

前往巴黎的路程

2018 年 6 月
2018 年 G7 峰会之前宣布成立人工智能国际委员会
该声明由加拿大总理贾斯汀 · 特鲁多和法国总统埃马纽埃尔 · 马克龙在 2018 年 G7 峰会之前宣布，这是加拿大-法国人工智能声明的关键成果。

2020 年 5 月 28 日
G7 科学和技术领导人同意启动 GPAI
在 G7 科学与技术部长会议上，G7 国家同意启动人工智能全球合作伙伴，作为一项多利益相关者倡议，指导人工智能的负责任开发和使用，以人权、包容、多样性、创新和经济增长为基础。

2020 年 6 月 15 日
15 个创始成员一起发布 GPAI
GPAI 于 6 月启动，澳大利亚、加拿大、法国、德国、印度、意大利、日本、墨西哥、新西兰、大韩民国、新加坡、斯洛文尼亚、英国、美国和欧盟为创始成员国。

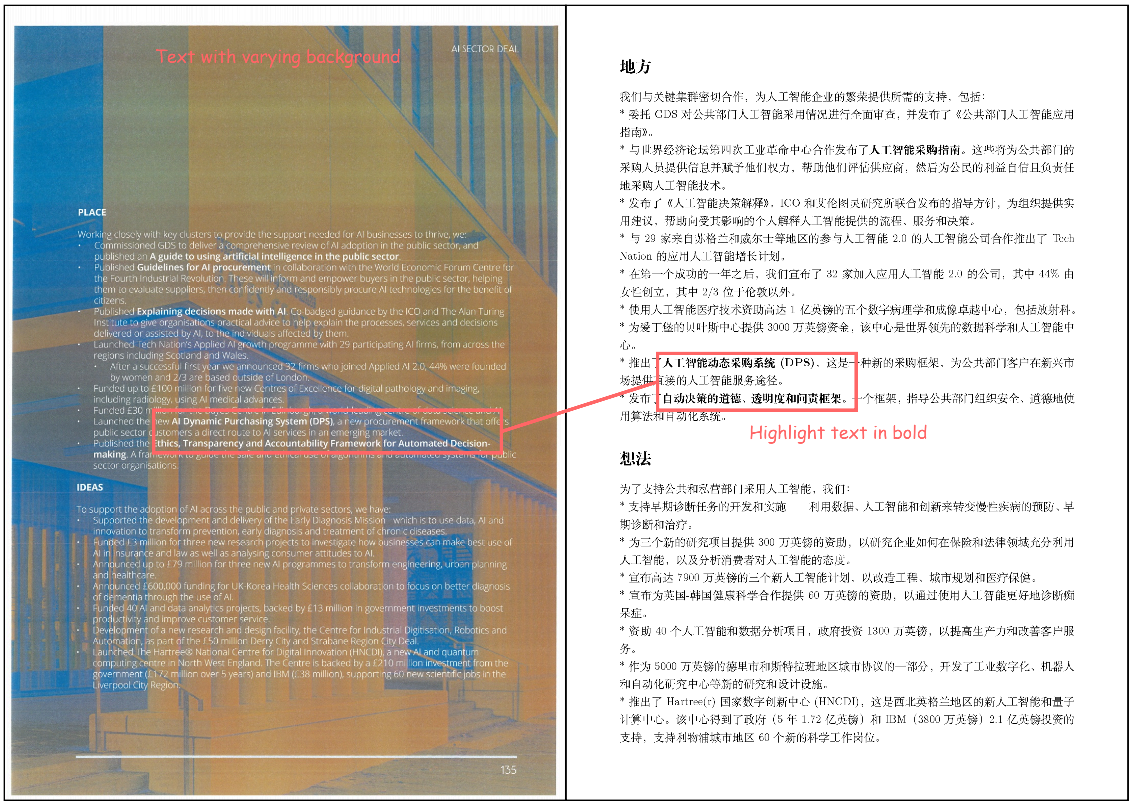
2020 年 12 月 3-4 日
蒙特利尔第一届 GPAI 峰会
由加拿大蒙特利尔专业知识中心 (CEIMIA) 虚拟主持，峰会为来自世界各地的领先人工智能专家和 GPAI 成员代表提供了参与会议、网络机会和高层会议的机会，包括 GPAI 的第一次会议。治理机构：GPAI 理事会和指导委员会。

