Computational Humanities Research

www.cambridge.org/chr

Fine-tuning vision–language models as OCR systems for low-resource languages: A case study of Manchu

Yan Hon Michael Chung[1] and Donghyeok Choi[2]

[1]Division of Humanities, The Hong Kong University of Science and Technology, Hong Kong and [2]Department of History, Hong Kong Baptist University, Hong Kong

Short Article

Cite this article: Chung Yan Hon Michael and Donghyeok Choi. 2026. "Fine-tuning vision–language models as OCR systems for low-resource languages: A case study of Manchu." *Computational Humanities Research*, **2**:e20, https://doi.org/10.1017/chr.2026.10042

Received: 24 July 2025 UTC
Revised: 28 April 2026 UTC
Accepted: 24 June 2026 UTC

Keywords:
low-resource languages; Manchu; OCR; vision–language model

Corresponding author:
Yan Hon Michael Chung;
Email: hmichung@ust.hk

Abstract

Manchu, a low-resource language essential for understanding early modern Eastern Eurasian history, lacks effective optical character recognition (OCR) systems that can handle real-world historical documents. This study develops high-performing OCR systems by fine-tuning three open-source vision–language models (Llama-3.2-11B-Vision, Qwen2.5-VL-7B, and Qwen2.5-VL-3B) on 60,000 synthetic Manchu word images using parameter-efficient training. Llama-3.2-11B-Vision achieved exceptional performance with 98.6 percent word accuracy and a character error rate of 0.002 on synthetic data, while maintaining 87.4 percent word-level accuracy (95% CI: 85.1%–89.7%) on scanned word images from formal Qing administrative manuscripts and printed documents. Comparative evaluation reveals substantial advantages over traditional approaches: while a convolutional recurrent neural network baseline achieved 99.9 percent synthetic accuracy, it suffered severe degradation to 56.8 percent on scanned documents. Our approach demonstrates effective synthetic-to-real domain transfer, providing a cost-effective solution deployable on accessible infrastructure. This work establishes a transferable framework for endangered language OCR that reduces technical and financial barriers in digital humanities, enabling historians and linguists to process historical archives without specialized computing resources.

Plain language summary

The Manchu language is essential for understanding the political, social, and military history of the early modern Qing Empire, which governed much of Eastern Eurasia from the seventeenth to the early twentieth centuries. Today, however, Manchu is an endangered language, and most surviving materials exist only in handwritten manuscripts or woodblock-printed materials. Currently, Manchu has no widely available optical character recognition (OCR) tools that convert scanned pages into searchable and editable text. As a result, historians and linguists face significant challenges in accessing and studying these archives at scale.

This study addresses this gap by developing a new Manchu OCR system using recent advances in artificial intelligence known as vision–language models (VLMs). These models can analyze images and text together, making them especially suitable for reading historical scripts. We fine-tuned three open-source models – Llama-3.2-11B-Vision, Qwen2.5-VL-7B, and Qwen2.5-VL-3B – using 60,000 computer-generated images of Manchu words. These synthetic images mimic the appearance of real writing, allowing us to train accurate OCR systems even though very few annotated historical materials exist. Among the three models, Llama-3.2-11B-Vision performed best. It correctly recognized 98.6 percent of synthetic word images and achieved a very low character error rate. More importantly, when tested on 753 word images extracted from seven Qing manuscripts and printed texts, it maintained an accuracy of 87.4 percent. This demonstrates that the model can successfully apply what it learned from synthetic data to real-world historical materials.

To understand how much this matters, we compared our approach with a traditional OCR method called a convolutional recurrent neural network (CRNN). Although the CRNN achieved almost perfect accuracy on synthetic data, its performance dropped sharply to 56.8 percent when applied to scanned historical documents. This contrast highlights a core challenge in endangered-language OCR, where classical methods struggle to generalize once they encounter the noise and irregularity of real handwritten or printed archival materials. By contrast, VLMs are far more adaptable, enabling them to handle the diversity of script styles and visual conditions found in archival Manchu sources.

A major advantage of our approach is its accessibility. Since we fine-tune relatively lightweight open-source models and use efficient training techniques, the entire system can be trained and deployed on widely available computing hardware. This reduces the technical and financial barriers that have long limited digitization efforts for Manchu and other low-resource languages.

This article was awarded the Open Data badge for transparent practices. See the Data availability statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Overall, this project provides a practical, reproducible framework for building OCR systems for Manchu and other endangered or low-resource scripts. By enabling automatic transcription of historical documents, it supports language preservation and makes large-scale digital analysis possible for historians and linguists. The tools developed here open the door to new research possibilities across Qing studies and the broader digital humanities.

Introduction

In 1993, Pamela Kyle Crossley and Evelyn S. Rawski published a seminal article in the *Harvard Journal of Asiatic Studies* titled "A Profile of the Manchu Language in Ch'ing History," in which they argued that the Manchu language and its archival sources are essential for a comprehensive understanding of Qing political, administrative, cultural, and frontier histories. The Qing, the last imperial dynasty of China, was ruled by the Manchus, an ethnic minority whose native language, Manchu – classified as endangered by UNESCO in 2010 – once served as both a court and administrative language of the empire (UNESCO 2010). Challenging the long-standing assumption that Manchu-language materials merely duplicated their Chinese counterparts, Crossley and Rawski demonstrated that many Manchu documents contained unique content – particularly in sensitive political, military, and ideological contexts. With growing access to Manchu archival materials, they anticipated that "the relevance of the language (Manchu) to any particular category of research (of Qing history) is not as predictable as it was once assumed to be" (Crossley and Rawski 1993, 102). This call for a methodological shift was reinforced in Rawski's 1996 presidential address to the Association for Asian Studies, where she highlighted how a Manchu-centered perspective fundamentally altered prevailing narratives of the Qing (Rawski 1996, 829–30). Together, these interventions catalyzed the emergence of what is now known as the "New Qing History," a scholarly approach that repositions the Qing as a multiethnic empire rather than a purely sinicized Chinese dynasty.

That said, the language barrier has remained a major obstacle to the broader utilization of the vast Manchu archival corpus. Proficiency in Manchu remains limited to a small circle of specialists, and only a fraction of the available sources has been systematically cataloged or studied. Recent advances in vision–language models (VLMs) offer new possibilities for expanding access to these materials. This study introduces a high-performing optical character recognition (OCR) system for the Manchu language, developed by fine-tuning the Llama-3.2-11B-Vision model on a synthetically generated dataset comprising 60,000 Manchu word images. The resulting system achieves a remarkably low character error rate (CER) of 0.002 and a word-level accuracy of 98.6 percent on synthetic data, while maintaining a word-level accuracy of 87.4 percent (95% CI: 85.1%–89.7%) on scanned word images from formal Qing administrative manuscripts and printed documents.

This proposed Manchu OCR system not only facilitates the reading of Manchu-language texts but also helps promote their wider use in historical research. Additionally, this study presents a simple and transferable framework which requires only modest synthetic data and computing resources for building competitive OCR systems for low-resource languages. The recent advances in VLM AI thus open promising opportunities for the preservation and revitalization of many other historically significant, low-resource, and endangered languages that remain underserved by existing OCR technologies.

