

Unlimited OCR Works 2

Pioneering the OCR-3.0 Era: From Long- to Unlimited-Horizon Parsing

Baidu ERNIE Team

Abstract

Unlimited OCR ushers the era of long-horizon parsing, substantially mitigating the growth of working memory as document lengthens. However, as document length scales to hundreds of pages, the visual context grows substantially, sharply increasing the computational cost of each decoding step and limiting the scalability of ultra-long document parsing. Inspired by human book transcription via localized focus without re-scanning preceding pages, we introduce Unlimited OCR 2, a document parsing model that looks only at what it transcribes. Building on Unlimited OCR, we upgrade all the decoder attention modules to Sparse Reference Sliding Window Attention (SR-SWA), bounding active reference tokens to a fixed capacity and thus sustaining constant computational costs across arbitrarily long documents. By combining SR-SWA with Unlimited-Horizon Decoding Paradigm, Unlimited OCR 2 enables document parsing over virtually unlimited lengths without prohibitive memory overhead. At the hundred-page scale, Unlimited OCR 2 achieves 7.0 $\times$ and 5.6 $\times$ per-step speedups over Full Attention and R-SWA, respectively. Furthermore, it delivers state-of-the-art end-to-end performance on both MPDocBench and MarathonDocBench, our newly proposed benchmark targeting hundred-page document parsing. Ultimately, the capacity for uninterrupted, one-shot document parsing signifies a fundamental paradigm shift into the OCR 3.0 era.

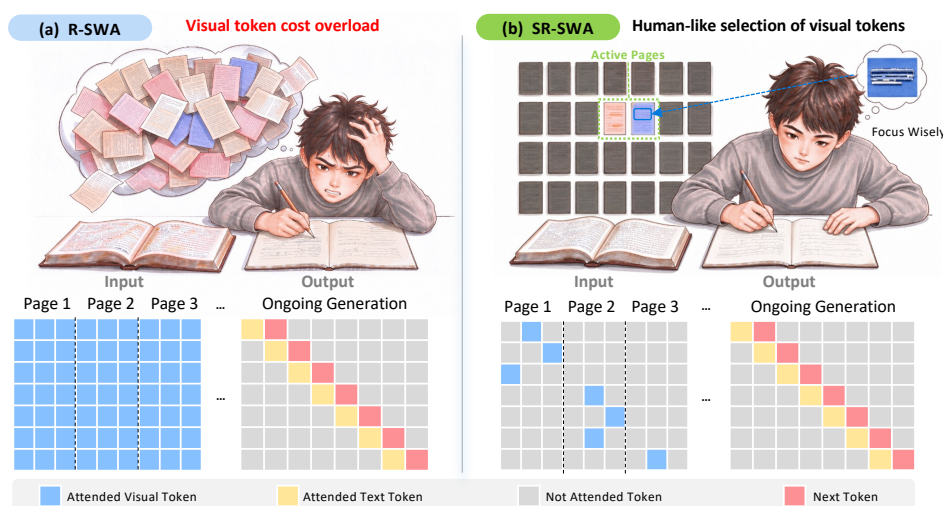


Figure 1 | R-SWA vs SR-SWA. R-SWA computes attention over all visual tokens across pages, inducing extreme visual redundancy; SR-SWA sparsely attends to relevant visual regions, aligning machine decoding with human reading intuition.

Contents

1	Introduction	3
2	Related Works	4
2.1	VLM-based Document Parsing	4
2.2	Efficient Long-Context Modeling	4
3	Methodology	4
3.1	Problem Setup	4
3.1.1	Unlimited-horizon Parsing	4
3.1.2	MarathonDocBench	5
3.2	Architecture	5
3.3	Sparse Reference Sliding Window Attention	6
3.3.1	Page Gate	6
3.3.2	Token Sieve	6
3.4	Unlimited-Horizon Decoding Paradigm	7
4	Experimental Settings	8
4.1	Data Engine	8
4.2	Implementation Details	8
5	Evaluation	9
5.1	Benchmarks and Metrics	9
5.2	Main Results	10
5.3	Scaling with Document Length	11
5.3.1	Feasibility of Unlimited-horizon Parsing	11
5.3.2	Length-Invariant Accuracy	11
5.3.3	Constant Per-Step Decoding Latency	12
6	Limitation and Future Work	12
7	Conclusion	12
8	Author List	13

1. Introduction

“Pixels are my body, and tokens are my blood. I have parsed over a thousand pages. Unknown to KV cache. Nor known to context limits. So, as I pray: Unlimited OCR Works.”

Snarcy on huggingface (Adapted from UBW)

Document parsing has undergone a major paradigm shift, transitioning from traditional multi-stage pipelines (OCR-1.0) [5, 6, 8, 20, 22, 26, 40] with handcrafted detection and recognition modules to end-to-end Vision-Language Models (OCR-2.0) [7, 18, 19, 21, 30, 31]. By formulating document understanding as a unified sequence-to-sequence transcription task, OCR-2.0 frameworks effectively eliminate cumulative pipeline errors. Our previous work, Unlimited OCR [37], pushed OCR-2.0 boundaries to tens of pages by using Reference Sliding Window Attention (R-SWA) to maintain a constant KV cache. Nevertheless, it still computes attention over all visual tokens across pages during decoding. As document length scales, this persistent attention to non-informative visual tokens severely escalates visual redundancy, compounding the difficulty of end-to-end inference. This visual attention bottleneck naturally prompts a fundamental question: how can we transition from length-limited parsing to an unlimited one-shot paradigm and genuinely inaugurate the OCR-3.0 era?

To answer this question, human transcription of lengthy books provides a clear physical intuition for parsing scaling: as we copy text page by page, we maintain localized visual focus, casting only momentary glances back when cross-page references are needed, rather than keeping the entire volume active in mind. This cognitive mechanism reveals that visual dependencies in multi-page documents are inherently page-sparse yet intra-page concentrated. However, standard dense global attention completely overlooks this sparse dynamic, incurring redundant attention interactions across all pages. This not only incurs prohibitive computational overhead, but also leads to severe memory bottlenecks as document length scales. To break through these efficiency and memory bottlenecks, we align machine perception with human reading patterns and present Unlimited OCR 2, pioneering an unlimited-horizon parsing framework for the OCR-3.0 era. Our main contributions are as follows:

- We introduce Sparse Reference Sliding Window Attention (SR-SWA), a hierarchical sparse attention mechanism designed for unlimited visual context. Specifically, Page Gate first narrows each query’s visual reference scope from the entire document to relevant pages while robustly modeling multi-page positional relationships. Token Sieve selectively incorporates top- k remote token buckets to capture long-range visual dependencies, thereby drastically reducing attention overhead while preserving comprehensive context.
- We introduce MarathonDocBench, a benchmark for hundred-page document parsing. It substantially extends existing evaluations in document length and cross-page structural complexity, enabling a comprehensive assessment of long-document parsing performance in the hundred-page regime.
- Comprehensive evaluations in long-horizon settings demonstrate that our approach maintains constant per-step decoding latency regardless of document length while achieving state-of-the-art end-to-end performance on multi-page benchmarks. Building on these advances, we establish an efficient unlimited-horizon parsing paradigm by integrating SR-SWA with the Unlimited-Horizon Decoding Paradigm, enabling document parsing over virtually unlimited lengths.