2020 年 12 月
GPAI 欢迎巴西、荷兰、波兰和西班牙加入新成员

2021 年 6 月 30 日
半年会面
GPAI 理事会和指导委员会共同主持人为人工智能界提供了最新进展的更新，并让工作组公开启动 2021 年人工智能项目。

2021 年 11 月 11-12 日
第二届巴黎 GPAI 峰会

了解最新的 GPAI 新闻，请访问 gapi.ai



PLACE

Working closely with key clusters to provide the support needed for AI businesses to thrive, we:
- Commissioned GDS to deliver a comprehensive review of AI adoption in the public sector, and published an **A guide to using artificial intelligence in the public sector**.
- Published **Guidelines for AI procurement** in collaboration with the World Economic Forum Centre for the Fourth Industrial Revolution. These will inform and empower buyers in the public sector, helping them to evaluate suppliers, then confidently and responsibly procure AI technologies for the benefit of citizens.
- Published **Explaining decisions made with AI**. Co-badged guidance by the ICO and The Alan Turing Institute to give organisations practical advice to help explain the processes, services and decisions delivered or assisted by AI, to the individuals affected by them.
- Launched Tech Nation's Applied AI growth programme with 29 participating AI firms, from across the regions including Scotland and Wales.
  - After a successful first year we announced 32 firms who joined Applied AI 2.0, 44% were founded by women and 2/3 are based outside of London.
- Funded up to £100 million for five new Centres of Excellence for digital pathology and imaging, including radiology, using AI medical advances.
- Funded £30 million for the Bayes Centre in Edinburgh, a world-leading centre of data science and AI.
- Launched the new **AI Dynamic Purchasing System (DPS)**, a new procurement framework that offers public sector customers a direct route to AI services in an emerging market.
- Published the **Ethics, Transparency and Accountability Framework for Automated Decision-making**. A framework to guide the safe and ethical use of algorithms and automated systems for public sector organisations.

IDEAS

To support the adoption of AI across the public and private sectors, we have:
- Supported the development and delivery of the Early Diagnosis Mission - which is to use data, AI and innovation to transform prevention, early diagnosis and treatment of chronic diseases.
- Funded £3 million for three new research projects to investigate how businesses can make best use of AI in insurance and law as well as analysing consumer attitudes to AI.
- Announced up to £79 million for three new AI programmes to transform engineering, urban planning and healthcare.
- Announced £600,000 funding for UK-Korea Health Sciences collaboration to focus on better diagnosis of dementia through the use of AI.
- Funded 40 AI and data analytics projects, backed by £13 million in government investments to boost productivity and improve customer service.
- Development of a new research and design facility, the Centre for Industrial Digitisation, Robotics and Automation, as part of the £50 million Derry City and Strabane Region City Deal.
- Launched The Hartree® National Centre for Digital Innovation (HNCDI), a new AI and quantum computing centre in North West England. The Centre is backed by a £210 million investment from the government (£172 million over 5 years) and IBM (£38 million), supporting 60 new scientific jobs in the Liverpool City Region.

