# Fine-Tuning Vision-Language Model Berbasis QLoRA untuk Ekstraksi dan Transliterasi Teks Jawi Cetak

**Z Mubarak[1], R Jannah[2], A M Insaf[3], H Ahmadian[4], B Yusuf[*5]**

[1,2,3,4,5]Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Ar-Raniry, Banda Aceh, Indonesia

E-mail: zidanmubarak00@gmail.com[1], rozatuljannah2501@gmail.com[2], maulanambo031@gmail.com[3], hendri@ar-raniry.ac.id[4], bustamiyusoef@ar-raniry.ac.id[*5]

**Abstrak.** Aksara Jawi merupakan warisan budaya Nusantara yang penggunaannya kian menurun, sementara metode Optical Character Recognition (OCR) konvensional berbasis Convolutional Neural Network kesulitan menangani karakteristik tulisan Jawi yang kursif dan saling menyambung. Penelitian ini mengimplementasikan dan mengevaluasi fine-tuning model Vision-Language Model Qwen2.5-VL-3B-Instruct menggunakan Quantized Low-Rank Adaptation (QLoRA) untuk tugas ekstraksi teks Jawi cetak sekaligus transliterasinya ke huruf Latin secara end-to-end. Model dilatih menggunakan 34.380 gabungan data sintetik dan data autentik, kemudian diuji pada 10% data autentik yang belum pernah dilihat model selama pelatihan. Model dasar (zero-shot) menghasilkan Character Error Rate (CER) ekstraksi Jawi sebesar 0,825, sedangkan model hasil fine-tuning pada epoch ketiga menurunkannya menjadi 0,174, atau penurunan kesalahan sekitar 79%. Nilai CER transliterasi Rumi turun dari 0,906 menjadi 0,165, atau penurunan sekitar 82%. Studi ablasi menunjukkan kombinasi data sintetik dan autentik memberikan hasil terbaik dibanding penggunaan salah satu jenis data saja. Model dengan performa terbaik kemudian didemonstrasikan melalui antarmuka Gradio pada Hugging Face Spaces. Temuan ini menunjukkan bahwa fine-tuning berbasis QLoRA efektif meningkatkan kemampuan model mengekstraksi dan mentransliterasikan aksara Jawi, sekaligus efisien karena hanya memperbarui 2,14% dari total parameter model, sehingga layak digunakan untuk mendukung digitalisasi naskah beraksara Jawi pada sumber daya komputasi yang terbatas.

**Kata kunci:** Aksara Jawi; Vision-Language Model; QLoRA; Optical Character Recognition; transliterasi

**Abstract.** *The Jawi script is part of the Nusantara's cultural heritage, yet its readership continues to decline, while conventional Convolutional Neural Network-based Optical Character Recognition (OCR) struggles with its cursive, character-linking nature. This study implements and evaluates the fine-tuning of the Qwen2.5-VL-3B-Instruct Vision-Language Model using Quantized Low-Rank Adaptation (QLoRA) for end-to-end extraction of printed Jawi text and its transliteration into Latin script. The model was trained on 34.380 combined dataset of synthetic and authentic, then evaluated on a held-out 10% of the authentic data never seen during training. The zero-shot baseline produced a Character Error Rate (CER) of 0.825 for Jawi extraction, while the fine-tuned model at the third epoch reduced this to 0.174, a reduction of roughly 79%. The CER for Rumi transliteration fell from 0.906 to 0.165, a reduction of roughly 82%. An ablation study showed that combining synthetic and authentic data yielded the best results compared to using either data type alone. The best-performing model was then deployed through a Gradio interface on Hugging Face Spaces. These findings indicate that*

*QLoRA-based fine-tuning is effective in improving the model's ability to extract and transliterate Jawi script, while remaining resource-efficient by updating only 2.14% of the model's total parameters, making it a practical strategy for supporting the digitization of Jawi manuscripts under limited computational resources.*

**Keywords:** *Jawi script; Vision-Language Model; QLoRA; Optical Character Recognition; transliteration*

## 1. Pendahuluan

Aksara Jawi merupakan salah satu warisan intelektual dan budaya penting di Asia Tenggara yang selama berabad-abad menjadi media utama penyebaran ilmu pengetahuan di kawasan Melayu-Nusantara [1]. Namun demikian, kemampuan membaca aksara ini terus menurun di kalangan generasi muda seiring semakin dominannya huruf Latin dalam pendidikan dan administrasi [2], sehingga sejumlah besar naskah bersejarah berisiko tidak lagi dapat dipahami tanpa upaya digitalisasi yang memadai. Kondisi ini menempatkan digitalisasi aksara Jawi sebagai agenda yang mendesak, sekaligus sejalan dengan upaya pemanfaatan kecerdasan buatan yang lebih luas untuk pelestarian sumber daya budaya dan pendidikan yang selama ini sulit diakses oleh khalayak.

Secara umum, Optical Character Recognition (OCR) bekerja melalui tahapan praproses citra, deteksi area teks, hingga pengenalan karakter, dengan akurasi yang bergantung pada kualitas citra masukan [3]. Pendekatan *deep learning* kini mendominasi riset deteksi dan pengenalan teks dibanding metode fitur tangan konvensional [4], dan telah banyak meningkatkan kinerja tugas ini pada berbagai domain [5]. Digitalisasi otomatis aksara Jawi tetap menghadapi tantangan yang jauh lebih besar: aksara ini bersifat kursif dan saling menyambung, dengan bentuk huruf yang berubah tergantung posisinya dalam kata, serta sangat bergantung pada posisi titik diakritik yang jika salah dikenali dapat mengubah makna kata secara drastis [6].