Literature review

Earlier OCR works

Efforts to develop an OCR system for the Manchu language began in the mid-2000s. Early research focused on the recognition of primitive Manchu characters at the character level. They proposed a stroke-based method that decomposed handwritten Manchu characters into basic stroke elements, matched these to a predefined stroke library, and reassembled them into Manchu-Roman code using fuzzy string matching (Zhang et al. 2004). However, this approach was constrained by the complexity of stroke combination rules and the high variability of handwriting styles, which led to recognition errors in cases of distorted or noisy inputs. Subsequent efforts continued along the character-level trajectory, developing recognition systems that extracted and classified Manchu character units (MCUs), a sub-character structural unit (Zhang, Li, and Wang 2006, 3339–40; Zhao et al. 2006, 802–3). While these models introduced some advances, they still relied on manual segmentation and struggled with compounded character forms and the lack of spacing between letters in the vertical Manchu script.

A significant methodological shift occurred around 2017 with the adoption of deep neural network architectures, particularly convolutional neural networks (CNNs). Researchers began to move away from character-level segmentation and instead adopted word-level, segmentation-free recognition frameworks (Huang et al. 2017, 46–9). Building on this direction, researchers further improved word-level recognition accuracy by training models on large synthetic Manchu word datasets, generated through augmentation techniques, such as rotation, blur, and elastic distortion. These datasets were derived from 2,135 word images representing 666 unique Manchu words extracted from the textbook *365 Sentences in Manchu*, with each word class expanded into 1,000 variants to create a corpus of 666,000 samples. These CNN models could achieve a peak accuracy of 97.68 percent and an average accuracy above 91 percent in printed fonts (Li et al. 2018; Zheng et al. 2018). A subsequent study conducted in 2021 adopted a similar segmentation-free approach using a deep CNN architecture and a sliding window mechanism. Their system was trained on a slightly expanded dataset of 671 classes, also with 1,000 samples per class, and achieved a peak accuracy of 98.84 percent (Zhang et al. 2021, 5–6). However, it is important to note that all these models were trained and evaluated on variations of a closed set of fewer than 700 Manchu words, and their performance on unseen vocabulary remains untested. Consequently, while these studies demonstrate strong performance in controlled environments, their generalizability to the full range of historical Manchu texts has yet to be established.

Beyond academic research, several institutions and individuals have also developed practical OCR systems for Manchu. The First Historical Archives of China in Beijing has reportedly created an OCR system tailored for scanned Manchu documents. However, as the system is restricted to internal institutional use, its technical specifications and recognition accuracy remain unavailable for public evaluation. In 2022, Zhang Zhuohui released an open-source OCR tool for Manchu developed in Python on GitHub (Zhang 2022). While the system demonstrates promising functionality in processing scanned Manchu texts, no formal assessment of its recognition accuracy has yet been published.

Recent efforts have begun to integrate linguistic reasoning with visual feature extraction to overcome some of the limitations of purely vision-based OCR systems for low-resource scripts. Among these approaches is the visual–language framework for Manchu

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press



Figure 1. manju gisun, meaning "Manchu language."

word recognition (VLMR), which links visual and textual representations through a self-teacher network that refines both modalities via self-knowledge distillation. Trained on 21,245 woodblock-printed and 6,721 handwritten Manchu word samples, VLMR reported 98.24 percent character recognition accuracy (CRA) and 93.84 percent word recognition accuracy (WRA) on the Woodblock Manchu Word (WMW) dataset, and 95.10 percent CRA and 84.08 percent WRA on the more challenging Handwritten Manchu Word (HMW) dataset (Wang et al. 2024). These results demonstrate that integrating visual and semantic cues, especially in a multimodal learning framework, can yield competitive performance even with constrained training data.

Complementing this direction, recent experiments have begun to fine-tune general-purpose VLMs for OCR tasks in similarly underserved linguistic contexts. A notable case is Miguel Escobar Varela's adaptation of Qwen2-VL-2B-Instruct model for recognizing historical Malay texts in the Jawi script, an Arabic-derived orthography characterized by cursive ligatures, letter ambiguity, and sparse digital resources. Despite training on just 911 annotated samples, the model achieved a CER of 8.66 percent and a word error rate (WER) of 25.50 percent (Varela 2024). This pioneering result demonstrates that even compact, generalist VLMs can be effectively fine-tuned for specialized OCR tasks, offering scalable and functional solutions for historically significant but digitally marginalized scripts.

Manchu script and romanization

The Manchu script, illustrated in Figure 1, is generally regarded as an alphabetic syllabary derived from the vertically written Mongolian script (Li 2000, 13). It consists of individual letters representing both vowels and consonants. These letters vary in shape depending on their position within a syllable – initial, medial, or final – and are visually connected in flowing vertical strokes. While not joined into complete words as in Latin cursive, the syllables are linked vertically, giving the appearance of uninterrupted script. Over time, diacritics and modified letter forms were introduced to improve phonetic clarity and standardization (Li 2000, 21–7).

To transliterate the Manchu script into Latin script, scholars have developed a variety of romanization systems, each reflecting different phonological emphases and typographic conventions. This study adopts the modified Möllendorff system as presented by Roth Li, which strikes a balance between phonetic precision and typographic simplicity (Li 2000, 16). Diacritics are avoided when possible, and Latin letters are selected to closely approximate Manchu phonemes while maintaining accessibility for learners.

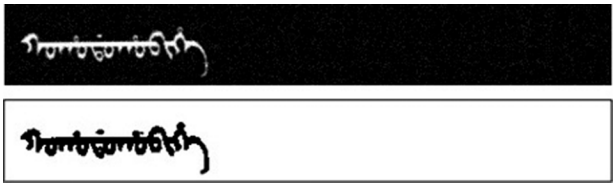


Figure 2. Original (top) and preprocessed (bottom) synthetic Manchu word images from the Zhang Zhuohui dataset (Zhang 2022).

Dataset

This study adopts the open-source synthetic Manchu dataset developed by Zhang Zhuohui, with substantial additional processing and formatting to prepare it for VLM training (Zhang 2022). Zhang's dataset was created through a modular pipeline that systematically generates and preprocesses word-level image samples for OCR training. The foundation of the dataset is a curated comprehensive dictionary of 130,917 Manchu words. From this lexicon, individual words were randomly sampled and rendered as images using multiple distinct Manchu typefaces. Each font contains a set of pre-rendered grayscale images corresponding to Manchu words. These images were retrieved and paired with their textual labels to ensure accurate alignment between visual input and semantic content. A total of 750,000 training samples were synthesized, along with an evaluation set of 25,000 samples.

For the present study, a random subset of 60,000 synthetic word images from the full dataset was selected as the training split and subjected to additional preprocessing tailored to the requirements of VLM architecture. Specifically, the images, originally rendered as white glyphs on a black background with synthetic noise, were color-inverted to a black-on-white format, denoised, and contrast-enhanced to improve clarity and reduce background artifacts. All images were also resized to fixed dimensions, ensuring uniformity across the dataset. As shown in Figure 2, this transformation produced cleaner, higher-contrast inputs, facilitating more effective model learning and improving OCR accuracy.

