Leveraging Vision Language Model (VLM) for Traditional Chinese Table Recognition

[1,*] *Yu-Zhuang Huang* (黃淯莊), [1]*Chiou-Shann Fuh* (傅楸善),

[1]Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan,
[*]E-mail: r13922144@ntu.edu.tw fuh@csie.ntu.edu.tw

ABSTRACT

Table recognition refers to the set of tasks of identification, extraction, and interpretation of tabular data from documents, images, or scanned materials. This task presents significant challenges due to the wide variability in table structures, formats, and visual presentations across different documents. Tables can contain complex hierarchical relationships, merged cells, nested tables, and various alignment patterns, making them difficult to parse accurately.

Modern table recognition tasks serve as an upstream task for downstream tasks, typically Large Language Model (LLM)-based document querying and understanding that requires formatted output.

Traditional approaches leveraging tools like Tesseract and PaddleOCR (Optical Character Recognition) employ multi-stage pipelines that separately handle text detection, recognition, and structural analysis. These methods require further processing to produce formatted results in Hyper Text Markup Language (HTML) or LaTeX [3].

Vision-Language Models (VLMs) have shown promising results in addressing these limitations by processing visual and textual information simultaneously. Their holistic approach enables better contextual understanding of table structures while interpreting Traditional Chinese characters within their visual context [1], [5], [13], [18]. However, VLMs still face several challenges when applied to Traditional Chinese table recognition. They require substantial amounts of annotated training data, which are scarce for Traditional Chinese documents [10], [9]. Performance degradation is observed when handling historical documents with classical Chinese characters or specialized terminology. Additionally, current VLMs struggle with very complex table structures as the table data possess quite simple structure.

In this project, we focus on generating diverse structured table data and fine-tuning existing pretrained VLM-OCR tools [2], [5], [13].

1. INTRODUCTION

Vision-Language Models (VLMs) are multimodal systems designed to learn joint representations of visual and textual information. One of the most influential architectures in this category is CLIP (Contrastive Language–Image Pretraining), which aligns images and text in a shared embedding space through contrastive learning [1]. VLMs like CLIP and its successors have demonstrated strong generalization capabilities across a wide range of vision-language tasks, including image classification, captioning, and retrieval [1], [13], [18], [19]. Their ability to connect visual semantics with natural language makes them increasingly relevant for tasks like OCR (character recognition).

OCR, at its core, is a sequence-to-sequence problem: it involves converting a visual input (image) into a coherent textual output, often preserving structure and semantics. This formulation aligns naturally with the strengths of VLMs, particularly those incorporating encoder-decoder architectures inspired by transformers [5], [13]. Unlike traditional OCR pipelines that decouple text detection, recognition, and layout analysis, VLMs can unify these steps by leveraging global context and language priors during inferencing [18].

A VLM-based OCR tool typically consists of the following components:

1. **Visual Encoder** – often based on ViT (Vision Transformer), this extracts rich spatial features from the input image [7], [8], [20].
2. **Language Decoder** – a transformer-based decoder that generates textual sequences conditioned on visual embeddings, enabling natural sequence modeling.

Each component contributes to end-to-end document understanding: the visual encoder preserves spatial and stylistic cues; the decoder handles language modeling; and the fusion and task modules ensure adaptability to complex layouts and domain-specific constraints. By treating OCR as a sequence-to-sequence learning task, VLMs offer a powerful, unified alternative to fragmented traditional systems [2], [5].

Another key advantage of prompt-based VLM OCR systems is their adaptability to varying output representations. In many real-world applications, the same visual content—such as a table—can be represented in multiple textual formats, including HTML, Markdown, CSV, or even plain natural language summaries. By modifying the prompt to specify the desired output format (e.g., "extract table as HTML" or "output Markdown table"), the model can adjust its decoding strategy accordingly. This eliminates the need for post-processing or separate conversion modules, streamlining the OCR pipeline [2], [5]. Such flexibility is particularly valuable when integrating OCR systems into downstream applications like web automation, database entry, or document conversion, where format consistency is critical.