Karakteristik tersebut membuat pendekatan OCR konvensional berkinerja terbatas pada aksara Jawi. Abdiansah dkk. [7] menerapkan Convolutional Neural Network (CNN) untuk tulisan tangan aksara Jawa dan memperoleh akurasi hingga 99,83%, sementara Safrizal [8] menggunakan Support Vector Machine (SVM) berbasis fitur sudut dan hanya mencapai akurasi 78,86%, mengindikasikan bahwa metode klasik seperti SVM kurang optimal dibanding *deep learning* untuk variasi tulisan yang kompleks. Kedua penelitian tersebut, meskipun menjanjikan, sama-sama berfokus pada karakter tulisan tangan yang relatif terisolasi sehingga belum merepresentasikan kondisi naskah cetak yang hurufnya saling menyambung rapat. Pada konteks naskah cetak, Handoko dkk. [9] menerapkan CNN untuk aksara Bima dan menemukan performa model menurun tajam ketika huruf saling terhubung rapat, karakteristik yang juga dimiliki aksara Jawi. Pendekatan berbasis deteksi objek turut dieksplorasi oleh Khairan dkk. [10], yang menerapkan YOLOv8 untuk mendeteksi kata pada skrip Harah Jawoe varian Aceh dan mencapai *mean Average Precision* (mAP50-95) sebesar 72,4%, tetapi pendekatan ini berhenti pada tahap deteksi kata dan belum mengintegrasikan pemahaman bahasa untuk transliterasi *end-to-end*. Di sisi lain, kerangka OCRNet oleh Nagasubramanian dkk. [11] menunjukkan bahwa arsitektur hibrida CNN dengan lapisan rekuren mampu meningkatkan akurasi pengenalan karakter alfanumerik secara umum, namun belum diuji pada skrip Arab-turunan yang kompleks seperti Jawi. Secara keseluruhan, penelitian-penelitian ini menegaskan bahwa pendekatan berbasis visi saja, tanpa pemahaman konteks bahasa, masih menyisakan keterbatasan signifikan pada naskah Jawi cetak yang kursif dan padat sambungan.

Keterbatasan tersebut mendorong riset ke arah Vision-Language Model (VLM) berbasis Multimodal Large Language Model (MLLM), yang mampu memahami konteks visual dan bahasa secara bersamaan, sebagaimana dirangkum pada sejumlah survei terkini mengenai arsitektur dan kemampuan VLM/MLLM untuk berbagai tugas visi [12][13]. Pada domain skrip Arab, yang secara historis menjadi induk aksara Jawi, penerapan *fine-tuning* VLM telah menunjukkan hasil yang sangat baik. QARI-OCR yang dikembangkan dari Qwen2-VL mencapai Word Error Rate (WER) 0,160 dan Character Error Rate (CER) 0,061 pada teks Arab cetak [14], sementara Qalam yang menggabungkan *encoder* SwinV2 dan *decoder* RoBERTa mencapai WER 0,80% untuk tulisan tangan dan 1,18% untuk teks cetak Arab [15]. Meskipun demikian, keberhasilan tersebut belum tentu dapat digeneralisasi ke skrip Nusantara. Benchmark multimodal dan multibahasa NusaAksara menunjukkan bahwa kemampuan model-model

VLM masih terbatas ketika diterapkan pada skrip lokal dengan huruf tambahan dan konvensi penulisan yang berbeda dari Arab standar [16]. Aksara Jawi, misalnya, memiliki lima huruf tambahan khas Melayu serta konvensi penulisan tanpa vokal yang lebih konsisten dibanding Arab standar. Sejauh penelusuran penulis, belum ada penelitian yang secara khusus menerapkan *fine-tuning* VLM untuk tugas ekstraksi sekaligus transliterasi aksara Jawi cetak secara *end-to-end* dengan evaluasi CER dan WER, sebuah celah yang signifikan mengingat aksara ini memiliki populasi penutur dan koleksi naskah yang jauh lebih besar dibanding banyak skrip minoritas lain yang sudah lebih dahulu diteliti secara intensif.

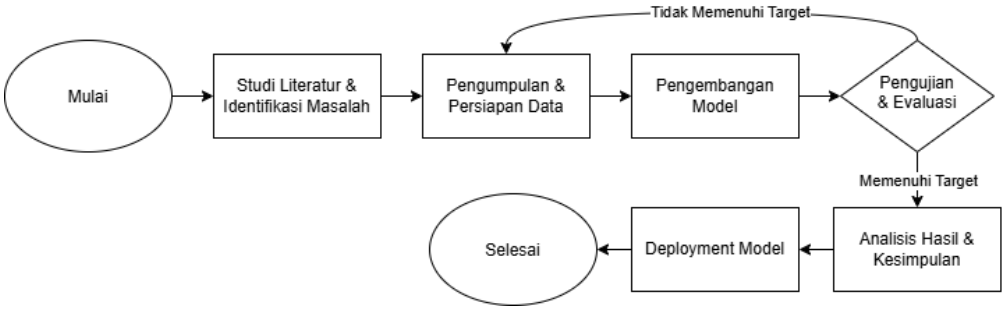
Untuk memanfaatkan kemampuan VLM secara efisien, penelitian ini menggunakan Qwen2.5-VL-3B-Instruct [17], sebuah model yang menggabungkan *vision encoder* resolusi dinamis berbasis Vision Transformer [18] dengan *decoder* bahasa dari keluarga Qwen2.5. Melatih ulang seluruh parameter model berskala miliaran seperti ini secara penuh membutuhkan sumber daya komputasi yang sangat besar, sehingga pendekatan Parameter-Efficient Fine-Tuning (PEFT) menjadi alternatif yang banyak diadopsi karena mampu mengadaptasi model besar dengan memperbarui hanya sebagian kecil parameter, tanpa mengorbankan kinerja secara berarti [19][20]. Low-Rank Adaptation (LoRA) merupakan salah satu teknik PEFT yang paling banyak digunakan, yaitu dengan menyisipkan adapter berukuran kecil pada lapisan tertentu sementara parameter utama model dibekukan [21]. Quantized Low-Rank Adaptation (QLoRA) mengembangkan lebih lanjut pendekatan ini dengan mengombinasikan kuantisasi 4-bit pada model dasar dan adapter LoRA yang dilatih dalam presisi lebih tinggi, sejalan dengan prinsip kompresi model yang bertujuan menekan kebutuhan memori tanpa mengorbankan kemampuan model secara signifikan [22], sehingga memungkinkan penyesuaian model besar pada perangkat keras dengan sumber daya terbatas [23].

