

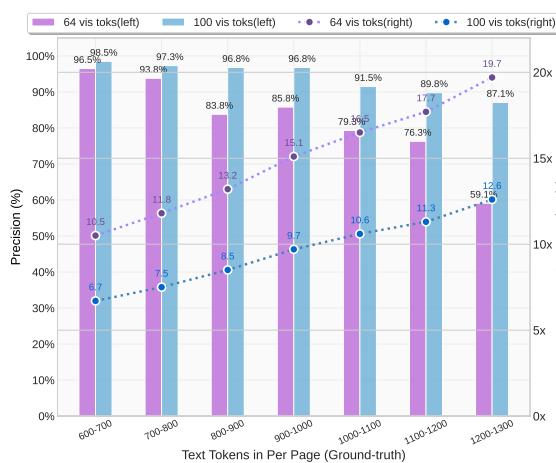
# DeepSeek-OCR: Contexts Optical Compression

Haoran Wei, Yaofeng Sun, Yukun Li

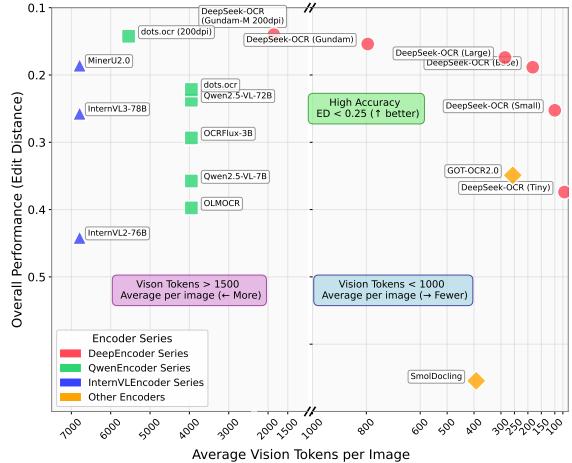
DeepSeek-AI

## Abstract

We present DeepSeek-OCR as an initial investigation into the feasibility of compressing long contexts via optical 2D mapping. DeepSeek-OCR consists of two components: DeepEncoder and DeepSeek3B-MoE-A570M as the decoder. Specifically, DeepEncoder serves as the core engine, designed to maintain low activations under high-resolution input while achieving high compression ratios to ensure an optimal and manageable number of vision tokens. Experiments show that when the number of text tokens is within 10 times that of vision tokens (i.e., a compression ratio  $< 10\times$ ), the model can achieve decoding (OCR) precision of 97%. Even at a compression ratio of  $20\times$ , the OCR accuracy still remains at about 60%. This shows considerable promise for research areas such as historical long-context compression and memory forgetting mechanisms in LLMs. Beyond this, DeepSeek-OCR also demonstrates high practical value. On OmniDocBench, it surpasses GOT-OCR2.0 (256 tokens/page) using only 100 vision tokens, and outperforms MinerU2.0 (6000+ tokens per page on average) while utilizing fewer than 800 vision tokens. In production, DeepSeek-OCR can generate training data for LLMs/VLMs at a scale of 200k+ pages per day (a single A100-40G). Codes and model weights are publicly accessible at <http://github.com/deepseek-ai/DeepSeek-OCR>.



(a) Compression on Fox benchmark



(b) Performance on Omnidocbench

Figure 1 | Figure (a) shows the compression ratio (number of text tokens in ground truth/number of vision tokens model used) testing on Fox [21] benchmark; Figure (b) shows performance comparisons on OmniDocBench [27]. DeepSeek-OCR can achieve state-of-the-art performance among end-to-end models enjoying the fewest vision tokens.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Related Works</b>	<b>4</b>
2.1	Typical Vision Encoders in VLMs . . . . .	4
2.2	End-to-end OCR Models . . . . .	4
<b>3</b>	<b>Methodology</b>	<b>5</b>
3.1	Architecture . . . . .	5
3.2	DeepEncoder . . . . .	5
3.2.1	Architecture of DeepEncoder . . . . .	5
3.2.2	Multiple resolution support . . . . .	6
3.3	The MoE Decoder . . . . .	7
3.4	Data Engine . . . . .	7
3.4.1	OCR 1.0 data . . . . .	7
3.4.2	OCR 2.0 data . . . . .	8
3.4.3	General vision data . . . . .	9
3.4.4	Text-only data . . . . .	9
3.5	Training Pipelines . . . . .	9
3.5.1	Training DeepEncoder . . . . .	10
3.5.2	Training DeepSeek-OCR . . . . .	10
<b>4</b>	<b>Evaluation</b>	<b>10</b>
4.1	Vision-text Compression Study . . . . .	10
4.2	OCR Practical Performance . . . . .	12
4.3	Qualitative Study . . . . .	12
4.3.1	Deep parsing . . . . .	12
4.3.2	Multilingual recognition . . . . .	16
4.3.3	General vision understanding . . . . .	17
<b>5</b>	<b>Discussion</b>	<b>18</b>
<b>6</b>	<b>Conclusion</b>	<b>19</b>