In this work, we employ GOT-OCR as our baseline model to build a table recognition pipeline for Traditional Chinese documents. GOT-OCR is a lightweight yet capable vision-language model designed for general-purpose OCR tasks [2]. Its encoder-decoder architecture and support for prompt-based inference make it a suitable foundation for our needs. By extending it with customized prompts and format-aware decoding strategies, we adapt the model to handle table extraction tasks with a focus on Traditional Chinese language, complex layouts, and varied output requirements. This setup serves as a practical demonstration of how VLM-based OCR frameworks can be tailored for domain-specific document understanding [2], [4].

2. RELATED WORK

2.1 GOT-OCR

GOT-OCR (General OCR Theory) introduces a unified, prompt-based vision-language architecture designed to consolidate multiple OCR subtasks—including detection, recognition, and understanding—into a single model [2]. It formulates OCR as a sequence-to-sequence problem and employs a Transformer-based decoder conditioned on task-specific prompts. This design enables flexible adaptation to various OCR objectives without explicitly retraining for each individual task. Unlike traditional modular systems, GOT-OCR is inherently extensible and more data-efficient in multi-task settings [2], [5].

However, its original training focuses primarily on English-centric documents, leaving a gap in performance and generalization when applied to languages with larger character sets, such as Traditional Chinese. Our work aims to bridge this gap by fine-tuning GOT-OCR on domain-specific datasets and enhancing its ability to parse highly structured content like tables [2], [4].

2.2 PaddleOCR

PaddleOCR is a high-performance OCR toolkit developed with practical deployment in mind. It adopts a modular two-stage architecture consisting of a text detector and a recognition head, supplemented by optional layout analysis modules [3]. This design allows flexible configuration and has been optimized for both Latin and CJK scripts, particularly Simplified and Traditional Chinese. PaddleOCR has been widely used in industry due to its support for lightweight models, real-time inference, and mobile deployment [3].

However, its architecture is fundamentally pipeline-based, leading to higher latency and potential error propagation between modules. Furthermore, it lacks native support for decoding structured document elements, such as table schemas, which limits its applicability in document understanding scenarios. In contrast, end-to-end models like GOT-OCR offer a more unified solution that can seamlessly handle both visual recognition and structure prediction [2].

2.3 PubTabNet

PubTabNet is a large-scale benchmark dataset aimed at advancing table structure recognition. It contains over 500,000 table images annotated with HTML strings that describe both layout and cell content [4]. The dataset has fueled the development of models like TableMaster and TableFormer, which combine visual encoders (e.g., ResNet, Swin Transformer) with sequence decoders to translate image features into structured outputs [12]. These models are trained to capture row/column relations, spanning cells, and nested structures.

However, most existing models focus on English-language scientific documents and rely on significant computational resources, making them less suitable for deployment in real-world, multilingual, or mobile settings. Additionally, generating high-quality table annotations—especially for Traditional Chinese documents—is time-consuming and labor-intensive. This motivates our exploration of LLM-assisted table synthesis methods to augment training data diversity and improve model robustness in low-resource languages [11], [10].

2.4 Donut:

Donut proposes a document understanding model that removes the need for explicit text detection and instead directly decodes structured representations (e.g., JSON, Markdown) from raw document images [5]. It follows an encoder-decoder Transformer architecture, where the encoder processes the visual features and the decoder, trained via teacher forcing, generates the corresponding structured text. Donut has demonstrated strong performance on diverse document tasks, including key-value extraction, form understanding, and table parsing [5]. It achieves this by leveraging synthetic pretraining followed by task-specific fine-tuning.

Nevertheless, Donut is relatively large in size and requires significant GPU memory, which restricts its practicality in low-resource environments. Moreover, although its decoding is structurally aware, errors such as missing delimiters in tables remain a challenge. Compared to Donut, GOT-OCR maintains a more compact architecture and enables prompt conditioning, which can make it more suitable for constrained or domain-specific deployments such as Traditional Chinese form parsing [2], [5].

2.5 CLIP

CLIP (Contrastive Language–Image Pretraining) represents a major advancement in vision-language modeling by training on 400 million image-text pairs using a contrastive objective [1]. It learns a joint embedding space where images and corresponding textual descriptions are aligned, enabling zero-shot performance on a wide range of vision tasks. CLIP's architecture comprises a vision encoder (e.g., ViT) and a text encoder (Transformer), which are trained to maximize cosine similarity between matched pairs [1], [7].

