DeepSeek-OCR

Contexts Optical Compression

Document AI Seminar

Presented by: Zhaokun Wang

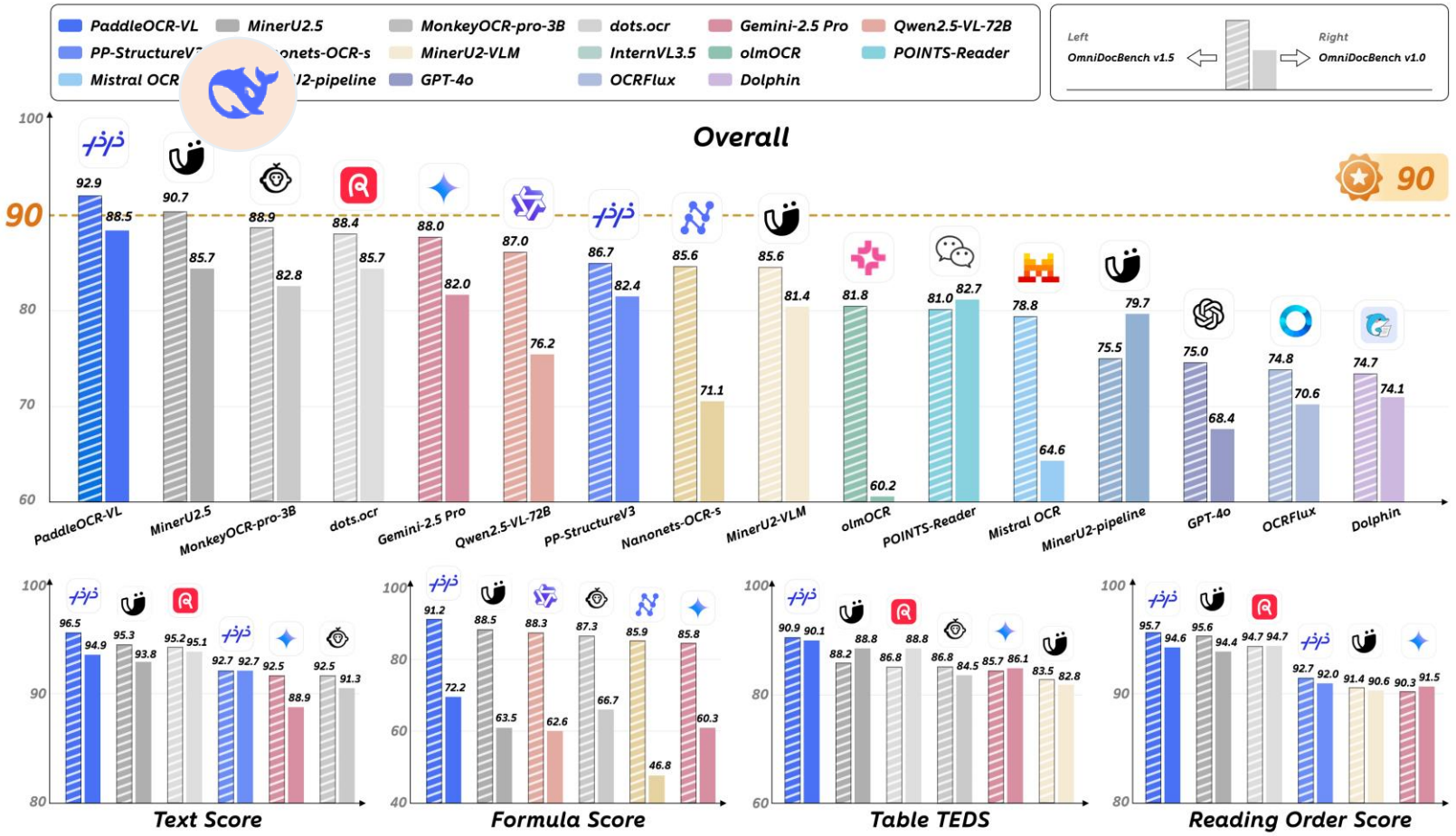
16.01.2026

Contents

Part I

Introduction

The OCR Benchmark



Downloads last month
3,152,827

HIGHLIGHTS

Intelligence: GPT-5.2 (xhigh) and Claude Opus 4.5 are the highest intelligence models, followed by Gemini 3 Pro Preview (high) & GPT-5.1 (high).

Output Speed (tokens/s): Granite 3.3 8B (671 t/s) and Gemini 2.5 Flash-Lite (Sep) (615 t/s) are the fastest models, followed by Gemini 2.5 Flash-Lite (Sep) & Nova Micro.

Latency (seconds): Apriel-v1.5-15B-Thinker (0.18s) and Llama Nemotron Super 49B v1.5 (0.21s) are the lowest latency models, followed by DeepSeek-OCR & NVIDIA Nemotron Nano 9B V2.

Price ($ per M tokens): Gemma 3n E4B ($0.03) and DeepSeek-OCR ($0.05) are the cheapest models, followed by Llama 3.2 1B & Llama 3.2 3B.

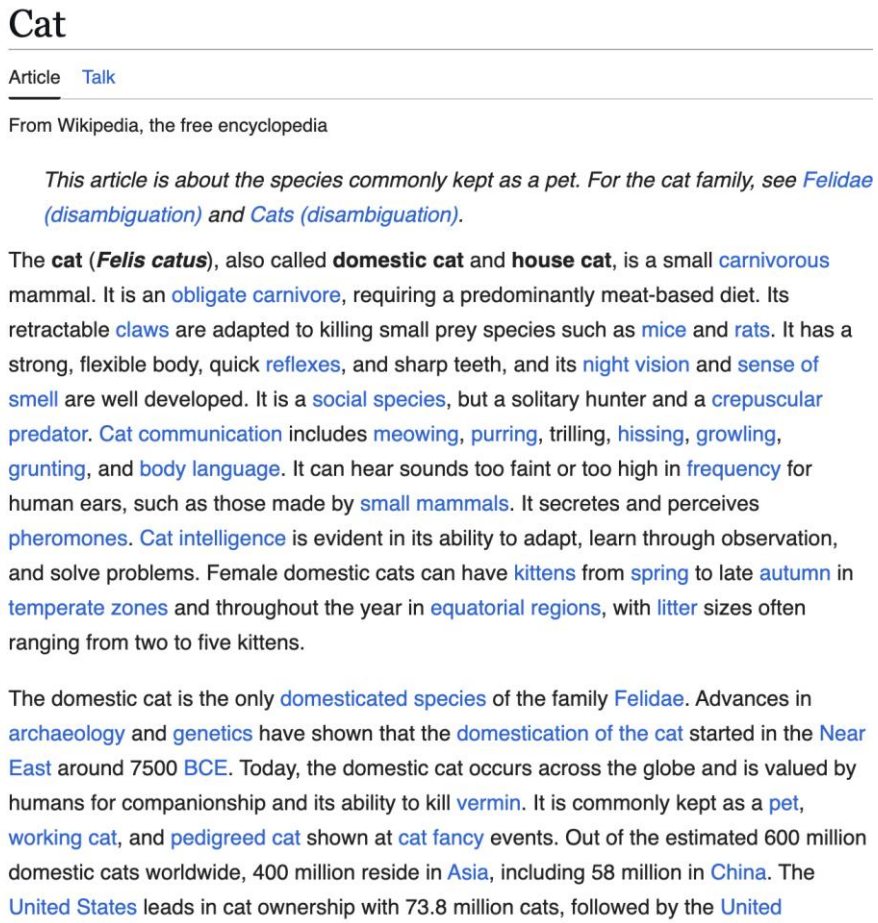
Context Window: Llama 4 Scout (10m) and Grok 4 Fast (2m) are the largest context window models, followed by Grok 4.1 Fast & Grok 4.1 Fast.

Figure 1: Overall OCR Benchmark Performance from PaddleOCR-VL paper and download times from Hugging Face(Comparison with SOTA Models)

Is a picture worth a thousand words?