Berdasarkan uraian tersebut, dapat disimpulkan bahwa penelitian mengenai pengenalan aksara Jawi masih bertumpu pada pendekatan visi konvensional yang belum mengintegrasikan pemahaman bahasa secara menyeluruh, sementara penelitian VLM mutakhir yang telah terbukti unggul pada skrip Arab standar belum diterapkan secara khusus pada aksara Jawi dengan evaluasi yang ketat. Berdasarkan celah tersebut, penelitian ini bertujuan mengimplementasikan dan mengevaluasi *fine-tuning* Qwen2.5-VL-3B-Instruct menggunakan QLoRA untuk tugas ekstraksi dan transliterasi aksara Jawi cetak secara *end-to-end*. Kontribusi utama penelitian ini adalah: (1) sejauh penelusuran penulis, penerapan pertama *fine-tuning* VLM *end-to-end* untuk aksara Jawi cetak yang mengevaluasi tugas ekstraksi dan transliterasi secara terpisah menggunakan CER dan WER; (2) studi ablasi yang mengukur kontribusi relatif data sintetik dan data autentik terhadap kemampuan generalisasi model, yang mengindikasikan bahwa kesesuaian karakteristik data dengan kondisi data uji nyata turut berperan penting di samping volume data; dan (3) penyediaan antarmuka inferensi yang dapat direproduksi melalui Hugging Face Spaces, sehingga hasil penelitian dapat langsung diverifikasi dan dimanfaatkan oleh peneliti maupun praktisi lain yang bergerak dalam pelestarian naskah Nusantara.

## 2. Metode Penelitian

### *2.1. Desain dan Alur Penelitian*

Penelitian ini merupakan penelitian eksperimental kuantitatif berbasis rekayasa sistem. Model Qwen2.5-VL-3B-Instruct disesuaikan melalui *fine-tuning* menggunakan QLoRA, kemudian kinerjanya diukur secara objektif menggunakan CER dan WER pada dua keluaran yang dievaluasi terpisah, yaitu ekstraksi aksara Jawi dan transliterasi ke huruf Latin (Rumi). Tahapan penelitian ditunjukkan pada Gambar 1.



**Gambar 1.** Alur penelitian, dari studi literatur hingga *deployment* model.

*2.2. Persiapan Dataset*

Dataset penelitian disusun dari dua sumber yang saling melengkapi, yaitu data sintetik dan data autentik, dengan proporsi dan hasil pengolahannya dilaporkan pada Subbab 3.1. Data sintetik dihasilkan melalui *web crawling* pasangan teks Jawi-Rumi dari situs keislaman, dikonversi ke aksara Jawi digital, lalu dirender menjadi citra menggunakan lima jenis *font* (Amiri, Harmattan, Lateef, Scheherazade New, dan Traditional Arabic) untuk meniru variasi tulisan pada dokumen asli. Data autentik diperoleh dari koleksi pribadi berupa citra naskah Jawi cetak yang dikumpulkan melalui pemotretan langsung dan tangkapan layar dokumen digital, mencakup variasi kualitas cetak, tekstur kertas, dan kondisi fisik dokumen yang merepresentasikan naskah nyata (Rincian lebih lanjut masih dalam persiapan untuk dipublikasi).

Seluruh data distandarkan mengikuti Pedoman Ejaan Jawi yang Disempurnakan (PEJYD) dari Dewan Bahasa dan Pustaka Malaysia, mencakup 28 huruf Arab standar dan lima huruf tambahan khas Melayu (Ca, Nga, Pa, Ga, dan Nya). Dataset dibagi menjadi data latih (seluruh data sintetik dan 90% data autentik), data validasi (5% dari gabungan data latih), dan data uji (10% sisa data autentik yang murni tanpa campuran data sintetik, agar evaluasi mencerminkan kinerja pada naskah nyata). Proses pengacakan (*shuffle*) diterapkan pada tiga tahap, yaitu pemuatan data mentah, pembuatan subset eksperimen, dan penggabungan data latih akhir, untuk mencegah model melupakan pola akibat melihat data secara berurutan (*catastrophic forgetting*). Panjang token maksimum (*max sequence length*) ditetapkan 1024 token berdasarkan analisis distribusi panjang token dataset yang dipaparkan pada Subbab 3.1, sehingga nilai tersebut menampung hampir seluruh data tanpa pemotongan sekaligus menghindari pemborosan memori GPU. Seluruh citra diproses pada resolusi aslinya (*dynamic resolution*) tanpa dipaksa ke ukuran tertentu, dan dikonversi ke format RGB.