The evaluation dataset for the current study consists of two parts. First, 15,000 synthetic word images, which are distinct from the 60,000 synthetic images used for training, served as a validation split during training and were evaluated every 1,000 steps to monitor model performance. For the final synthetic assessment, we randomly sampled 1,000 Manchu word images from this 15,000-image validation split and computed all evaluation metrics on that subset. Because this subset is drawn from the same pool used for checkpoint selection, the resulting synthetic metrics should be understood as checkpoint-selection-set performance rather than independent test-set estimates, and are therefore likely optimistic. This applies equally to both the VLM and convolutional recurrent neural network (CRNN) evaluations, so relative comparisons on synthetic data remain internally consistent. Second, we built a real-world test set of 753 Manchu word images drawn from seven scanned sources (five handwritten and two printed). The handwritten materials consist of Qing court documents – the *Old Manchu Archives from the Grand Secretariat* (Wu et al. 2009), the *Veritable Records of the Taizong Emperor, Manchu Edition* (Veritable Records n.d.), and three memorials to the throne written by different officials in the eighteenth and nineteenth centuries (Disangga 1736; Furdan 1728; Peiceng 1849). The printed materials are the *General Gazetteer of the Eight Banners* (Ortai et al 1739) and an imperial edict reproduced in Thomas Taylor Meadows's *Translations from the Manchu: With the Original Texts, Prefaced by*

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press



Figure 3. Sample Manchu word image from *Old Manchu Archives from the Grand Secretariat* (Wu et al. 2009).

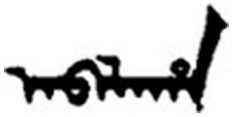


Figure 4. Sample Manchu word image after processing for evaluation (Wu et al. 2009).

an Essay on the Language (Meadows 1849). Details of the sources are provided in Appendix D. Images were identified and extracted using a custom OpenCV-based preprocessing and segmentation pipeline that performs binarization, denoising, and glyph-segment isolation, as shown in Figure 3. After segmentation, we applied light post-processing, including background inversion to black-on-white, contrast enhancement, and resizing to fixed dimensions, in order to normalize the word images against the synthetic training data, as shown in Figure 4.

Methodology

Model preparation

We selected three open-source VLMs with varying parameter sizes and architectures to assess the effect of model scale and design on OCR accuracy: Qwen2.5-VL-3B, Qwen2.5-VL-7B (both from Alibaba), and Llama-3.2-11B-Vision (Meta). This selection was motivated by several factors: (1) Parameter diversity spanning 3B to 11B parameters allows systematic investigation of scale effects on OCR performance, as recent studies indicate that while larger models generally provide superior OCR capabilities (Ding et al. 2024, 1), well-optimized smaller models can achieve competitive results through architectural improvements (Xu et al. 2024, 1). (2) Both Qwen2.5-VL and Llama-3.2, despite sharing similar multimodal capabilities (Meta AI 2024; Qwen Team 2025), offer different parameter scales and training optimizations that provide insights into low-resource script recognition. (3) Both model families have demonstrated state-of-the-art performance on vision–language benchmarks, with Qwen2.5-VL consistently achieving top scores on OCR tasks (Qwen Team 2025) and Llama-3.2-11B-Vision excelling in document VQA and high-resolution image processing up to 1120 × 1120 pixels (Meta AI 2024).

Each model was fine-tuned using the `FastVisionModel` wrapper from the *Unsloth* framework (Han et al. 2023), which supports efficient parameter-efficient fine-tuning (PEFT) of both visual and language components. Our training objective was to generate structured OCR output containing both the Manchu script and its romanized transliteration. To achieve this, we designed a consistent instruction format and applied a shared base configuration across all models, with model-specific adjustments to learning rate, warm-up schedule, and epoch count (see Table 1), to ensure fair comparison. The detailed prompt design and training configuration are described in the following sections.

Table 1. Training hyperparameters for VLM fine-tuning

Hyperparameter	Value
Number of epochs	15 (3B), 10 (7B), 5 (11B)
Optimizer	AdamW (8-bit)
Learning rate	2×10^{-4} (3B), 2×10^{-4} (7B), 1×10^{-4} (11B)
LR scheduler	Cosine with restarts
Warmup steps	3,000 (3B), 2,000 (7B), 1,000 (11B)
Per-device batch size	4
Gradient accumulation steps	2
Effective batch size	8
Mixed precision	Float32
Weight decay	0.01
Random seed	3,407
LoRA rank	32
LoRA alpha	64
LoRA dropout	0.05

Prompt design

Prompt design played a critical role in adapting general-purpose VLMs to the OCR task. We designed a structured prompt to guide the model to generate outputs in a dual-script format. Each input image was paired with the following prompt:

```
You are an expert OCR system for Manchu script.
Extract the text from the provided image with
perfect accuracy. Format your answer exactly
as follows: first line with 'Manchu:' followed
by the Manchu script, then a new line with
'Roman:' followed by the romanized
transliteration.
```

This instruction was applied uniformly across all training examples without distinction between system and user prompts, treating it as a single unified instruction for the OCR task. The target output was formatted to include both the Unicode-based Manchu script (rendered using Mongolian Unicode due to script similarity) and its corresponding Roman transliteration, encouraging the model to learn both visual decoding and phonetic transduction.

For example, given an image containing the Manchu word for "behind" or "after, later," the expected output format would be:

```
Manchu:ᠠᠮᠠᠯᠠ
Roman:AMALA
```

The prompt was held constant across all models to control for variation in instruction interpretation. We did not conduct prompt ablation or variation studies in this work, focusing instead on evaluating the effect of model architecture and size under a fixed instruction format.

Because the model generates both Manchu script and romanization jointly from a single visual input, the romanization output serves as an auxiliary training signal that reinforces script learning. However, since Manchu-to-Roman transliteration follows a deterministic mapping under the modified Möllendorff system, romanization accuracy closely tracks script recognition accuracy. Our

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press

evaluation therefore focuses primarily on Manchu script metrics, as this represents the core visual recognition challenge.

Training

Hyperparameter configuration

To optimize training efficiency while maintaining model performance, we employed differentiated epoch schedules based on model size: Qwen2.5-VL-3B for 15 epochs, Qwen2.5-VL-7B for 10 epochs, and Llama-3.2-11B-Vision for 5 epochs. This inverse relationship between model size and training epochs is motivated by empirical findings that larger models are more sample efficient and converge faster than smaller counterparts, while training beyond 4 epochs yields diminishing returns and can lead to performance degradation (Zhang et al. 2024). The graduated schedule allows smaller models to benefit from extended training while preventing overfitting in larger models, optimizing computational efficiency across different parameter scales. All models used the AdamW optimizer with an 8-bit precision variant (`paged_adamw_8bit`) to reduce memory consumption. The learning rate was set to 2×10^{-4} for the Qwen models and 1×10^{-4} for Llama-3.2-11B-Vision, with cosine learning rate scheduling with restarts.

Training was performed with a per-device batch size of 4 and gradient accumulation over 2 steps to effectively simulate a batch size of 8. Training was conducted in full precision (float32) to maintain numerical stability. We used weight decay of 0.01.

All models were fine-tuned using the *Unsloth* training framework's `SFTTrainer`, which wraps HuggingFace's `transformers` Trainer with additional support for multimodal instruction tuning. Training was conducted on an NVIDIA RTX A6000 Ada GPU with Ubuntu 24.04 LTS; detailed hardware and software specifications are provided in Appendix A. We employed LoRA-based PEFT for both vision and language components, as LoRA enables faster training and significantly reduced memory requirements compared to full fine-tuning while maintaining competitive performance (Hu et al. 2022). The LoRA configuration used a rank of 32, $\alpha = 64$, and a dropout rate of 0.05. These hyperparameters were selected through preliminary experiments to balance training efficiency and model performance, following common practices in multimodal fine-tuning literature.

