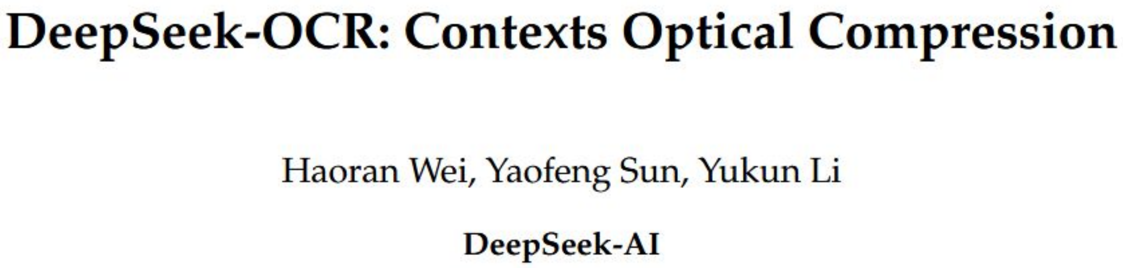
ArXiv 2025



DeepSeek-OCR: Contexts Optical Compression

Haoran Wei, Yaofeng Sun, Yukun Li

DeepSeek-AI

- Problem / objective
  - LLM의 **Long Context** 처리 한계 (연산 비용, 메모리 한계)
- Contribution / Key idea
  - **Optical Compression**: 텍스트를 이미지(비전 토큰)로 압축해 훨씬 적은 토큰 수로 문서 정보를 표현

전유진

- **Preliminaries (Vision Encoder)**

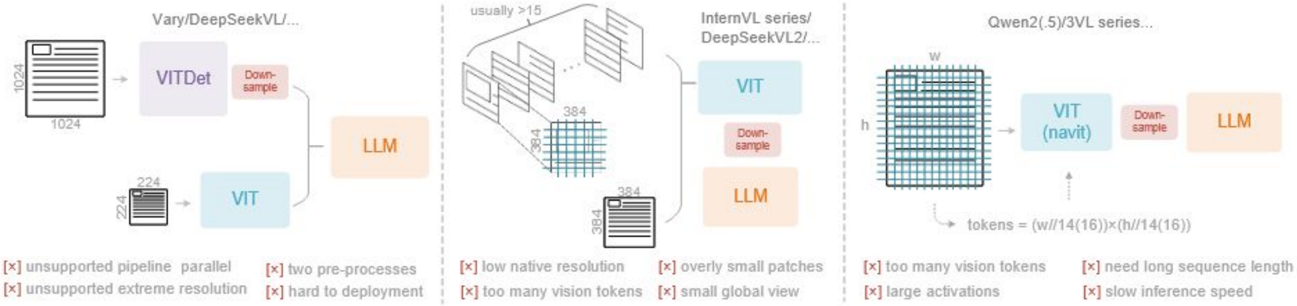


Figure 2 | Typical vision encoders in popular VLMs. Here are three types of encoders commonly used in current open-source VLMs, all of which suffer from their respective deficiencies.