In summary, SR-SWA fundamentally breaks the computational and memory bottlenecks of ultra-long visual context. Operating at constant per-step decoding latency with length-invariant accuracy across hundreds of pages, Unlimited OCR 2 unlocks scalable long-document parsing and paves the way for the OCR-3.0 era of unlimited-horizon understanding.

2. Related Works

2.1. VLM-based Document Parsing

Document parsing has increasingly been reshaped by advances in vision-language models, evolving from pipeline-based OCR-1.0 systems [5, 6, 8, 20, 22, 26, 40], which combine specialized page analysis modules with VLM-based decoding, to unified VLM-based OCR-2.0 systems that directly generate structured representations from document images. OCR-2.0 encompasses both single-page parsers [7, 18, 19, 21, 30, 31] and Unlimited OCR [37], which extends end-to-end parsing to multi-page documents by jointly modeling their visual content within a shared context. However, although Unlimited OCR improves long-document efficiency with R-SWA, it still densely attends to all remote visual tokens, causing the visual context accessible to each decoding query to expand with document length and creating a fundamental scalability bottleneck at the hundred-page scale. Unlimited OCR 2 overcomes this limitation by explicitly bounding remote visual access, enabling attention-efficient end-to-end parsing of hundred-page documents and advancing document parsing into the OCR-3.0 era.

2.2. Efficient Long-Context Modeling

Efficient long-context modeling has been explored through both attention replacement and sparse attention. Linear or recurrent alternatives [11, 15, 17, 35, 36] reduce long-context cost by summarizing historical information into compact states. Document parsing, however, requires long-range memory to preserve fine-grained visual evidence for accurate recognition and cross-page retrieval, which may not be fully retained within such compact representations.

Sparse attention retains direct access to selected historical tokens or KV blocks while avoiding dense all-to-all computation. Early methods use predefined patterns such as local windows and global tokens [2, 38], whereas recent approaches adopt query-dependent retrieval. Hard-selection methods, such as Compressed Sparse Attention(CSA) [33] and MiniMax Sparse Attention(MSA) [16], rely on discrete top- k routing and typically require auxiliary supervision or warmup. In contrast, DashAttention [13] and HiLS [12] incorporate differentiable or forward-integrated routing scores, enabling more direct end-to-end optimization.

3. Methodology

3.1. Problem Setup

3.1.1. Unlimited-horizon Parsing

Unlimited-horizon parsing aims to enable end-to-end processing of documents of arbitrary lengths, from single-page forms to massive multi-page documents, with bounded computational and memory complexities. Formally, multi-page document parsing generates a transcription $\mathcal{Y} = (y_1, \dots, y_L)$ from an ordered document $\mathcal{D} = (P_1, \dots, P_N)$ with visual tokens $\mathcal{V}_{1:N}$. At decoding step t , the model operates on:

$$\mathcal{X}_t = \text{Concat}(\mathcal{V}_1, \dots, \mathcal{V}_N, y_{<t}). \quad (1)$$

While Unlimited OCR constrained text history to a fixed sliding window via R-SWA, it still requires attending over all preceding visual tokens $\mathcal{V}_{1:N}$. In the hundred-page regime, this page-dependent visual expansion becomes the primary bottleneck. Decoupling visual context from page count N to maintain a bounded computational and memory footprint directly motivates our proposed Sparse Reference Sliding Window Attention (SR-SWA).

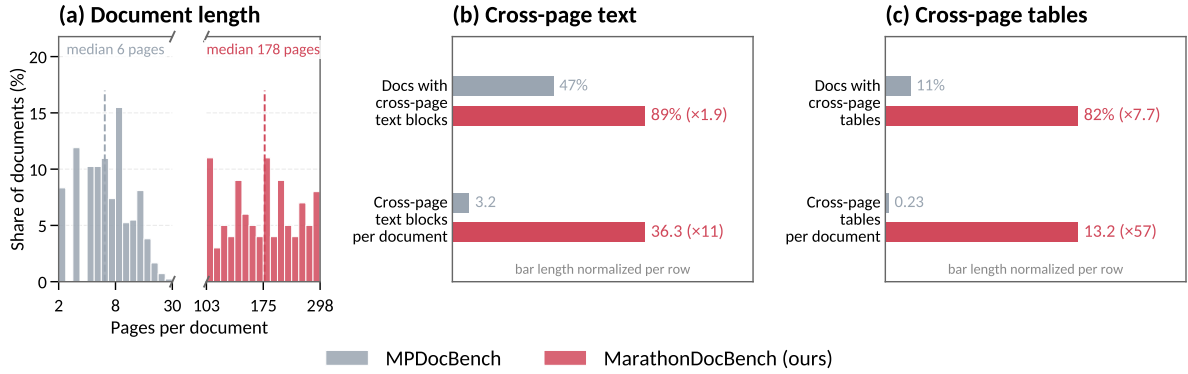


Figure 2 | MarathonDocBench versus MPDocBench. MarathonDocBench extends document length from tens to hundreds of pages (a), with substantially richer cross-page structures as shown in (b) and (c).

3.1.2. MarathonDocBench

Existing multi-page document parsing benchmarks neither extend to the hundred-page scale nor sufficiently capture the rich cross-page structures that emerge in long documents. To address this gap, we introduce MarathonDocBench, the first benchmark specifically designed for hundred-page document parsing. It contains 100 documents spanning 103–300 pages with 205k annotated layout elements, extending the document-length range by roughly an order of magnitude beyond MPDocBench [41].

Beyond document length, MarathonDocBench contains substantially richer cross-page structures than MPDocBench, with cross-page text and tables appearing much more frequently and densely across documents, as summarized in Figure 2. This makes MarathonDocBench a substantially more challenging testbed for evaluating whether end-to-end parsers can preserve document-level structure while scaling from tens to hundreds of pages. To ensure reliable evaluation, structural annotations, including bounding boxes, block type, hierarchy, reading order, and cross-page continuations of text and tables, are initialized from a strong page-by-page parser and carefully refined by human annotators. Content transcriptions are initialized from the same parser outputs and verified through an iterative VLM-based refinement pipeline, with flagged disagreements adjudicated by human annotators.

