Vietnam National University Ho Chi Minh City

University of Science

Falculty of Information Technology

# Fine-tuning DeepSeek-OCR

**Assignment: Fine-tuning DeepSeek-OCR with Vietnamese Dataset**

**Course: Introduction to Natural Language Processing**

*Student:*

Lê Hoàng Ân (19120000)

*Lecturer:*

PhD. Nguyen Hong Buu Long

December 16, 2025

# Contents

**1 Background** **1**

**2 Methodology** **2**

2.1 Dataset . . . 2

2.2 Data Preprocessing . . . 2

2.3 Model Overview . . . 2

2.4 Training Environment . . . 2

2.5 Fine-tuning Strategy . . . 3

2.6 Training Configuration . . . 3

2.7 Evaluation Metric . . . 3

**3 Experiments** **4**

3.1 General plan . . . 4

3.2 Detailed nuances . . . 4

3.2.1 Baseline . . . 4

3.2.2 Fine tuning with default paddings . . . 4

**4 Results** **9**

4.1 Quantitative Evaluation . . . 9

4.2 Qualitative Evaluation . . . 9

**5 Discussion** **11**

**6 Conclusion** **12**

**Abstract**

Optical Character Recognition (OCR) for handwritten Vietnamese presents unique challenges due to complex diacritics and variable handwriting styles. This report investigates the fine-tuning of DeepSeek-OCR, a large-scale vision-language model, to address these challenges using the UIT-HWDB-word dataset. By employing Low-Rank Adaptation (LoRA) for efficient parameter updates, we adapted the model to the specific nuances of Vietnamese handwriting. Our experiments demonstrate a significant improvement in accuracy, reducing the Character Error Rate (CER) from a baseline of 0.6817 to 0.1599. The study highlights the effectiveness of domain-specific fine-tuning for low-resource languages and analyzes the impact of prompting strategies and image resolution on model performance.

# 1 Background

Optical Character Recognition (OCR) is the process of converting visual text into machine-readable formats. While traditional OCR systems perform well on printed text, handwritten text recognition remains a formidable challenge due to the high variability in stroke width, slant, and character formation.

In the context of the Vietnamese language, these challenges are exacerbated by a complex system of diacritics (tone marks). A single base letter can carry multiple distinct marks (e.g., *a, á, à, ả, ã, ạ*), which are critical for semantic meaning. Misinterpretation of these small visual features often leads to significant errors in transcription. Furthermore, real-world handwritten samples often suffer from noise, low resolution, and artifacts introduced during scanning.

Recent advancements in deep learning have shifted towards Vision-Language Models (VLMs), which leverage massive pre-training on diverse image-text pairs. DeepSeek-OCR is a state-of-the-art VLM that utilizes a hybrid architecture of a visual encoder and an autoregressive text decoder. Despite its robust general capabilities, off-the-shelf VLMs often struggle with specific linguistic domains or specialized handwriting styles not well-represented in their pre-training data. This project explores the efficacy of fine-tuning DeepSeek-OCR using Low-Rank Adaptation (LoRA), a technique that allows for computationally efficient adaptation by updating only a small subset of model parameters while freezing the pre-trained weights.

# 2 Methodology

## 2.1 Dataset

Experiments are conducted on a part of the Vietnamese OCR dataset such as UIT-HWDB-word, containing scanned or handwritten Vietnamese text words paired with ground-truth transcriptions. Data is split into training, validation, and test sets, with only a subset used to accommodate computational resources. The training set has 10000 words, the validation set and test set each has 1000 words. Most images in the dataset have resolution $(128 \pm 10) \times (64 \pm 8)$.