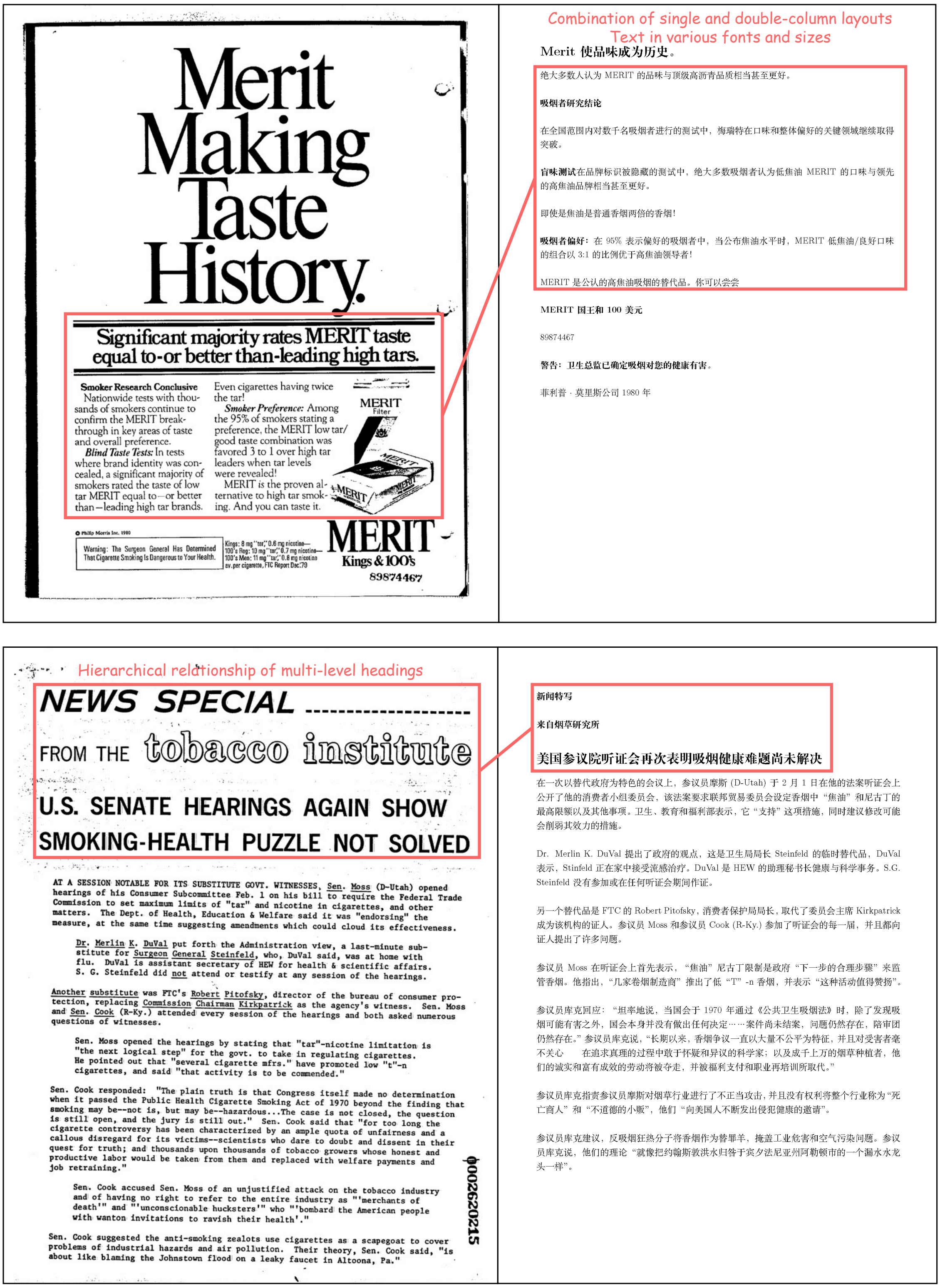
地方

我们与关键集群密切合作，为人工智能企业的繁荣提供所需的支持，包括：
* 委托 GDS 对公共部门人工智能采用情况进行全面审查，并发布了《公共部门人工智能应用指南》。
* 与世界经济论坛第四次工业革命中心合作发布了**人工智能采购指南**。这些将为公共部门的采购人员提供信息并赋予他们权力，帮助他们评估供应商，然后为公民的利益自信且负责任地采购人工智能技术。
* 发布了《人工智能决策解释》。ICO 和艾伦图灵研究所联合发布的指导方针，为组织提供实用建议，帮助向受其影响的个人解释人工智能提供的流程、服务和决策。
* 与 29 家来自苏格兰和威尔士等地区的参与人工智能 2.0 的人工智能公司合作推出了 Tech Nation 的应用人工智能增长计划。
* 在第一个成功的一年之后，我们宣布了 32 家加入应用人工智能 2.0 的公司，其中 44% 由女性创立，其中 2/3 位于伦敦以外。
* 使用人工智能医疗技术资助高达 1 亿英镑的五个数字病理学和成像卓越中心，包括放射科。
* 为爱丁堡的贝叶斯中心提供 3000 万英镑资金，该中心是世界领先的数据科学和人工智能中心。
* 推出了**人工智能动态采购系统** (DPS)，这是一种新的采购框架，为公共部门客户在新兴市场提供直接的人工智能服务途径。
* 发布了**自动决策的道德、透明度和问责框架**。一个框架，指导公共部门组织安全、道德地使用算法和自动化系统。