3.2. Architecture

Unlimited OCR 2 inherits the fundamental backbone from Unlimited OCR, leveraging DeepSeek OCR [28] as its foundation. Specifically, it employs the DeepEncoder paired with a Mixture-of-Experts (MoE) decoder architecture featuring 3B total parameters and 500M activated parameters per token. Building upon this base, we replace the baseline R-SWA with our proposed SR-SWA as shown in Figure 3. While R-SWA bounds local interactions, its visual context scope still expands with growing document length, leading to noticeable latency accumulation over ultra-long horizons. By introducing a selective remote visual routing mechanism, SR-SWA strictly bounds the active context scope, enabling Unlimited OCR 2 to maintain a constant KV cache footprint across both text and visual modalities during unlimited-horizon decoding. We detail the inner mechanics of SR-SWA in the following subsections.

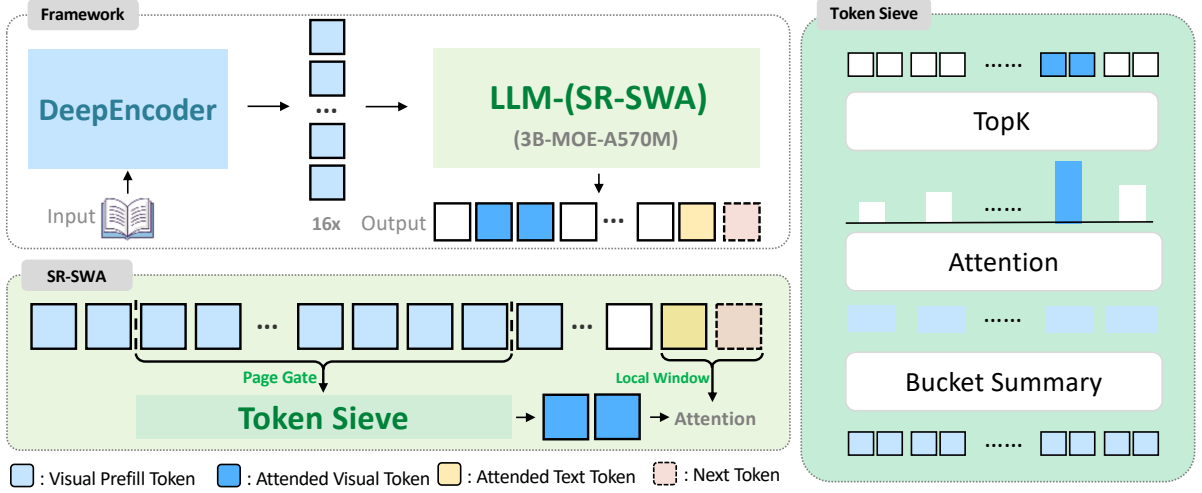


Figure 3 | Inspired by the process of humans copying books, we propose Unlimited OCR 2. SR-SWA combines Page Gate and Token Sieve to construct a fixed-length set of working reference visual tokens, substantially reducing computational and memory overhead while keeping both independent of document length throughout generation.

3.3. Sparse Reference Sliding Window Attention

Following standard block-sparse attention [13, 16, 32, 33], we partition X_t into fixed-size blocks $\{\mathcal{B}_b\}_{b=1}^B$ and preserve a fixed local window $\mathcal{L}(q)$ for each query q . SR-SWA constructs $\mathcal{S}(q)$ hierarchically, using Page Gate for page-level organization and Token Sieve for fixed-budget selection.

3.3.1. Page Gate

To eliminate irrelevant visual tokens and concentrate focus on the current page, Page Gate restricts the visual attention scope for a query q associated with page $g(q)$ by retaining only blocks from designated target pages:

$$C_{\text{page}}(q) = \{b \mid \mathcal{B}_b \cap \mathcal{V}_{g(q)} \neq \emptyset\}. \quad (2)$$

Since block boundaries are independent of page boundaries, a candidate block may straddle adjacent pages. Meanwhile, to prevent positional explosion as page count scales, Page Gate factorizes each token position into a bounded page-level coordinate (following an alternating 0-1 pattern) and a monotonic intra-page coordinate:

$$\mathbf{p}_{i,j} = \left[\mathbf{p}_i^{\text{page}}, \mathbf{p}_{i,j}^{\text{intra}} \right], \quad (3)$$

ensuring both stable length extrapolation and efficient cross-page visual routing.

3.3.2. Token Sieve

While Page Gate restricts the routing space across pages, efficient fixed-budget routing requires a principled block selection strategy. Conventional sparse routers typically rely on dense-attention distillation, which lacks direct alignment with the language modeling objective. Inspired by recent end-to-end routing designs [12, 13], Token Sieve directly optimizes fixed-budget block selection under the language modeling objective.

Given the candidate set $C_{\text{page}}(q)$ from Page Gate, Token Sieve summarizes each candidate block b into a compact key representation $\bar{\mathbf{k}}_b$ and computes its relevance to query q . The router then forms a sparse routing distribution and retains only the top- k blocks:

$$s_b(q) = \frac{\tilde{\mathbf{q}}^\top \bar{\mathbf{k}}_b}{\sqrt{d}}, \quad \boldsymbol{\pi}(q) = \alpha\text{-entmax} \left(\{s_b(q)\}_{b \in C_{\text{page}}(q)} \right), \quad (4)$$

$$\mathcal{S}(q) = \text{TopK}(\boldsymbol{\pi}(q), k).$$

Here, d denotes the hidden dimension. By selectively attending to a fixed number of top- k remote buckets, Token Sieve captures long-range visual dependencies with minimal computational overhead. To preserve relative preferences among selected blocks, the routing probabilities are converted into a mean-centered log-prior:

$$\beta_b(q) = \log \pi_b(q) - \frac{1}{|\mathcal{S}(q)|} \sum_{b' \in \mathcal{S}(q)} \log \pi_{b'}(q), \quad b \in \mathcal{S}(q). \quad (5)$$

With the selected remote blocks $\mathcal{S}(q)$ and the fixed local window $\mathcal{L}(q)$, each query attends to a length-agnostic retained token set defined as:

$$\mathcal{R}(q) = \mathcal{L}(q) \cup \bigcup_{b \in \mathcal{S}(q)} \mathcal{B}_b. \quad (6)$$

The attention weights over $\mathcal{R}(q)$ are computed as:

$$\alpha_j(q) = \text{Softmax} \left(\frac{q^\top k_j}{\sqrt{d}} + \mathbb{I}[b(j) \in \mathcal{S}(q)] \beta_{b(j)}(q) \right), \quad (7)$$

where the softmax is normalized over all tokens in $\mathcal{R}(q)$, $b(j)$ denotes the block containing token j , and k_j denotes its key projection. Tokens within the local window use standard attention logits, while tokens from selected remote blocks additionally incorporate the corresponding routing bias. Since the routing bias directly participates in these logits, gradients from the parsing objective can propagate to the routing module, enabling end-to-end optimization without an auxiliary KL-divergence distillation loss.

