DeepSeek-OCR

Heng-Shiou Sheu
2025 Oct.

Why DeepSeek-OCR Went Viral

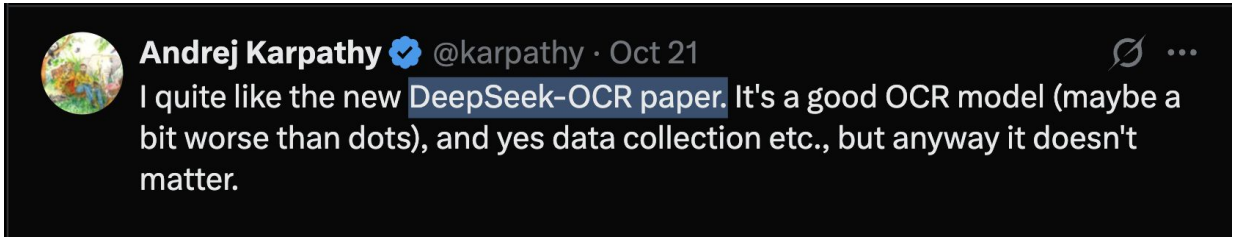
Frequently discussed on social media for its concept of Optical Compression.

deepseek

DeepSeek-OCR: Contexts Optical Compression

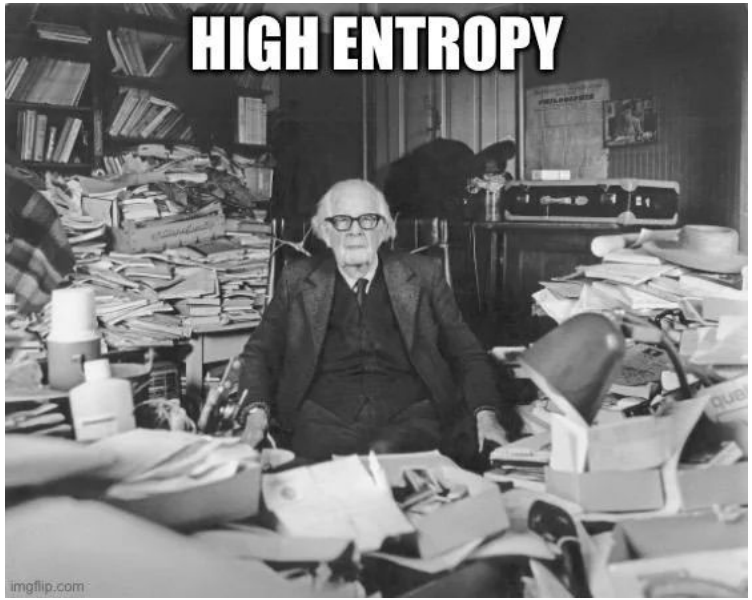
Haoran Wei, Yaofeng Sun, Yukun Li

DeepSeek-AI



Information Entropy

How humans compress infinite meanings into finite symbols



船

boat

舟 八 口

vessel *eight* *people*

The Discovery of Genesis, C.H. Kang and Ethel Nelson, p. 55

Language is compression.

Tokenizer

It’s how symbols begin to imitate meaning

Table 1: Tokenizer comparisons between original LLaMA and Chinese LLaMA.

	Length	Content
Original Sentence	28	人工智能是计算机科学、心理学、哲学等学科融合的交叉学科。
Original Tokenizer	35	‘_’, ‘人’, ‘工’, ‘智’, ‘能’, ‘是’, ‘计’, ‘算’, ‘机’, ‘科’, ‘学’, ‘、’, ‘心’, ‘理’, ‘学’, ‘、’, ‘0xE5’, ‘0x93’, ‘0xB2’, ‘学’, ‘等’, ‘学’, ‘科’, ‘0xE8’, ‘0x9E’, ‘0x8D’, ‘合’, ‘的’, ‘交’, ‘0xE5’, ‘0x8F’, ‘0x89’, ‘学’, ‘科’, ‘。’
Chinese Tokenizer	16	‘_’, ‘人工智能’, ‘是’, ‘计算机’, ‘科学’, ‘、’, ‘心理学’, ‘、’, ‘哲学’, ‘等’, ‘学科’, ‘融合’, ‘的’, ‘交叉’, ‘学科’, ‘。’

A balance between structure and understanding.

