DIMT25@ICDAR2025：HW-TSC 的端到端文档图像机器翻译系统借助大型视觉语言模型

Zhanglin Wu, Tengfei Song, Ning Xie, Weidong Zhang,
Pengfei Li, Shuang Wu, Chong Li, Junhao Zhu, and Hao Yang

Huawei Translation Service Center, Nanjing, China
{ wuzhanglin2, songtengfei2, nicolas.xie, zhangweidong17,
lipengfei203, wushuang42, august.li, zhujunhao, yanghao30 } @huawei.com

Abstract. 本文介绍了华为翻译服务中心（HW-TSC）为第 19 届国际文档分析与识别大会（DIMT25@ICDAR2025）上的“复杂布局文档图像端到端机器翻译”竞赛所提出的技术解决方案。通过利用最先进的开源大型视觉语言模型（LVLM），我们引入了一个将多任务学习与感知链式思维相结合的训练框架，以开发一个综合的端到端文档翻译系统。在推理阶段，我们应用最小贝叶斯解码和后处理策略，以进一步增强系统的翻译能力。我们的解决方案在一个统一框架内独特地解决了基于 OCR 和无 OCR 的文档图像翻译任务。本文系统地详细介绍了训练方法、推理策略、LVLM 基本模型、训练数据、实验设置和结果，展示了文档图像机器翻译的有效方法。

1 介绍

文件图像机器翻译 (DIMT) [1,2] 结合了文本识别和翻译来处理像手册、报告和档案这样的多语言文件 [3] 。尽管深度学习 [4] 取得了进展，但建立一个能够高精度处理各种布局的端到端 DIMT 系统仍然具有挑战性。

为了促进该领域的创新，第十九届国际文档分析与识别会议（ICDAR 2025）专门设立了一项名为“复杂布局的端到端文档图像机器翻译”的竞赛。为应对这一挑战，我们团队（HW-TSC）开发了一种基于最先进的大型视觉-语言模型（LVLMs）的创新框架 [5,6] 。该框架结合了多任务学习（MTL）[7] 和感知链式思维（PCOT）[8] 训练方法，使模型能够同时理解视觉布局和语言内容。在推理阶段，系统进一步采用最小贝叶斯风险（MBR）解码 [9] 和后处理策略来优化输出质量。

本研究的主要创新在于将基于 OCR 和非 OCR 的翻译任务融合到一个统一框架内，消除了构建独立流程的需求。这一设计显著增强了系统的适应性，使其能够灵活处理在现实场景中光学字符识别（OCR）[10] 可能适用或不适用的情况。对测试集的评估验证了该方法在处理复杂结构文档时的有效性。

本文系统地展示了完整的技术解决方案，强调可重复性和透明性。后续部分详细阐述了：1）训练方法和推理策略，2）训练数据的组成和 LVLM 基础模型，3）实验设置和结果。这种结构化的呈现有助于全面理解我们的研究贡献和创新。

2 方法描述

2.1 训练方法

随着大型语言模型（LLM）[11,12] 技术的进步，其“继续预训练（CPT）[13] - 监督微调（SFT）[14] - 从人类反馈中强化学习（RLHF）[15] ”的三阶段训练范式已成功应用于 LVLMs [16,17] 。这种范式使开源 LVLMs 在跨模态任务中实现了卓越的性能 [18] 。基于此，我们提出了一种创新的训练策略：将 MTL 与 PCOT 相结合，在最先进的 LVLM [6] 上进行 SFT，以提高其在 DIMT25 任务上的性能。图 1 展示了训练数据的格式。

多任务学习 MTL [7] 是一种机器学习范式 [19] ，通过在不同任务间共享表示来增强模型的泛化能力。具体来说，MTL 在单个模型中同时优化多个相关任务，利用潜在的任务间相关性，有效缓解传统单任务学习 [20] 中常见的问题，例如由于数据隔离导致的模型过拟合或表示能力有限。尤其是在 LVLM 的 SFT 过程中，MTL 通过参数共享机制和联合损失函数优化促进任务间的协同训练。该方法不仅增强了模型提取多模态特征的能力，还显著提升了在复杂场景中语义理解的鲁棒性。

PCOT 链式思维（Chain-of-Thought, CoT）[8] 方法通过模拟人类逐步推理过程，在复杂的认知任务（如数学推理、逻辑推导和多模态理解）中展现出显著优势。与传统的端到端训练范式相比，CoT 强调显式的逐步推理机制，这种结构化处理增强了模型决策的透明性。为了解决 DIMT25 任务的独特挑战，我们将标准的 CoT 扩展为 PCOT。它采用分层处理流程：第一阶段专注于图像中文本内容的精确检测和识别，而第二阶段执行专门的跨语言转换。这种两阶段设计有效解决了传统方法在混合文本-图像场景中的语义碎片化问题。通过在视觉感知和语言理解之间建立深度耦合机制，它显著提高了翻译结果的准确性和一致性。