Algorithm 1 Cross-Chunk KV Reuse for Unlimited-Horizon Decoding

Require: Pages $P_1, \dots, P_N$; instruction q ; chunk size W ; tail length T

Ensure: Transcription Y

Notation: C is a KV cache; $\oplus$ concatenates KV in order;

$\text{ENC}(X \mid C)$ returns the KV of input X computed with C as context.

- 1: Partition N pages into K overlapping chunks $C_1, \dots, C_K$ with a 1-page boundary overlap $\triangleright$
 $K=1$ indicates unsegmented
 - 2: $C \leftarrow \text{ENC}(C_1 \oplus q)$ $\triangleright$ only the first chunk carries q
 - 3: $y_1 \leftarrow \text{DECODE}(C)$; $Y \leftarrow y_1$
 - 4: **for** $k = 2$ **to** K **do**
 - 5: $C \leftarrow C_{\text{overlap}}^{\text{img}} \oplus \text{ENC}(C_k \mid C_{\text{overlap}}^{\text{img}}) \oplus C_{\text{tail}}^{\text{txt}}$
 - 6: $y_k \leftarrow \text{DECODE}(C)$; $Y \leftarrow Y \oplus y_k$
 - 7: **end for**
 - 8: **return** Y
-

3.4. Unlimited-Horizon Decoding Paradigm

While SR-SWA bounds the number of reference tokens accessed by each query, processing extremely long multi-page documents remains constrained by the GPU memory required to

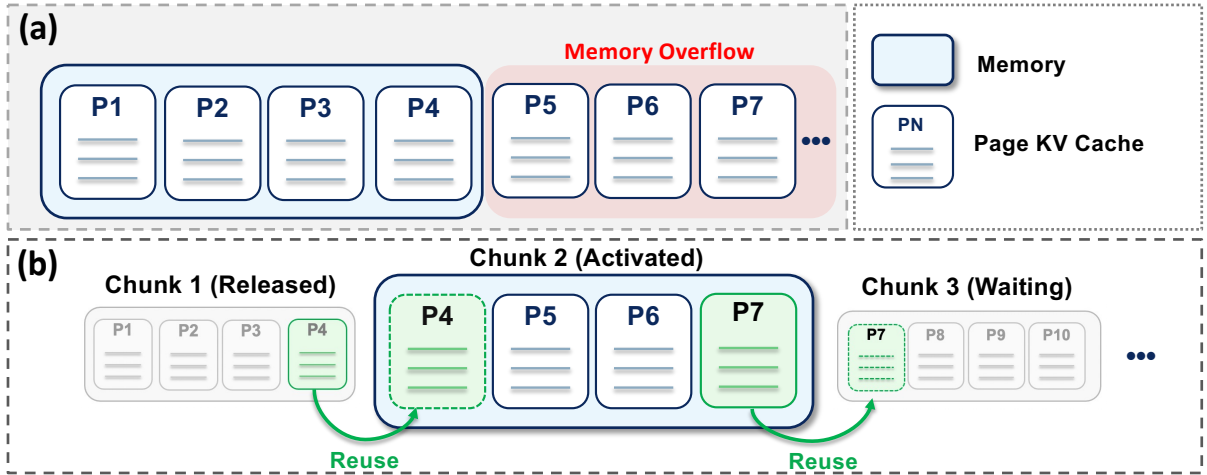


Figure 4 | Comparison of (a) standard decoding and (b) our Unlimited-Horizon Decoding paradigm. Standard decoding retains visual KV caches for all pages, causing GPU memory usage to grow linearly with document length and eventually overflow. Our paradigm divides long documents into manageable chunks and processes them sequentially, reusing the KV cache of shared boundary pages and releasing inactive caches to keep GPU memory usage bounded while preserving cross-chunk continuity.

prefill and retain their visual KV caches. As illustrated in Figure 4, conventional all-at-once prefill keeps all page-level visual KV caches in memory, causing the memory footprint to grow linearly with document length. To overcome this limitation, we introduce the Unlimited-Horizon Decoding Paradigm, which partitions the document into overlapping page-level chunks and processes them sequentially. Adjacent chunks share their boundary page, whose KV cache is directly reused to preserve cross-chunk continuity, while visual KV caches outside the active chunk are released. Consequently, only a fixed-size chunk remains resident in GPU memory, bounding the memory footprint independently of the total document length. The complete procedure is formally described in Algorithm 1.

4. Experimental Settings

4.1. Data Engine

To train Unlimited OCR 2 for versatile document parsing, we construct a large-scale multimodal corpus comprising 2M multi-page and 10M single-page documents. Pseudo-labels across both single- and multi-page subsets are extracted from MinerU2.5-Pro [26] and PaddleOCR-VL-v1.6 [40], tailored to their respective domain strengths. To enrich visual and layout diversity, we additionally incorporate versatile synthetic document generation pipelines. Across both real and synthetic data, we leverage VLM-based multi-dimensional feature analysis to conduct balance filtering, ensuring uniform coverage over complex layout distributions. Finally, to explicitly fortify structural hierarchy and cross-page continuity in multi-page scenarios, we apply MinerU-Popo [34] refinement alongside dedicated heuristic post-processing rules to reconcile layout inconsistencies across adjacent page boundaries.

4.2. Implementation Details

We initialize Unlimited OCR 2 from the Unlimited OCR checkpoint and perform post-training for 1 epoch on the curated dataset. Training is conducted on a cluster of 25×8 GPU NVIDIA H800 nodes with a global batch size of 1,600 and a maximum sequence length of 64K tokens.

During training, we freeze both the DeepEncoder and the aligner, optimizing only the LLM parameters. Optimization is carried out using the AdamW optimizer with a cosine learning rate decay schedule, starting at an initial learning rate of 2.5×10^{-4} . The entire training workflow is implemented using the MS-Swift framework. For deployment, the Unlimited-Horizon Decoding Paradigm is integrated into SGLang to facilitate efficient long-document inference.