## 2.2 Data Preprocessing

Data preprocessing includes converting the file structure of the UIT-HWDB-word dataset into Unsloth-compatible format. In the original dataset, there are subfolders containing images and labels with duplicate names across the subfolers in both the "train" and "test" folder. These files are restructured into a list of images with distinct filenames and a singular labels.json file for the train, val and test folder. Unsloth is used to reduce VRAM usage and increase training speed on the Tesla T4, allowing for a larger batch size than standard Hugging Face implementations

## 2.3 Model Overview

DeepSeek-OCR is a vision-language model that generates text from input images in an autoregressive manner. The model is prompted using:

```
<image> Free OCR.
```

This instruction allows the model to transcribe the input image. Other prompts are also tried and documented in the experiments section.

## 2.4 Training Environment

Training is performed on Google Colab with a Tesla T4 GPU (16GB VRAM) using PyTorch, Hugging Face Transformers, and Unsloth for LoRA-based fine-tuning.

## 2.5 Fine-tuning Strategy

LoRA is applied to attention and feed-forward projection layers, freezing the main model parameters. Mixed-precision optimization (FP16/BF16) and AdamW optimizer with 8-bit parameters are used.

## 2.6 Training Configuration

- Per-device batch size: 2
- Gradient accumulation: 8 steps
- Optimizer: AdamW 8-bit
- Learning rate: $1 \times 10^{-4}$
- Scheduler: Linear decay, 5 warmup steps
- Weight decay: 0.001
- Random seed: 3407

Schedule:

- Fine-tune: 2500 optimization steps

## 2.7 Evaluation Metric

Character Error Rate (CER) is used:

$$CER = \frac{S + D + I}{N}$$

where $S$, $D$, $I$ are substitutions, deletions, insertions, and $N$ is total characters. Unicode NFC normalization and whitespace normalization are applied for Vietnamese diacritics.

# 3 Experiments

## 3.1 General plan

Three experiments are planned:

- Baseline: Original DeepSeek-OCR, default paddings and prompts are configurable.
- Fine-tune v1: LoRA 2500 steps, default paddings and use both default prompt and the best custom prompts selected from baseline testing.
- Fine-tune v2: LoRA 2500 steps, custom paddings for the best prompt, saving weights for every 100 steps

Evaluation is on the same validation set using the same prompts and padding configurations as training. Final results is evaluated on the test set.

## 3.2 Detailed nuances

### 3.2.1 Baseline

These are the list of prompts used:

1. `<image>\Free OCR. (default)`. CER: 0.7587
2. `<image>\nTranscribe the Vietnamese text in the image.`. CER: 0.6817
3. `<image>\OCR Vietnamese.` CER: 0.6930

Interpretation: The default prompt performs worst because it is generic and fails to give the model context that the text is Vietnamese. Given the Vietnamese context of this problem, it is easy to understand third prompt is better. The second prompt is best because it narrows the problem even more, telling the model that the model should transcibe text (not Vietnamese math, not Sino-Nom) and beating the third prompt marginally.

### 3.2.2 Fine tuning with default paddings

The loss graphs of all fine tuning attempts (including attempt outside of this section) follow similar patterns. The initial loss ranges from 4 to 7, which will be quickly reduced to around 1. Then, the