The **cat** (***Felis catus***), also called **domestic cat** and **house cat**, is a small carnivorous mammal. It is an obligate carnivore, requiring a predominantly meat-based diet. Its retractable claws are adapted to killing small prey species such as mice and rats. It has a strong, flexible body, quick reflexes, and sharp teeth, and its night vision and sense of smell are well developed. It is a social species, but a solitary hunter and a crepuscular predator. Cat communication includes meowing, purring, trilling, hissing, growling, grunting, and body language. It can hear sounds too faint or too high in frequency for human ears, such as those made by small mammals. It secretes and perceives pheromones. Cat intelligence is evident in its ability to adapt, learn through observation, and solve problems. Female domestic cats can have kittens from spring to late autumn in temperate zones and throughout the year in equatorial regions, with litter sizes often ranging from two to five kittens.
The domestic cat is the only domesticated species of the family Felidae. Advances in archaeology and genetics have shown that the domestication of the cat started in the Near East around 7500 BCE. Today, the domestic cat occurs across the globe and is valued by humans for companionship and its ability to kill vermin. It is commonly kept as a pet, working cat, and pedigreed cat shown at cat fancy events. Out of the estimated 600 million domestic cats worldwide, 400 million reside in Asia, including 58 million in China. The United States leads in cat ownership with 73.8 million cats, followed by the United Kingdom with approximately 10.9 million cats. It also ranges freely as a feral cat, avoiding human contact. Pet abandonment contributes to increasing of the global feral cat population, which has driven the decline of bird, mammal, and reptile species. Population control includes spaying and neutering.

VS



Cat

Article Talk

From Wikipedia, the free encyclopedia

This article is about the species commonly kept as a pet. For the cat family, see Felidae (disambiguation) and Cats (disambiguation).

The **cat** (***Felis catus***), also called **domestic cat** and **house cat**, is a small carnivorous mammal. It is an obligate carnivore, requiring a predominantly meat-based diet. Its retractable claws are adapted to killing small prey species such as mice and rats. It has a strong, flexible body, quick reflexes, and sharp teeth, and its night vision and sense of smell are well developed. It is a social species, but a solitary hunter and a crepuscular predator. Cat communication includes meowing, purring, trilling, hissing, growling, grunting, and body language. It can hear sounds too faint or too high in frequency for human ears, such as those made by small mammals. It secretes and perceives pheromones. Cat intelligence is evident in its ability to adapt, learn through observation, and solve problems. Female domestic cats can have kittens from spring to late autumn in temperate zones and throughout the year in equatorial regions, with litter sizes often ranging from two to five kittens.

The domestic cat is the only domesticated species of the family Felidae. Advances in archaeology and genetics have shown that the domestication of the cat started in the Near East around 7500 BCE. Today, the domestic cat occurs across the globe and is valued by humans for companionship and its ability to kill vermin. It is commonly kept as a pet, working cat, and pedigreed cat shown at cat fancy events. Out of the estimated 600 million domestic cats worldwide, 400 million reside in Asia, including 58 million in China. The United States leads in cat ownership with 73.8 million cats, followed by the United

Figure 2: Screenshot vs. Textual Description

For a document containing 1000 words, how many vision tokens are at least needed for decoding?

Research Motivation

The Problem

LLMs face quadratic scaling challenges with long contexts, creating computational bottlenecks for document analysis and multi-turn conversations.

$O(n^2)$

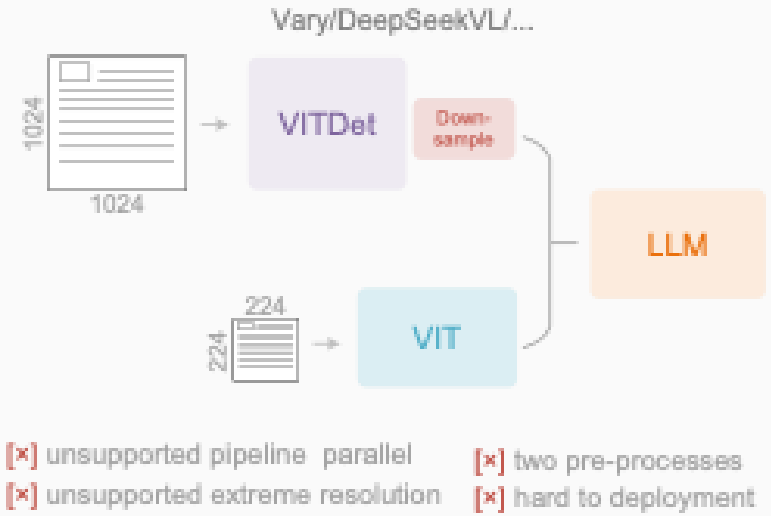
Attention Complexity

10M+

Computations for 10k tokens

This research explores whether a picture is worth a thousand words in the context of AI processing.

Current Vision Encoder Limitations

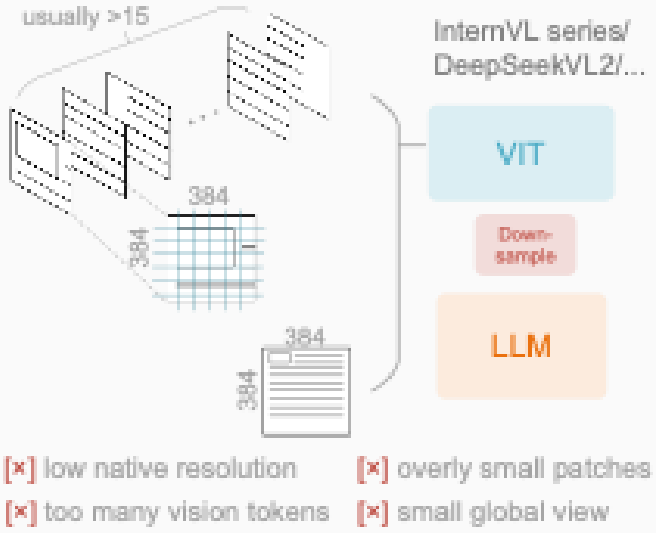


Method	Vision Backbone	Average	General				Knowledge				OCR & Chart				Vision-Centric			
			MME^P	MMB	$SEED^I$	GQA	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	CV-Bench^{2D}	CV-Bench^{3D}
SigLIP+DINOv2		51.61	1,432.02	61.28	65.99	63.30	68.82	35.69	29.40	60.01	43.00	35.70	60.40	37.54	30.00	53.99	55.52	53.58
SigLIP+DINOv2+ConvNext		54.52	1,503.51	63.83	67.97	63.95	70.40	35.99	29.30	60.69	48.20	36.90	64.97	45.53	34.67	58.69	55.74	60.33
SigLIP+DINOv2+ConvNext+CLIP		54.74	1,479.46	63.32	67.63	64.04	71.39	35.49	29.10	59.88	50.24	39.60	64.55	46.12	32.67	58.95	58.54	60.42
SigLIP+ConvNext		54.53	1,494.97	64.60	67.98	63.58	71.05	34.90	29.80	60.85	50.64	38.00	64.53	46.52	32.00	57.91	58.83	56.58
CLIP+ConvNext		54.45	1,511.08	63.83	67.41	63.63	70.80	35.09	30.40	59.91	51.32	35.00	64.45	47.88	33.33	57.25	56.32	59.08
SigLIP+DINOv2+ConvNext-L		53.78	1,450.64	63.57	67.79	63.63	71.34	34.80	30.20	61.04	49.32	37.70	64.05	45.83	30.00	56.21	58.08	54.33
SigLIP+CLIP+ConvNext-L		54.53	1,507.28	63.23	68.64	63.63	71.10	35.89	30.90	59.97	52.36	38.50	65.40	47.92	28.67	57.25	57.66	55.92

Table 3 | **All Benchmark Results for Model Ensemble with 1.2M Adapter Data + 737K Instruction Tuning Data.** Here, "SigLIP" = ViT-SO400M/14@384, "DINOv2" = ViT-L/14@518, "ConvNext" = OpenCLIP ConvNeXt-XXL@1024, and "CLIP" = OpenAI CLIP ViT-L/14@336.

Figure 3: Vision Backbone Benchmark Results (SigLIP, DINOv2, etc.)