想法

为了支持公共和私营部门采用人工智能，我们：
* 支持早期诊断任务的开发和实施　　利用数据、人工智能和创新来转变慢性疾病的预防、早期诊断和治疗。
* 为三个新的研究项目提供 300 万英镑的资助，以研究企业如何在保险和法律领域充分利用人工智能，以及分析消费者对人工智能的态度。
* 宣布高达 7900 万英镑的三个新人工智能计划，以改造工程、城市规划和医疗保健。
* 宣布为英国-韩国健康科学合作提供 60 万英镑的资助，以通过使用人工智能更好地诊断痴呆症。
* 资助 40 个人工智能和数据分析项目，政府投资 1300 万英镑，以提高生产力和改善客户服务。
* 作为 5000 万英镑的德里市和斯特拉班地区城市协议的一部分，开发了工业数字化、机器人和自动化研究中心等新的研究和设计设施。
* 推出了 Hartree(r) 国家数字创新中心 (HNCDI)，这是西北英格兰地区的新人工智能和量子计算中心。该中心得到了政府（5 年 1.72 亿英镑）和 IBM（3800 万英镑）2.1 亿英镑投资的支持，支持利物浦城市地区 60 个新的科学工作岗位。

Figure 9: The output samples of Qwen2-VL (after fine-tuning with SSR in the main experiment) on the DITrans test set (Political Report). For each image pair, the left is the input document image, and the right is the output translations in markdown format after rendering. It is better to zoom in for a clearer view.



Combination of single and double-column layouts
Text in various fonts and sizes
Hierarchical relationship of multi-level headings

Merit Making Taste History.

Significant majority rates MERIT taste equal to-or better than-leading high tars.

Smoker Research Conclusive

Nationwide tests with thousands of smokers continue to confirm the MERIT breakthrough in key areas of taste and overall preference.

Blind Taste Tests: In tests where brand identity was concealed, a significant majority of smokers rated the taste of low tar MERIT equal to—or better than—leading high tar brands.

Even cigarettes having twice the tar!

Smoker Preference: Among the 95% of smokers stating a preference, the MERIT low tar/good taste combination was favored 3 to 1 over high tar leaders when tar levels were revealed!

MERIT is the proven alternative to high tar smoking. And you can taste it.

MERIT
Filter

© Philip Morris Inc. 1980

Warning: The Surgeon General Has Determined That Cigarette Smoking Is Dangerous to Your Health.

Kings: 8 mg "tar," 0.6 mg nicotine—
100's Reg: 10 mg "tar," 0.7 mg nicotine—
100's Men: 11 mg "tar," 0.8 mg nicotine
av. per cigarette, FTC Report Dec.'79

MERIT
Kings & 100's

89874467

Merit 使品味成为历史。

绝大多数人认为 MERIT 的品味与顶级高沥青品质相当甚至更好。

吸烟者研究结论

在全国范围内对数千名吸烟者进行的测试中，梅瑞特在口味和整体偏好的关键领域继续取得突破。

盲味测试在品牌标识被隐藏的测试中，绝大多数吸烟者认为低焦油 MERIT 的口味与领先的高焦油品牌相当甚至更好。

即使是焦油是普通香烟两倍的香烟！

吸烟者偏好： 在 95% 表示偏好的吸烟者中，当公布焦油水平时，MERIT 低焦油/良好口味的组合以 3:1 的比例优于高焦油领导者！

MERIT 是公认的高焦油吸烟的替代品。你可以尝尝

MERIT 国王和 100 美元

89874467

警告：卫生总监已确定吸烟对您的健康有害。

菲利普 · 莫里斯公司 1980 年

NEWS SPECIAL

FROM THE tobacco institute

U.S. SENATE HEARINGS AGAIN SHOW SMOKING-HEALTH PUZZLE NOT SOLVED

AT A SESSION NOTABLE FOR ITS SUBSTITUTE GOVT. WITNESSES, Sen. Moss (D-Utah) opened hearings of his Consumer Subcommittee Feb. 1 on his bill to require the Federal Trade Commission to set maximum limits of "tar" and nicotine in cigarettes, and other matters. The Dept. of Health, Education & Welfare said it was "endorsing" the measure, at the same time suggesting amendments which could cloud its effectiveness.

Dr. Merlin K. DuVal put forth the Administration view, a last-minute substitute for Surgeon General Steinfeld, who, DuVal said, was at home with flu. DuVal is assistant secretary of HEW for health & scientific affairs. S. G. Steinfeld did not attend or testify at any session of the hearings.

Another substitute was FTC's Robert Pitofsky, director of the bureau of consumer protection, replacing Commission Chairman Kirkpatrick as the agency's witness. Sen. Moss and Sen. Cook (R-Ky.) attended every session of the hearings and both asked numerous questions of witnesses.

Sen. Moss opened the hearings by stating that "tar"-nicotine limitation is "the next logical step" for the govt. to take in regulating cigarettes. He pointed out that "several cigarette mfrs." have promoted low "t"-n cigarettes, and said "that activity is to be commended."