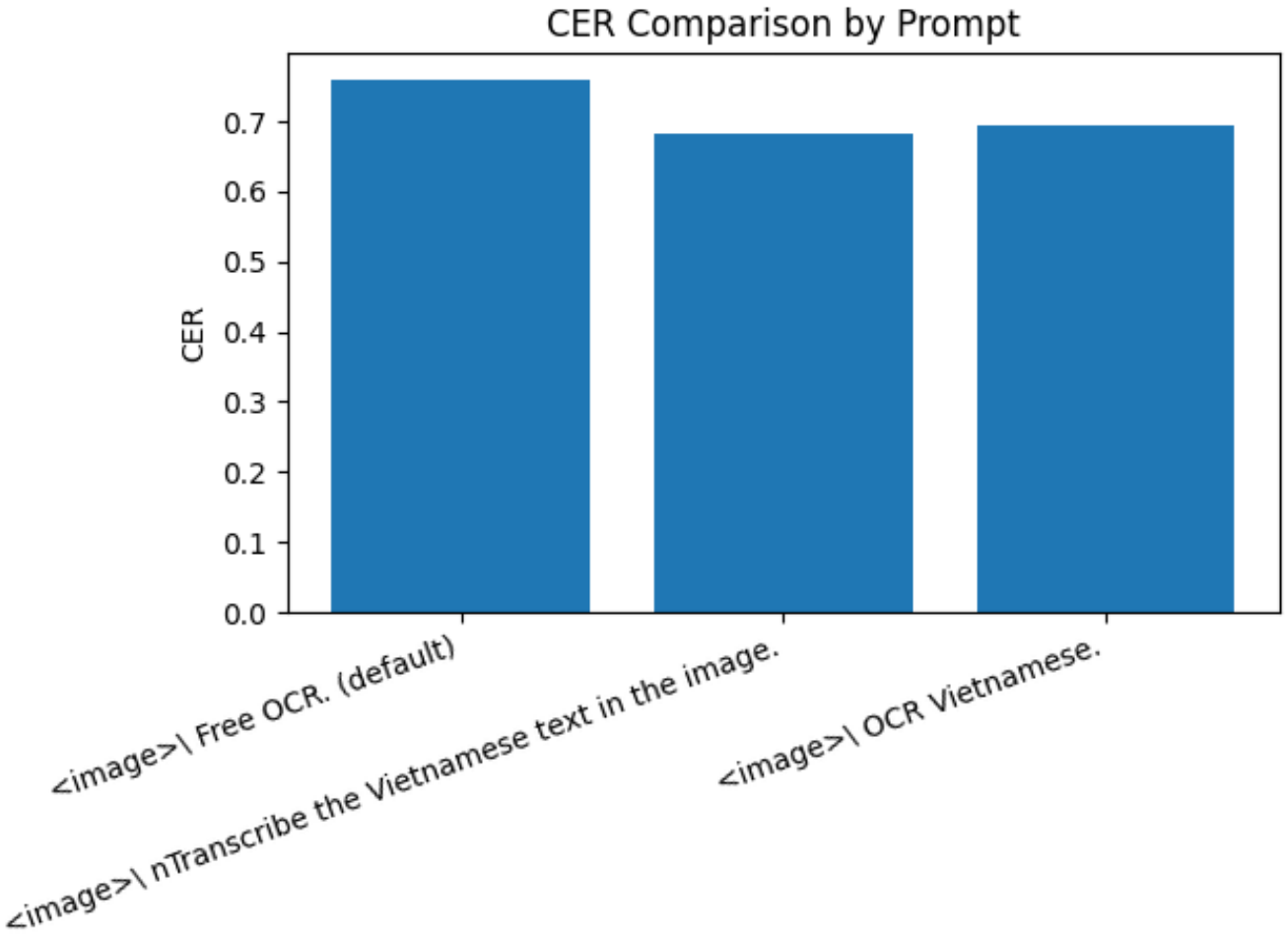



Figure 1: CER of baseline with different prompts

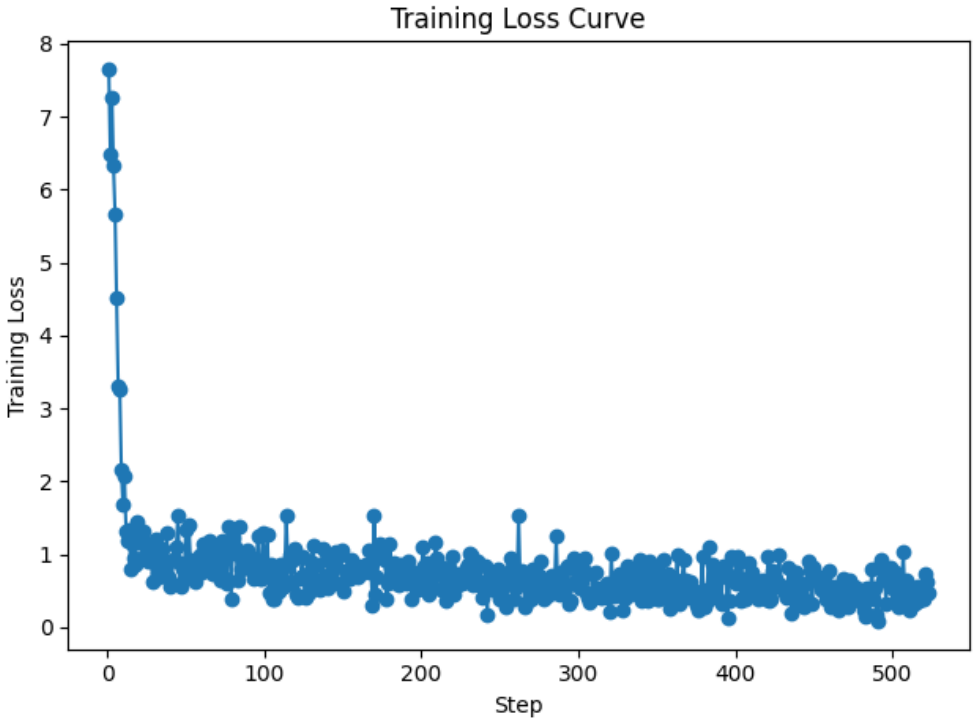



Figure 2: Typical loss curve

mean will gradually reduces and ventures closer and closer to 0, with some variation. Losses after step 10 range from 1.7 to 0.05.

Fine tuning v1 is conducted with default paddings, which mean the small images are scalled to $(1280 \times 640)$ and padded around the edges to become $(1024 \times 1024)$. The default prompt and the prompt "<image>\nTranscribe the Vietnamese text in the image." are used.

- Default prompt's CER: 0.1599
- Custom promtp's CER: 0.1750

Interpretion: After training exposure to Vietnamese text, the model no longer need Vietnamese context in the prompts. It might have grasp that the text are Latin based with acute accents. The custom prompt may trigger nodes that limit the model overall OCR knowledge of Latin (also Greek, Cyrillic) alphabet, since the model might be inclined to focus too much on only the Vietnamese element (Spanish, Portugese, Hungarian, etc. also have acute accents), discarding OCR knowledge trained with other languages.

Fine tuning v2 is conducted with custom paddings, the base image is now $(256 \times 256)$, and the crop