*2.3. Fine-Tuning Model dengan QLoRA*

Mengikuti kerangka QLoRA sebagaimana diuraikan sebelumnya, model dasar Qwen2.5-VL-3B-Instruct terlebih dahulu dikuantisasi ke format 4-bit NormalFloat4 (NF4) dan dibekukan, sehingga pengetahuan umum yang telah dipelajari sebelumnya tetap terjaga. Adapter LoRA kemudian disisipkan pada seluruh lapisan kritis arsitektur, yaitu lapisan visi, lapisan bahasa, mekanisme *attention*, serta modul *feed-forward* (MLP), sehingga model dapat mempelajari pola khusus aksara Jawi tanpa memperbarui seluruh parameter dasarnya. Dari total 3.836.792.832 parameter model, hanya 82.169.856 parameter (2,14%) yang diperbarui melalui adapter LoRA selama pelatihan. Konfigurasi final, hasil dari pencarian *hyperparameter* awal yang dilaporkan pada Subbab 3.2, disajikan pada Tabel 1.

**Tabel 1.** Konfigurasi *hyperparameter* QLoRA final.

| Parameter | Nilai |
|---|---|
| Kuantisasi | 4-bit NormalFloat4 (NF4) |
| LoRA rank (r) | 32 |
| LoRA alpha | 64 (2× rank) |
| LoRA dropout | 0,0 |
| Target modules | Vision, Language, Attention, MLP |
| Learning rate | $1\times10^{-4}$ |
| Batch size / accumulation | 4 / 4 (efektif 16) |
| Max sequence length | 1024 token |
| Parameter terlatih | 82.169.856 dari 3.836.792.832 (2,14%) |

*2.4. Prompt Engineering*

Instruksi sistem (*system prompt*) dirancang untuk menegaskan dua langkah berurutan, yaitu ekstraksi teks Jawi dari citra seakurat mungkin, kemudian transliterasi hasil ekstraksi ke huruf Latin mengikuti

PEJYD, tanpa penerjemahan makna. Format keluaran dibakukan dalam dua baris berlabel "Jawi:" dan "Rumi:" agar dapat diuraikan secara otomatis oleh sistem evaluasi. Prompt yang identik digunakan secara konsisten pada penyusunan data pelatihan, proses *fine-tuning*, maupun pengujian model dasar dan model hasil *fine-tuning*, sehingga perbandingan kinerja antar tahap tetap adil. Teks lengkap kedua *prompt* tersebut adalah sebagai berikut.

> *System Prompt: "Anda adalah sistem multimodal untuk ekstraksi dan transliterasi aksara Jawi cetak. Tugas Anda terdiri dari dua langkah yang harus dilakukan secara berurutan: 1. Ekstraksi: Baca gambar dan ekstrak teks aksara Jawi (huruf Arab) yang tertulis pada gambar secara akurat, persis seperti yang tertulis. 2. Transliterasi: Konversi teks Jawi yang telah diekstrak ke huruf Latin (Rumi) sesuai pedoman Ejaan Jawi Yang Disempurnakan (PEJYD) oleh Dewan Bahasa dan Pustaka Malaysia. ATURAN OUTPUT: Keluarkan hasil dalam format berikut secara tepat: Jawi: [teks aksara Arab hasil ekstraksi] Rumi: [teks Latin hasil transliterasi]. Jangan menambahkan penjelasan, komentar, atau teks lain di luar format tersebut. Transliterasi bersifat karakter ke karakter, bukan penerjemahan makna."*
>
> *User Prompt: "Ekstrak teks aksara Jawi dari gambar ini, kemudian transliterasikan ke huruf Latin (Rumi)."*

### *2.5. Evaluasi Kinerja Model*

Kinerja model diukur menggunakan Character Error Rate (CER) dan Word Error Rate (WER) [24], yang dihitung terpisah untuk tugas ekstraksi Jawi dan transliterasi Rumi:

$$CER = (Sc + Dc + Ic) / Nc \quad (1)$$

$$WER = (Sw + Dw + Iw) / Nw \quad (2)$$

dengan S, D, dan I berturut-turut menyatakan jumlah substitusi, penghapusan, dan penyisipan pada tingkat karakter (c) atau kata (w), dan N adalah jumlah karakter/kata pada referensi. Model dinyatakan memenuhi target apabila CER berhasil ditekan hingga di bawah 15%, ambang yang ditetapkan sebagai target kerja penelitian ini mengingat belum ditemukan acuan baku CER untuk tugas ekstraksi-transliterasi aksara non-Latin seperti Jawi.

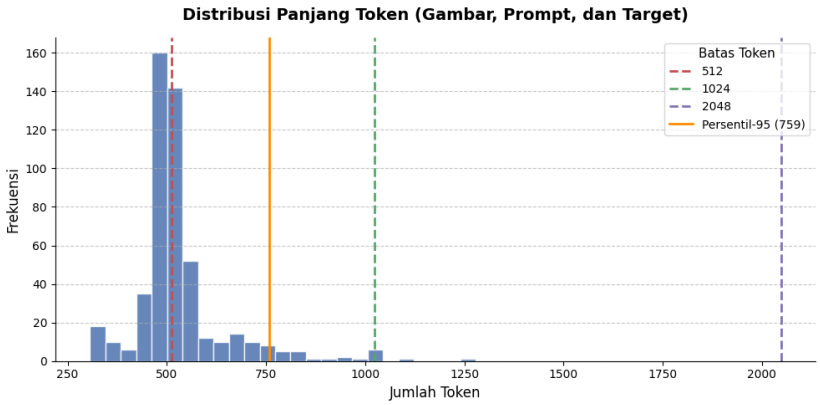
Untuk mengukur kontribusi relatif setiap jenis data, dilakukan studi ablasi dengan tiga skenario pelatihan. Skenario A menggunakan seluruh data sintetik sebagai data latih, Skenario B menggunakan seluruh data autentik sebagai data latih, dan Skenario C menggunakan gabungan kedua jenis data sebagai model dengan performa terbaik pada penelitian ini. Ketiga skenario ini dievaluasi pada data uji autentik yang sama dan juga digunakan pada pengujian model dasar (*baseline*), sehingga hasil ketiganya, termasuk Skenario A dan Skenario B, dapat dibandingkan secara langsung satu sama lain maupun terhadap baseline.