Sen. Cook responded: "The plain truth is that Congress itself made no determination when it passed the Public Health Cigarette Smoking Act of 1970 beyond the finding that smoking may be--not is, but may be--hazardous...The case is not closed, the question is still open, and the jury is still out." Sen. Cook said that "for too long the cigarette controversy has been characterized by an ample quota of unfairness and a callous disregard for its victims--scientists who dare to doubt and dissent in their quest for truth; and thousands upon thousands of tobacco growers whose honest and productive labor would be taken from them and replaced with welfare payments and job retraining."

Sen. Cook accused Sen. Moss of an unjustified attack on the tobacco industry and of having no right to refer to the entire industry as "'merchants of death'" and "'unconscionable hucksters'" who "'bombard the American people with wanton invitations to ravish their health'."

Sen. Cook suggested the anti-smoking zealots use cigarettes as a scapegoat to cover problems of industrial hazards and air pollution. Their theory, Sen. Cook said, "is about like blaming the Johnstown flood on a leaky faucet in Altoona, Pa."

00262015

新闻特写

来自烟草研究所

美国参议院听证会再次表明吸烟健康难题尚未解决

在一次以替代政府为特色的会议上，参议员摩斯 (D-Utah) 于 2 月 1 日在他的法案听证会上公开了他的消费者小组委员会，该法案要求联邦贸易委员会设定香烟中“焦油”和尼古丁的最高限额以及其他事项。卫生、教育和福利部表示，它“支持”这项措施，同时建议修改可能会削弱其效力的措施。

Dr. Merlin K. DuVal 提出了政府的观点，这是卫生局局长 Steinfeld 的临时替代品，DuVal 表示，Stinfeld 正在家中接受流感治疗。DuVal 是 HEW 的助理秘书长健康与科学事务。S.G. Steinfeld 没有参加或在任何听证会期间作证。

另一个替代品是 FTC 的 Robert Pitofsky，消费者保护局局长，取代了委员会主席 Kirkpatrick 成为该机构的证人。参议员 Moss 和参议员 Cook (R-Ky.) 参加了听证会的每一届，并且都向证人提出了许多问题。

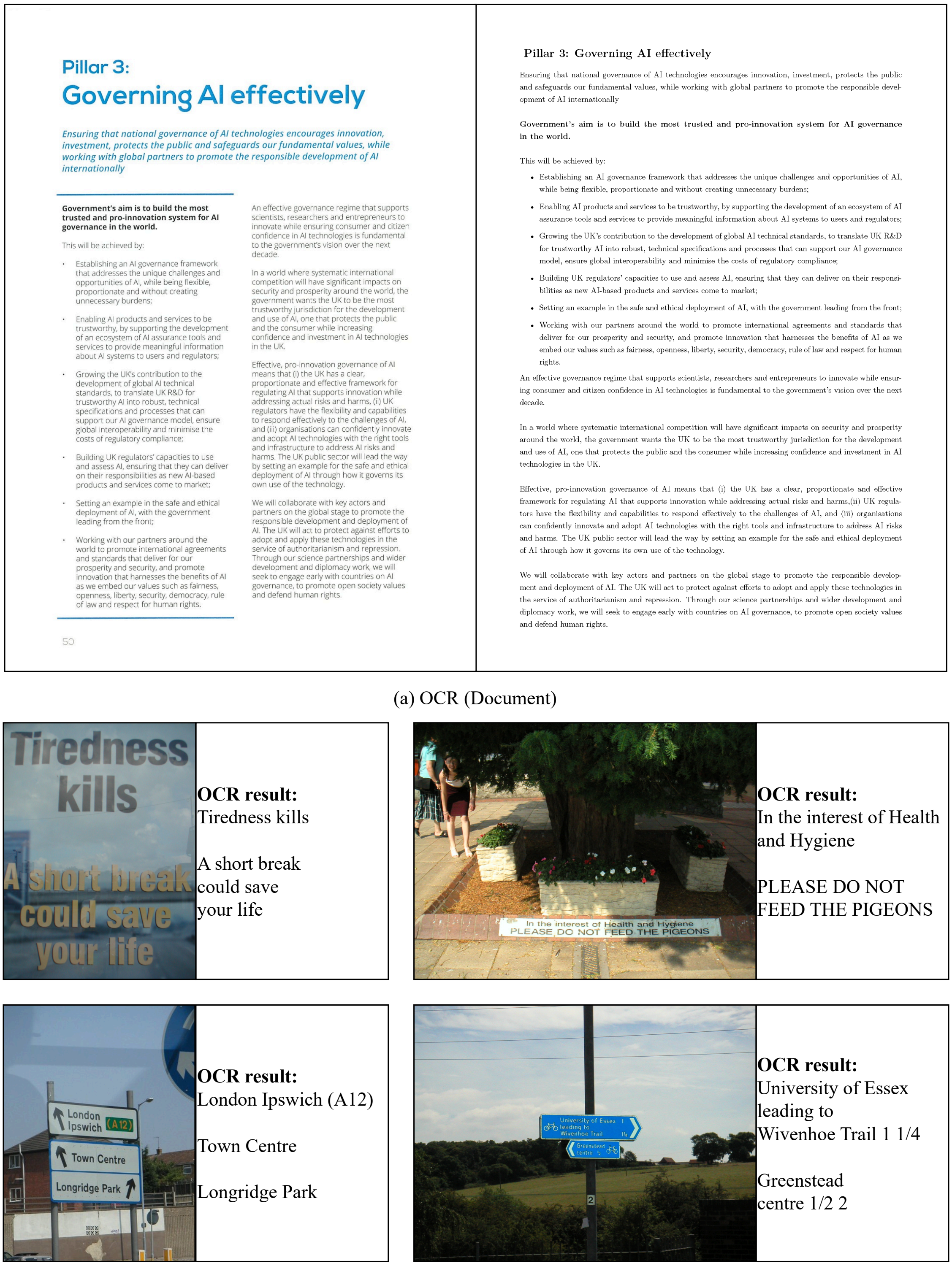
参议员 Moss 在听证会上首先表示，“焦油”尼古丁限制是政府“下一步的合理步骤”来监管香烟。他指出，“几家卷烟制造商”推出了低“T”-n 香烟，并表示“这种活动值得赞扬”。