Evaluation metrics

To assess OCR performance, we employed four primary evaluation metrics that capture different aspects of model accuracy and practical usability. All metrics were computed for each sample and aggregated over the evaluation set:

Character error rate: We calculated CER using the Levenshtein distance (edit distance) (Levenshtein 1966) between the predicted Manchu string and the reference, normalized by the total number of characters in the ground truth. This standard OCR metric captures insertion, deletion, and substitution errors at the character level.

Word accuracy: We measured exact-match accuracy at the word level, computing the percentage of test samples where the model's predicted Manchu script exactly matches the ground truth. This metric provides a direct assessment of the model's ability to correctly recognize complete Manchu words without any character-level errors.

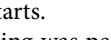
F1 score: We implemented a character-level F1 score based on sequence matching to balance precision and recall. We identified matching character blocks between predicted and ground truth sequences to calculate true positives. Precision was computed as the ratio of true positives to the total predicted characters, while recall was the ratio of true positives to the total ground truth characters. The F1 score was then calculated as the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (1)$$

For illustration, consider these scenarios:
Under-generation (deletion):

Ground truth:  (5 characters) .
Predicted:  (3 characters).
True positives: 3 (matching block).
Precision: 3/3 = 1.0.
Recall: 3/5 = 0.6.
F1 score: 0.75.

Over-generation (insertion):

Ground truth:  (5 characters).
Predicted:  (8 characters).
True positives: 5 (matching block).
Precision: 5/8 = 0.625.
Recall: 5/5 = 1.0.
F1 score: 0.77.

This metric effectively captures the model's ability to correctly identify characters while penalizing both over-generation and under-generation errors.

Inference time: We measured the average time required for each model to process a single image and generate the OCR output, including both the Manchu script and its romanized transliteration. This metric is crucial for assessing the practical feasibility of deploying the OCR system in real-world applications, where processing speed directly impacts user experience and system scalability. Detailed inference configuration settings are provided in Appendix B.

For all reported metrics, confidence intervals were computed using non-parametric bootstrap resampling with 1,000 iterations. For each bootstrap sample, we resampled with replacement from the per-sample metrics and computed the mean. The 95 percent CI bounds were determined using the percentile method (2.5th and 97.5th percentiles of the bootstrap distribution).

Results

Model selection and training stability

We saved model checkpoints at every 1,000 training steps to monitor training progression comprehensively. Following training completion, we evaluated all checkpoints on our synthetic validation dataset and selected those achieving the highest word accuracy performance. Because the synthetic evaluation metrics reported in Tables 2 and 3 are computed on a subset of this same validation pool, they reflect checkpoint-selection-set performance and should be interpreted as optimistic upper-bound estimates of synthetic accuracy. Crucially, the real-world test dataset of 753 archival images was held out entirely and used only for final evaluation of the selected checkpoints, so the real-world results are not affected by this limitation.

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press

Table 2. Manchu performance comparison across validation and test datasets. Bold values indicate the best performance for each metric within each dataset

Model	WA (%)	CER	F1	Time (s)
Validation dataset (1,000 synthetic samples)				
Llama-3.2-11B-Vision	**98.6 ± 0.7**	**0.00 ± 0.00**	**0.998 ± 0.001**	14.7 ± 0.2
Qwen2.5-VL-7B	86.3 ± 2.2	0.03 ± 0.01	0.974 ± 0.005	1.4 ± 0.0
Qwen2.5-VL-3B	83.5 ± 2.4	0.04 ± 0.01	0.968 ± 0.006	**1.3 ± 0.0**
Test dataset (753 real-world samples)				
Llama-3.2-11B-Vision	**87.4 ± 2.3**	**0.05 ± 0.01**	**0.967 ± 0.007**	9.2 ± 0.2
Qwen2.5-VL-7B	33.7 ± 3.5	0.35 ± 0.03	0.759 ± 0.017	1.0 ± 0.0
Qwen2.5-VL-3B	27.8 ± 3.1	0.39 ± 0.03	0.724 ± 0.017	**0.9 ± 0.0**

Table 3. Roman performance comparison across validation and test datasets

Model	WA (%)	CER	F1	Time (s)
Validation dataset (1,000 synthetic samples)				
Llama-3.2-11B-Vision	**98.4 ± 0.8**	0.00 ± 0.00	**0.998 ± 0.001**	14.7 ± 0.2
Qwen2.5-VL-7B	86.4 ± 2.2	0.03 ± 0.01	0.974 ± 0.005	1.4 ± 0.0
Qwen2.5-VL-3B	83.5 ± 2.4	0.04 ± 0.01	0.969 ± 0.005	**1.3 ± 0.0**
Test dataset (753 real-world samples)				
Llama-3.2-11B-Vision	**87.0 ± 2.4**	**0.05 ± 0.01**	**0.966 ± 0.007**	9.2 ± 0.2
Qwen2.5-VL-7B	33.7 ± 3.5	0.34 ± 0.03	0.763 ± 0.017	1.0 ± 0.0
Qwen2.5-VL-3B	27.6 ± 3.1	0.39 ± 0.03	0.725 ± 0.017	**0.9 ± 0.0**

Figure 5a reveals distinct convergence patterns across model architectures. The Llama-3.2-11B-Vision model demonstrated superior efficiency, achieving peak validation accuracy after only 33,000 training steps, while the Qwen variants required substantially longer training periods to reach their optimal performance points. Specifically, the selected optimal checkpoints were: Qwen2.5-VL-3B at step 107,000, Qwen2.5-VL-7B at step 74,000, and Llama-3.2-11B-Vision at step 33,000. These convergence differences reflect underlying architectural efficiencies, with the larger Llama model reaching optimal accuracy approximately two to three times faster than the smaller Qwen variants.

To further assess training stability, we monitored gradient norm throughout the training process as a key indicator of optimization health. Gradient norm serves as a critical diagnostic metric for understanding training dynamics: excessive values may signal unstable optimization or poor hyperparameter choices, while consistently low values indicate smooth convergence and well-conditioned parameter updates.

Figure 5b illustrates the gradient norm trajectories throughout training for all three models. The analysis reveals markedly different optimization behaviors across architectures. Llama-3.2-11B-Vision demonstrates the most stable gradient dynamics, with smooth convergence to low gradient norms, indicating well-conditioned parameter updates and effective optimization. The Qwen variants, particularly the 3B model, exhibit more volatile gradient trajectories with higher final gradient norms, suggesting less stable optimization landscapes that may require different hyperparameter configurations or longer training periods to achieve optimal convergence.

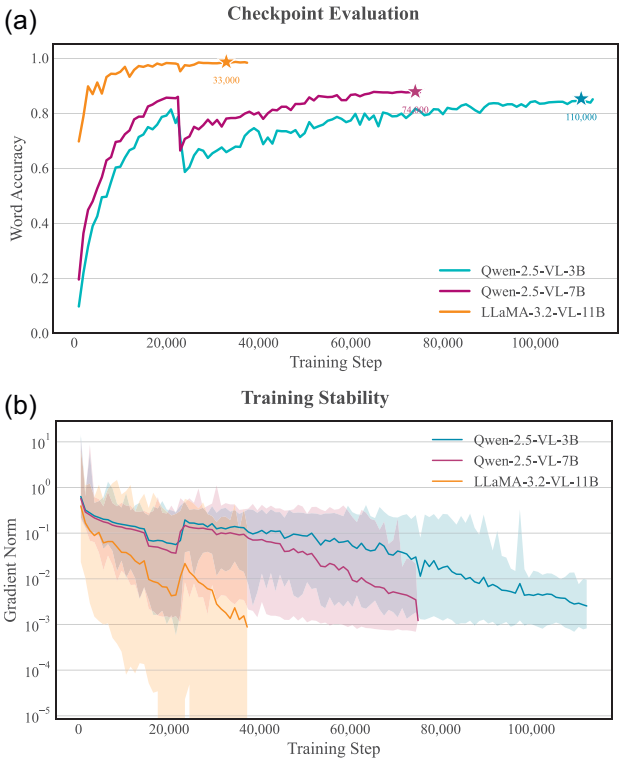


Figure 5. Training dynamics and checkpoint selection.