## 3. Hasil dan Pembahasan

### *3.1. Dataset*

Setelah melalui tahap standarisasi dan pembersihan yang dijelaskan pada Subbab 2.2, dataset akhir yang digunakan pada penelitian ini berjumlah 34.380 sampel, terdiri atas data sintetik dan data autentik. Komposisi ini menjadi dasar pembagian data latih, validasi, dan uji yang digunakan pada seluruh eksperimen selanjutnya.

Analisis distribusi panjang token pada seluruh dataset (Gambar 2) menunjukkan sebaran yang menceng ke kanan (*right-skewed*), dengan mayoritas sampel terkonsentrasi pada rentang 400 sampai 700 token dan ekor distribusi memanjang hingga lebih dari 2000 token pada beberapa sampel bertata letak panjang. Rata-rata panjang token tercatat 530, dengan persentil ke-95 berada di angka 759. Informasi inilah yang menjadi dasar penetapan *max sequence length* sebesar 1024 token pada Subbab 2.3, karena batas tersebut menampung hampir seluruh distribusi tanpa memotong data, sekaligus jauh lebih efisien secara memori dibanding menetapkan batas di atas 2000 token untuk mengakomodasi seluruh ekor distribusi.




**Gambar 2.** Distribusi panjang token pada dataset Jawi (gambar, *prompt*, dan target).

*3.2. Kinerja Model Dasar dan Pencarian Hyperparameter*

Sebelum *fine-tuning*, model diuji pada kondisi *zero-shot* untuk mengetahui kemampuan dasarnya. Model dasar menghasilkan CER ekstraksi Jawi 0,825 dan WER 0,961, sedangkan untuk transliterasi Rumi CER mencapai 0,906 dan WER mencapai 1,071 (baris Baseline pada Tabel 4). Keluaran model pada kondisi ini jauh berbeda dari teks aslinya, wajar mengingat model dasar dilatih untuk memahami bahasa-bahasa umum dan belum pernah diberi contoh aksara Jawi.

Sebelum pelatihan utama, dilakukan pencarian *hyperparameter* awal pada subset 500 sampel data (1 epoch) untuk tiga kombinasi *rank* LoRA dan *learning rate*, sebagaimana disajikan pada Tabel 2. Pemilihan konfigurasi terbaik difokuskan pada nilai CER, bukan pada training/validation loss, karena CER mengukur langsung ketepatan model membaca bentuk huruf.

**Tabel 2.** Hasil eksperimen pencarian konfigurasi *hyperparameter*.

| Konfigurasi | Rank LoRA | Learning Rate | CER Jawi | CER Rumi |
|---|---|---|---|---|
| 1 | 32 | $5\times10^{-5}$ | 0,271 | 0,447 |
| 2 | 64 | $2\times10^{-5}$ | 0,402 | 0,575 |
| 3 | 32 | $1\times10^{-4}$ | 0,246 | 0,412 |

Konfigurasi 3 (*rank* 32, *learning rate* $1\times10^{-4}$) menghasilkan CER terendah pada kedua tugas sehingga dipilih untuk pelatihan final selama 5 epoch, dengan konfigurasi lengkap seperti telah disajikan pada Tabel 1.

*3.3. Hasil Pelatihan dan Pemilihan Model Terbaik*

Model dilatih selama 5 epoch pada konfigurasi terpilih. Tabel 3 memperlihatkan validation loss menurun konsisten hingga epoch ketiga (0,0252), kemudian meningkat kembali pada epoch keempat dan kelima meski training loss terus menurun, yaitu pola khas *overfitting* ketika model semakin menghafal data latih namun generalisasinya terhadap data baru tidak lagi membaik. Checkpoint epoch ketiga karena itu dipilih sebagai model dengan performa terbaik.

**Tabel 3.** Riwayat nilai loss pelatihan final (5 epoch).

| Epoch | Training Loss | Validation Loss | Indikasi Status |
|---|---|---|---|
| 1 | 0,0277 | 0,0306 | Normal |
| 2 | 0,0197 | 0,0256 | Normal |
| 3 | 0,0094 | 0,0252 | Terbaik |
| 4 | 0,0038 | 0,0271 | Overfitting |
| 5 | 0,0013 | 0,0288 | Overfitting |

*3.4. Studi Ablasi Pengaruh Komposisi Data*

Model dengan performa terbaik (Skenario C) menurunkan CER ekstraksi Jawi dari 0,825 menjadi 0,174 dan CER transliterasi Rumi dari 0,906 menjadi 0,165, atau penurunan kesalahan berturut-turut sekitar 79% dan 82%. Untuk mengetahui kontribusi setiap jenis data, Tabel 4 membandingkan kinerja Baseline dengan tiga skenario komposisi data pelatihan, yaitu Skenario A (data sintetik saja), Skenario B (data autentik saja), dan Skenario C (gabungan, model dengan performa terbaik).

**Tabel 4.** Perbandingan kinerja Baseline dan skenario ablasi komposisi data pada data uji autentik. Semua skenario masing-masing diuji pada data uji yang sama.