Current Vision Encoder Limitations



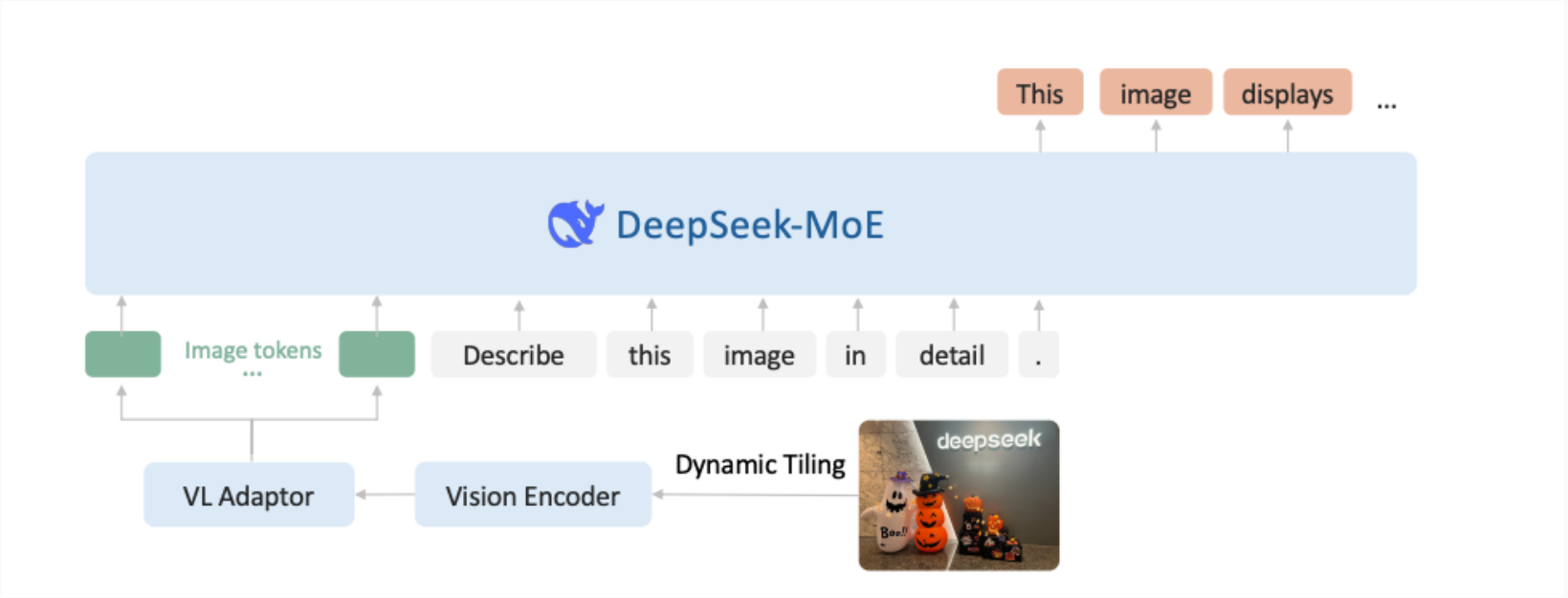
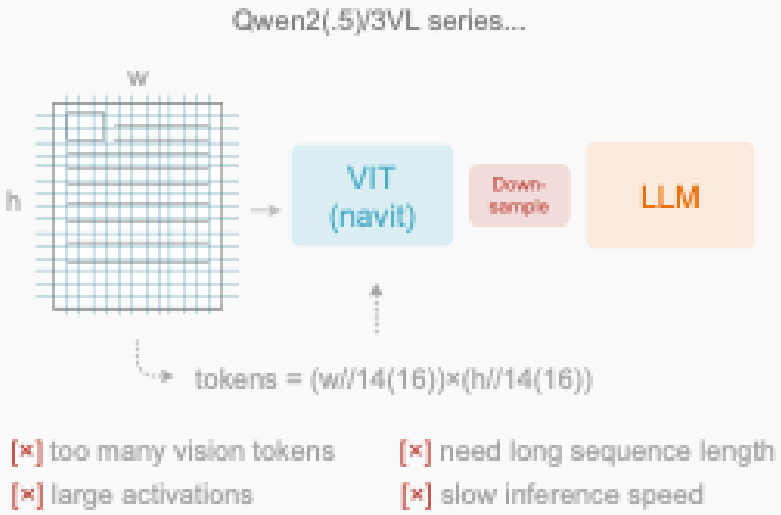


Figure 4: DS VL2'Mechanism

Current Vision Encoder Limitations



We present an alternative, NaViT. Multiple patches from different images are packed in a single sequence—termed Patch n' Pack—which enables variable resolution while preserving the aspect ratio (Figure 2). This

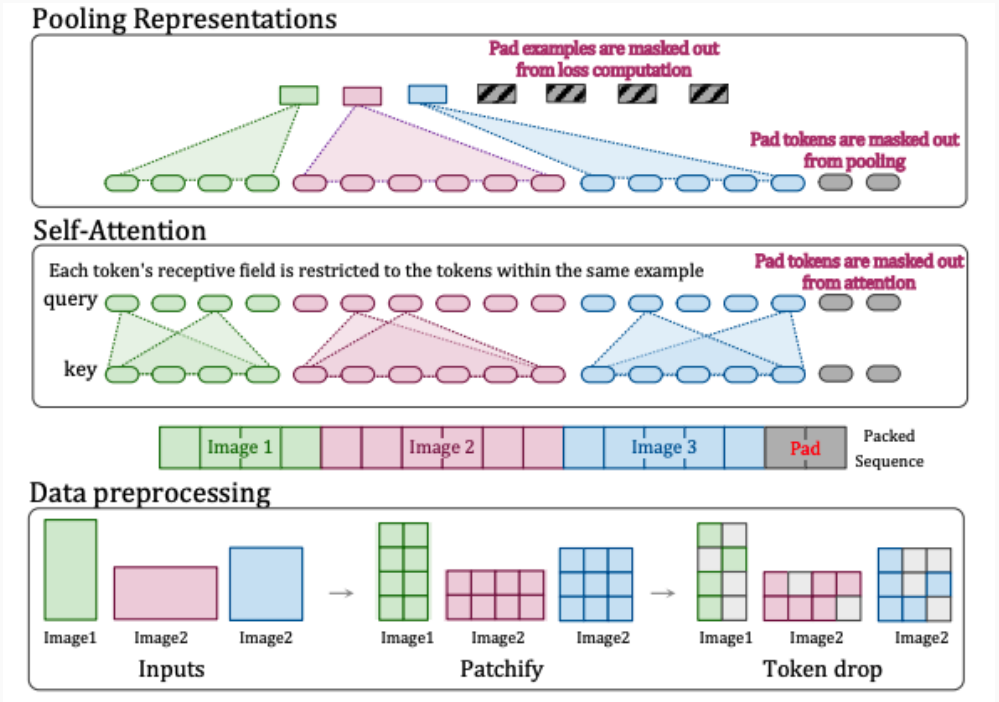


Figure 5: Patch n' Pack / NaViT Sequence Processing Mechanism

Current Vision Encoder Limitations

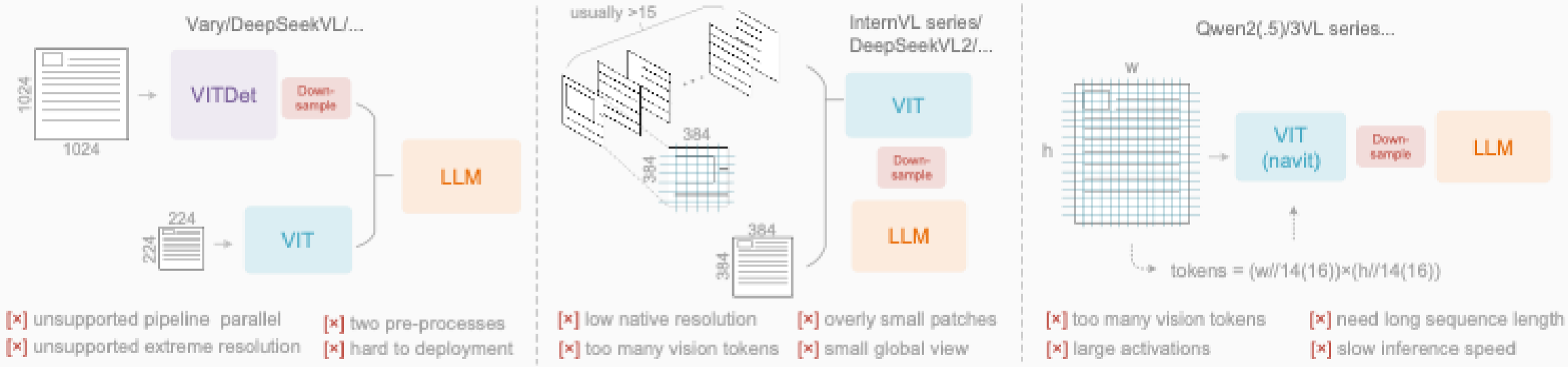


Figure 6: Comparison of Typical Vision Encoders in VLMs

These approaches can not achieve high-resolution, low-activation, few-token processing at the same time.