These gradient norm patterns align with the observed performance differences, where models with more stable gradient dynamics (Llama-3.2-11B) achieved superior OCR accuracy. The correlation between gradient stability and final performance underscores the importance of monitoring optimization health beyond traditional loss metrics when fine-tuning VLMs for specialized tasks.

Overall performance

Evaluation was conducted on 1,000 randomly selected samples from the synthetic validation dataset comprising diverse glyph sequences, as well as the real-world test dataset of 753 Manchu word images drawn from seven historical sources. The evaluation metrics included Word Accuracy, CER, F1 score, and Inference Time for Manchu script recognition. Tables 2 and 3 summarize performance across both datasets.

On the synthetic validation dataset (noting that these figures represent checkpoint-selection-set performance, as discussed in the "Model selection and training stability" section), Llama-3.2-11B-Vision achieved 98.6 percent word accuracy, 0.998 F1 score, and a CER of 0.002. The Qwen variants demonstrated competitive but lower performance: Qwen2.5-VL-7B reached 86.3 percent accuracy with 0.974 F1 score, while Qwen2.5-VL-3B achieved 83.5 percent accuracy with 0.968 F1 score. However, the Qwen models offered significant speed advantages, processing samples in 1.3–1.4 seconds compared to Llama's 14.7 seconds per sample.

The performance landscape changes dramatically on real-world test data. Llama-3.2-11B-Vision maintained competitive performance with 87.4 percent word accuracy (95% CI: 85.1%–89.7%), representing an 11.2 percentage point accuracy decline from validation. In contrast, the Qwen models suffered severe performance degradation: Qwen2.5-VL-7B dropped to 33.7 percent accuracy (95% CI: 30.2%–37.2%), a 52.6 point decline, and Qwen2.5-VL-

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press

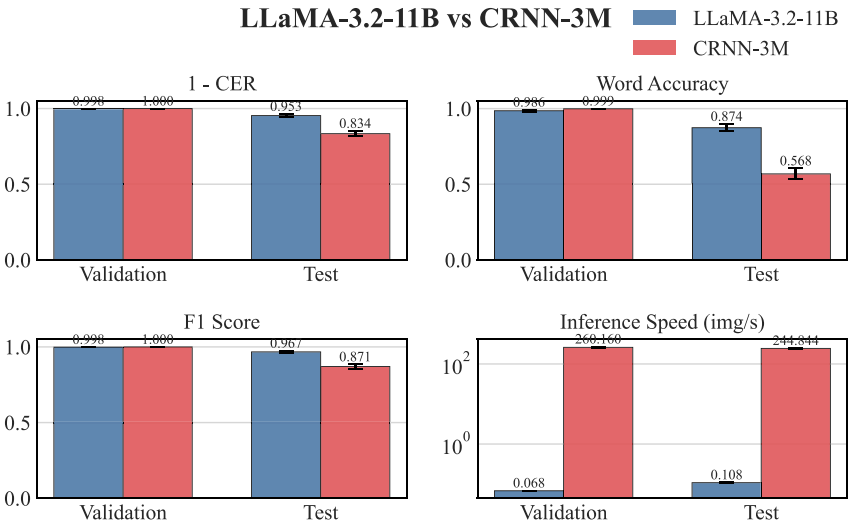


Figure 6. Performance comparison between CRNN and Llama-3.2-11B-Vision models.

3B fell to 27.8 percent accuracy (95% CI: 24.7%–30.9%), a 55.7 point decline. These results reveal fundamental differences in generalization capability, with Llama's superior architecture enabling more effective transfer from synthetic training data to real historical documents, while the Qwen models exhibit substantial overfitting to synthetic patterns.

Comparison with traditional OCR approaches

To contextualize the VLM performance, we conducted a comprehensive comparative evaluation against a traditional CRNN-based OCR system, representing the established approach for sequence recognition tasks in computer vision. The CRNN architecture combines CNNs for feature extraction with recurrent neural networks for sequence modeling, followed by connectionist temporal classification (CTC) decoding for alignment-free recognition. Detailed experimental configurations for the CRNN baseline are provided in Appendix C. Following the same rigorous checkpoint selection methodology as the VLM experiments, we evaluated all training checkpoints and identified epoch 48 as the optimal checkpoint for the CRNN model based on validation performance.

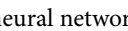
On the synthetic validation split, the CRNN model demonstrated exceptional specialization for Manchu script recognition, achieving 99.9 percent word accuracy, 0.9999 F1 score, and remarkably low CER of 0.0001. This performance substantially outperformed the Llama model on synthetic data, highlighting the effectiveness of task-specific architectural design when training and test distributions are well-matched. The CRNN's superior performance on controlled synthetic data reflects its optimized feature extraction pipeline, specifically tuned for the visual patterns present in rendered Manchu script.

However, this advantage reversed dramatically when evaluated on real-world test data. On the 753 real-world samples, CRNN performance deteriorated significantly to 56.8 percent word accuracy, 0.871 F1 score, and 0.166 CER, representing a substantial decrease of 43 percentage points. Although the CRNN model still outperformed both Qwen variants on real-world data (56.8% vs. 33.7% and 27.8%), it is substantially inferior to Llama-3.2-11B-Vision's 87.4 percent accuracy. This comparison reveals that while architectural specialization can achieve superior performance on matched distributions, the multimodal understanding and pre-trained representations of large VLMs provide crucial advantages for cross-domain generalization. Figure 6 illustrates this performance comparison, highlighting the superior generalization capability of the VLM approach and the fundamental trade-offs between specialized optimization and robust adaptability.

Findings and discussion

Error analysis

Detailed examination of successful predictions demonstrates the models' capacity for complex sequence recognition. The longest correctly predicted synthetic string by Llama-3.2-11B-Vision spans 13 characters:

Manchu: 
Roman: FERGUWEBURAKŪNGGE

This example illustrates the model's ability to maintain accuracy across extended character sequences, including complex consonant clusters and vowel harmonies characteristic of Manchu phonology.

Nevertheless, our analysis reveals systematic error patterns stemming from morphological similarity between Manchu glyphs. Table 4 and Figure 7 show that all models exhibit systematic vowel confusion, with Llama-3.2-11B-Vision demonstrating the highest error concentration (47.5% on test data), indicating more systematic rather than random failure patterns. The universal difficulty with ᡳ (A) across all models – appearing as the top error in half of all cases – reflects the inherent morphological complexity of Manchu script. Most failures occur with characters sharing similar or even identical stroke patterns and positional variants, particularly among vowels (ᡳ (A), ᡴ (E), ᡶ (I)). In some cases, these glyphs are visually indistinguishable in isolation and can only be correctly identified by considering the preceding and succeeding characters, as well as subtle diacritical marks, stroke orientations, and their position within the syllable (initial, medial, or final).