Compression

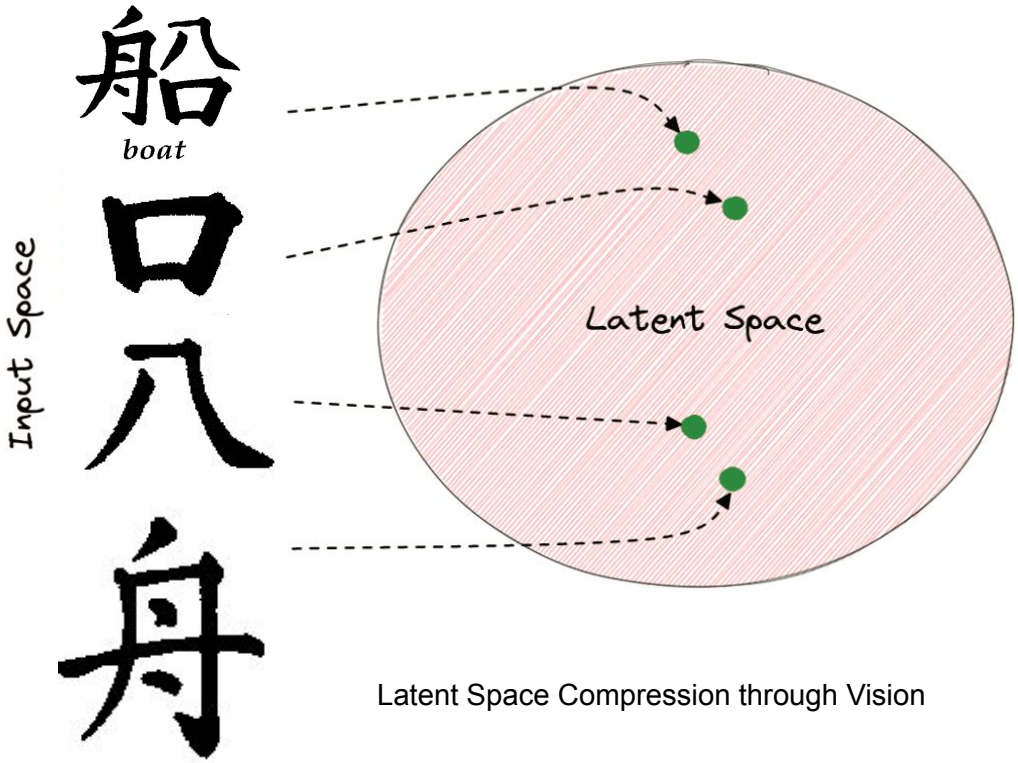
What happens when we move beyond text?

Text Tokens	Vision Tokens =64		Vision Tokens=100		
	Precision	Compression	Precision	Compression	Pages
600-700	96.5%	10.5×	98.5%	6.7×	7
700-800	93.8%	11.8×	97.3%	7.5×	28
800-900	83.8%	13.2×	96.8%	8.5×	28
900-1000	85.9%	15.1×	96.8%	9.7×	14
1000-1100	79.3%	16.5×	91.5%	10.6×	11
1100-1200	76.4%	17.7×	89.8%	11.3×	8
1200-1300	59.1%	19.7×	87.1%	12.6×	4

Exploring the mapping between visual tokens and text tokens.

Beyond the Limits of Language Compression

What if we could encode meaning—not in words—but in the latent space of vision?



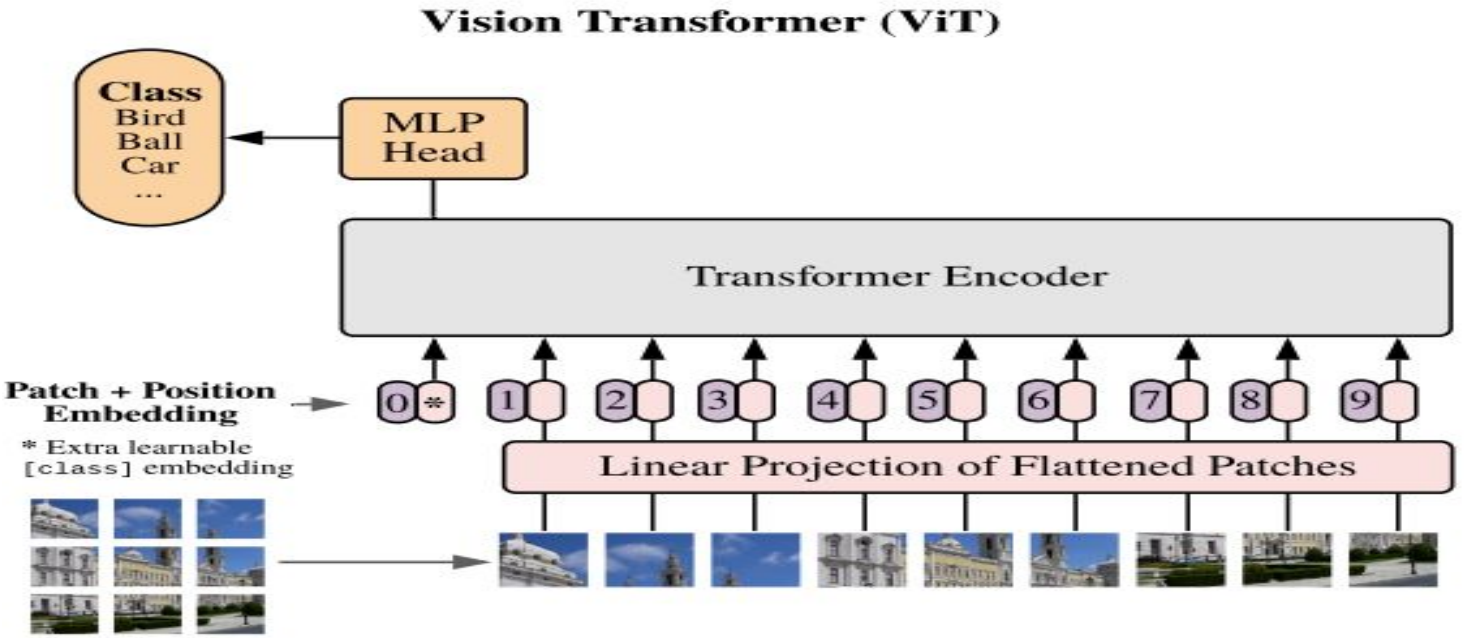
Latent Space Compression through Vision

How can a model “see” an image?