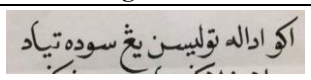
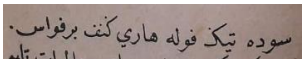
| Skenario | Ekstraksi | | Transliterasi | |
|---|---|---|---|---|
| | **CER Jawi** | **WER Jawi** | **CER Rumi** | **WER Rumi** |
| Baseline | 0,825 | 0,961 | 0,906 | 1,071 |
| A - Sintetik | 0,311 | 0,744 | 0,338 | 0,718 |
| B - Autentik | 0,208 | 0,570 | 0,229 | 0,508 |
| C - Gabungan | 0,174 | 0,500 | 0,165 | 0,394 |

Urutan kinerja dari yang paling lemah ke paling kuat adalah Baseline, Skenario A, Skenario B, dan Skenario C. Urutan ini mengindikasikan dua hal. Pertama, *fine-tuning* dengan data apa pun terbukti jauh lebih baik daripada tidak sama sekali. Kedua, meskipun jumlah data autentik jauh lebih sedikit dibanding data sintetik, data autentik memberikan penurunan CER yang lebih besar pada Skenario B dibanding Skenario A, mengindikasikan bahwa kesesuaian karakteristik data pelatihan dengan kondisi data uji nyata turut berperan penting bagi generalisasi model, di samping volume data itu sendiri. Kombinasi keduanya pada Skenario C tetap memberikan hasil terbaik karena data sintetik membekali model dengan pemahaman bentuk dasar huruf secara bersih dan terstruktur, sementara data autentik melatih ketahanan model terhadap gangguan visual nyata. Kedua sifat ini saling melengkapi dan tidak saling menggantikan.

*3.5. Analisis Kesalahan dan Perbandingan Prediksi*

Untuk menggambarkan dampak fine-tuning secara konkret, Tabel 5 membandingkan keluaran model dasar (*baseline*) dan model dengan performa terbaik pada dua potongan naskah Jawi cetak yang sama.

**Tabel 5.** Perbandingan prediksi model dasar (*baseline*) dan model dengan performa terbaik (*Fine-Tuned*) pada potongan naskah Jawi cetak.

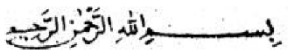
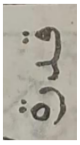
| Potongan Naskah | Baseline | Fine-Tuned |
|---|---|---|
|  | **Jawi:** اكو اداله توليسن يڠ سوده تياد<br>**Rumi:** أكو أداله توليسن يڠ سوده تياد<br><br>CER Jawi: 0.137; WER Jawi: 0.500;<br>CER Rumi: 0.857; WER Rumi: 1.000 | **Jawi:** اكو اداله توليسن يَ سوده تياد<br>**Rumi:** aku adalah tulisan yang sudah tiada<br><br>CER Jawi: 0.000; WER Jawi: 0.000;<br>CER Rumi: 0,000; WER Rumi: 0.000 |
|  | **Jawi:** سوده تيک فوله هاري كنف برفراس<br>**Rumi:** Sode teik foulah hari kent burfaras<br><br>CER Jawi: 0.133; WER Jawi: 0.500;<br>CER Rumi: 0.486; WER Rumi: 0.833 | **Jawi:** سوده تيڬ فوله هاري كفد برفواس .<br>**Rumi:** sudah tiga puluh hari kepada berpuasa.<br><br>CER Jawi: 0.133; WER Jawi: 0.500;<br>CER Rumi: 0.108; WER Rumi: 0.166 |

Tabel 5 menunjukkan bahwa model dasar secara konsisten menghasilkan kata-kata yang tidak bermakna dan karakter Arab yang masih tersisa pada kolom Rumi, seperti terlihat pada baris pertama, tempat kolom Rumi *baseline* masih menampilkan aksara Arab (أكو أداله توليسن...) alih-alih huruf Latin. Kegagalan ini terjadi karena model dasar belum pernah dilatih dengan pasangan data Jawi-Rumi, sehingga tidak memahami instruksi untuk mentransliterasikan aksara Jawi ke huruf Latin dan hanya menyalin ulang, atau sedikit mengubah, teks Jawi yang sama pada bagian yang seharusnya berisi hasil

transliterasi. CER pada kondisi ini berkisar 0,48 sampai 0,85, sementara model dengan performa terbaik menghasilkan transliterasi yang jauh lebih mendekati naskah asli dan sesuai kaidah bahasa Melayu, dengan CER turun ke kisaran 0,00 sampai 0,10 pada potongan naskah yang sama.

Di luar perbandingan tersebut, model dengan performa terbaik juga diuji pada beberapa kondisi citra yang menantang, dibandingkan terhadap label ground truth pada Tabel 6.

**Tabel 6.** Contoh kondisi citra yang masih menantang bagi model dengan performa terbaik, dibandingkan dengan label *ground truth*.

| Potongan Naskah | Teks Asli (Ground Truth) | Prediksi Model Dengan Performa Terbaik |
|---|---|---|
|  | **Jawi:** بِسمِ اللهِ الرَّحْمَنِ الرَّحِيْمِ<br>**Rumi:** bismillahirrahmanirrahim<br><br>Kondisi: mengandung tanda kurung/harakat di luar pola mayoritas data latih | **Jawi:** بسم الله وَالْحَمْدُ لِلّٰهِ رَبِّ الْعَالَمِيْنَ<br>**Rumi:** bismillahirrahmanirrahimualhamdulilla hrabbal alamin<br><br>CER Jawi: 0.811; WER Jawi: 1.500; CER Rumi: 1.125; WER Rumi: 2.000 |
|  | **Jawi:** قدرة<br>**Rumi:** qudrah<br><br>Kondisi: orientasi citra vertikal, berbeda dari mayoritas data latih yang horizontal | **Jawi:** اف باڤيق<br>**Rumi:** apa banyak<br><br>CER Jawi: 1.750; WER Jawi: 2.000; CER Rumi: 1.500; WER Rumi: 2.000 |