5. Evaluation

5.1. Benchmarks and Metrics

To comprehensively evaluate Unlimited OCR 2 on document parsing across varying lengths, we conduct experiments on two challenging multi-page benchmarks: MPDocBench [41] (up to 30 pages) and our proposed MarathonDocBench (ranging from 100 to 300 pages). Following the official evaluation protocol of MPDocBench, we report Edit Distance for textual content and reading order recognition, TEDS for table structure recognition, CDM for mathematical formula recognition, and F1 score for figure region detection. To further evaluate cross-page consistency, we include T-TextEdit and T-TEDS to assess text and table parsing across page boundaries, respectively. An overall score aggregated across these fine-grained structural categories is also computed to summarize overall parsing performance.

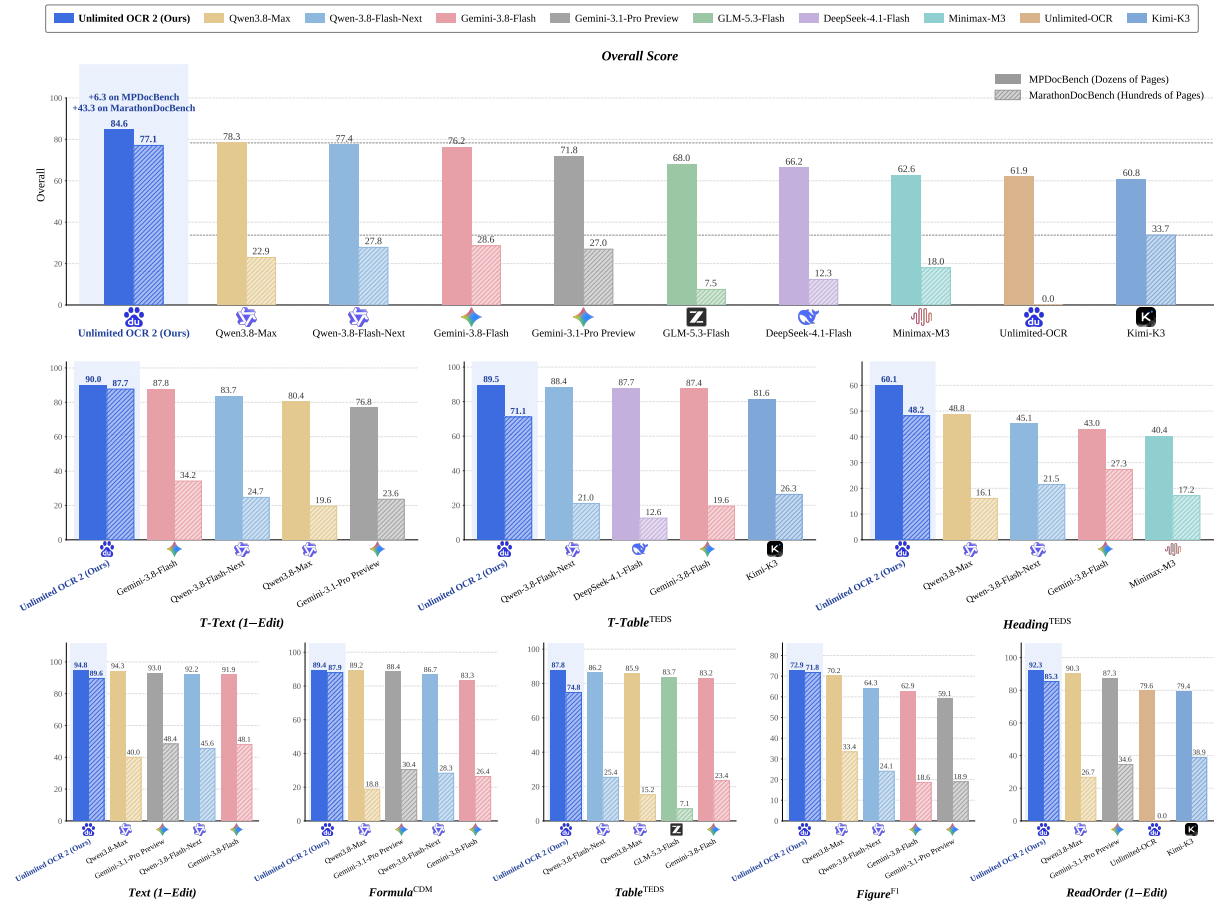


Figure 5 | Performance comparison on multi-page benchmarks. Unlimited OCR 2 (Ours) consistently achieves state-of-the-art performance across all evaluation categories.

Table 1 | Performance comparison on MPDocBench. All models in the table are end-to-end VLM-based architectures.

Method	Overall↑	Text ^{Edit} ↓	T-Text ^{Edit} ↓	Formula ^{CDM} ↑	Table ^{TEDS} ↑	T-Table ^{TEDS} ↑	Figure ^{F1} ↑	ReadOrder ^{Edit} ↓	Heading ^{TEDS} ↑
<i>General VLMs</i>									
Gemini-3.1-Pro Preview [9]	71.79	0.070	0.232	88.37	81.68	61.26	59.08	0.127	26.94
Qwen3.8-Max [25]	78.30	0.057	0.196	89.24	85.89	67.18	70.19	0.097	48.80
Kimi-K3 [24]	60.77	0.162	0.551	41.16	75.45	81.57	40.47	0.206	39.30
Qwen3.6-plus [25]	72.14	0.096	0.245	88.77	82.99	60.50	64.80	0.179	32.06
Qwen3-VL-235B [1]	73.98	0.087	0.187	84.71	81.35	61.25	63.25	0.136	42.34
InternVL-3.5-38B [27]	57.01	0.133	0.515	84.30	69.86	51.69	7.97	0.199	27.01
Minimax-M3 [16]	62.65	0.167	0.422	81.30	77.91	76.51	7.18	0.231	40.39
Gemini-3.8-Flash [10]	76.25	0.081	0.122	83.27	83.15	87.41	62.86	0.294	43.05
Qwen-3.8-Flash-Next [25]	77.36	0.078	0.163	86.70	86.24	88.36	64.31	0.277	45.09
DeepSeek-4.1-Flash [32]	66.25	0.114	0.451	82.12	80.14	87.72	35.55	0.312	32.10
GLM-5.3-Flash [39]	68.02	0.111	0.344	81.86	83.70	70.71	43.09	0.288	39.09
<i>Specialized VLMs (OCR-2.0)</i>									
Unlimited-OCR [37]	61.89	0.167	0.605	73.71	73.51	60.76	53.19	0.204	31.58
<i>Specialized VLMs (OCR-3.0)</i>									
Unlimited-OCR 2 (Ours)	84.59	0.052	0.100	89.39	87.78	89.54	72.91	0.077	60.10

Table 2 | Performance comparison on MarathonDocBench. Unlimited OCR is not reported on MarathonDocBench due to its 40-page input limitation.