2.2 推理策略

MBR 解码 MBR 解码 [9] 的核心目标是通过优化期望效用函数来选择最优输出，该函数量化候选假设与参考文本之间的相似性。之前的研究 [21–23] 已经证明了 MBR 解码在机器翻译任务中的有效性，在这些任务中通常使用自动评估指标 [24, 25] 作为相似性度量。我们将这种方法应用于 DIMT25 任务。具体来说，我们同时收集由两种解码策略生成的候选输出，包括来自束搜索的确定性输出和通过温度和核采样（附带 $t = 0.7$ 和 $p = 0.95$ ）生成的 10 个多样样本。随后，我们使用 BLEU [26] 计算候选集内的成对相似性评分，并选择具有最高相似性评分的样本作为最终系统输出。

后处理 对模型输出的分析表明，当处理图像中重复出现的特殊符号（如-，…，_，* 等）或内容繁多的表格时，模型往往会过度翻译。为了解决这个问题，我们制定了两个简单的处理规则：将超过 10 个连续出现的特殊符号减少到 10 个，并删除对过于复杂表格的翻译输出。此外，对于中文翻译结果，我们在进行 jieba 分词后实行连续空格规范化，将多个连续空格替换为单个空格，以使输出更符合人类的阅读习惯。

3 实现细节

3.1 训练数据

如表 2 所示，DIMT25 竞赛的多模态数据集是我们用于模型训练和评估的数据来源，其中包含两个独立的轨道用于不同的任务场景。

3.2 LVLM 基础模型

我们选择了 InternVL2.5-8B-MPO [1]（大型模型）和 InternVL2.5-1B-MPO [2]（小型模型）作为我们不同模型规模段的 LVLM 基准模型，这两者在开放源码的 LVLM 评估基准如 MOTBench [27] 上表现优异，并在它们上进行 SFT。

4 实验

4.1 设置

我们使用开源的 InternVL 框架 [3] 对 InternVL2.5-MPO 模型进行全参数微调，利用 DeepSpeed 的 zero3-offload 技术加速训练过程。详细的训练配置如下：我们在 8 个 GPU 上进行并行训练，每个 GPU 的批量大小为 4，并且梯度累积步数为 4。初始学习率设定为 4e-5。最大序列长度配置为 8,192 个标记符，模型训练 5 个周期。

4.2 结果

表 1 展示了三个主要发现：(1) 与单任务直接 SFT 相比，MTL-PCOT 结合 SFT 方法在 LVLM 的 DIMT25 任务上表现出更显著的性能提升；(2) MBR 和后处理推理策略提供了互补的性能增强；(3) 模型能力随着参数规模而扩大，使用相同方法时，较大型模型的表现优于较小型模型。

合规声明:
我们团队确认所有使用的数据均符合竞赛规则以及相关数据隐私、版权和伦理标准。

References

1. Zhang Z, Zhang Y, Liang Y, et al. LayoutDIT: Layout-aware end-to-end document image translation with multi-step conductive decoder[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. 2023: 10043-10053.
2. Liang Y, Zhang Y, Ma C, et al. Document image machine translation with dynamic multi-pre-trained models assembling[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024: 7084-7095.

[1] https://huggingface.co/OpenGVLab/InternVL2_5-8B-MPO
[2] https://huggingface.co/OpenGVLab/InternVL2_5-1B-MPO
[3] https://github.com/OpenGVLab/InternVL

Table 1. 每个细粒度 DIMT25 子轨的 BLEU 分数

Model	track1				track2	
	Valid-OCR	Valid-MT	Test-OCR	Test-MT	Valid-MT	Test-MT
InternVL2.5-1B-MPO Zero-Shot	/	/	/	/	/	/
InternVL2.5-1B-MPO SFT	/	67.21	/	/	59.81	/
InternVL2.5-1B-MPO MTL-PCOT SFT	/	70.81	94.63	62.16	62.17	57.35
+ MBR	/	/	96.50	64.08	/	59.06
+ Post-processing	/	/	97.00	66.16	/	59.56
InternVL2.5-8B-MPO Zero-Shot	/	30.71	/	/	29.53	/
InternVL2.5-8B-MPO SFT	/	72.74	/	/	64.24	/
InternVL2.5-8B-MPO MTL-PCOT SFT	/	75.72	94.89	65.32	65.77	58.57
+ MBR	/	/	97.16	68.26	/	60.33
+ Post-processing	/	/	97.66	70.48	/	60.78