Vision-Text Compression Paradigm

The idea: Use visual modality as an efficient compression medium for textual information.

A single image can represent rich information using fewer tokens and achieve 7-20× compression ratios .

This can reduce computational cost while maintaining semantic integrity.

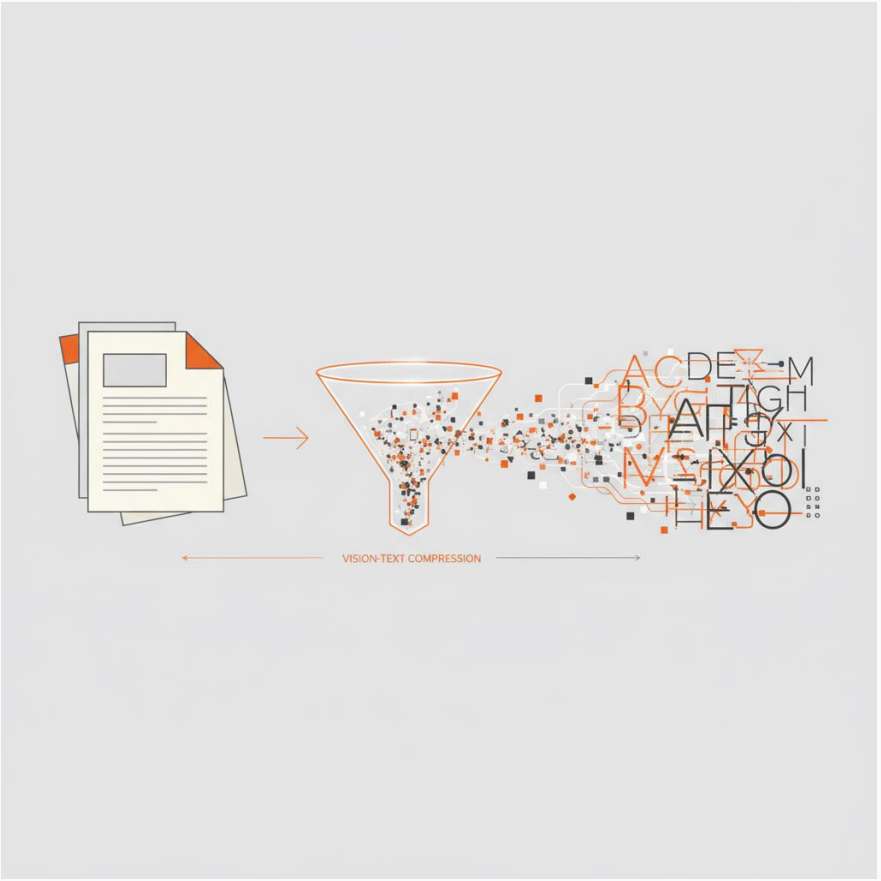


Figure 7: Vision-Text Compression Paradigm

Part II

Methodology

DeepSeek-OCR Architecture

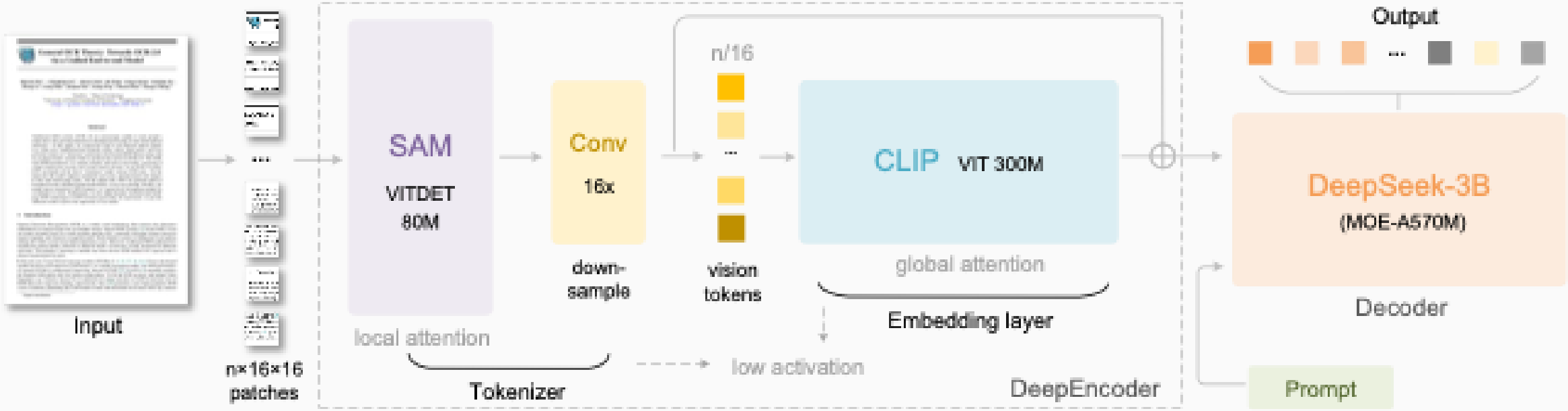


Figure 8: DeepSeek-OCR System Architecture (End-to-End)

DeepSeek-OCR Architecture

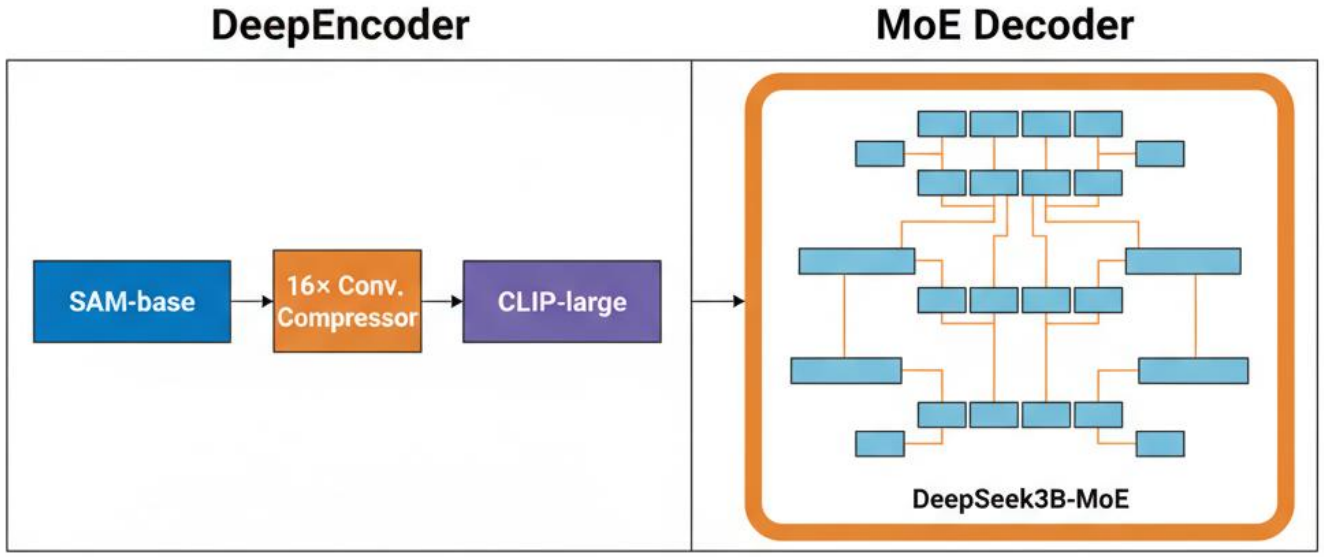


Figure 9: DeepSeek-OCR Module Details (Encoder & Decoder)

DeepEncoder

380M parameters combining SAM-base (80M) and CLIP-large (300M) with a 16× convolutional compressor.

16× Compressor

2-layer convolutional module that reduces vision tokens by 16× before global attention processing.

MoE Decoder

DeepSeek-3B-MoE with 570M activated parameters, using 6 routed experts and 2 shared experts.