Method	Overall↑	Text ^{Edit} ↓	T-Text ^{Edit} ↓	Formula ^{CDM} ↑	Table ^{TEDS} ↑	T-Table ^{TEDS} ↑	Figure ^{F1} ↑	ReadOrder ^{Edit} ↓	Heading ^{TEDS} ↑
<i>General VLMs</i>									
Gemini-3.1-Pro Preview [9]	26.97	0.516	0.764	30.44	28.48	21.71	18.94	0.654	9.68
Qwen3.8-Max [25]	22.93	0.600	0.804	18.82	15.20	13.71	33.39	0.733	16.09
Kimi-K3 [24]	33.74	0.475	0.664	40.50	27.80	26.30	21.38	0.611	28.90
Minimax-M3 [16]	17.97	0.659	0.818	14.20	17.79	14.84	3.67	0.763	17.16
Gemini-3.8-Flash [10]	28.61	0.519	0.658	26.35	23.40	19.56	18.62	0.687	27.34
Qwen-3.8-Flash-Next [25]	27.83	0.544	0.753	28.31	25.38	21.04	24.07	0.679	21.50
DeepSeek-4.1-Flash [32]	12.34	0.740	0.933	10.45	12.74	12.59	6.95	0.843	7.62
GLM-5.3-Flash [39]	7.51	0.860	0.936	4.50	7.12	6.65	7.10	0.923	6.55
<i>Specialized VLMs (OCR-3.0)</i>									
Unlimited-OCR 2 (Ours)	77.06	0.104	0.123	87.93	74.81	71.11	71.79	0.147	48.20

5.2. Main Results

As visually summarized in Figure 6 and further detailed in Table 1 and Table 2, Unlimited OCR 2 takes a significant step toward the OCR 3.0 era, advancing document parsing along three key dimensions. First, it establishes the practical feasibility of hundred-page end-to-end parsing, achieving a score of 77.06 on MarathonDocBench and demonstrating its ability to handle documents at a scale beyond conventional multi-page settings. Second, it maintains robust parsing performance as document length increases. Through end-to-end modeling with SR-SWA, Unlimited OCR 2 filters redundant visual tokens while preserving relevant cross-page information, achieving the highest Overall scores and state-of-the-art performance on both cross-page text and table metrics, T-Text and T-Table, respectively, from dozens of pages on MPDocBench to hundreds of pages on MarathonDocBench. Third, it establishes a new state-of-the-art in long-document parsing, outperforming leading closed-source models and top-tier open-source MLLMs across these benchmarks. Furthermore, Unlimited OCR 2 retains strong single-page recognition capabilities, achieving an Overall score of 94.02% on OmniDocBench in Table 3, confirming that its sparse attention design preserves fine-grained local recognition.

Table 3 | Comparison on OmniDocBench v1.6. All models in the table are end-to-end VLM-based architectures.

Model	Size	Overall $\uparrow$	Text ^{Edit} $\downarrow$	Formula ^{CDM} $\uparrow$	Table ^{TEDs} $\uparrow$	Table ^{TEDS-S} $\uparrow$	ReadOrder ^{Edit} $\downarrow$
HunyuanOCR [23]	1B	89.95	0.088	87.68	91.01	92.23	0.171
DeepSeek-OCR 2 [29]	3B-A0.5B	90.25	0.050	91.84	83.89	87.75	0.144
dots.ocr [19]	3B	90.77	0.048	89.95	87.18	90.58	0.138
FireRed-OCR [31]	2B	93.26	0.037	95.44	88.04	91.06	0.131
ABot-OCR [14]	2B	93.30	0.037	94.86	88.69	91.87	0.137
Logics-Parsing-v2 [4]	4B	93.33	0.041	95.65	88.42	91.98	0.137
Qianfan-OCR [7]	4B	93.90	0.040	95.08	90.53	93.31	0.130
OvisOCR2 [21]	0.8B	96.47	0.027	97.49	94.58	96.98	0.112
TeleOCR [3]	1.2B	96.91	0.027	96.60	96.82	98.18	0.118
Unlimited-OCR [37]	3B-A0.5B	93.92	0.042	95.79	90.16	93.32	0.129
Unlimited-OCR 2	3B-A0.5B	94.02	0.038	96.29	89.60	92.70	0.128

Table 4 | Feasibility evaluation of the Unlimited-Horizon Decoding Paradigm. Sequential chunk processing reduces peak KV cache memory at $B = 1$ and improves end-to-end throughput at $B = 16$, enabling hundred-page parsing under constrained GPU memory.

Method	Peak KV Memory ($B = 1$)	TPS ($B = 16$)
Standard Decoding	47.00 GB	360
Unlimited-Horizon Decoding	1.77 GB	1315

5.3. Scaling with Document Length

5.3.1. Feasibility of Unlimited-horizon Parsing

Benefiting from SR-SWA and the Unlimited-Horizon Decoding Paradigm, Unlimited OCR 2 enables efficient parsing of extremely long documents under constrained GPU memory. We evaluate its feasibility on the hundred-page MarathonDocBench benchmark by measuring peak KV cache memory and end-to-end throughput, as reported in Table 4. Compared with standard decoding, page-level chunking with boundary-page KV cache reuse reduces peak KV cache memory from 47.00 GB to 1.77 GB under single-sequence inference ($B = 1$). Meanwhile, end-to-end throughput increases from 360 to 1,315 TPS under concurrent inference with $B = 16$. These results demonstrate that sequential chunk processing with KV cache reuse substantially reduces memory requirements while improving throughput, making hundred-page document parsing more practical under limited GPU memory.

5.3.2. Length-Invariant Accuracy

To verify whether extending the document horizon compromises parsing fidelity, we evaluate the performance trajectories of Unlimited OCR 2 across continuously increasing document lengths. As illustrated in Figure 6 (a), frontier Vision-Language Models (VLMs) typically experience a noticeable performance degradation as the sequence length grows, primarily caused by attention diffusion and context dilution over ultra-long horizons. In contrast, Unlimited OCR 2 exhibits remarkable length-invariant accuracy. By leveraging SR-SWA to dynamically isolate and retrieve only the most relevant visual tokens, our model maintains consistently high parsing precision from short sequences up to extreme horizons of 300 pages.

