

MinerU-Diffusion: Rethinking Document OCR as Inverse Rendering via Diffusion Decoding

Hejun Dong^{1*}, Junbo Niu^{2,1*}, Bin Wang^{1†,‡}, Weijun Zeng², Wentao Zhang^{2†},
and Conghui He^{1†}

¹ Shanghai Artificial Intelligence Laboratory, OpenDataLab, Shanghai, China

² Peking University, Beijing, China

heconghui@pjlab.org.cn

Abstract. Optical character recognition (OCR) has evolved from line-level transcription to structured document parsing, requiring models to recover long-form sequences containing layout, tables, and formulas. Despite recent advances in vision-language models, most existing systems rely on autoregressive decoding, which introduces sequential latency and amplifies error propagation in long documents. In this work, we revisit document OCR from an inverse rendering perspective, arguing that left-to-right causal generation is an artifact of serialization rather than an intrinsic property of the task. Motivated by this insight, we propose MinerU-Diffusion, a unified diffusion-based framework that replaces autoregressive sequential decoding with parallel diffusion denoising under visual conditioning. MinerU-Diffusion employs a block-wise diffusion decoder and an uncertainty-driven curriculum learning strategy to enable stable training and efficient long-sequence inference. Extensive experiments demonstrate that MinerU-Diffusion consistently improves robustness while achieving up to 3.2× faster decoding compared to autoregressive baselines. Evaluations on the proposed Semantic Shuffle benchmark further confirm its reduced dependence on linguistic priors and stronger visual OCR capability.

Keywords: Document OCR · Masked Diffusion Language Models · Vision-Language Models

1 Introduction

In recent years, Vision-Language Models (VLMs) [1, 6, 11, 16, 44, 56] have become the dominant paradigm for document optical character recognition (OCR) [7, 17, 18, 29, 47]. These models encode textual images into visual representations and generate structured text through left-to-right autoregressive decoding [41], achieving strong performance across benchmarks. Despite architectural unification and scaling, the decoding process remains strictly sequential. This design introduces efficiency and reliability bottlenecks when parsing long documents

*Equal contribution. †Corresponding author. ‡Project leader.

and complex layouts, particularly in highly structured scenarios such as tables and formulas [32].

From the perspective of task formulation, a high-quality OCR system should primarily depend on authentic visual evidence to perform character-level recognition, rather than relying on semantic completion from a language model. However, autoregressive formulations implicitly cast OCR as language-conditioned reconstruction, where textual outputs are generated under strong linguistic priors. When visual signals are weak or semantic constraints are disrupted, models tend to over-rely on these priors, leading to semantic hallucinations and cumulative errors. Experiments on the Semantic Shuffle benchmark confirm that disrupting semantic structure causes substantial performance degradation in AR-based OCR systems. This fragility stems not merely from data or training strategy, but from the causal factorization inherent in autoregressive decoding.

In contrast, Diffusion Language Models (DLMs) [10, 14, 28, 37, 51], based on discrete diffusion processes, provide a modeling paradigm better aligned with the structural characteristics of document OCR tasks. Masked diffusion models assume conditional independence among tokens given partially observed sequences and visual inputs [52], which is reasonable in OCR scenarios where the mapping between image content and text is largely deterministic with limited semantic ambiguity. This property allows models to exploit the local consistency of visual signals and perform parallel decoding of long textual segments while maintaining global coherence and accuracy. Compared with autoregressive generation for open-ended text tasks [20], diffusion decoding naturally matches the deterministic nature of document understanding [32]. Moreover, DLMs support parallel multi-token updates, significantly improving inference efficiency for long-document parsing. Although current diffusion-based VLMs [5, 21, 49, 52] still encounter issues such as instability in long sequences, repetition, and hallucination in high-resolution settings, these limitations can be progressively alleviated through improved model design and training strategies, making masked diffusion a promising and principled alternative for accurate and efficient OCR modeling.