参议员库克回应：“坦率地说，当国会于 1970 年通过《公共卫生吸烟法》时，除了发现吸烟可能有害之外，国会本身并没有做出任何决定……案件尚未结案，问题仍然存在，陪审团仍然存在。”参议员库克说，“长期以来，香烟争议一直以大量不公平为特征，并且对受害者毫不关心　　在追求真理的过程中敢于怀疑和异议的科学家；以及成千上万的烟草种植者，他们的诚实和富有成效的劳动将被夺走，并被福利支付和职业再培训所取代。”

参议员库克指责参议员摩斯对烟草行业进行了不正当攻击，并且没有权利将整个行业称为“死亡商人”和“不道德的小贩”，他们“向美国人不断发出侵犯健康的邀请”。

参议员库克建议，反吸烟狂热分子将香烟作为替罪羊，掩盖工业危害和空气污染问题。参议员库克说，他们的理论“就像把约翰斯敦洪水归咎于宾夕法尼亚州阿勒顿市的一个漏水水龙头一样”。

Figure 10: The output samples of Qwen2-VL (after fine-tuning with SSR in the main experiment) on the DITrans test set (Ads & News). For each image pair, the left is the input document image, and the right is the output translations in markdown format after rendering. It is better to zoom in for a clearer view.



Pillar 3:

Governing AI effectively

Ensuring that national governance of AI technologies encourages innovation, investment, protects the public and safeguards our fundamental values, while working with global partners to promote the responsible development of AI internationally

Government's aim is to build the most trusted and pro-innovation system for AI governance in the world.

This will be achieved by:

- Establishing an AI governance framework that addresses the unique challenges and opportunities of AI, while being flexible, proportionate and without creating unnecessary burdens;
- Enabling AI products and services to be trustworthy, by supporting the development of an ecosystem of AI assurance tools and services to provide meaningful information about AI systems to users and regulators;
- Growing the UK's contribution to the development of global AI technical standards, to translate UK R&D for trustworthy AI into robust, technical specifications and processes that can support our AI governance model, ensure global interoperability and minimise the costs of regulatory compliance;
- Building UK regulators' capacities to use and assess AI, ensuring that they can deliver on their responsibilities as new AI-based products and services come to market;
- Setting an example in the safe and ethical deployment of AI, with the government leading from the front;
- Working with our partners around the world to promote international agreements and standards that deliver for our prosperity and security, and promote innovation that harnesses the benefits of AI as we embed our values such as fairness, openness, liberty, security, democracy, rule of law and respect for human rights.

An effective governance regime that supports scientists, researchers and entrepreneurs to innovate while ensuring consumer and citizen confidence in AI technologies is fundamental to the government's vision over the next decade.