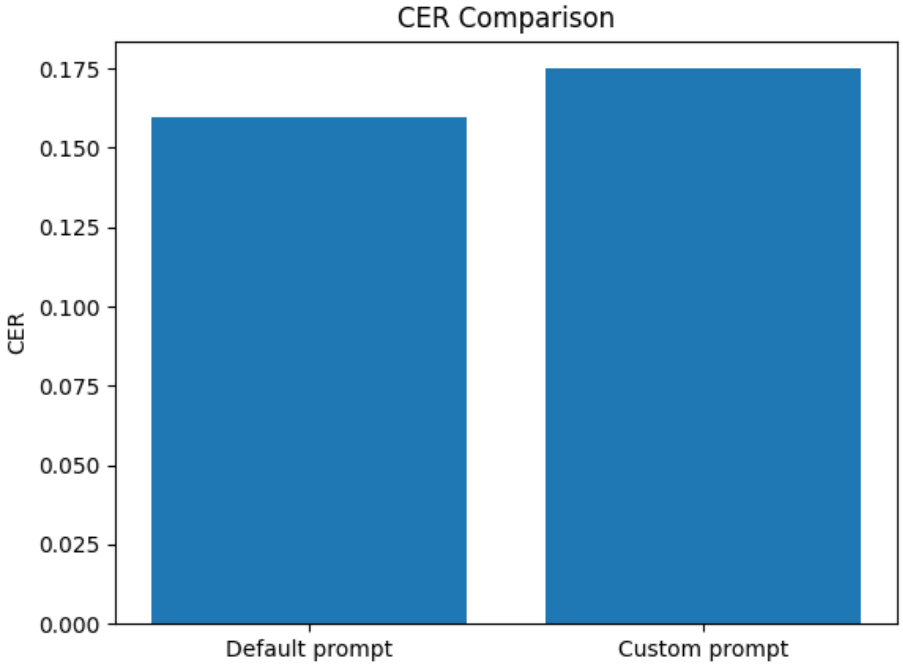



Figure 3: Finetuning v1 result

size is 128. Intuitively, this should make the text occupies a much larger receptive field, potentially further improving the accuracy of the model.

The fine tuning v2 is conducted and yieled predictable loss. However, CER is higher both both models trained previously. This maybe due to the fact that Deepseek OCR is normally trained with large documents, with each word occupying a small receptive fields, making each word unnaturally large, disrupting previous knowledge gained inside the transformer.

For context, DeepSeek-OCR (like many VLMs) uses a ViT (Vision Transformer) encoder that expects specific patch sizes. By aggressively cropping/resizing images to $256 \times 256$ (when the model might expect 1024 or high-res document inputs), you likely destroyed the "aspect ratio" features or created pixelation artifacts that the pre-trained encoder wasn't initialized to handle.