High Resolution
Up to 1280×1280

Low Activation
Memory efficient

Few Tokens
64-400 tokens

Multi-Resolution
Dynamic support

Multi-Resolution Support

Tiny
512×512 (64)

Small
640×640 (100)

Base
1024×1024 (256)

Large
1280×1280 (400)

Gundam
Dynamic (n×100+256)

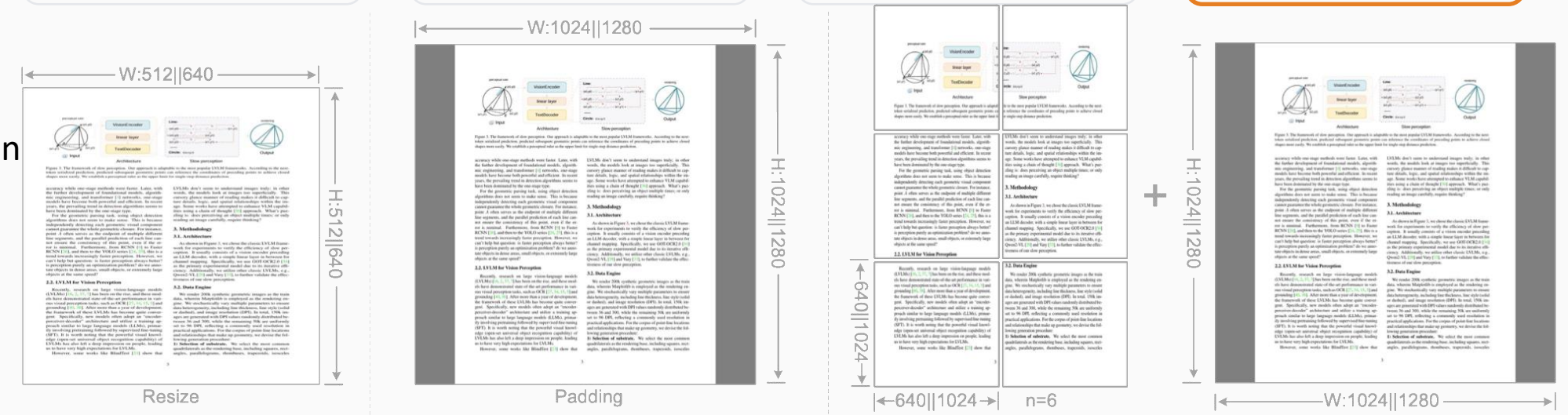


Figure 10: Multi-Resolution Support: Native vs. Dynamic Modes

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

Table 1: DeepEncoder Specifications: Resolution Modes & Token Counts

Data Engine Overview

70%
OCR Data
Core capability

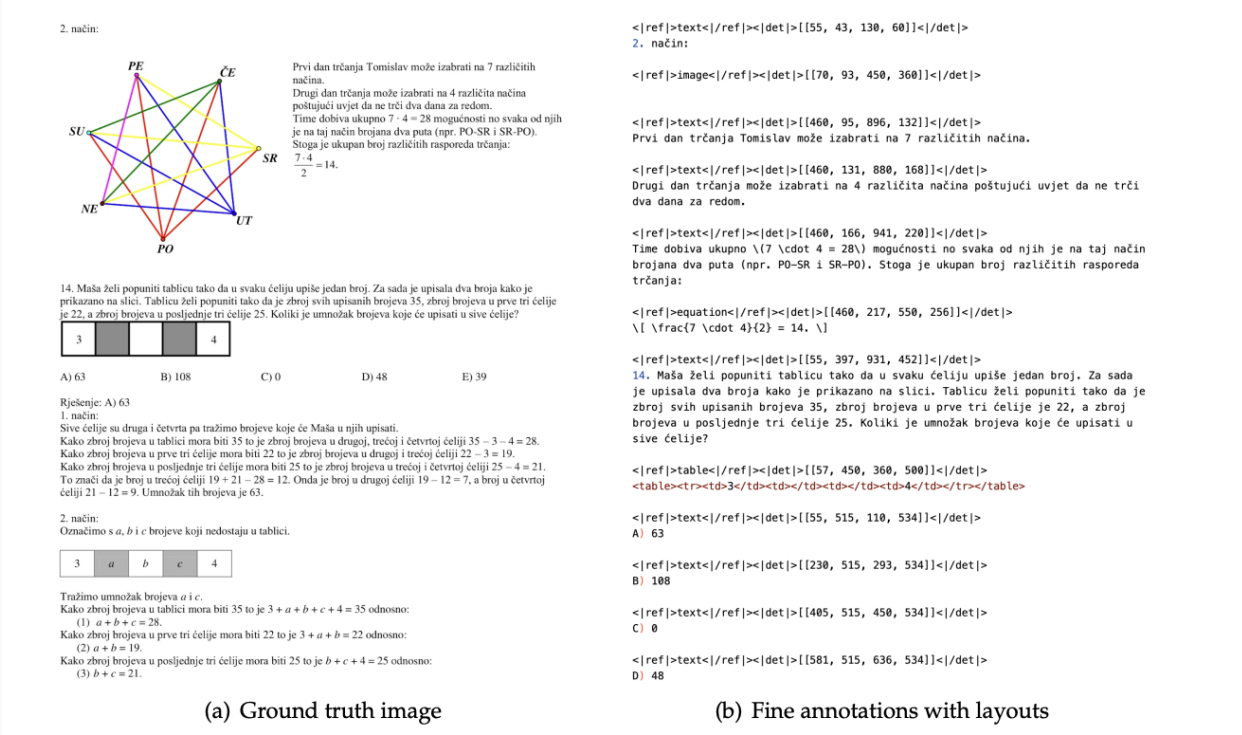
20%
Vision Data
General interface

10%
Text Data
Language skills

100+
Languages
Multilingual support

OCR 1.0 Data

30M PDF pages, 3M Word docs, 20M scene images covering traditional OCR tasks with both coarse and fine annotations.



(a) Ground truth image


```
<|ref|>text<|/ref|><|det|>[[55, 43, 130, 60]]<|/det|>
2. način:

<|ref|>image<|/ref|><|det|>[[70, 93, 450, 360]]<|/det|>


<|ref|>text<|/ref|><|det|>[[460, 95, 896, 132]]<|/det|>
Prvi dan trčanja Tomislav može izabrati na 7 različitih načina.

<|ref|>text<|/ref|><|det|>[[460, 131, 880, 168]]<|/det|>
Drugi dan trčanja može izabrati na 4 različita načina poštujući uvjet da ne trči
dva dana za redom.

<|ref|>text<|/ref|><|det|>[[460, 166, 941, 220]]<|/det|>
Time dobiva ukupno \(7 \cdot 4 = 28\) mogućnosti no svaka od njih je na taj način
brojana dva puta (npr. PO-SR i SR-PO). Stoga je ukupan broj različitih rasporeda
trčanja:

<|ref|>equation<|/ref|><|det|>[[460, 217, 550, 256]]<|/det|>
\[ \frac{7 \cdot 4}{2} = 14. \]

<|ref|>text<|/ref|><|det|>[[55, 397, 931, 452]]<|/det|>
14. Maša želi popuniti tablicu tako da u svaku ćeliju upiše jedan broj. Za sada
je upisala dva broja kako je prikazano na slici. Tablicu želi popuniti tako da je
zbroj svih upisanih brojeva 35, zbroj brojeva u prve tri ćelije je 22, a zbroj
brojeva u posljednje tri ćelije 25. Koliki je umnožak brojeva koje će upisati u
sive ćelije?

<|ref|>table<|/ref|><|det|>[[57, 450, 360, 500]]<|/det|>
<table><tr><td>3</td><td></td><td></td><td></td><td>4</td></tr></table>

<|ref|>text<|/ref|><|det|>[[55, 515, 110, 534]]<|/det|>
A) 63

<|ref|>text<|/ref|><|det|>[[230, 515, 293, 534]]<|/det|>
B) 108

<|ref|>text<|/ref|><|det|>[[405, 515, 450, 534]]<|/det|>
C) 0

<|ref|>text<|/ref|><|det|>[[581, 515, 636, 534]]<|/det|>
D) 48
```