Vision Tokens — the bridge connecting visual input and language understanding

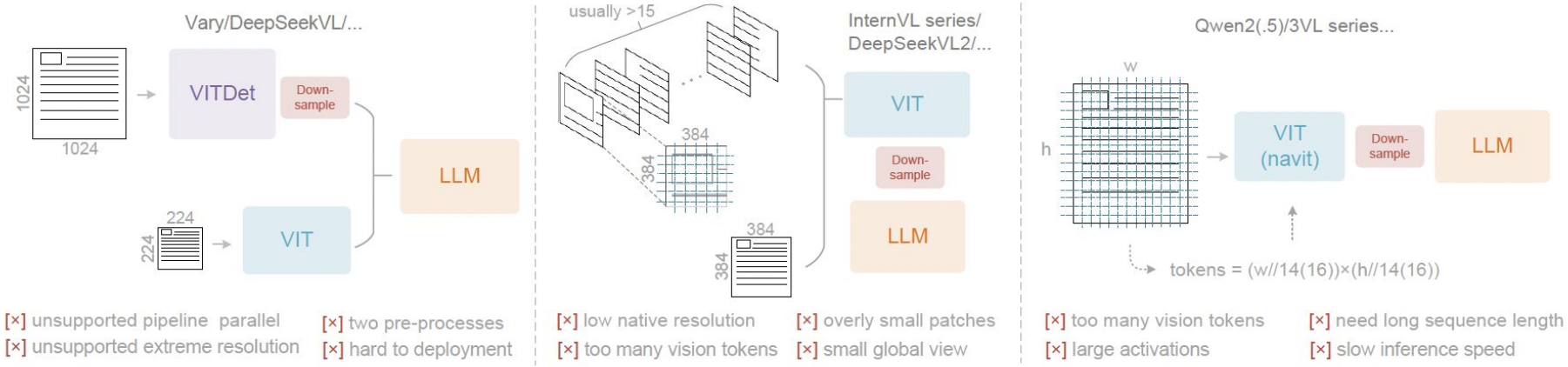
How can a model “see” an image?

Vision Tokens — the bridge connecting visual input and language understanding



The Trade-off of Vision Tokens

Sequence Explosion and the Compression Bottleneck



Fewer tokens. Same meaning.

A two-stage compression architecture

A two-stage compression framework that enables visual informationto retain its full semantic meaning with the fewest possible tokens.

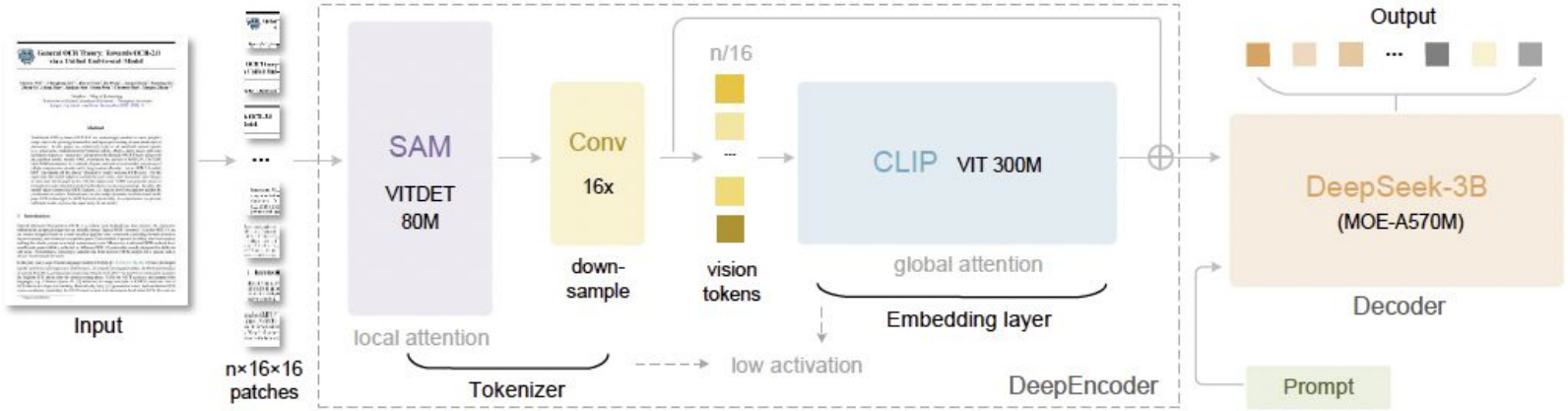
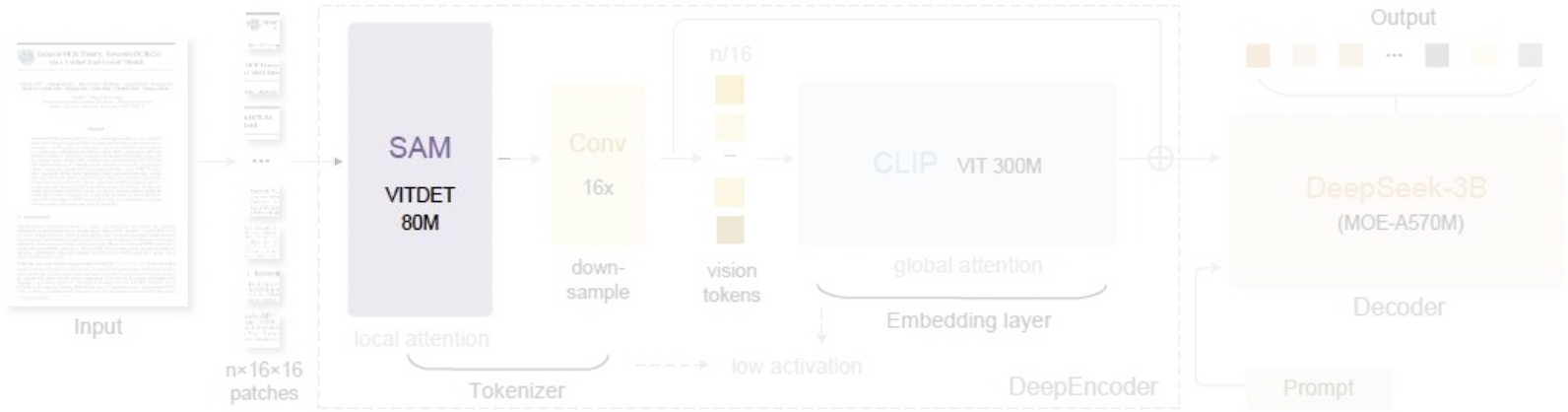