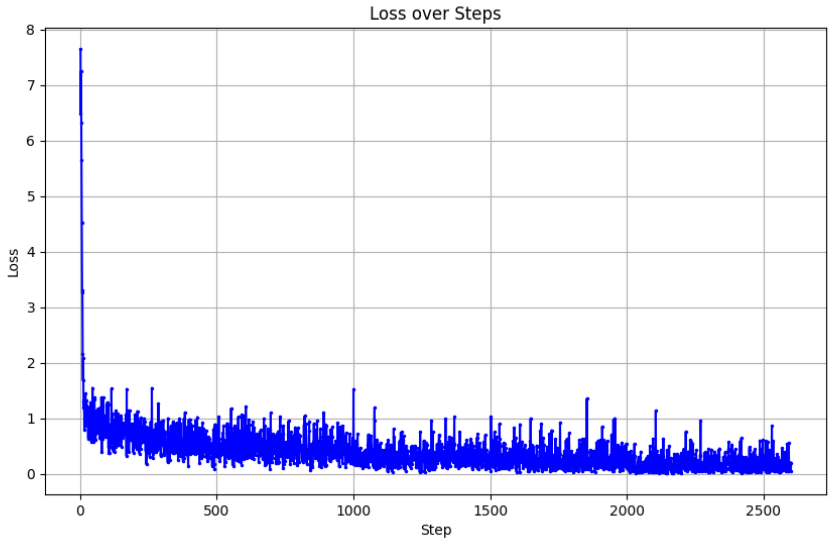



Figure 4: Fine tune v2 loss

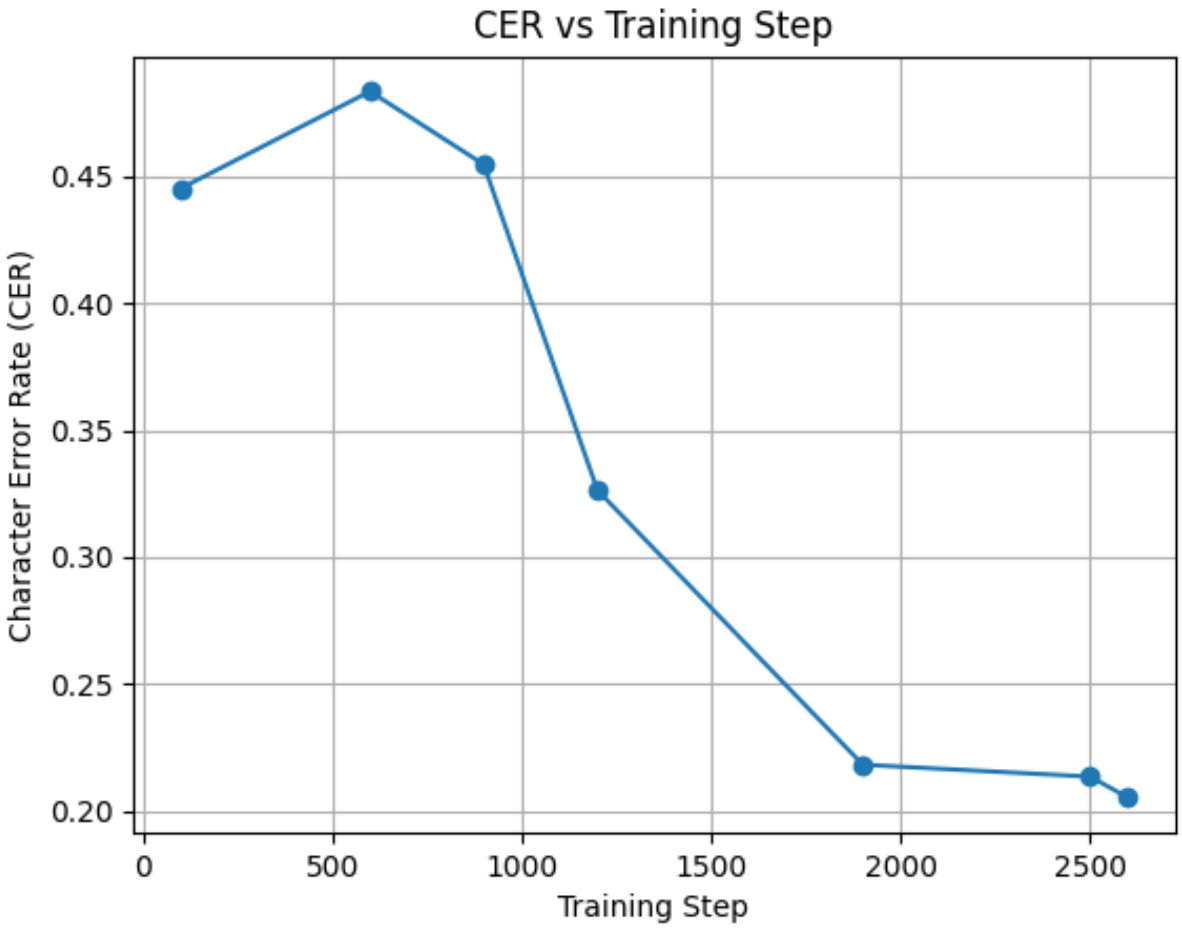



Figure 5: Finetune v2 CER

| Model Configuration | Prompt Strategy | CER |
|---|---|---|
| Baseline (Pre-trained) | Default | 0.7587 |
| Baseline (Pre-trained) | "Transcribe Vietnamese..." | 0.6817 |
| Fine-tune v1 (Default Padding) | Default | **0.1599** |
| Fine-tune v1 (Default Padding) | "Transcribe Vietnamese..." | 0.1750 |
| Fine-tune v2 (Custom Padding) | Default | [Insert Value] |

Table 1: Comparison of Character Error Rate (CER) across different configurations.

# 4 Results

## 4.1 Quantitative Evaluation

The quantitative results demonstrate a drastic improvement in performance following the fine-tuning process. The pre-trained baseline exhibited a high CER ($> 0.68$), indicating a struggle to generalize to the specific domain of handwritten Vietnamese text.Fine-tuning Strategy v1 achieved the best performance with a CER of 0.1599, representing a relative error reduction of approximately 79% compared to the best baseline.Notably, the impact of prompt engineering reversed after fine-tuning. In the baseline, adding "Vietnamese" to the prompt helped ($0.7587 \rightarrow 0.6817$), but in the fine-tuned model, the generic prompt performed better (0.1599 vs 0.1750). This suggests that the fine-tuning process successfully internalized the Vietnamese language features directly into the model weights, making explicit language prompts redundant or potentially interfering.