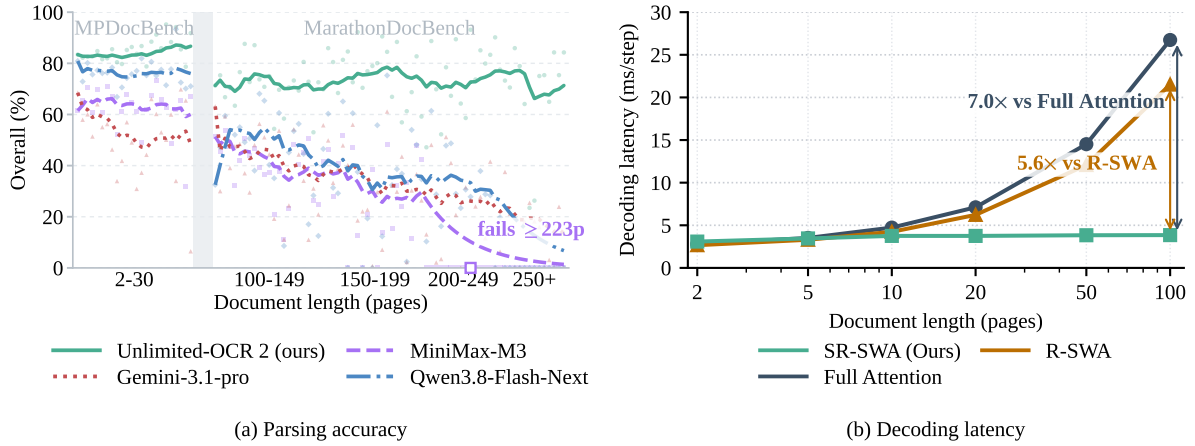


Figure 6 | Performance and efficiency comparison across varying document lengths. Unlimited-OCR 2 demonstrates (a) robust parsing accuracy and (b) near-constant decoding latency (ms/step) as document length scaling.

5.3.3. Constant Per-Step Decoding Latency

Figure 6 (b) examines how decoding efficiency scales with document length. All measurements are conducted in Gundam mode with a batch size of 3 and a fixed generation length of 1,024 tokens per page. As the document length grows, Full Attention incurs rapidly growing per-step decoding latency because each query must attend to a linearly expanding visual context. While R-SWA mitigates this overhead by bounding local window, its remote visual scope still scales with sequence length, leading to noticeable latency accumulation over ultra-long horizons. In contrast, SR-SWA strictly bounds the effective per-query attention scope, achieving near-constant per-step decoding latency from short contexts up to hundred-page documents. Consequently, its latency and throughput advantage over Full Attention and naive windowed baselines becomes increasingly pronounced as document length scales, confirming the computational viability of our paradigm for real-world long-document parsing.

6. Limitation and Future Work

Although Unlimited OCR 2 enables efficient parsing of hundred-page documents, its training data remains limited in diversity, and further data refinement will improve its robustness and generalization. More importantly, the current sparse-attention mechanism does not preserve global document information, limiting its understanding of document-level structures such as heading hierarchies. Future work may introduce architectures such as linear attention to retain global information with constant per-token computational cost, further improving document-level structure understanding.

7. Conclusion

In this technical report, we presented Unlimited OCR 2, a next-generation architecture tailored for unlimited-horizon document parsing. Specifically, Unlimited OCR 2 introduces Sparse Reference Sliding Window Attention (SR-SWA) to break the long-context visual computation bottleneck, guaranteeing constant per-token computational complexity throughout decoding. Additionally, integrated with an optimized page-split decoding paradigm, the model significantly mitigates GPU memory consumption, enabling robust unlimited-horizon inference on resource-constrained hardware. Extensive evaluations and analysis confirm that Unlimited OCR 2 maintains superior parsing accuracy alongside exceptional efficiency across all scales, thereby pioneering the OCR-3.0 era of unlimited-horizon document understanding.

8. Author List

Core Contributors: Shu Wei, Jingjing Wu, Qunyi Xie

Contributors: Lingshu Zhang, Xiangcheng Du, Yannian Fu, Mengjie Wang, Zuyi Wang, Zhichao Liu, Yangliu Xu, Youyang Yin, Huanhuan Liu, YY, Yi Yang

References

- [1] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-vl technical report. arXiv preprint arXiv:2511.21631, 2025.
- [2] I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150, 2020.
- [3] P. Cai, Z. Zou, S. Liu, Y. Wang, J. Tang, K. Yang, M. Tong, M. Jiang, Z. He, and H. Sun. Teleocr: Navigating document parsing across digital and camera-captured documents. arXiv preprint arXiv:2608.12898, 2026. URL <https://arxiv.org/abs/2608.12898>.
- [4] X. Chen, S. Li, X. Zhu, Y. Chen, F. Yang, C. Fang, L. Qu, X. Xu, H. Wei, and M. Wu. Logics-parsing technical report. arXiv preprint arXiv:2509.19760, 2025.
- [5] C. Cui, T. Sun, S. Liang, T. Gao, Z. Zhang, J. Liu, X. Wang, C. Zhou, H. Liu, M. Lin, et al. Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. arXiv preprint arXiv:2510.14528, 2025.
- [6] C. Cui, T. Sun, S. Liang, T. Gao, Z. Zhang, J. Liu, X. Wang, C. Zhou, H. Liu, M. Lin, et al. Paddleocr-vl-1.5: Towards a multi-task 0.9 b vlm for robust in-the-wild document parsing. arXiv preprint arXiv:2601.21957, 2026.
- [7] D. Dong, M. Zheng, D. Xu, C. Luo, B. Zhuang, Y. Li, R. He, H. Wang, W. Zhang, W. Wang, et al. Qianfan-ocr: A unified end-to-end model for document intelligence. arXiv preprint arXiv:2603.13398, 2026.
- [8] S. Duan, Y. Xue, W. Wang, Z. Su, H. Liu, S. Yang, G. Gan, G. Wang, Z. Wang, S. Yan, et al. Glm-ocr technical report. arXiv preprint arXiv:2603.10910, 2026.
- [9] Google DeepMind. Gemini 3.1. <https://deepmind.google/models/gemini/pro/>, 2026.
- [10] Google DeepMind. Gemini 3.8. <https://deepmind.google/models/gemini/flash/>, 2026.
- [11] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023.
- [12] X. Hu, X. Wei, H. Gu, M. Zhang, T. Liang, H. Li, L. Zhu, Y. Wang, S. Han, Y. Bai, et al. Hierarchical sparse attention done right: Toward infinite context modeling. arXiv preprint arXiv:2607.02980, 2026.
- [13] Y. Huang, N. M. Gonçalves, F. Alvetreti, L. Li, X. Han, E. M. Ponti, A. F. Martins, and M. V. Treviso. Dashattention: Differentiable and adaptive sparse hierarchical attention. arXiv preprint arXiv:2605.18753, 2026.