Motivated by these observations, we formulate document OCR explicitly as an inverse rendering problem under visual conditioning, shifting from autoregressive causal decoding to diffusion-based decoding. We propose MinerU-Diffusion, a unified diffusion-based parsing framework tailored for document OCR. Centered on block-wise diffusion decoding [2, 4, 46] and coupled with an uncertainty-driven curriculum learning strategy, MinerU-Diffusion enables global parallel reconstruction for document OCR. While maintaining high recognition accuracy, MinerU-Diffusion significantly improves long-sequence inference efficiency and effectively mitigates semantic hallucination and cumulative error propagation observed in autoregressive decoding. Extensive experiments demonstrate that MinerU-Diffusion achieves performance on par with state-of-the-art approaches across multiple challenging document parsing benchmarks and semantic perturbation settings, while attaining a more favorable balance among recognition accuracy, robustness, and decoding efficiency.

2 Related Works

Vision-Language Models for Document OCR. Driven by large-scale pre-training, document OCR has evolved from traditional modular pipelines [7, 23, 33, 43] toward end-to-end Vision-Language Models (VLMs) that generate structured text directly from pixels [1, 6, 11, 17, 18, 26, 34, 47]. Representative systems such as MinerU2.5 [29] and PaddleOCR-VL [7] formulate document OCR as sequence generation and rely on autoregressive (AR) decoders to produce text token by token. While this unified paradigm simplifies traditional pipelines and improves cross-domain generalization, it inherits structural limitations from causal left-to-right decoding. Inference latency scales linearly with output length, limiting efficiency in long-document scenarios. Moreover, the strong coupling between generation order and linguistic context encourages reliance on language priors, which may compromise robustness when visual evidence is ambiguous or semantic structure is disrupted. Recent acceleration techniques for autoregressive generation, such as speculative decoding [15], can reduce wall-clock latency by verifying multiple draft tokens in parallel. However, they retain the left-to-right token commitment and causal dependence of AR decoding. Our work instead explores diffusion-based decoding as an alternative formulation for document OCR, enabling iterative refinement under visual conditioning and reducing reliance on linguistic priors.

Masked Diffusion Language Models. Diffusion Language Models (DLMs) [10, 14, 28, 37, 51] offer a non-autoregressive generative framework based on discrete diffusion processes. In masked diffusion, tokens in a clean sequence x_0 are progressively replaced by mask tokens [MASK] under a continuous corruption schedule $t \in [0, 1]$, yielding a noised sequence x_t . The forward process is defined as:

$$q(x_t | x_0) = \prod_{i=1}^n \text{Cat}\left(x_t^i; (1-t)\delta_{x_0^i} + t\delta_{[\text{MASK}]}\right). \quad (1)$$

The corresponding training objective can be derived from maximum likelihood estimation [31, 35, 36, 54], resulting in an evidence lower bound (ELBO) on $\log p_\theta(x)$:

$$\mathcal{J}_{\text{full}}(x_0, Q, \theta) = \int_0^1 \frac{1}{t|x_0|} \mathbb{E}_{q(x_t|x_0)} \left[\sum_{i: x_t^i = [\text{MASK}]} \log p_\theta(x_0^i | x_t, Q) \right] dt, \quad (2)$$

where Q denotes the prompt and $|x_0|$ is the sequence length.

To enhance scalability, block diffusion models [2, 3] introduce block-wise attention mechanisms that balance the optimization stability of autoregressive (AR) training with the parallel sampling efficiency of diffusion-based generation. Their structured attention patterns naturally enable KV-cache reuse, alleviating the inference latency commonly observed in full-attention DLMs [2, 12, 22, 24, 48, 53].

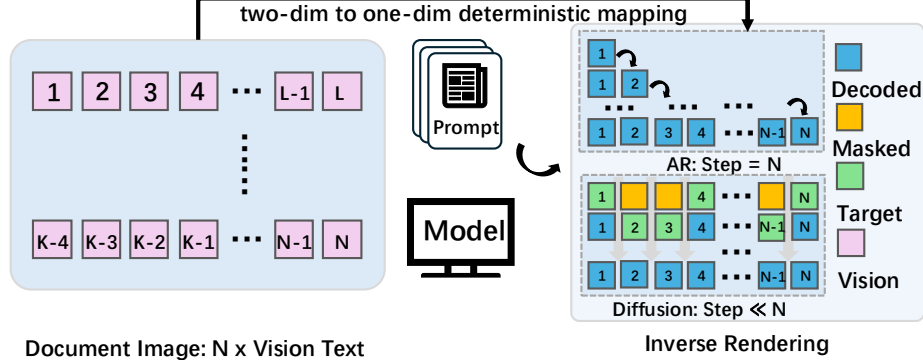


Fig. 1: Overview of the document OCR inverse rendering process via different decoding methods. The model maps a 2D document structure to a 1D token sequence for decoding through autoregressive and diffusion-based methods.