Performance comparison between VLM and CRNN

The substantial generalization advantage of VLMs highlights the contextual understanding capabilities inherent in VLMs. While traditional CNN-based approaches excel at pattern recognition

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press



Figure 7. Top misrecognized characters in Manchu predictions.

Table 4. Character-level error analysis on test and validation data

Model	Top 3 characters	Share
Validation dataset (1,000 synthetic samples)		
Llama-3.2-11B-Vision	(E) (21.6%), (A) (13.5%), (B) (8.1%)	43.2%
Qwen2.5-VL-7B	(E) (12.6%), (A) (12.4%), (I) (11.5%)	36.5%
Qwen2.5-VL-3B	(A) (13.1%), (E) (11.4%), (I) (10.6%)	35.1%
Test dataset (753 real-world samples)		
Llama-3.2-11B-Vision	(A) (22.2%), (I) (16.7%), (E) (8.6%)	47.5%
Qwen2.5-VL-7B	(A) (15.5%), (I) (14.1%), (E) (11.3%)	40.8%
Qwen2.5-VL-3B	(I) (14.9%), (A) (12.5%), (E) (11.4%)	38.9%

within constrained domains, VLMs leverage linguistic knowledge to maintain coherence even when encountering visually ambiguous or degraded characters. The dramatic performance gap between validation and test sets for CRNN (99.9% vs. 56.8%) compared to Llama (98.6% vs. 87.4%) underscores the fundamental advantage of multimodal language models for low-resource OCR applications.

The CRNN's overfitting to synthetic training patterns demonstrates the limitations of traditional approaches when faced with real-world document variability. In contrast, the robust cross-domain performance of VLMs suggests that their pre-trained linguistic representations provide valuable inductive biases for handling script variations and degradation typical in historical documents.

Limitations and future development

Despite the promising results reported above, several limitations constrain the current OCR system and point toward directions for future work. First, the model is designed to operate as part of an end-to-end pipeline in which images are normalized before OCR. Our evaluation assumes a standardized preprocessing workflow (binarization, denoising, contrast enhancement, and resizing), and performance is noticeably sensitive to deviations from these conditions. In particular, severe tilt, low-resolution or heavily compressed scans, uneven illumination, and strong background noise can all degrade recognition accuracy.

Second, the current system is trained and evaluated at the word level on carefully segmented images. Layout analysis, line segmentation, and word boundary detection for full-page archival scans remain outside the scope of this study. In practical applications, errors in upstream segmentation will propagate to the OCR stage, and the overall system accuracy will therefore depend on the robustness of the full pipeline rather than the recognition model alone. A related limitation is that our real-world evaluation focuses on formal Qing administrative manuscripts and printed documents and does not yet cover informal manuscripts or marginal annotations.

Third, inference speed presents a practical bottleneck for large-scale archival processing. Llama-3.2-11B-Vision requires 9–15 seconds per word image on an NVIDIA RTX A6000 Ada GPU, which makes processing full documents or entire archival collections time-intensive without further optimization, such as quantization or batched inference. The hardware itself also poses an accessibility constraint: the 48 GB VRAM required by the A6000 represents a significant cost that may limit adoption by individual researchers or smaller institutions, despite our emphasis on open-source and locally deployable infrastructure. While newer and more affordable GPU generations may alleviate this barrier over time, the current setup is not as widely accessible as consumer-grade hardware.

Finally, although Llama-3.2-11B-Vision achieves a promising 87.4 percent word-level accuracy on the real-world test set, this means that approximately one in eight words is incorrectly recognized. For the stringent accuracy requirements of humanities scholarship, where even small transcription errors can substantially affect interpretation and analysis, this error rate necessitates

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press

manual post-correction by domain experts before the outputs can be used for scholarly purposes.

These limitations suggest several concrete avenues for future development. First, we plan to fine-tune the model on a larger and more diverse dataset composed of both synthetic word images and scanned Manchu texts, encompassing a broader range of hands, print styles, and document genres. By exposing the model to greater real-world visual variation, we aim to improve accuracy and reduce sensitivity to scan quality and stylistic differences. Second, we intend to use the outputs of the current OCR system to fine-tune a paragraph-level correction model that performs post-processing on longer sequences. We expect such a model to correct systematic glyph confusions, enforce lexical and morphological coherence, and integrate information across word and line boundaries. Finally, we are extending the pipeline by developing robust image preprocessing, segmentation, and normalization modules, together with systematic evaluation of romanization quality. These components are essential steps toward a fully integrated Manchu OCR framework suitable for large-scale archival digitization.

Conclusion

This study demonstrates that state-of-the-art OCR for low-resource and endangered languages is now achievable using open-source VLMs that can be fine-tuned and deployed locally with modest computational resources. By leveraging synthetic data and parameter-efficient training, our Llama-3.2-11B-Vision-based system achieves a CER of 0.002 and 98.6 percent word-level accuracy on synthetic benchmarks. It also demonstrates promising potential in generalizing to real-world Manchu archival texts.

A key finding of this work is that high-quality synthetic datasets can be used to train OCR systems that perform well on real historical materials. This is especially crucial for digital humanities researchers, who often lack the resources or access to build large annotated datasets from real-world sources. Our results show that with careful preprocessing and model design, synthetic data can bridge the gap between resource constraints and the need for accurate, scalable OCR solutions.

Equally important, our approach can be run entirely on a single workstation or in accessible environments such as Google Colab, making advanced OCR technology available to humanities scholars, archivists, and independent researchers regardless of institutional resources. This local, cost-effective, and fully open-source solution addresses critical barriers in digital humanities, where data privacy, sustainability, and reproducibility are paramount.

This Manchu OCR system will significantly lower the technical barrier for historians and linguists working with Manchu archival materials. The complete implementation, including model weights, training scripts, and evaluation code, is publicly available at https://github.com/mic7ch1/ManchuAI-OCR/ to facilitate reproducibility and enable further research in low-resource language OCR. We hope this work will inspire further collaboration and innovation at the intersection of artificial intelligence and the digital humanities, contributing to the preservation and revitalization of the world's linguistic heritage.

Data availability statement. The synthetic Manchu dataset used for training is available from https://github.com/tyotakuki/ManchuOCR. Model weights, training scripts, and evaluation code are available at https://github.com/mic7ch1/ManchuAI-OCR/.

Acknowledgements. The authors thank the editor, the associate editor, and the two anonymous reviewers for their constructive comments and suggestions. The authors are also grateful to Dr. Javier Cha, Dr. Yumeng Hou, Dr. Eric Chow, and Mr. Solomon Ho of the Big Data Studies Lab, University of Hong Kong, for their valuable input during the preparation and revision of this article. Finally, the authors would like to thank Dr. Zhuohui Zhang for generously sharing the Manchu synthetic dataset. During the preparation of this manuscript, the authors made limited use of generative AI tools to help refine the wording of selected passages; all such content was subsequently reviewed and edited by the authors, who take full responsibility for it.

Author contributions. Conceptualization: Y.H.M.C.; Data curation: Y.H.M.C.; Formal analysis: Y.H.M.C. and D.C.; Investigation: Y.H.M.C. and D.C.; Methodology: Y.H.M.C. and D.C.; Project administration: Y.H.M.C.; Resources: Y.H.M.C.; Software: D.C.; Supervision: Y.H.M.C.; Validation: Y.H.M.C. and D.C.; Visualization: D.C.; Writing – original draft: Y.H.M.C. and D.C.; Writing – review and editing: Y.H.M.C. and D.C. Both authors approved the final submitted draft.