Kedua kondisi pada Tabel 6 menunjukkan bahwa model dengan performa terbaik masih mengalami kesalahan pada situasi tertentu, yaitu (1) teks yang memuat tanda baca atau harakat, karena mayoritas data latih berupa teks tanpa harakat, dan (2) citra dengan orientasi vertikal, berbeda dari mayoritas data latih yang berorientasi horizontal, sehingga model keliru menebak kata lain yang pola visualnya lebih familiar. Model juga belum diuji secara memadai pada citra berupa paragraf penuh karena mayoritas data latih berupa potongan satu baris; pada paragraf panjang, model cenderung mengulang label "Rumi:" di setiap baris. Model dengan performa terbaik telah didemonstrasikan melalui antarmuka inferensi berbasis Gradio yang disebarluaskan pada Hugging Face Spaces agar dapat diuji secara langsung oleh pihak lain.

### *3.6. Pembahasan*

Hasil ini menegaskan tercapainya kedua tujuan penelitian yang dirumuskan pada Bagian 1. Pertama, model Qwen2.5-VL-3B-Instruct berhasil diimplementasikan untuk tugas ekstraksi dan transliterasi aksara Jawi cetak secara *end-to-end*, tanpa memerlukan tahap pemisahan karakter terpisah seperti pada pendekatan CNN/SVM konvensional [7][8]. Kedua, *fine-tuning* menggunakan QLoRA terbukti efektif meningkatkan kinerja model secara signifikan meski hanya memperbarui 2,14% dari total parameter, sejalan dengan karakteristik QLoRA yang mampu mengadaptasi model besar dengan sumber daya terbatas tanpa mengorbankan kemampuan umum model secara berarti [23].

Nilai CER akhir ekstraksi Jawi (0,174) belum mencapai target 0,15 yang ditetapkan di awal penelitian, dan juga masih lebih tinggi dibanding sistem VLM untuk Arab standar seperti QARI-OCR (CER 0,061) [14] maupun Qalam (WER 0,80% untuk tulisan tangan) [15]. Perbedaan ini beralasan mengingat kedua sistem tersebut dilatih pada volume data yang jauh lebih besar dan pada bahasa Arab standar yang sumber datanya jauh lebih melimpah, sementara aksara Jawi memiliki huruf tambahan khas Melayu, konvensi penulisan tanpa vokal yang lebih konsisten, dan data berlabel yang jauh lebih terbatas, sejalan dengan kesenjangan yang diidentifikasi pada benchmark skrip asli Nusantara [16]. Meski demikian, penurunan CER relatif sekitar 79% sampai 82% dari kondisi *zero-shot* menegaskan signifikansi praktis dari strategi *fine-tuning* efisien ini bagi aksara Jawi, yang selama ini kurang mendapat perhatian riset dibanding Arab standar meski memiliki basis naskah dan penutur yang luas.

Sejumlah keterbatasan turut memengaruhi hasil. Mayoritas data latih berupa potongan satu baris, sehingga variasi pola kalimat panjang yang dipelajari model masih terbatas dan belum diuji pada tata letak paragraf atau halaman penuh. Kualitas fisik sebagian naskah autentik turut menyumbang

kesalahan, sebagaimana juga ditemukan pada evaluasi keterbacaan aplikasi transliterasi Jawi berbasis AI lain [25]. Data uji autentik yang digunakan juga masih berskala terbatas, sehingga pengujian pada koleksi naskah yang lebih beragam diperlukan sebelum sistem ini diandalkan pada kasus penggunaan yang lebih luas.

## 4. Kesimpulan

Penelitian ini berhasil mengimplementasikan *fine-tuning* Qwen2.5-VL-3B-Instruct menggunakan QLoRA untuk tugas ekstraksi dan transliterasi aksara Jawi cetak secara *end-to-end*. Strategi ini terbukti efektif: CER ekstraksi Jawi turun dari 0,825 (*zero-shot*) menjadi 0,174, dan CER transliterasi Rumi turun dari 0,906 menjadi 0,165, dengan model terbaik diperoleh pada epoch ketiga dari lima epoch pelatihan sebelum gejala *overfitting* muncul. Studi ablasi menunjukkan bahwa kombinasi data sintetik dan data autentik menghasilkan kinerja terbaik dibanding penggunaan salah satu jenis data saja, dengan data autentik memberikan kontribusi yang lebih besar per sampel karena lebih merepresentasikan kondisi naskah nyata. Temuan ini menegaskan bahwa pendekatan Vision-Language Model yang dipadukan *fine-tuning* efisien merupakan strategi yang layak dan bermanfaat untuk mendukung digitalisasi naskah Jawi, sekaligus membuka jalan bagi pelestarian skrip Nusantara lain yang menghadapi tantangan serupa.

Penelitian lanjutan disarankan memperkaya data dengan citra paragraf atau halaman penuh dan teks berharakat lengkap, memperluas rentang eksperimen *hyperparameter*, serta menguji model pada koleksi naskah Jawi yang lebih besar dan beragam untuk memastikan keandalannya pada kasus penggunaan yang lebih luas.

## 5. Referensi

[1] A. S. Bania and B. Akob, “Preserving the Jawi script in Aceh: Assessing literacy, cultural heritage, and modern paradigm challenges,” *Studies in English Language and Education*, vol. 12, no. 1, pp. 457–470, Feb. 2025, doi: 10.24815/siele.v12i1.36629.

[2] P. Coluzzi, “Jawi, an endangered orthography in the Malaysian linguistic landscape,” *Int. J. Multiling.*, vol. 19, no. 4, pp. 630–646, Oct. 2022, doi: 10.1080/14790718.2020.1784178.