In a world where systematic international competition will have significant impacts on security and prosperity around the world, the government wants the UK to be the most trustworthy jurisdiction for the development and use of AI, one that protects the public and the consumer while increasing confidence and investment in AI technologies in the UK.

Effective, pro-innovation governance of AI means that (i) the UK has a clear, proportionate and effective framework for regulating AI that supports innovation while addressing actual risks and harms, (ii) UK regulators have the flexibility and capabilities to respond effectively to the challenges of AI, and (iii) organisations can confidently innovate and adopt AI technologies with the right tools and infrastructure to address AI risks and harms. The UK public sector will lead the way by setting an example for the safe and ethical deployment of AI through how it governs its own use of the technology.

We will collaborate with key actors and partners on the global stage to promote the responsible development and deployment of AI. The UK will act to protect against efforts to adopt and apply these technologies in the service of authoritarianism and repression. Through our science partnerships and wider development and diplomacy work, we will seek to engage early with countries on AI governance, to promote open society values and defend human rights.

50

Pillar 3: Governing AI effectively

Ensuring that national governance of AI technologies encourages innovation, investment, protects the public and safeguards our fundamental values, while working with global partners to promote the responsible development of AI internationally

Government's aim is to build the most trusted and pro-innovation system for AI governance in the world.

This will be achieved by:

- Establishing an AI governance framework that addresses the unique challenges and opportunities of AI, while being flexible, proportionate and without creating unnecessary burdens;
- Enabling AI products and services to be trustworthy, by supporting the development of an ecosystem of AI assurance tools and services to provide meaningful information about AI systems to users and regulators;
- Growing the UK's contribution to the development of global AI technical standards, to translate UK R&D for trustworthy AI into robust, technical specifications and processes that can support our AI governance model, ensure global interoperability and minimise the costs of regulatory compliance;
- Building UK regulators' capacities to use and assess AI, ensuring that they can deliver on their responsibilities as new AI-based products and services come to market;
- Setting an example in the safe and ethical deployment of AI, with the government leading from the front;
- Working with our partners around the world to promote international agreements and standards that deliver for our prosperity and security, and promote innovation that harnesses the benefits of AI as we embed our values such as fairness, openness, liberty, security, democracy, rule of law and respect for human rights.

An effective governance regime that supports scientists, researchers and entrepreneurs to innovate while ensuring consumer and citizen confidence in AI technologies is fundamental to the government's vision over the next decade.

In a world where systematic international competition will have significant impacts on security and prosperity around the world, the government wants the UK to be the most trustworthy jurisdiction for the development and use of AI, one that protects the public and the consumer while increasing confidence and investment in AI technologies in the UK.

Effective, pro-innovation governance of AI means that (i) the UK has a clear, proportionate and effective framework for regulating AI that supports innovation while addressing actual risks and harms,(ii) UK regulators have the flexibility and capabilities to respond effectively to the challenges of AI, and (iii) organisations can confidently innovate and adopt AI technologies with the right tools and infrastructure to address AI risks and harms. The UK public sector will lead the way by setting an example for the safe and ethical deployment of AI through how it governs its own use of the technology.

We will collaborate with key actors and partners on the global stage to promote the responsible development and deployment of AI. The UK will act to protect against efforts to adopt and apply these technologies in the service of authoritarianism and repression. Through our science partnerships and wider development and diplomacy work, we will seek to engage early with countries on AI governance, to promote open society values and defend human rights.

(a) OCR (Document)

OCR result:
Tiredness kills

A short break
could save
your life

OCR result:
In the interest of Health and Hygiene

PLEASE DO NOT FEED THE PIGEONS

OCR result:
London Ipswich (A12)