(b) Fine annotations with layouts

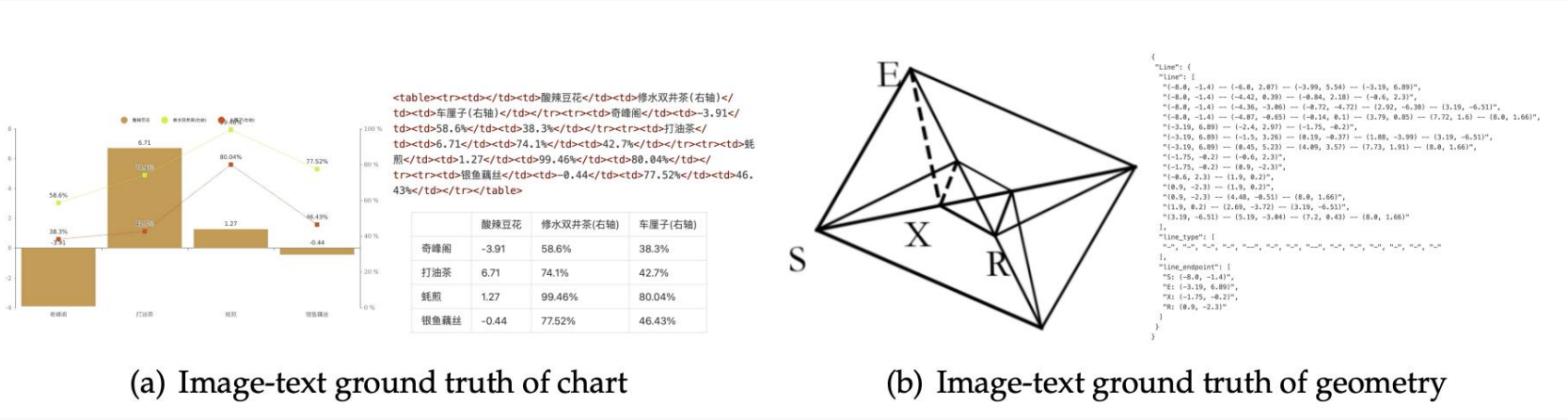
Figure 11a (Data): Training Data: Fine Annotation Examples (OCR 1.0)

Figure 11b (Data Formatting): Label Formatting: HTML Tables & Geometric Dictionaries

Data Engine Overview

OCR 2.0 Data

10M charts, 5M chemical formulas, 1M geometry problems for complex parsing tasks requiring structured understanding.



<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) — (4.08, -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

Figures 12 (Chart/Geometry Data): OCR 2.0 Data Generation: Charts & Plane Geometry

General Data

100M LAION samples for general vision understanding and interface preservation, plus text-only data for language capabilities.

Two-Stage Training Pipeline

1 DeepEncoder Training

Data: All OCR 1.0/2.0 data + 100M general samples for comprehensive feature learning.

Framework: Next token prediction with compact language model as decoder proxy.

Schedule: 2 epochs, batch size 1280, cosine annealing with LR 5e-5.

Length: 4096 sequence length for efficient training of vision-text alignment.

2 End-to-End Training

Platform: HAI-LLM with pipeline parallelism (4 parts) for efficient large model training.

Setup: 20 nodes (8 A100-40G each), DP=40, global batch size 640.

Speed: 90B text tokens/day, 70B multimodal tokens/day processing capability.

Optimization: AdamW with step-based scheduler, initial LR 3e-5.

Part III

Experiments

Vision-Text Compression Study

Table 2: Quantitative Analysis: Vision-Text Compression Ratios (Fox Benchmark)

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

Key Findings

Can achieve 97%+ precision within 10× compression ratio.

While 60% precision maintained even at 20× compression.

Near-lossless compression is achievable.

OmniDocBench Performance

Table 3: Performance Comparison on OmniDocBench (State-of-the-Art)

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

Competitive Advantages

- Surpasses GOT-OCR2.0 with 100 vs 256 tokens (2.56× efficiency).
- Matches SOTA performance with 400 tokens vs 6000+ for competitors.
- Outperforms MinerU2.0 with <800 tokens vs 6000+ tokens (7.5× efficiency).
- Demonstrates higher token efficiency across all document types.

OmniDocBench Performance

Document Type Performance

- Slides: 64 tokens is sufficient
- Books/Reports: 100 tokens optimal
- Newspapers: Gundam mode required for high text density

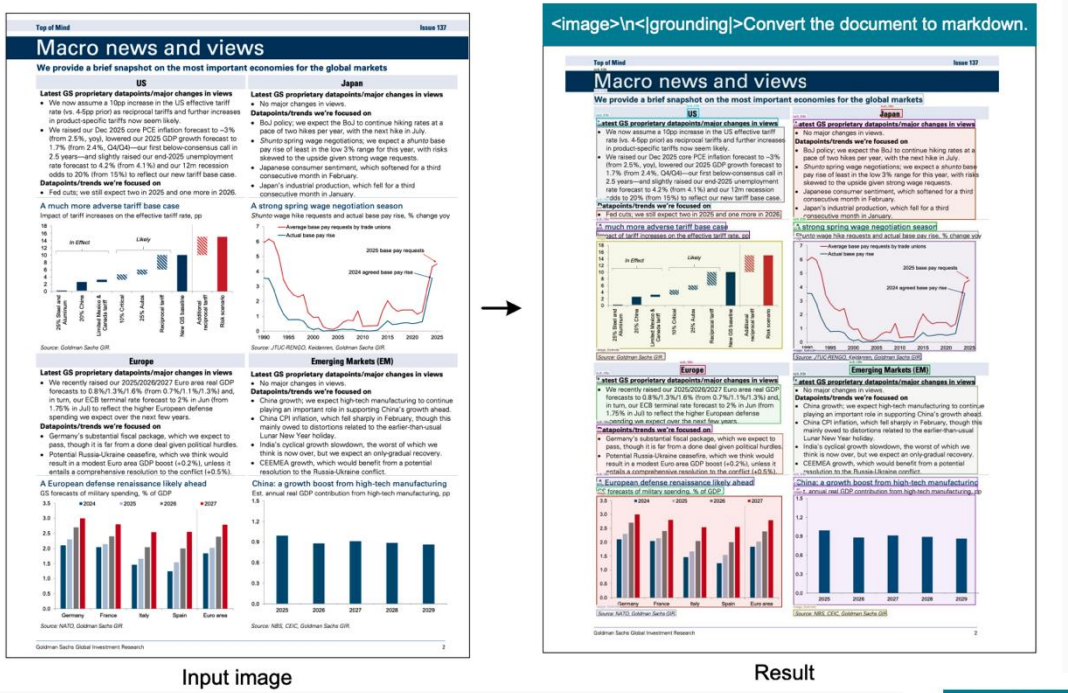
Table 4: Edit Distance Performance Across Document Categories

Mode \ Type	Book	Slides	Financial Report	Textbook	Exam Paper	Magazine	Academic Papers	Notes	Newspaper	Overall
Tiny	0.147	0.116	0.207	0.173	0.294	0.201	0.395	0.297	0.94	0.32
Small	0.085	0.111	0.079	0.147	0.171	0.107	0.131	0.187	0.744	0.205
Base	0.037	0.08	0.027	0.1	0.13	0.073	0.052	0.176	0.645	0.156
Large	0.038	0.108	0.022	0.084	0.109	0.06	0.053	0.155	0.353	0.117
Gundam	0.035	0.085	0.289	0.095	0.094	0.059	0.039	0.153	0.122	0.083
Guandam-M	0.052	0.09	0.034	0.091	0.079	0.079	0.048	0.1	0.099	0.077