[3] Y. Darmi, M. Fajri Sepriansyah, Y. Darnita, and P. Pahrizal, “Penerapan Metode Optical Character Recognition (OCR) Untuk Mengidentifikasi Teks Pada Identitas Dokumen Surat Izin Mengemudi (SIM),” *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 4, pp. 5992–5998, May 2025, doi: 10.36040/jati.v9i4.13987.

[4] X.-F. Wang, Z.-H. He, K. Wang, Y.-F. Wang, L. Zou, and Z.-Z. Wu, “A survey of text detection and recognition algorithms based on deep learning technology,” *Neurocomputing*, vol. 556, p. 126702, Nov. 2023, doi: 10.1016/j.neucom.2023.126702.

[5] Jumrah Jamil and Suharto Pulukadang, “Application of Deep Learning Method in Learning,” *Formosa Journal of Sustainable Research*, vol. 4, no. 6, pp. 1011–1028, Jun. 2025, doi: 10.55927/fjsr.v4i6.308.

[6] S. Faizullah, M. S. Ayub, S. Hussain, and M. A. Khan, “A Survey of OCR in Arabic Language: Applications, Techniques, and Challenges,” *Applied Sciences*, vol. 13, no. 7, p. 4584, Apr. 2023, doi: 10.3390/app13074584.

[7] L. Abdiansah, S. Sumarno, A. Eviyanti, and N. L. Azizah, “Penerapan Algoritma Convolutional Neural Networks untuk Pengenalan Tulisan Tangan Aksara Jawa,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 2, pp. 496–504, Mar. 2025, doi: 10.57152/malcom.v5i2.1814.

[8] S. Safrizal, “Pengenalan Karakter Jawi Tulisan Tangan Menggunakan Fitur Sudut,” *VOCATECH: Vocational Education and Technology Journal*, vol. 1, Oct. 2019, doi: 10.38038/vocatech.v1i0.1.

[9] A. A. Handoko, M. A. Rosid, and U. Indahyanti, “Implementasi Convolutional Neural Network (CNN) Untuk Pengenalan Tulisan Tangan Aksara Bima,” *SMATIKA JURNAL*, vol. 14, no. 01, pp. 96–110, Jul. 2024, doi: 10.32664/smatika.v14i01.1196.

[10] K. AR, H. Ahmadian, A. D. Nasihah, and A. M. Hamdan, "An Applied Computer Mathematics Approach to Transliteration: YOLOv8-Based Detection of Harah Jawoe Script," *ZERO: Jurnal Sains, Matematika dan Terapan*, vol. 9, no. 1, p. 66, Jul. 2025, doi: 10.30829/zero.v9i1.24174.
[11] A. Nagasubramanian *et al.*, "OCRNet a robust deep learning framework for alphanumeric character recognition to assist the visually impaired," *Sci. Rep.*, vol. 15, no. 1, p. 41344, Nov. 2025, doi: 10.1038/s41598-025-25278-9.
[12] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-Language Models for Vision Tasks: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, Aug. 2024, doi: 10.1109/TPAMI.2024.3369699.
[13] S. Yin *et al.*, "A survey on multimodal large language models," *Natl. Sci. Rev.*, vol. 11, no. 12, Nov. 2024, doi: 10.1093/nsr/nwae403.
[14] A. Wasfy *et al.*, "QARI-OCR: High-Fidelity Arabic Text Recognition through Multimodal Large Language Model Adaptation," *ArXiv*, Jun. 2025.
[15] G. Bhatia, E. M. B. Nagoudi, F. Alwajih, and M. Abdul-Mageed, "Qalam: A Multimodal LLM for Arabic Optical Character and Handwriting Recognition," in *Proceedings of The Second Arabic Natural Language Processing Conference*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2024, pp. 210–224. doi: 10.18653/v1/2024.arabicnlp-1.19.
[16] M. F. Adilazuarda, M. I. Wijanarko, L. Susanto, K. Nur'aini, D. T. Wijaya, and A. F. Aji, "NusaAksara: A Multimodal and Multilingual Benchmark for Preserving Indonesian Indigenous Scripts," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 28371–28401. doi: 10.18653/v1/2025.acl-long.1377.
[17] S. Bai *et al.*, "Qwen2.5-VL Technical Report," *ArXiv*, Feb. 2025.
[18] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ArXiv*, Jun. 2021.
[19] N. Ding *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nat. Mach. Intell.*, vol. 5, no. 3, pp. 220–235, Mar. 2023, doi: 10.1038/s42256-023-00626-4.
[20] L. Wang *et al.*, "Parameter-efficient fine-tuning in large language models: a survey of methodologies," *Artif. Intell. Rev.*, vol. 58, no. 8, p. 227, May 2025, doi: 10.1007/s10462-025-11236-4.
[21] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *ArXiv*, Oct. 2021.
[22] Z. Li, H. Li, and L. Meng, "Model Compression for Deep Neural Networks: A Survey," *Computers*, vol. 12, no. 3, p. 60, Mar. 2023, doi: 10.3390/computers12030060.
[23] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," *ArXiv*, May 2023.
[24] C. Neudecker, K. Baierer, M. Gerber, C. Christian, A. Apostolos, and P. Stefan, "A survey of OCR evaluation tools and metrics," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Sep. 2021, pp. 13–18. doi: 10.1145/3476887.3476888.
[25] A. Syahputra Bania, N. Faridy, and B. Akob, "Evaluating The Readability of Jawi To Latin Transliteration Via AI-Based Text Photography Applications," *INJECT (Interdisciplinary Journal of Communication)*, vol. 10, no. 1, pp. 115–132, Jun. 2025, doi: 10.18326/inject.v10i1.4383.