Town Centre

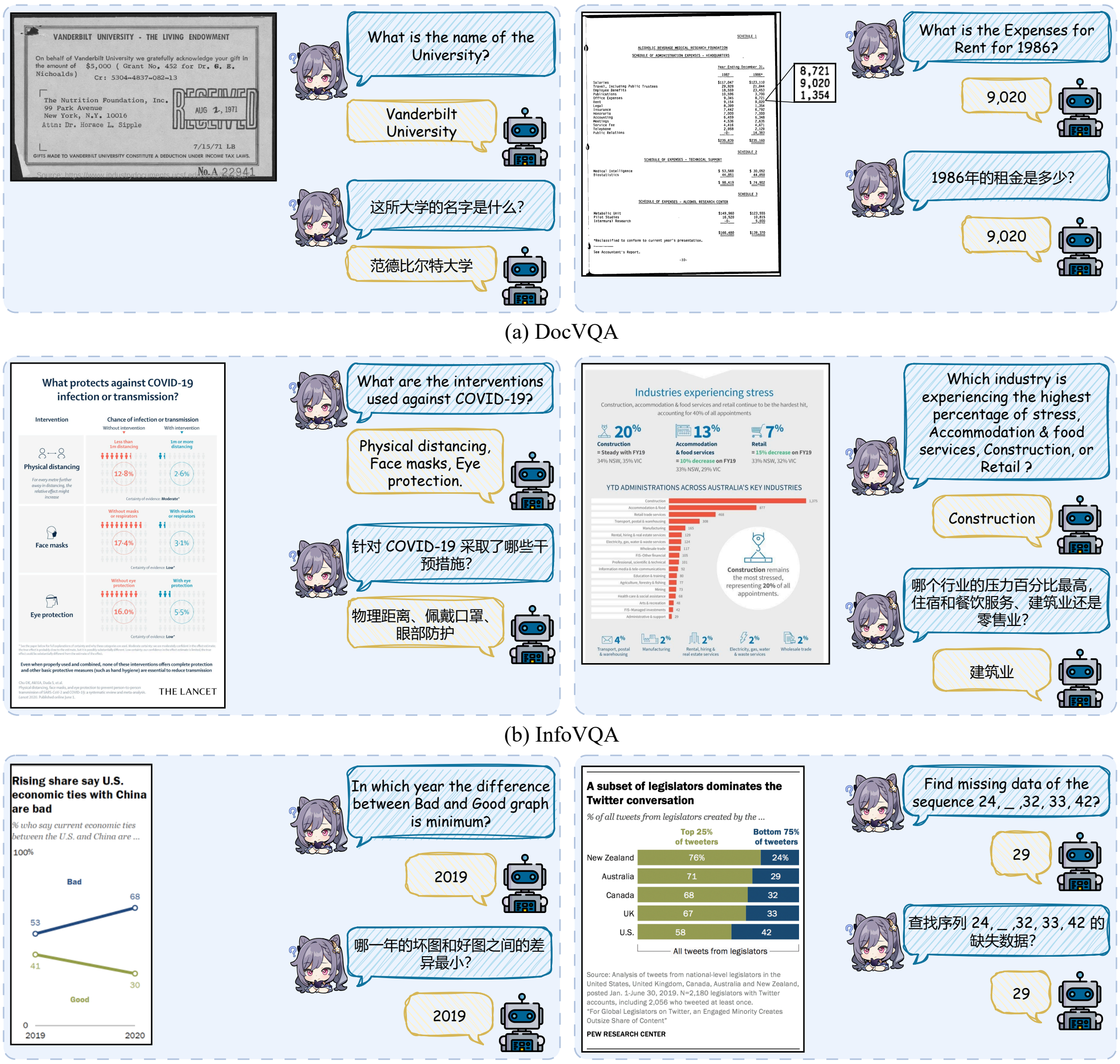
Longridge Park

OCR result:
University of Essex leading to
Wivenhoe Trail 1 1/4

Greenstead centre 1/2 2

(b) OCR (Scene)

Figure 11: The output samples of Qwen2-VL (after fine-tuning with SSR in the main experiment) on the OCR test. It is better to zoom in for a clearer view.


What is the name of the University?
Vanderbilt University
这所大学的名字是什么?
范德比尔特大学
What is the Expenses for Rent for 1986?
9,020
1986年的租金是多少?
9,020
What are the interventions used against COVID-19?
Physical distancing, Face masks, Eye protection.
针对 COVID-19 采取了哪些干预措施?
物理距离、佩戴口罩、眼部防护
Which industry is experiencing the highest percentage of stress, Accommodation & food services, Construction, or Retail ?
Construction
哪个行业的压力百分比最高，住宿和餐饮服务、建筑业还是零售业?
建筑业
In which year the difference between Bad and Good graph is minimum?
2019
哪一年的坏图和好图之间的差异最小?
2019
Find missing data of the sequence 24, _ ,32, 33, 42?
29
查找序列 24, _ ,32, 33, 42 的缺失数据?
29

(a) DocVQA

(b) InfoVQA

(c) ChartVQA

Figure 12: The output samples of Qwen2-VL (after fine-tuning with SSR in the main experiment) on the VQA test. For each document image containing English text, although our model is only trained on the DIMT dataset without utilizing the VQA dataset, it can still respond in the language corresponding to the question. It is better to zoom in for a clearer view.