

# 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>