- [14] K. Jiang, R. Gong, X. Cheng, K. Niu, T. Li, and M. Xu. Abot-ocr technical report. arXiv preprint arXiv:2605.27978, 2026.
- [15] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In International conference on machine learning, pages 5156–5165. PMLR, 2020.
- [16] X. Lai, W. Xu, Y. Yang, Q. Chen, Y. Xu, L. Zeng, X. Li, H. Sun, H. Zhu, V. Zhang, et al. Minimax sparse attention. arXiv preprint arXiv:2606.13392, 2026.
- [17] A. Li, B. Gong, B. Yang, B. Shan, C. Liu, C. Zhu, C. Zhang, C. Guo, D. Chen, D. Li, et al. Minimax-01: Scaling foundation models with lightning attention. arXiv preprint arXiv:2501.08313, 2025.
- [18] G. Li, X. Wan, S. Peng, W. Wang, H. Feng, Y. Du, B. Wu, Z. Ruan, Z. Lu, L. Wu, et al. Hunyuanocr-1.5: Making lightweight ocr vlms faster and better. arXiv preprint arXiv:2607.04884, 2026.
- [19] Y. Li, G. Yang, H. Liu, B. Wang, and C. Zhang. dots. ocr: Multilingual document layout parsing in a single vision-language model. arXiv preprint arXiv:2512.02498, 2025.
- [20] Z. Li, Y. Liu, Q. Liu, Z. Ma, Z. Zhang, S. Zhang, Z. Guo, J. Zhang, X. Wang, and X. Bai. MonkeyOCR: Document parsing with a structure-recognition-relation triplet paradigm. arXiv preprint arXiv:2506.05218, 2025.
- [21] S. Lu, Y. Li, Y. Xia, Y. Chen, A.-Y. Ji, J.-P. Jiang, Q.-G. Chen, J. Zhao, E. Lin, H. Li, et al. Ovisocr2 technical report. arXiv preprint arXiv:2607.13639, 2026.
- [22] J. Niu, Z. Liu, Z. Gu, B. Wang, L. Ouyang, Z. Zhao, T. Chu, T. He, F. Wu, Q. Zhang, et al. MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics–Industry Track, 2025.
- [23] H. V. Team, P. Lyu, X. Wan, G. Li, S. Peng, W. Wang, L. Wu, H. Shen, Y. Zhou, C. Tang, et al. Hunyuanocr technical report. arXiv preprint arXiv:2511.19575, 2025.
- [24] K. Team, T. Bai, Y. Bai, Y. Bao, S. Cai, Y. Cao, Y. Charles, H. Che, C. Chen, G. Chen, et al. Kimi k2. 5: Visual agentic intelligence. arXiv preprint arXiv:2602.02276, 2026.
- [25] TongYi. TongYi. <https://www.aliyun.com/product/tongyi>, 2025.
- [26] B. Wang, T. He, L. Ouyang, F. Wu, Z. Zhao, T. Chu, Y. Qu, Z. Jin, W. Zeng, Z. Miao, et al. Mineru2. 5-pro: Pushing the limits of data-centric document parsing at scale. arXiv preprint arXiv:2604.04771, 2026.
- [27] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265, 2025.
- [28] H. Wei, Y. Sun, and Y. Li. Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234, 2025.
- [29] H. Wei, Y. Sun, and Y. Li. Deepseek-ocr 2: Visual causal flow. arXiv preprint arXiv:2601.20552, 2026.

- [30] S. Wei, J. Wu, L. Zhang, Q. Xie, H. Zou, L. Xiang, X. Fan, Y. Xu, M. Lin, X. Ma, et al. Hpd-parsing: Hierarchical parallel document parsing. arXiv preprint arXiv:2607.18839, 2026.
- [31] H. Wu, H. Lou, X. Li, Z. Zhong, Z. Sun, P. Chen, X. Zhou, K. Zuo, Y. Chen, X. Tang, et al. Firered-ocr technical report. arXiv preprint arXiv:2603.01840, 2026.
- [32] A. Xu, B. Li, B. Lin, B. Xue, B. Xian, B. Xu, B. Wu, B. Zhang, B. Deng, C. Yu, et al. Deepseek-v4. 1-flash: Pushing the limits of kv cache compression. arXiv preprint arXiv:2609.19969, 2026.
- [33] A. Xu, B. Lin, B. Xue, B. Wang, B. Xu, B. Wu, B. Zhang, C. Lin, C. Dong, C. Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. arXiv preprint arXiv:2606.19348, 2026.
- [34] B. Xu, Z. Miao, X. Zhou, Y. Lin, Z. Tang, X. Zhao, F. Wu, C. Tan, B. Wang, and C. He. Mineru-popo: Universal post-processing model for structured document parsing. arXiv preprint arXiv:2605.24973, 2026.
- [35] S. Yang, B. Wang, Y. Shen, R. Panda, and Y. Kim. Gated linear attention transformers with hardware-efficient training. arXiv preprint arXiv:2312.06635, 2023.
- [36] S. Yang, B. Wang, Y. Zhang, Y. Shen, and Y. Kim. Parallelizing linear transformers with the delta rule over sequence length. Advances in neural information processing systems, 37: 115491–115522, 2024.
- [37] Y. Yin, H. Liu, Q. Xie, C. Liu, S. Yang, S. Wang, Z. Liu, H. Zou, J. Chen, S. Wei, et al. Unlimited ocr works. arXiv preprint arXiv:2606.23050, 2026.
- [38] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al. Big bird: Transformers for longer sequences. Advances in neural information processing systems, 33:17283–17297, 2020.
- [39] Z.AI. GLM-5.3-Flash. <https://huggingface.co/zai-org/GLM-5.3-Flash>. Model card, accessed 2026-09-25.
- [40] Z. Zhang, H. Liu, S. Liang, Y. Zhang, Y. Xiang, J. Liu, T. Sun, M. Lin, Y. Zhang, C. Zhou, et al. Paddleocr-vl-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. arXiv preprint arXiv:2606.03264, 2026.
- [41] B. Zhou, H. Xing, Y. Chen, J. Xu, Q. Zheng, F. Gao, Z. Yang, S. Bai, M. Yan, J. Ye, et al. Mpdobench-parse: Benchmarking practical multi-page document parsing. arXiv preprint arXiv:2605.22100, 2026.