- **Architecture**

I.E., Deep Encoder + MoE Decoder

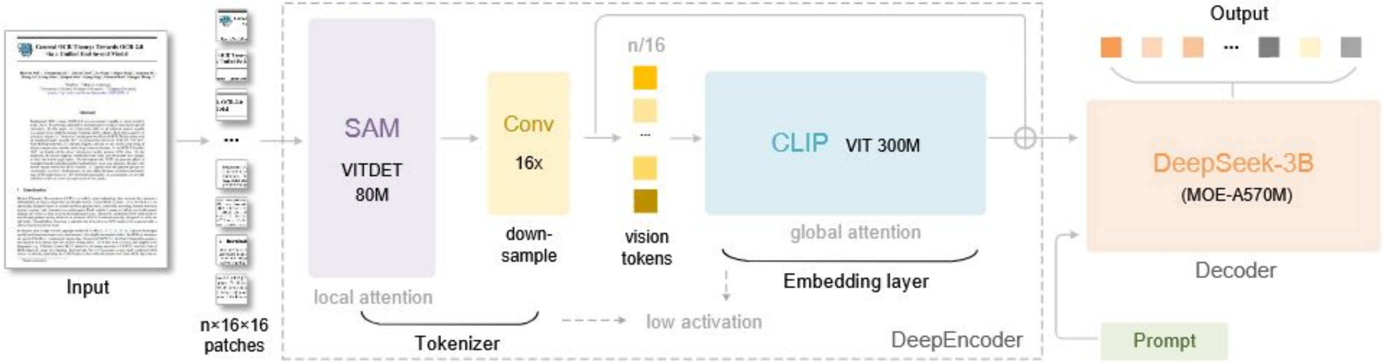


Figure 3 | The architecture of DeepSeek-OCR. DeepSeek-OCR consists of a DeepEncoder and a DeepSeek-3B-MoE decoder. DeepEncoder is the core of DeepSeek-OCR, comprising three components: a SAM [17] for perception dominated by window attention, a CLIP [29] for knowledge with dense global attention, and a 16× token compressor that bridges between them.

- **Deep Encoder**

1. Visual perception feature extraction (window attention) + Visual knowledge feature extraction (global attention)

2. SAM-B -> Compression module -> CLIP-L

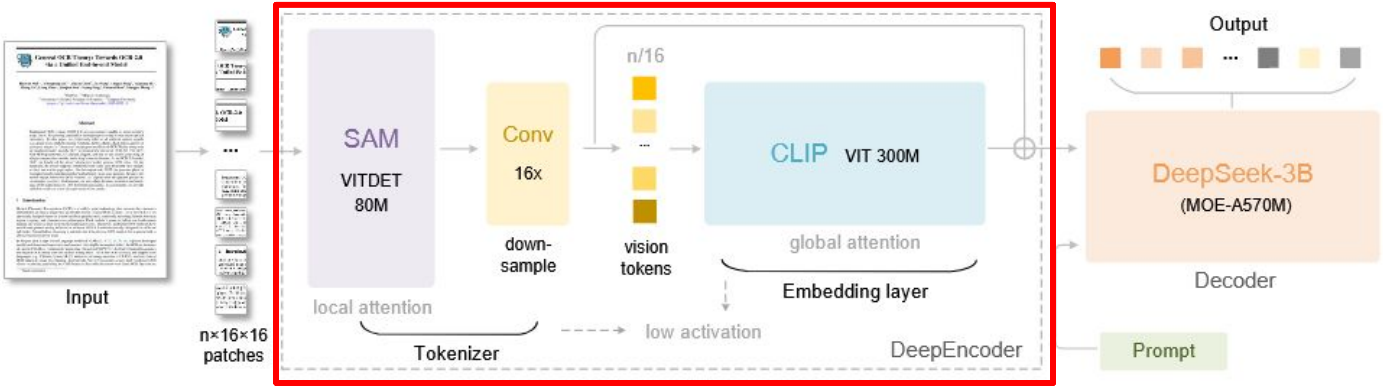
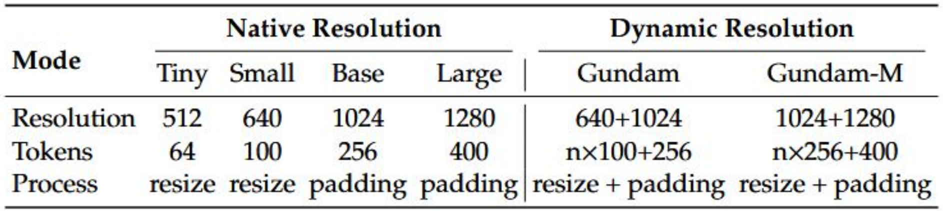


Figure 3 | The architecture of DeepSeek-OCR. DeepSeek-OCR consists of a DeepEncoder and a DeepSeek-3B-MoE decoder. DeepEncoder is the core of DeepSeek-OCR, comprising three components: a SAM [17] for perception dominated by window attention, a CLIP [29] for knowledge with dense global attention, and a 16× token compressor that bridges between them.