Figure 3 | The architecture of DeepSeek-OCR. DeepSeek-OCR consists of a DeepEncoder and a DeepSeek-3B-MoE decoder. DeepEncoder is the core of DeepSeek-OCR, comprising three

keeping meaning intact, even with minimal visual tokens.

DeepEncoder

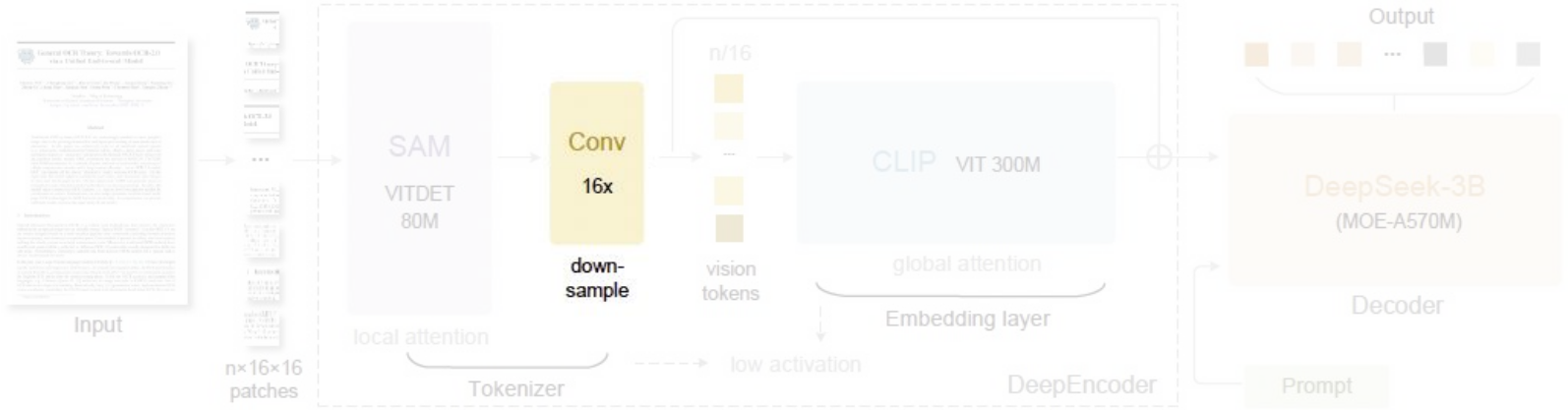
Local Information: Detail Extraction with SAM-base



- Windowed Attention → Focus on patch-level details
- 1024×1024 image → 16×16 patches
- 4096 Vision Tokens generated

DeepEncoder

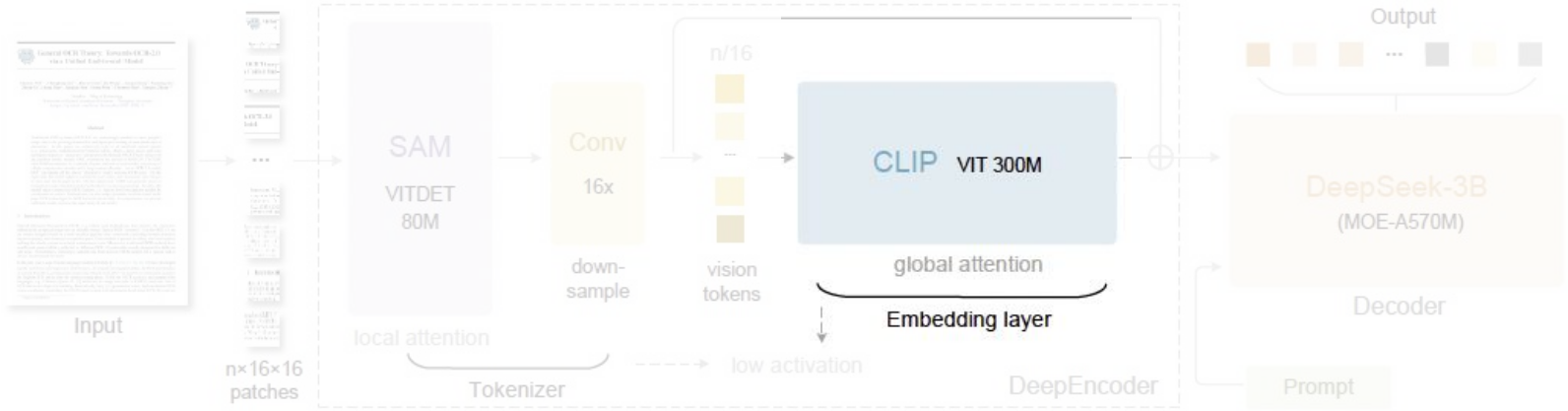
16× Convolutional Compressor: Efficient Dimensionality Reduction



- Using 3×3 kernels (stride 2, padding 1),
- it expands the feature channels from 256 to 1024,
- compressing 4096 patch tokens into just 256.