Qualitative Study: Deep Parsing

Chart Parsing
HTML table extraction

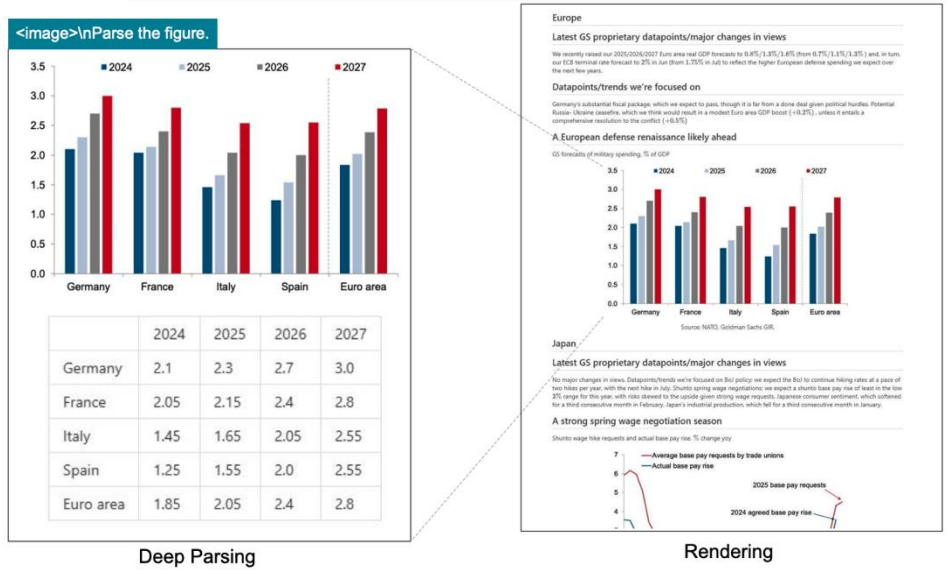
Natural Images
Dense captioning



Input image

Result

Figure 13 (Results): Deep Parsing: Structural Extraction of Financial Charts



	2024	2025	2026	2027
Germany	2.1	2.3	2.7	3.0
France	2.05	2.15	2.4	2.8
Italy	1.45	1.65	2.05	2.55
Spain	1.25	1.55	2.0	2.55
Euro area	1.85	2.05	2.4	2.8

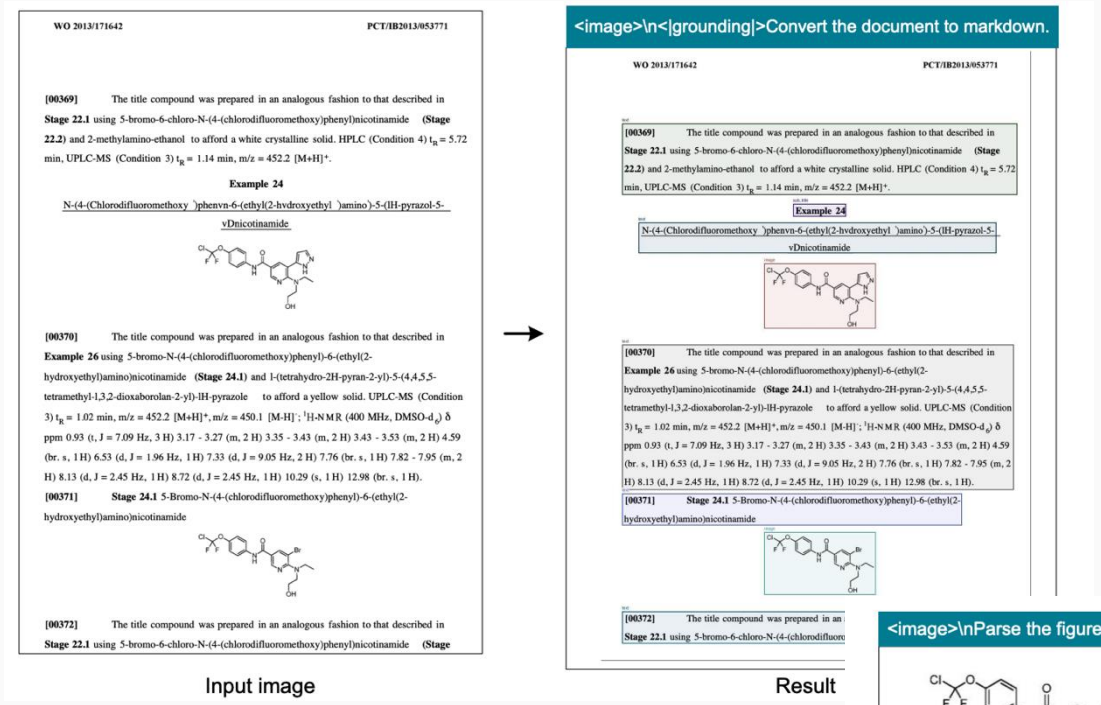
Deep Parsing

Rendering

Qualitative Study: Deep Parsing

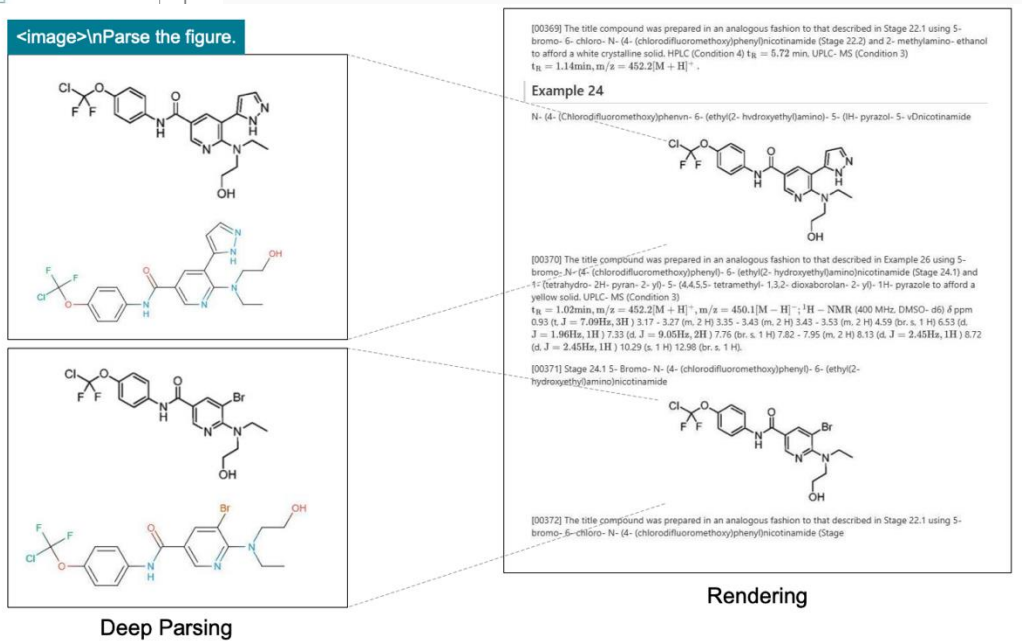
Chemical Formulas
SMILES format output

Geometry
Structured line segments



Input image

Result



Deep Parsing

Rendering

Figure 14: Deep Parsing: Chemical Formula Recognition to SMILES

Multilingual Recognition

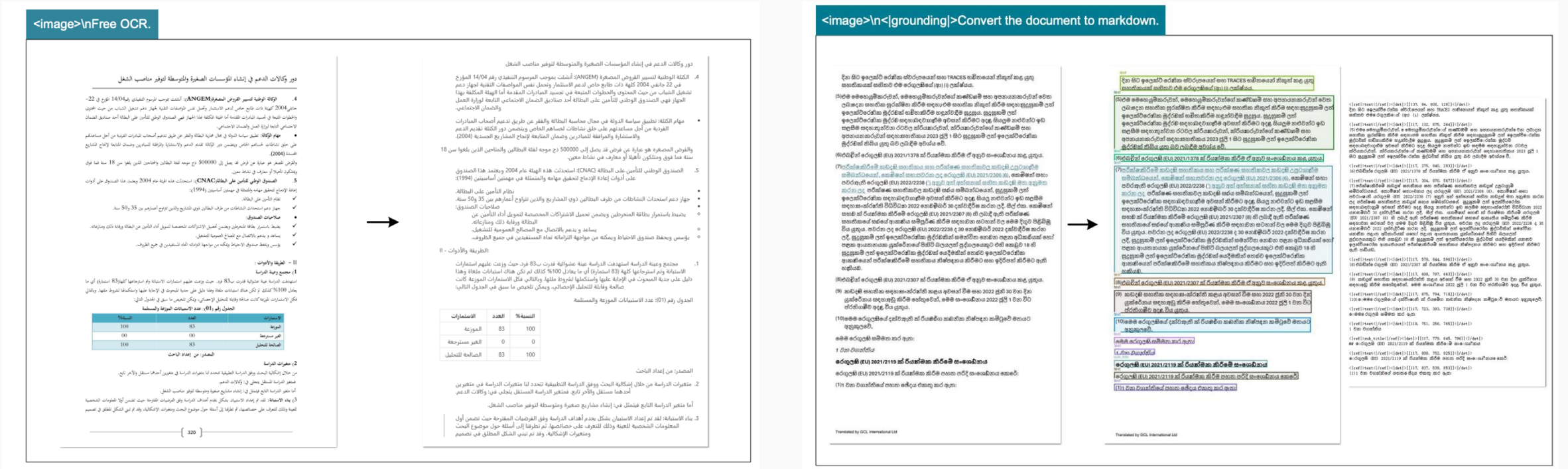


Figure 15: Multilingual Recognition Capabilities (100+ Languages)

Supports nearly 100 languages for PDF processing, crucial for training multilingual LLMs and global document processing.