- **Deep Encoder**

3. Native resolution + Dynamic resolution

Table 1 | Multi resolution support of DeepEncoder. For both research and application purposes, we design DeepEncoder with diverse native resolution and dynamic resolution modes.



Mode	Native Resolution				Dynamic Resolution	
	Tiny	Small	Base	Large	Gundam	Gundam-M
Resolution	512	640	1024	1280	640+1024	1024+1280
Tokens	64	100	256	400	n×100+256	n×256+400
Process	resize	resize	padding	padding	resize + padding	resize + padding

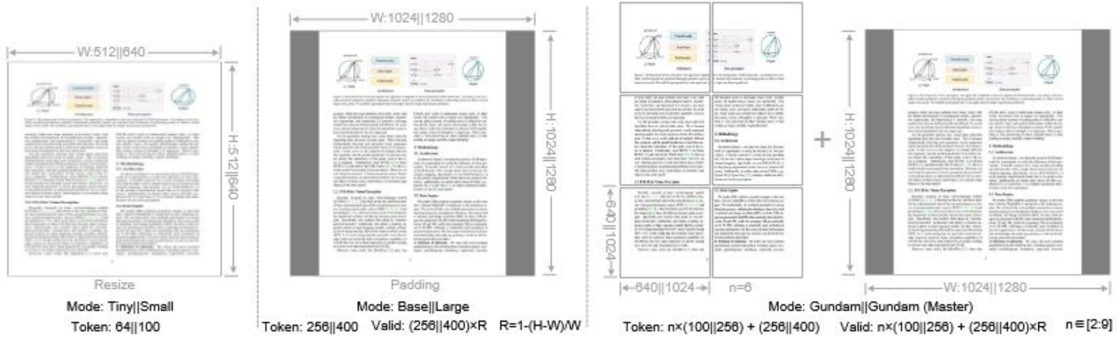


Figure 4 | To test model performance under different compression ratios (requiring different numbers of vision tokens) and enhance the practicality of DeepSeek-OCR, we configure it with multiple resolution modes.

- **MoE Decoder**

Reconstruct the original text representation from the compressed latent vision tokens of DeepEncoder

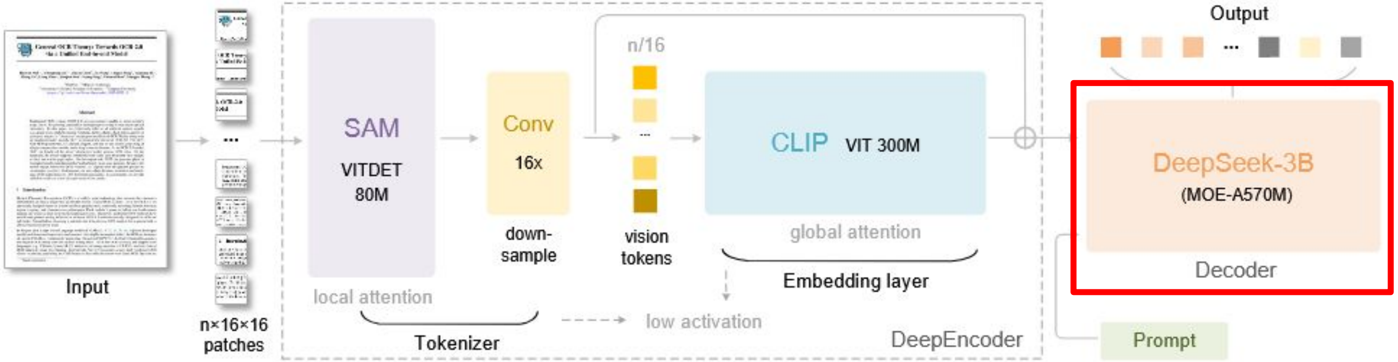


Figure 3 | The architecture of DeepSeek-OCR. DeepSeek-OCR consists of a DeepEncoder and a DeepSeek-3B-MoE decoder. DeepEncoder is the core of DeepSeek-OCR, comprising three components: a SAM [17] for perception dominated by window attention, a CLIP [29] for knowledge with dense global attention, and a 16× token compressor that bridges between them.