DeepEncoder

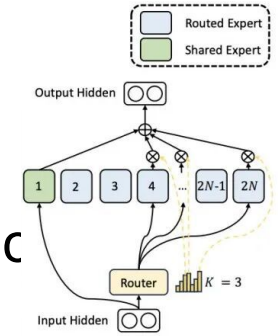
Global Semantic Integration (CLIP-Large)

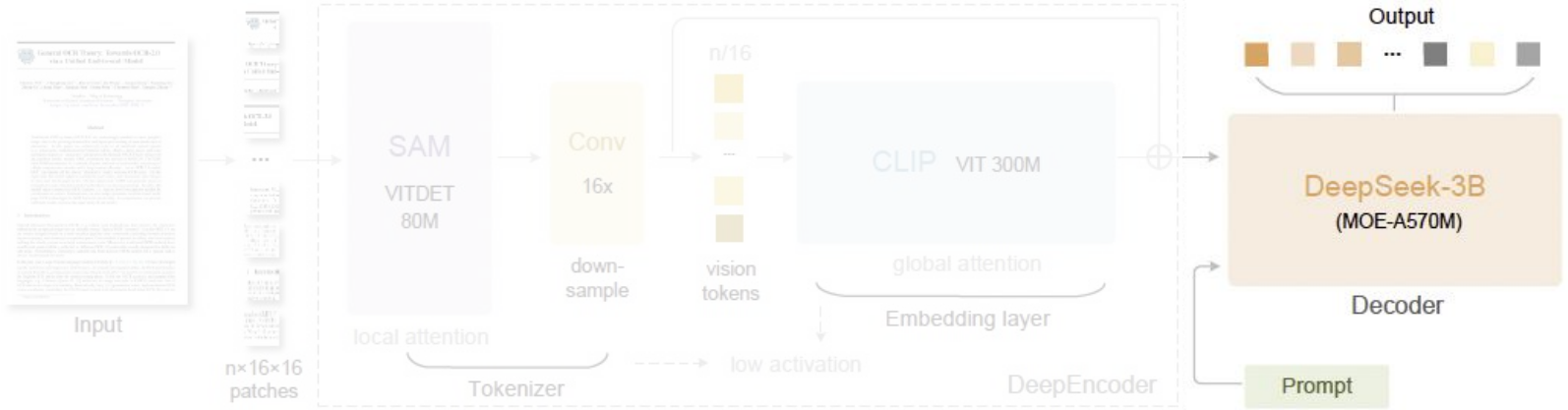


- Extracts global semantics & context
- Removes patch embedding layer
- Dense global attention for holistic meaning

DeepDecoder

Achieving large-model expressiveness with small-model effic

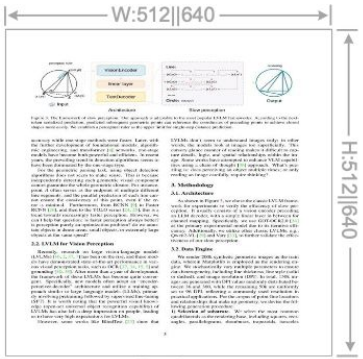




- Built on the **DeepSeekMoE-3B-MoE** architecture
- Activates **6 routing experts** and **2 shared experts**,
  for a total of approximately **570 million active parameters**
- $f_{\text{dec}} : \mathbb{R}^{n \times d_{\text{latent}}} \rightarrow \mathbb{R}^{N \times d_{\text{text}}}; \quad \hat{\mathbf{X}} = f_{\text{dec}}(\mathbf{Z}) \quad \text{where } n \leq N$

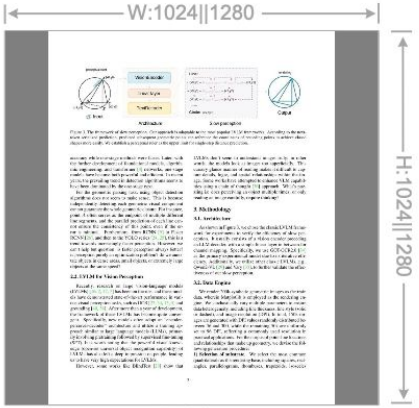
Multi-Resolution Capability

The DeepSeek-OCR model supports multiple input resolutions.



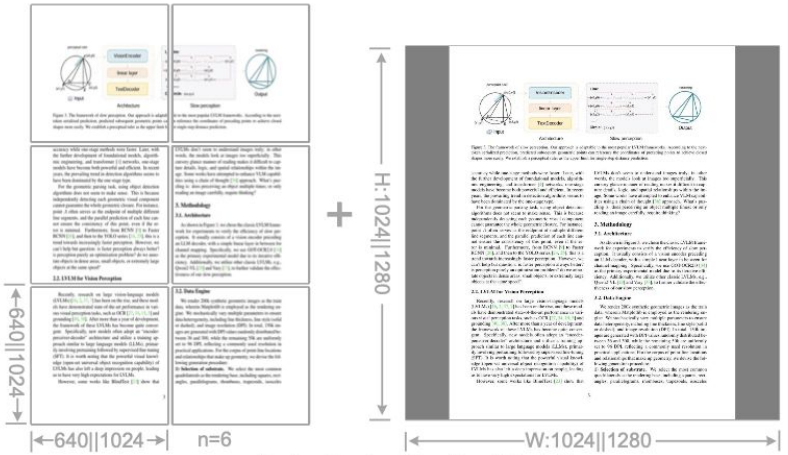
Mode: Tiny||Small

Token: 64||100



Mode: Base||Large

Token: 256||400 Valid: (256||400)×R R=1-(H-W)/W



Mode: Gundam||Gundam (Master)