Its success has inspired many downstream models, including BLIP, Flamingo, and even general OCR frameworks [13], [19], [18]. However, CLIP is not directly designed for sequence-level generation and lacks fine-grained token supervision, which is essential for dense tasks like OCR and structured document recognition. Its output is typically a class-level or phrase-level prediction, making it unsuitable for recovering exact textual content from images. Nevertheless, CLIP's powerful image encoder remains valuable as a feature extractor, and its training paradigm has influenced the prompt-based learning strategy adopted by GOT-OCR and others [2], [14].

3. DATASET

3.1 Adapting PubTabNet with Traditional Chinese Content

To construct a high-quality dataset tailored for Traditional Chinese table recognition, we adopt PubTabNet as our structural foundation [4]. PubTabNet provides a large-scale corpus of table images annotated with corresponding HTML representations, which describe cell content, row and column spans, and layout structure. While PubTabNet is rich in structural variety and widely used in table recognition tasks, its content is primarily in English and largely derived from scientific literature, making it less suitable for training models in Chinese contexts.

To address this, we preserve the HTML structure and layout annotations of each table but replace the original English text content within each cell with Traditional Chinese phrases. The substitution process ensures that the cell-level token length distribution remains comparable to the original dataset, preserving visual and structural realism. We draw Chinese replacement content from a curated pool of common words, domain-specific terms (e.g., finance, education, news), and syntactically correct phrases to maintain semantic plausibility. This approach enables us to retain the layout complexity and structural diversity of PubTabNet while adapting its content domain to Traditional Chinese [2].

3.2 Artificial Table Data using LLMs and Web Corpora

To further increase content diversity and better simulate real-world usage scenarios, we augment our dataset by generating artificial Traditional Chinese tables using a hybrid strategy that combines a large language model (ChatGPT) and curated text corpora from Hugging Face datasets [11]. We begin by sampling paragraphs, sentences, and named entities from large-scale Chinese corpora such as Taiwanese news datasets, Wikipedia Chinese, and CC100-zh [9], [10], extracting plausible table topics such as statistical reports, academic metadata, and product listings.

Then, we use ChatGPT to generate synthetic table structures, complete with headers, multi-row entries, and realistic numeric or categorical values [11]. These tables are output in HTML and Markdown formats to mimic the style of existing benchmarks. To maximize variability, we inject randomness into the schema design (e.g., varying number of columns, hierarchical headers, rowspan/colspan usage), and apply post-processing rules to ensure structural correctness. This artificially generated dataset supplements our fine-tuning pipeline with diverse, domain-relevant, and well-formed Traditional Chinese tables, especially in topics not well-represented in PubTabNet [4].

By combining structural reuse from PubTabNet with content replacement and LLM-assisted generation, we create a highly diverse and domain-adapted dataset suitable for training and evaluating GOT-OCR on Traditional Chinese table recognition tasks [2].

4. METHOD

4.1. Architecture of GOT-OCR and Three-Stage Training Strategy

GOT-OCR adopts a streamlined encoder-decoder architecture tailored for generalized OCR tasks [2]. The vision encoder is a modified ViTDet-base model (~80M parameters) that compresses a 1024×1024 image into a dense grid of 256×1024 image tokens [8]. These are projected via a linear layer into the embedding space of the language decoder. The decoder is based on Qwen-0.5B, a lightweight yet multilingual language model with 500M parameters and an extended context length (up to 8K tokens), enabling dense document and table decoding [6].

Training GOT-OCR is done in three progressive stages, each designed to instill specific capabilities into the model [2]:

- **Stage 1: Encoder Pretraining**

The vision encoder is trained independently using a lightweight decoder (OPT-125M) on synthetic and real OCR data, covering both scene and document images [2]. This establishes foundational perceptual abilities across layout styles and languages (including Chinese and English) [2], [6].

- **Stage 2: Joint Training**

The pretrained encoder is attached to the Qwen-0.5B decoder [6]. The model is then jointly trained on more diverse data—formulas, sheet music, tables, and structured content—ensuring that the decoder learns rich multimodal representations while preserving the encoder's visual capacity [2], [14].