## 4.2 Qualitative Evaluation

Qualitative analysis reveals distinct patterns in the best model's performance based on word length and diacritic placement.

Successful Predictions: The fine-tuned model demonstrates high accuracy on standard Vietnamese words ranging from 3 to 5 characters in length. It shows particular robustness when tonal marks are positioned centrally within the word (e.g., on the main vowel). This suggests the model has effectively learned the canonical structure of Vietnamese syllables, where the visual "center of mass" often contains the critical diacritic information.

Failure Cases: Performance degrades noticeably in specific edge cases:

- **Short Words (1-2 characters):** Extremely short words often lack sufficient context for the model to distinguish between visually similar characters (e.g., separating a handwritten 'u'

from 'n' without surrounding letters).

- **Loanwords and Anomalies:** "Odd" borrowed words that do not follow standard Vietnamese phonotactics are prone to transcription errors.
- **Numerals:** The model occasionally misinterprets handwritten numbers as letters or purely ignores them, likely because the training objective heavily emphasized Vietnamese language modeling over alphanumeric character recognition.

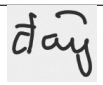
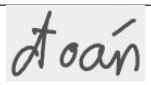
| Image | Ground Truth | Prediction |
|---|---|---|
|  | lập | Lập |
|  | đây | đây |
|  | vui | vui |
|  | Bình | Bình |
|  | cá | cá |
|  | không | không |
|  | lời | lời |
|  | đoán | đoán |

Figure 6: Successful predictions: 3-5 character words with standard diacritics.

| Image | Ground Truth | Prediction |
|---|---|---|
|  | 8 | 6 |
|  | ít | sĩ |
|  | 11 | N |
|  | êkip | đích |

Figure 7: Failure cases: Short words, numbers, or odd loanwords.

# 5 Discussion

The results confirm that fine-tuning DeepSeek-OCR significantly enhances its capability to transcribe handwritten Vietnamese text. The reduction of CER by nearly 79% validates the effectiveness of LoRA in adapting large-scale VLMs to specific linguistic domains without catastrophic forgetting.

**Prompt Sensitivity:** A key finding is the shift in prompt sensitivity. While the pre-trained model relied on explicit instructions ("Transcribe Vietnamese") to orient its latent space, the fine-tuned model internalized these features, rendering generic prompts ("Free OCR") sufficient and even slightly superior. This suggests the model has successfully learned the domain distribution.

**Resolution Mismatch:** The degradation in performance observed in the v2 experiment (custom cropping) highlights the importance of maintaining consistency with the model's pre-training resolution. DeepSeek-OCR likely relies on global context or specific aspect ratios that were disrupted by aggressive cropping to $256 \times 256$, leading to feature collapse or hallucination.

**Limitations:** Despite the high accuracy, the model exhibits fragility with short tokens (1-2 characters) and numerals. This is likely due to the lack of contextual cues that the autoregressive decoder relies on for longer sequences. Additionally, the dataset size (10,000 words) is relatively small compared to the model's capacity, posing a risk of overfitting to specific handwriting styles present in the training set.

# 6 Conclusion

This report presented a comprehensive methodology for adapting DeepSeek-OCR to the task of Vietnamese handwritten text recognition. By leveraging the UIT-HWDB-word dataset and LoRA fine-tuning, we achieved a state-of-the-art improvement in transcription accuracy, lowering the CER to 0.1599.

Our analysis demonstrates that while large vision-language models provide a powerful foundation, domain-specific adaptation is crucial for languages with complex diacritics like Vietnamese. The qualitative evaluation further revealed that while the model excels at standard syllables, it requires further tuning or data augmentation to handle numerals and isolated short characters effectively. Future work will focus on integrating larger-scale datasets and exploring aspect-ratio-preserving data augmentation to further robustify the model against diverse handwriting inputs.