- Arabic, Sinhala, and other minority languages with complex scripts
- Both layout and non-layout output formats via prompt control

General Visual Understanding

Figure 16: General Vision Understanding (Detection & Grounding)

Part IV

Applications

Production Deployment

200k+
Pages/Day
Single A100-40G

33M
Pages/Day
20 nodes (8 GPUs each)

100+
Languages
Production ready

24/7
Operation
Scalable infrastructure

Scalability Features

- **High throughput:** Massive data generation for LLM/VLM training at scale
- **Cost efficiency:** Minimal computational resources per page
- **Quality assurance:** Consistent high accuracy across document types

Document AI Applications

Historical Documents

Large-scale preservation and digitization of historical archives.

Multilingual Processing

100+ languages support for global document processing.

Memory Efficient

Dialogue history processing with 10× compression.

The system can generate 200k+ pages of training data per day on a single A100-40G GPU.

Memory Forgetting Mechanisms

Think about simulating human-like memory forgetting through visual compression.

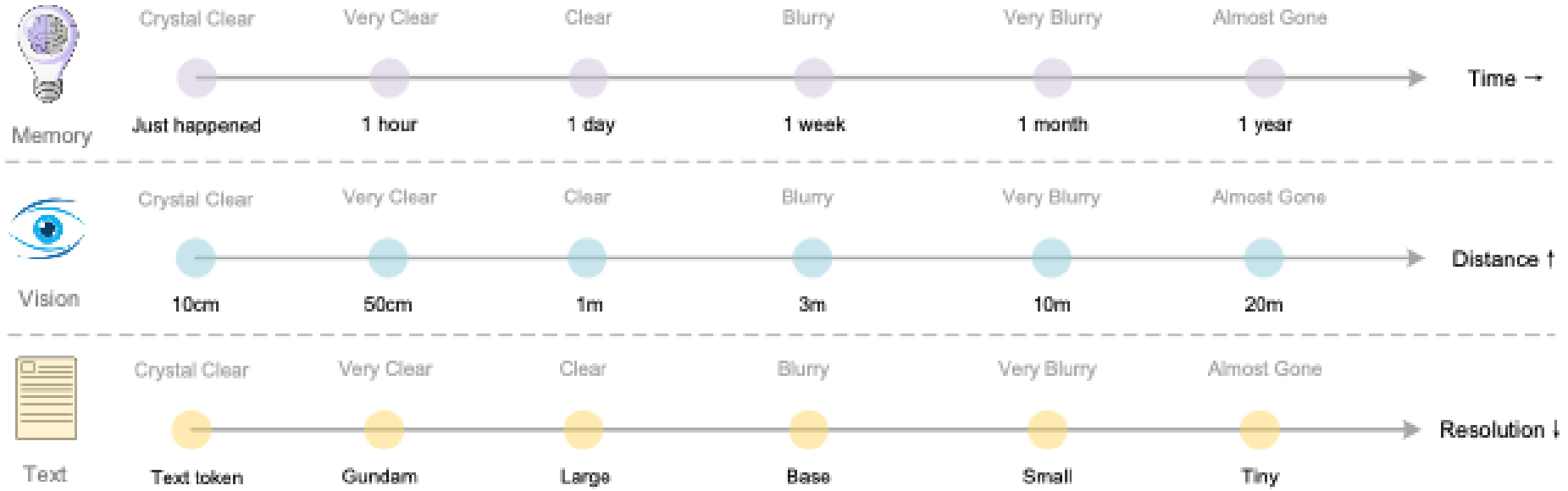


Figure 17: Future Direction: Optical Context Compression & Forgetting Mechanisms

Part V

Discussion

Look Back: Technical Innovation

Architecture Innovations

- Serial window and global attention connection through compression
- 16× token compressor bridging local and global processing
- Multi-resolution training flexibility for single model
- Dynamic positional encoding interpolation for resolution support

Efficiency Features

- MoE architecture for efficient inference (570M active of 3B)
- Dynamic resolution support through intelligent tiling
- Balanced capability development across modalities
- Memory-efficient processing with activation control

Training Pipeline

- 70% OCR data for core capability development
- 20% general vision data for interface preservation
- 10% text-only data for language skill maintenance
- Two-stage training with progressive complexity

Data Engineering

- 30M pages diverse PDF data from internet sources
- 100 languages coverage with specialized handling
- Sophisticated annotation pipeline with model flywheel
- Balanced multilingual representation

Limitations & Future Work

Current Limitations

- **OCR-focused validation:** Only demonstrates text compression, not general context compression
- **Resolution constraints:** Limited by image resolution for very long text sequences
- **Training scope:** No SFT stage for chatbot capabilities

Future Research Directions

- **Digital-optical pretraining:** Interleaved text-image training for general context
- **Needle-in-haystack testing:** Long-context retrieval and reasoning evaluation
- **Progressive compression:** Multi-level forgetting mechanisms with controllable decay
- **General context compression:** Beyond OCR to dialogue, code, and structured data

Impact on Document AI Field

Paradigm Shift
From text to vision compression

Practical Impact
Real-world deployment ready

Research Direction
New avenue for efficiency

Theoretical Contributions

- Compression limits
- Forgetting mechanisms: Biologically-inspired
- Cross-modal learning
- Efficiency paradigms

Practical Implications

- Long-context LLMs
- Memory systems
- Document processing

References

1. DeepSeek-OCR: Contexts Optical Compression H. Wei, Y. Sun, Y. Li. *arXiv preprint arXiv:2510.18234*, 2025.
2. PaddleOCR-VL: Boosting Multilingual Document Parsing via a 0.9B Ultra-Compact Vision-Language Model C. Cui, et al. *arXiv preprint arXiv:2510.14528*, 2025.
3. Qwen3 Technical Report Qwen Team. *arXiv preprint arXiv:2505.09388*, 2025.
4. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding Z. Lu, et al. *arXiv preprint arXiv:2412.10302*, 2024.
5. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs S. Tong, et al. *arXiv preprint arXiv:2406.16860*, 2024.
6. NExT-GPT: Any-to-Any Multimodal LLM S. Wu, et al. *ICML*, 2024.
7. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites (InternVL 1.2) Z. Chen, et al. *arXiv preprint arXiv:2404.16821*, 2024.
8. DeepSeek-VL: Towards Real-World Vision-Language Understanding H. Lu, et al. *arXiv preprint arXiv:2403.05525*, 2024.
9. Vary: Scaling up the Vision Vocabulary for Large Vision-Language Models H. Liu, et al. *arXiv preprint arXiv:2312.06109*, 2023.
10. Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution M. Dehghani, et al. *NeurIPS*, 2023.
11. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows Z. Liu, et al. *ICCV*, 2021.
12. Learning Transferable Visual Models From Natural Language Supervision (CLIP) A. Radford, et al. *ICML*, 2021.
13. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (ViT) A. Dosovitskiy, et al. *ICLR*, 2021.

References

•**Figure 1**: From **[2] PaddleOCR-VL ,Hugging Face, https://artificialanalysis.ai/leaderboards/models**.
•**Figure 2**: From a screenshot of Wikipedia.
•**Figure 3**: From **[5] Cambrian-1** (Benchmark results for vision backbones like SigLIP/DINOv2).
•**Figure 4**: From **[4] DeepSeek-VL2**.
•**Figure 5**: From **[10] NaViT** (Patch n' Pack mechanism).
•**Figure 7**: Generated by Gemini.
•**Figure 6 , Figures 8–17**: From the current paper **[1] DeepSeek-OCR** (Includes research motivation, model architecture, data engineering, and experimental results).

Thank you!