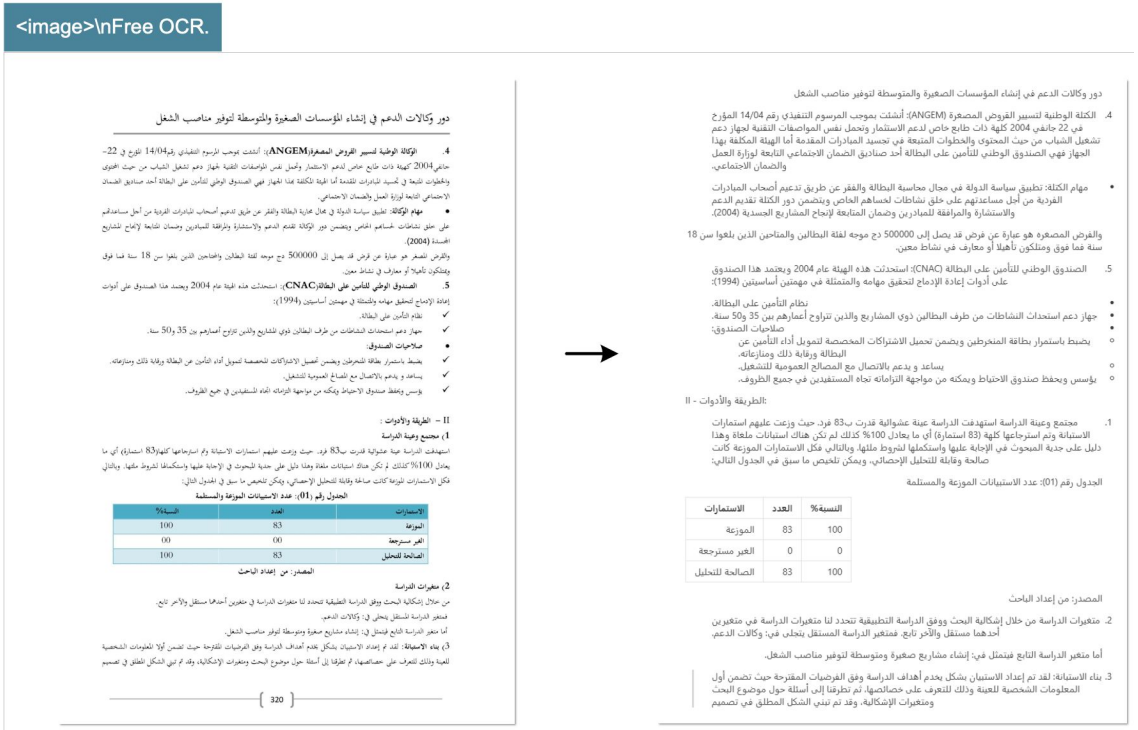
Token: n×(100||256) + (256||400) Valid: n×(100||256) + (256||400)×R n∈[2:9]

- Tiny(512×512, 64 token),
- Small(640×640, 100 token),
- Base(1024×1024, 256 token),
- Large(1280×1280, 400 token)

- Dynamic Modes: *Gundam* / *Gundam-Master*
- Ultra-high-resolution inputs
- Tiling strategy
- Local details + Global context

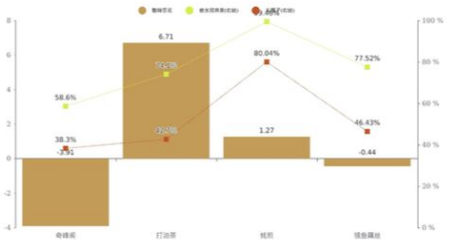
Real-World Application Scenarios

Traditional OCR Use Case: Based on the OCR1 dataset



Real-World Application Scenarios

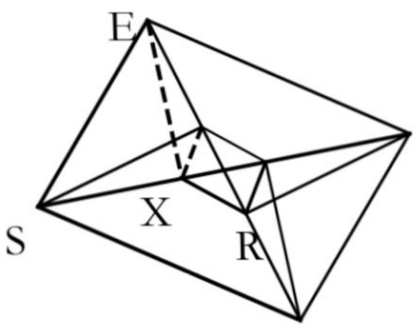
Advanced OCR Scenarios: OCR2 Dataset



<table><tr><td></td><td>酸辣豆花</td><td>修水双井茶(右轴)</td><td>车厘子(右轴)</td></tr><tr><td>奇峰阁</td><td>-3.91</td><td>58.6%</td><td>38.3%</td></tr><tr><td>打油茶</td><td>6.71</td><td>74.1%</td><td>42.7%</td></tr><tr><td>蚝煎</td><td>1.27</td><td>99.46%</td><td>80.04%</td></tr><tr><td>银鱼藕丝</td><td>-0.44</td><td>77.52%</td><td>46.43%</td></tr></table>

	酸辣豆花	修水双井茶(右轴)	车厘子(右轴)
奇峰阁	-3.91	58.6%	38.3%
打油茶	6.71	74.1%	42.7%
蚝煎	1.27	99.46%	80.04%
银鱼藕丝	-0.44	77.52%	46.43%