- **Stage 3: Decoder Fine-Tuning**

The pretrained encoder is attached to the Qwen-0.5B decoder [6]. The model is then jointly trained on more diverse data—formulas, sheet music, tables, and structured content—ensuring that the decoder learns rich multimodal representations while preserving the encoder's visual capacity [2], [14].

4.2 Fine-Tuning

In this work, we adopt only the Stage 3 fine-tuning phase of GOT-OCR, which updates the decoder parameters while freezing the pretrained vision encoder [2]. This choice is motivated by both computational efficiency and the suitability of the pretrained encoder for our target domain.

The GOT-OCR encoder has been trained on a broad range of scene and document text, including English and Chinese, using a ViTDet-based architecture designed for high-resolution visual compression [2], [8]. It demonstrates strong generalization in perceiving dense optical structures such as full-page documents, character slices, and tabular layouts. Since table recognition is primarily a layout-driven task, the encoder already captures sufficient visual information, including grid structures, cell boundaries, and alignment cues. Notably, these layout patterns are largely language-agnostic.

Given this, the adaptation required for Traditional Chinese table recognition lies chiefly in the decoding process—specifically, mapping visual tokens to Chinese sequences and formatting them into structured outputs such as Markdown or HTML. Stage 3 fine-tuning effectively enables this by training the decoder to align encoded visual representations with domain-specific output styles, without modifying the general-purpose vision encoder [2], [5].

This strategy allows us to adapt GOT-OCR to our task with significantly reduced computational cost, while preserving the encoder's general visual capacity. Our experiments confirm that this lightweight fine-tuning approach is sufficient to achieve high performance on Traditional Chinese structured table OCR [2].

4.3 Distributed Training using MS-SWiFT

In this work, we adopt only the Stage 3 fine-tuning phase of the GOT-OCR training pipeline. This stage involves updating the decoder parameters while keeping the vision encoder fixed [2]. Our choice is motivated by both practical resource considerations and the observation that the pretrained encoder already encodes sufficiently rich visual representations to support effective decoding for Traditional Chinese table recognition [2], [5].

The GOT-OCR vision encoder is based on a ViTDet backbone and is pretrained using both scene-text and document-style OCR data [2], [8]. Its architecture is designed to operate on high-resolution inputs (up to 1024×1024) and to compress them efficiently into dense image tokens via local attention mechanisms. This enables the encoder to preserve critical layout and spatial information—such as character positioning, table grids, cell divisions, and structural alignment—while minimizing token size [2].

Despite our target domain involving Traditional Chinese text, which contains a large and complex character set, the encoder proves general enough to capture the layout-invariant and language-agnostic features that are essential for recognizing table structure [2].

While Traditional Chinese OCR introduces challenges such as character similarity, high token diversity, and stylistic variation, our specific task—structured table recognition—relies more heavily on accurate layout understanding than on fine-grained character discrimination alone. The encoder's pretraining on Chinese document data equips it with sufficient perceptual capability to detect and distinguish individual cells, boundaries, and alignment cues, even in dense multi-row tabular contexts [2], [6].

These visual cues provide the decoder with enough grounding to generate semantically and structurally correct outputs. Accordingly, our fine-tuning focuses solely on adapting the language decoder, which is responsible for translating the encoded visual features into structured outputs such as Markdown or HTML [2].

- The linguistic and lexical patterns of Traditional Chinese text,
- The formatting logic needed to correctly express table structures,
- And the long-range dependencies between content elements such as headers and multi-cell spans.

By limiting training to the decoder, Stage 3 fine-tuning offers a highly efficient adaptation strategy. It avoids the computational expense and risk of overfitting associated with full-model training, while preserving the pretrained encoder's strong generalization capabilities. Our empirical results demonstrate that this lightweight yet targeted approach is sufficient to produce high-quality outputs for Traditional Chinese table recognition tasks.

5. Experimental Result

To evaluate the performance of our model on Traditional Chinese table recognition, we adopt edit distance as the primary evaluation metric. Edit distance provides a robust measurement of sequence-level similarity between the predicted and ground-truth outputs, capturing both structural and lexical accuracy [2].

We report the results of our fine-tuned GOT-OCR model on our constructed dataset in Table 1. The table summarizes the edit distance across different evaluation sets, demonstrating the model's effectiveness in accurately recognizing and formatting table content in Traditional Chinese [2], [4], [11].