3. Thendral, R., et al. "Document Image Analysis for Text Extraction and Translation." 2023 4th International Conference on Intelligent Technologies (CONIT). IEEE, 2024.
4. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.
5. Chen, Zhe, et al. "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling." arXiv preprint arXiv:2412.05271 (2024).
6. Wang, Weiyun, et al. "Enhancing the reasoning ability of multimodal large language models via mixed preference optimization." arXiv preprint arXiv:2411.10442 (2024).
7. Zhang, Yu, and Qiang Yang. "A survey on multi-task learning." IEEE transactions on knowledge and data engineering 34.12 (2021): 5586-5609.
8. Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in neural information processing systems 35 (2022): 24824-24837.
9. Farinhas, António, José GC de Souza, and André FT Martins. "An empirical study of translation hypothesis ensembling with large language models." arXiv preprint arXiv:2310.11430 (2023).
10. Mithe, Ravina, Supriya Indalkar, and Nilam Divekar. "Optical character recognition." International journal of recent technology and engineering (IJRTE) 2.1 (2013): 72-75.
11. Chang, Yupeng, et al. "A survey on evaluation of large language models." ACM transactions on intelligent systems and technology 15.3 (2024): 1-45.
12. Kasneci, Enkelejda, et al. "ChatGPT for good? On opportunities and challenges of large language models for education." Learning and individual differences 103 (2023): 102274.
13. Gupta, Kshitij, et al. "Continual pre-training of large language models: How to (re) warm your model?." arXiv preprint arXiv:2308.04014 (2023).
14. Dong, Guanting, et al. "How abilities in large language models are affected by supervised fine-tuning data composition." arXiv preprint arXiv:2310.05492 (2023).
15. Bai, Yuntao, et al. "Training a helpful and harmless assistant with reinforcement learning from human feedback." arXiv preprint arXiv:2204.05862 (2022).
16. Bai, Shuai, et al. "Qwen2. 5-vl technical report." arXiv preprint arXiv:2502.13923 (2025).

17. Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." arXiv preprint arXiv:2312.11805 (2023).
18. Shin, Andrew, Masato Ishii, and Takuya Narihira. "Perspectives and prospects on transformer architecture for cross-modal tasks with language and vision." International journal of computer vision 130.2 (2022): 435-454.
19. Lampropoulos, Aristomenis S., and George A. Tsihrintzis. "Machine learning paradigms." Applications in recommender systems. Switzerland: Springer International Publishing (2015).
20. Marquet, Thomas, and Elisabeth Oswald. "A comparison of multi-task learning and single-task learning approaches." International Conference on Applied Cryptography and Network Security. Cham: Springer Nature Switzerland, 2023.
21. Wu, Zhanglin, et al. "HW-TSC' s Submission to the CCMT 2024 Machine Translation Tasks." China Conference on Machine Translation. Singapore: Springer Nature Singapore, 2024.
22. Wu, Zhanglin, et al. "Choose the Final Translation from NMT and LLM hypotheses Using MBR Decoding: HW-TSC's Submission to the WMT24 General MT Shared Task." arXiv preprint arXiv:2409.14800 (2024).
23. Wu, Zhanglin, et al. "Improving the Quality of IWLST 2024 Cascade Offline Speech Translation and Speech-to-Speech Translation via Translation Hypothesis Ensembling with NMT models and Large Language Models." Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024). 2024.
24. Rei, Ricardo, et al. "COMET: A Neural Framework for MT Evaluation." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.
25. Rei, Ricardo, et al. "COMET-22: Unbabel-IST 2022 submission for the metrics shared task." Proceedings of the Seventh Conference on Machine Translation (WMT). 2022.
26. Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
27. Wu, Zhanglin, et al. "Evaluating Menu OCR and Translation: A Benchmark for Aligning Human and Automated Evaluations in Large Vision-Language Models" arXiv preprint arXiv:2504.13945 (2025).

A 训练数据格式

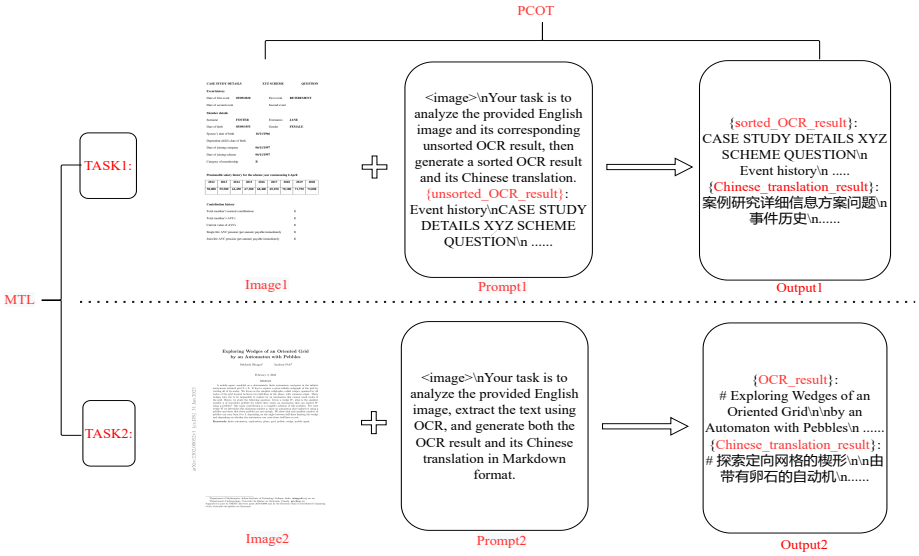

Fig. 1. 我们方法的训练数据组织结构。

B 训练数据规模

Table 2. DIMT25 训练数据的统计信息

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