(a) Image-text ground truth of chart




```
{
 "Line": {
  "line": [
   "(-8.0, -1.4) — (-6.0, 2.07) — (-3.99, 5.54) — (-3.19, 6.89)",
   "(-8.0, -1.4) — (-4.42, 0.39) — (-0.84, 2.18) — (-0.6, 2.3)",
   "(-8.0, -1.4) — (-4.36, -3.06) — (-0.72, -4.72) — (2.92, -6.38) — (3.19, -6.51)",
   "(-8.0, -1.4) — (-4.07, -0.65) — (-0.14, 0.1) — (3.79, 0.85) — (7.72, 1.6) — (8.0, 1.66)",
   "(-3.19, 6.89) — (-2.4, 2.97) — (-1.75, -0.2)",
   "(-3.19, 6.89) — (-1.5, 3.26) — (0.19, -0.37) — (1.88, -3.99) — (3.19, -6.51)",
   "(-3.19, 6.89) — (0.45, 5.23) — (4.09, 3.57) — (7.73, 1.91) — (8.0, 1.66)",
   "(-1.75, -0.2) — (-0.6, 2.3)",
   "(-1.75, -0.2) — (0.9, -2.3)",
   "(-0.6, 2.3) — (1.9, 0.2)",
   "(0.9, -2.3) — (1.9, 0.2)",
   "(0.9, -2.3) — (4.48, -0.51) — (8.0, 1.66)",
   "(1.9, 0.2) — (2.69, -3.72) — (3.19, -6.51)",
   "(3.19, -6.51) — (5.19, -3.04) — (7.2, 0.43) — (8.0, 1.66)"
  ],
  "line_type": [
   "-","-","-","-","-","-","-","-","-","-","-","-","-","-","-"
  ],
  "line_endpoint": [
   "S: (-8.0, -1.4)",
   "E: (-3.19, 6.89)",
   "X: (-1.75, -0.2)",
   "R: (0.9, -2.3)"
  ]
 }
}
```


(b) Image-text ground truth of geometry

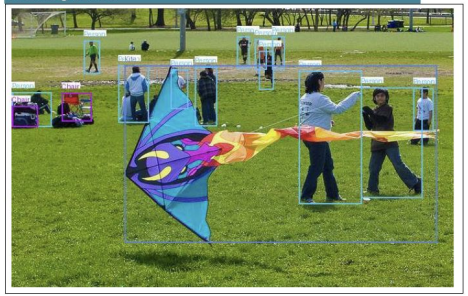
- Chart analysis: Converts visual charts to structured HTML table format
- Chemical formula recognition: Transforms molecular structures to SMILES notation
- Geometric parsing: Processes mathematical figures into structured coordinate systems
- Beyond what traditional OCR can do.

Real-World Application Scenarios

DeepSeek-OCR retains VLM-level generalization using the OCR3 dataset.

<image>\nLocate <|ref|>11-2=<|/ref|> in the image.

<image>\nIdentify all objects in the image and output them in bounding boxes.



<image>\nDescribe this image in detail.

A green plastic container labeled for food storage purposes. The lid is yellow and appears slightly worn out.

On the front label:
- At the top left corner, some Chinese characters.
- In bold black letters near the center-top reads "6月香" (June Fragrance).
- Below that, smaller white font states "鲜甜咸香" which translates roughly to "Fresh Sweet Salty Fragrant."
- Dominating the middle part of the label is large lettering stating "豆瓣酱," meaning "Bean Paste Sauce."

Additional Information includes:
- A small circular logo indicating quality assurance ("优质品").
- An illustration showing what seems like a dish made using bean paste sauce.
- Text below the main title says "专注25年，只做原酿酱" translating to "Focusing since 1998, only making authentic fermented bean paste."
- On the bottom right side, additional product specifications include "净含量：500g".

Overall Style/Type:
It looks like commercial packaging designed specifically for preserving food items within the jar. It has clear labeling intended for easy identification and understanding among consumers who read Mandarin.

Note: There might be slight variations due to handwriting styles but overall maintains readability regarding brand identity and contents description.

<image>\nLocate <|ref|>the teacher<|/ref|> in the image.

<image>\n这是一张

照片，展示了一辆红色的消防栓。消防栓上有一个笑脸，显得非常友好和亲切。消防栓的顶部有一个黑色的盖子，周围有一些金属铆钉。在消防栓的底部，有一个粉红色的贴纸，上面写着“bowtie”。背景中可以看到一条街道，街道上有几辆停放的汽车和一些树木。整体画面给人一种温馨和友好的感觉。

君不见，黄河之水天上来

，奔流到海不复回。君不见，高堂明镜悲白发，朝如青丝暮成雪。人生得意须尽欢，莫使金樽空对月。天生我材必有用，千金散尽还复来。烹羊宰牛且为乐，会须一饮三百杯。岑夫子，丹丘生，将进酒，杯莫停。与君歌一曲，请君为我倾耳听。钟鼓馔玉不足贵，但愿长醉不愿醒。古来圣贤皆寂寞，惟有饮者留其名。陈王昔时宴平乐，斗酒十千恣欢谑。主人何为言少钱，径须沽取对君酌。五花马，千金裘，呼儿将出换美酒，与尔同销万古愁。