Method	Edit distance
GOT-OCR	1.332
Fine-tuned (ours)	0.188

Table 1. The fine-tuned version result

6. Conclusion

In this work, we adapt the GOT-OCR model to the task of Traditional Chinese table recognition through efficient Stage 3 fine-tuning [2]. By leveraging the pretrained vision encoder—which captures rich, layout-aware visual features—we are able to specialize the decoder for structured output generation in Traditional Chinese, without retraining the full model [2], [8].

To support this, we construct a diverse training dataset by combining structurally rich samples from PubTabNet with content substitution, and by generating synthetic tables using large language models and Chinese corpora [4], [11], [10], [9].

Our experimental results, evaluated using edit distance, demonstrate that the fine-tuned GOT-OCR model performs effectively on this task [2]. The findings validate that decoder-only adaptation, combined with high-quality table data, is sufficient for achieving strong performance in low-resource, language-specific structured OCR scenarios.

This work highlights the practical applicability of lightweight VLM adaptation for multilingual document understanding and opens the door for further research in extending general OCR models to other languages and structured formats [1], [2], [18].

7. Reference

[1] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[2] H. Wei, C. Liu, J. Chen, L. Kong, et al., "General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model," *arXiv preprint*, arXiv:2409.01704, 2024.

[3] Y. Du, C. Li, R. Guo, et al., "PP-OCRv2: Bag of Tricks for Ultra Lightweight OCR System," *arXiv preprint*, arXiv:2109.03144, 2021.

[4] X. Zhong, J. Tang, and A. Jimeno-Yepes, "PubTabNet: Structured Table Recognition using Conditional Attention," *European Conference on Computer Vision (ECCV)*, 2020.

[5] G. Kim, M. Seo, and J. Shin, "Donut: Document Understanding Transformer without OCR," *European Conference on Computer Vision (ECCV)*, 2022.

[6] J. Bai, S. Bai, S. Yang, P. Wang, et al., "Qwen Technical Report," *arXiv preprint*, arXiv:2309.16609, 2023.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations (ICLR)*, 2021.

[8] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring Plain Vision Transformer Backbones for Object Detection," *European Conference on Computer Vision (ECCV)*, 2022.

[9] A. Conneau, K. Khandelwal, N. Goyal, et al., "Unsupervised Cross-lingual Representation Learning at Scale," *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[10] Q. Lhoest, A. Villanova del Moral, Y. Jernite, et al., "Datasets: A Community Library for Natural Language Processing," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2021, pp. 175–184.

[11] OpenAI, "GPT-4 Technical Report," *arXiv preprint*, arXiv:2303.08774, 2023.

[12] M. E. Khan, A. Swaminathan, and T. Q. Nguyen, "TableFormer: End-to-End Table Parsing using Pre-trained Language Models," *arXiv preprint*, arXiv:2205.12443, 2022.

[13] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with

Frozen Image Encoders and Large Language Models," *arXiv preprint*, arXiv:2301.12597, 2023.

[14] H. Wei, L. Kong, J. Chen, Z. Ge, et al., "Vary: Scaling up the Vision Vocabulary for Large Vision-Language Models," *arXiv preprint*, arXiv:2312.06109, 2023.

[15] B. Shi, C. Yao, M. Yang, et al., "ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17)," *International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

[16] C. Zhang, G. Peng, Y. Tao, et al., "ShopSign: A Diverse Scene Text Dataset of Chinese Shop Signs," *arXiv preprint*, arXiv:1903.10412, 2019.

[17] A. Veit, T. Matera, L. Neumann, et al., "COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images," *arXiv preprint*, arXiv:1601.07140, 2016.

[18] W. Wang, Z. Liu, J. Lin, et al., "A Survey of Vision-Language Pretrained Models," *arXiv preprint*, arXiv:2101.09675, 2021.

[19] J. Alayrac, J. Donahue, P. Huang, et al., "Flamingo: A Visual Language Model for Few-Shot Learning," *arXiv preprint*, arXiv:2204.14198, 2022.

[20] M. Touvron, H. Jegou, M. Douze, et al., "Training Data-efficient Image Transformers and Distillation through Attention," *International Conference on Machine Learning (ICML)*, 2021.