Funding statement. This work was supported by Schmidt Sciences through the Humanities and AI Virtual Institute under the project "Playing Heaven: Remapping Early Modern Neo-Confucian Worlds with AI."

Competing interests. The authors declare none.

Ethical standards. The authors affirm that this research did not involve human participants.

References

Ortai (鄂爾泰), et al. 1739. "General Gazetteer of the Eight Banners (Han i araha jakūn gūsai tung ji bithe; ⟨⟨八旗通志初集⟩⟩)." Kyoto Neifu block-print edition; Berlin State Library–Prussian Cultural Heritage Foundation digital edition, 2016. Vol. 178.

Crossley, P. K., and E. S. Rawski. 1993. "A Profile of the Manchu Language in Ch'ing History." *Harvard Journal of Asiatic Studies* 53, no. 1: 63–102.

Rang, M., Z. Bi, C. Liu, Y. Wang, and K. Han (2024). "An Empirical Study of Scaling Law for Scene Text Recognition." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15619–15629.

Disangga. 1736. "Manchu Palace Memorials, Qianlong Reign (⟨⟨宮中檔滿文奏摺・乾隆朝⟩⟩). Memorial by Disangga, Titled ⟨奏報雍正十三年比較堪看來年乾隆元年農祥雨水收成摺, Dated Yongzheng 13.12.29 (10 February 1736). NPM Archival ID: Gu-gong 157489, Item 1, Digital File K4D157489-0.pdf." Accessed 20 November 2025.

Furdan. 1728. "Manchu Palace Memorials, Guangxu Reign (⟨⟨宮中檔滿文奏摺・光緒朝⟩⟩). Memorial by Furdan, Titled ⟨奏謝天恩賞戴二眼花翎⟩; Dated Yongzheng 05.12.04 (14 January 1728). The NPM Catalog Assigns this Item to the Guangxu Reign, but the Memorial Date Corresponds to the Yongzheng Period; the Discrepancy Reflects the Archival Metadata in the Digital System. NPM Archival ID: Gu-gong 158156, Item 1, Digital File K4D158156-0.pdf." Accessed November 20, 2025.

Han, D., M. Han, and **Unsloth team**. 2023. "Unsloth: Finetune LLMs 2x Faster with 70% Less Memory. Open-Source Library for Efficient Fine-Tuning of Large Language Models."

Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. "LoRA: Low-Rank Adaptation of Large Language Models." In *Proceedings of the Tenth International Conference on Learning Representations (ICLR 2022)*. OpenReview.net. Virtual conference, April 25–29.

Huang, D., M. Li, R. Zheng, S. Xu, and J. Bi. 2017. "Synthetic Data and DAG-SVM Classifier for Segmentation-Free Manchu Word Recognition." In *2017 International Conference on Computing Intelligence and Information System (CIIS)*, 46–50. Piscataway, NJ: IEEE.

Levenshtein, V. I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals." *Soviet Physics Doklady* 10, no. 8: 707–10. Translated from Doklady Akademii Nauk SSSR, vol. 163, no. 4, pp. 845–848, 1965. Introduces the Levenshtein distance algorithm.

Li, G. R. 2000. *Manchu: A Textbook for Reading Documents*. Honolulu: University of Hawai'i Press.

Li, M., R. Zheng, S. Xu, Y. Fu, and D. Huang. 2018. "Manchu Word Recognition Based on Convolutional Neural Network with Spatial Pyramid Pooling." In *2018 11th International Congress on Image and Signal Processing,*

BioMedical Engineering and Informatics (CISP-BMEI), edited by Wei Li, Qingli Li, and Lipo Wang, 1–6. Piscataway, NJ: IEEE.

Meadows, T. T. 1849. *Translations from the Manchu: With the Original Texts, Prefaced by an Essay on the Language (⟨⟨清文敘略⟩⟩ Manju Gisun Be Majige Gisurehe Bithe)*. Canton: Press of S. Wells Williams. Digital edition available via Internet Archive, Accessed November 20, 2025.

Meta AI. 2024. "The Llama 3 Herd of Models." Preprint, arXiv:2407.21783.

Peiceng. 1849. "Manchu Palace Memorials, Xianfeng Reign (⟨⟨宮中檔滿文奏摺・咸豐朝⟩⟩). Memorial by Peiceng, Titled ⟨奏謝天恩並報到任日期⟩, Dated Daoguang 29.08.24 (10 October 1849). NPM Archival ID: Gugong 157901, Item 1, Digital File K4D157901-0.pdf." Accessed November 20, 2025.

Qwen Team. 2025. "Qwen2.5-VL: Vision-Language Model." https://huggingface.co/Qwen/Qwen2.5-VL

Rawski, E. S. 1996. "Presidential Address: Reenvisioning the Qing: The Significance of the Qing Period in Chinese History." *The Journal of Asian Studies* 55, no. 4: 829–50.

UNESCO. 2010. *Atlas of the World's Languages in Danger*, 3rd ed. Paris: UNESCO Publishing.

Varela, M. E. 2024. "Qwen for Jawi v1: A Model for Jawi OCR." Model created at National University of Singapore. https://huggingface.co/mevsg/qwen-for-Jawi-v1.

Veritable Records. n.d. "Daicing Gurun i taidzung genggiyen šu huwangdi i yargiyan kooli (⟨⟨大清太宗文皇帝滿文實錄卷五:天聰三年正月至十二月⟩⟩). Manchu-language text, vol. 5, section for Tiancong 3rd year, 8th month. Archival ID 故官012620 (件10); digital file K4A012620-009.pdf." 2–4. Accessed November 20, 2025.

Wang, Z., S. Lu, X. Wei, R. Su, Y. Qi, and W. Lu. 2024. "Learn More Manchu Words with a New Visual-Language Framework." *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, no. 8: 1–18.

Wu, Y., Y. Zhang, and **Editorial Committee of Neige cangben Manwen laodang**. 2009. *Old Manchu Archives from the Grand Secretariat (*内閣藏本滿文老檔*)*, Vol. 20. Shenyang: Liaoning minzu chubanshe.

Xu, W., Y. Liu, L. He, X. Huang, and L. Jiang. 2024. "Xmodel-VLM: A Simple Baseline for Multimodal Vision Language Model." Preprint, arXiv:2405.09215.

Zhang, D., Y. Liu, Z. Wang, and D. Wang. 2021. "OCR with the Deep CNN Model for Ligature Script-Based Languages like Manchu." *Scientific Programming* 2021, no. 1: 5520338.

Zhang, G..J. Li, R. He, and A. Wang. 2004. "An Offline Recognition Method of Handwritten Primitive Manchu Characters Based on Strokes." In *Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR-9)*, edited by F. Kimura and H. Fujisawa, 432–7. Los Alamitos, CA: IEEE Computer Society.

Zhang, G., J. Li, and A. Wang. 2006. "A New Recognition Method for the Handwritten Manchu Character Unit." In *2006 International Conference on Machine Learning and Cybernetics*, 3339–44. Piscataway, NJ: IEEE.

Zhang, P., G. Zeng, T. Wang, and W. Lu. 2024. "TinyLlama: An Open-Source Small Language Model." Preprint, arXiv:2401.02385.