Performance Evaluation

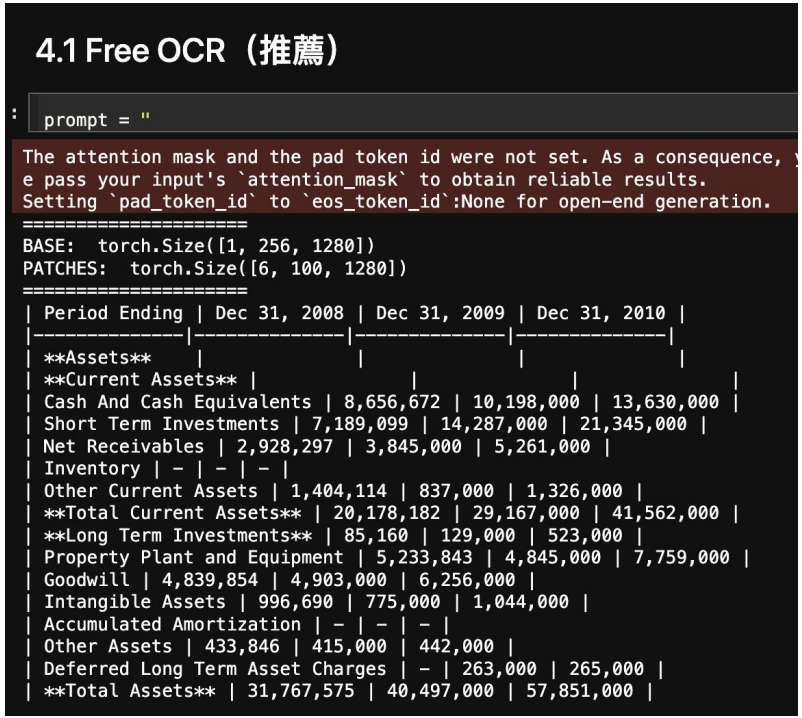
OCR Task Results: Achieving Efficiency Beyond SOTA

Model	Tokens	English					Chinese				
		overall	text	formula	table	order	overall	text	formula	table	order
Pipline Models											
Dolphin [11]	-	0.356	0.352	0.465	0.258	0.35	0.44	0.44	0.604	0.367	0.351
Marker [1]	-	0.296	0.085	0.374	0.609	0.116	0.497	0.293	0.688	0.678	0.329
Mathpix [2]	-	0.191	0.105	0.306	0.243	0.108	0.364	0.381	0.454	0.32	0.30
MinerU-2.1.1 [34]	-	0.162	0.072	0.313	0.166	0.097	0.244	0.111	0.581	0.15	0.136
MonkeyOCR-1.2B [18]	-	0.154	0.062	0.295	0.164	0.094	0.263	0.179	0.464	0.168	0.243
PPstructure-v3 [9]	-	0.152	0.073	0.295	0.162	0.077	0.223	0.136	0.535	0.111	0.11
End-to-end Models											
Nougat [6]	2352	0.452	0.365	0.488	0.572	0.382	0.973	0.998	0.941	1.00	0.954
SmolDocling [25]	392	0.493	0.262	0.753	0.729	0.227	0.816	0.838	0.997	0.907	0.522
InternVL2-76B [8]	6790	0.44	0.353	0.543	0.547	0.317	0.443	0.29	0.701	0.555	0.228
Qwen2.5-VL-7B [5]	3949	0.316	0.151	0.376	0.598	0.138	0.399	0.243	0.5	0.627	0.226
OLMOCR [28]	3949	0.326	0.097	0.455	0.608	0.145	0.469	0.293	0.655	0.652	0.277
GOT-OCR2.0 [38]	256	0.287	0.189	0.360	0.459	0.141	0.411	0.315	0.528	0.52	0.28
OCRFlux-3B [3]	3949	0.238	0.112	0.447	0.269	0.126	0.349	0.256	0.716	0.162	0.263
GPT4o [26]	-	0.233	0.144	0.425	0.234	0.128	0.399	0.409	0.606	0.329	0.251
InternVL3-78B [42]	6790	0.218	0.117	0.38	0.279	0.095	0.296	0.21	0.533	0.282	0.161
Qwen2.5-VL-72B [5]	3949	0.214	0.092	0.315	0.341	0.106	0.261	0.18	0.434	0.262	0.168
dots.ocr [30]	3949	0.182	0.137	0.320	0.166	0.182	0.261	0.229	0.468	0.160	0.261
Gemini2.5-Pro [4]	-	0.148	0.055	0.356	0.13	0.049	0.212	0.168	0.439	0.119	0.121
MinerU2.0 [34]	6790	0.133	0.045	0.273	0.15	0.066	0.238	0.115	0.506	0.209	0.122
dots.ocr$^{\dagger 200dpi}$ [30]	5545	0.125	**0.032**	0.329	**0.099**	**0.04**	0.16	**0.066**	0.416	0.092	**0.067**
DeepSeek-OCR (end2end)											
Tiny	**64**	0.386	0.373	0.469	0.422	0.283	0.361	0.307	0.635	0.266	0.236
Small	100	0.221	0.142	0.373	0.242	0.125	0.284	0.24	0.53	0.159	0.205
Base	256(182)	0.137	0.054	0.267	0.163	0.064	0.24	0.205	0.474	0.1	0.181
Large	400(285)	0.138	0.054	0.277	0.152	0.067	0.208	0.143	0.461	0.104	0.123
Gundam	795	0.127	0.043	0.269	0.134	0.062	0.181	0.097	0.432	0.089	0.103
Gundam-M$^{\dagger 200dpi}$	1853	**0.123**	0.049	**0.242**	0.147	0.056	**0.157**	0.087	**0.377**	**0.08**	0.085

- Small mode: Outperforms GOT-OCR 2.0 using only 100 vision tokens (vs. 256).
- Gundam mode: Outperforms MinerU 2.0 with under 800 tokens (vs. ~7,000).

Tutorial

Github Colab



- https://github.com/Heng-xiu/all-things-llm/blob/main/talks/DeepSeek%20OCR.ipynb