Zhang, Z. 2022. "ManchuOCR: A OCR System for the Manchu Script." Accessed: May 29, 2025. https://github.com/tyotakuki/ManchuOCR

Zhao, J., J. Li, G. Zhang, and J. Wan. 2006. "Design and Implementation of off Line Handwritten Document Recognition System of Manchu Manuscript." *Pattern Recognition and Artificial Intelligence* 19, no. 6: 801–5.

Zheng, R., M. Li, J. He, J. Bi, and B. Wu. 2018. "Segmentation-Free Multi-Font Printed Manchu Word Recognition Using Deep Convolutional Features and Data Augmentation." In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, edited by Wei Li, Qingli Li, and Lipo Wang, 1–6. Piscataway, NJ: IEEE.

Appendices

Hardware and software specifications

Computing infrastructure

All experiments were conducted on a high-performance workstation. Table A1 summarizes the hardware configuration used for model training and inference.

Table A1. Hardware specifications for model training

Component	Specification
GPU	NVIDIA RTX A6000 Ada (48GB VRAM)
CPU	Intel i9-13900KS (32 cores)
RAM	192GB DDR5
Storage	4TB SSD

Table A2. Software environment specifications

Component	Version
Operating system	Ubuntu 24.04
CUDA	12.8
Python	3.11
PyTorch	2.9
Transformers	4.57
Datasets	4.3
Unsloth	2025.11.6

Software environment

Table A2 lists the software stack and library versions used throughout this study.

VLM inference and evaluation

To ensure reproducibility and fair comparison across models, we standardized the inference configuration for all VLM evaluations. Table B1 details the generation parameters, model loading settings, and text normalization procedures used during evaluation.

All models employed greedy decoding (`do_sample=False`) with default temperature (1.0) and top-p (1.0) values to ensure deterministic outputs. We set `max_new_tokens` to 1,536 to accommodate the longest possible Manchu word sequences while preventing excessive generation. Models were loaded in FP16 precision without quantization to maintain accuracy, and inference was performed on CUDA-enabled GPUs.

For output processing, we applied minimal text normalization: whitespace was stripped from generated outputs for parsing the structured response format. Importantly, no Unicode normalization was applied to preserve the integrity of Manchu Unicode characters, which can be affected by normalization forms. No explicit

Table B1. VLM inference settings

Parameter	Value
max_new_tokens	1,536
temperature	1.0 (default)
top_p	1.0 (default)
do_sample	False (greedy)
use_cache	True
pad_token_id	eos_token_id

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press

stop sequences were used, allowing models to naturally terminate generation based on their training.

The evaluation protocol was consistent across all experiments to ensure fair comparison. For synthetic validation, we used a 1,000-sample subset, while the real-world test set consisted of 753 Manchu word images. All metrics (word accuracy, character error rate (CER), and F1 score) were computed using the same implementations across both VLM and CRNN assessments. Note that the CRNN baseline is excluded from this comparison, as it is a character-level sequence model that outputs Manchu script directly without a romanization component.

CRNN baseline training

Model architecture

The CRNN baseline implements a deep convolutional recurrent architecture optimized for sequence recognition tasks. The model comprises three main components: a 9-layer convolutional feature extractor, a 4-layer bidirectional LSTM sequence modeler, and a linear classification layer with CTC decoding. Table C1 provides an overview of the architecture.

Convolutional feature extractor: The CNN backbone processes 64 × 480 pixel inputs through nine convolutional layers with progressive channel expansion (3 →64 →128 →256 →512). Each convolutional block includes batch normalization, ReLU activation, and light dropout (5% for CNN layers). Strategic max pooling operations reduce spatial dimensions while preserving sequence length: initial pooling reduces height to 32 ×240, then 16 ×120, followed by asymmetric pooling operations (2 ×1 kernel) that compress height to 4 ×120 while maintaining width for sequence modeling. The final layers use adaptive average pooling to produce 512-dimensional features per sequence position.

Recurrent sequence modeling: Feature maps are reshaped and fed into a 4-layer bidirectional LSTM with 256 hidden units per direction, producing 512-dimensional contextual representations. The deep LSTM architecture with 30 percent dropout between layers enables robust sequence modeling for variable-length Manchu word recognition.

Classification and decoding: A final linear layer with dropout projects the bidirectional LSTM outputs to character class probabilities, followed by CTC decoding to handle alignment-free sequence prediction without explicit character segmentation.

Table C1. CRNN model architecture overview

Component	Specification
CNN feature extractor	
Layers	9 convolutional layers
Channel progression	3 → 64 → 128 → 256 → 512
Input size	64 × 480 pixels
Output features	512-dimensional per position
LSTM sequence modeler	
Architecture	4-layer bidirectional LSTM
Hidden units	256 per direction (512 total)
Dropout rate	30% between layers
Output layer	
Type	Linear + CTC decoding
Function	Character class probabilities

Table C2. CRNN training configuration

Parameter	Value
Dataset	60,000 training samples
Optimizer	AdamW
Initial learning rate	1e-3
Weight decay	0.05
LR scheduler	CosineAnnealingWarmRestarts
Input image size	64 × 480 pixels
Hidden size	256
Dropout	0.3
Normalization	[0, 1]
Training epochs	100
Warmup epochs	5
Mixed precision	True
Gradient clipping	Max norm 1.0
Batch size	16

Training configuration

Table C2 details the training hyperparameters and configuration for the CRNN baseline.

Construction of the 753-item real-world evaluation dataset

The real-world evaluation dataset consists of 753 Manchu word images drawn from seven scanned sources, including five handwritten archival materials and two printed works. To avoid oversampling frequently occurring terms within the same source, we applied a uniform extraction rule: for each source, only the first appearance of each Manchu word type was included. When a word occurred multiple times, its earliest instance, determined by page, line, and top-to-bottom reading order, was selected. This procedure ensures that the dataset captures a diverse set of word forms rather than repeating identical tokens. After assembling the full set, two images were removed due to severe scan artifacts, resulting in the final total of 753 items. Table D1 presents the distribution of unique word images contributed by each source.

Table D1. Distribution of the 753 real-world Manchu word images by source

Source	Word images
General Gazetteer of the Eight Banners (1739)	144
Translations from the Manchu (Meadows, 1849)	147
Old Manchu Archives from the Grand Secretariat	134
Veritable Records of the Taizong Emperor (Manchu ed.)	159
Gongzhongdang Manwen Zou zhe – Qianlong reign	48
Gongzhongdang Manwen Zou zhe – Xianfeng reign	49
Gongzhongdang Manwen Zou zhe – Guangxu reign	72
Total	**753**

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press

Digitization quality varies among the seven sources, and these differences informed our preprocessing approach. One source, the *Old Manchu Archives from the Grand Secretariat*, derives from a modern published edition that underwent editorial cleaning before scanning; its images are therefore more regular and typographically standardized. All other sources were extracted directly from unedited color scans of handwritten or printed originals. For these materials, we employed a standardized semi-automated segmentation and preprocessing pipeline developed for this project. This pipeline includes binarization, denoising, glyph-segment isolation using OpenCV, and spatial clustering to reconstruct valid Manchu word units. Light normalization was applied to produce consistent input quality across the dataset and to match the conditions used in the OCR system.

https://doi.org/10.1017/chr.2026.10042 Published online by Cambridge University Press