Masked diffusion models are structurally well aligned with the characteristics of document OCR tasks. In document OCR, the target text typically exhibits a near-deterministic mapping to the textual content present in the image, with limited semantic ambiguity. Under this setting, the conditional independence assumption underlying masked diffusion—that each token can be predicted independently given the input and partially observed sequence—becomes considerably more reasonable than in open-ended language generation [25, 28, 55]. This alignment allows the model to decode long-range text spans in parallel without sacrificing consistency.

Therefore, diffusion-based decoding is not merely an efficiency-oriented alternative to autoregressive methods. Instead, it constitutes a modeling paradigm that is inherently better matched to the structural properties of OCR, offering both theoretical justification and practical advantages for large-scale text recognition.

3 Method

3.1 Problem Formulation: Inverse Rendering via Diffusion

We model document OCR [32] as the inverse rendering of a unified structured token sequence:

$$y = (y^{(1)}, \dots, y^{(L)}) \in \mathcal{V}^L, \quad (3)$$

where $\mathcal{V}$ is a shared vocabulary encompassing text symbols, layout markers, table delimiters, and mathematical operators. This unified representation enables the encoding of heterogeneous document elements—such as paragraphs, tables, formulas, and reading order—within a single sequential interface.

Although serialized as a one-dimensional sequence, y corresponds to an underlying two-dimensional document structure. The statistical dependencies between tokens arise primarily from spatial arrangement, layout regularities, and

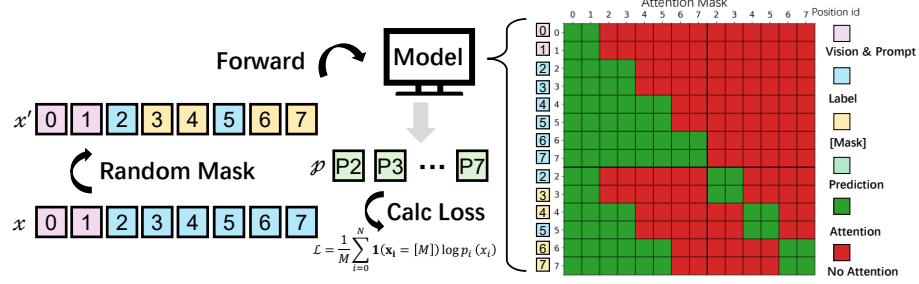


Fig. 2: Training process of MinerU-Diffusion.

formatting constraints, rather than from an intrinsic causal generation order. Therefore, the serialization order should be viewed as an implementation artifact introduced for representation convenience, rather than a fundamental property of the document generation process. In this sense, OCR output is more naturally modeled as a spatially coupled discrete random field, rather than a strictly directional sequence.

Document OCR can be framed as posterior inference over a latent structured token sequence, where the input document x serves as partial and noisy evidence, constraining both token identities and spatial positioning. Traditional OCR systems typically parameterize the posterior through autoregressive decompositions [7, 17, 29, 47], which impose a fixed causal order and limit the ability to model document structure globally. In contrast, as illustrated in Fig. 1, diffusion-based decoding methods [2, 28] introduce a discrete diffusion process that avoids a fixed causal ordering, enabling global iterative refinement under visual conditioning that naturally aligns with the structural properties of OCR tasks. The conditional independence assumption inherent in masked diffusion models [28], i.e., that each token can be independently predicted given the input and partially observed sequence, becomes particularly reasonable in OCR, where the target text has a one-to-one correspondence with the text in the image. Through multiple denoising iterations, diffusion models jointly update all tokens across the entire sequence, circumventing the single-pass update limitation of autoregressive decoding and providing a more structurally aligned approximation to the posterior distribution $p(y | x)$.

3.2 MinerU-Diffusion : Unified Diffusion Architecture for OCR

A straightforward implementation of discrete diffusion for OCR is to apply a full-attention dLM [28, 51, 53] over the entire token sequence at each denoising step. However, when scaling to long structured documents, such a design suffers from fundamental structural and computational limitations [48].

Full self-attention incurs quadratic complexity $O(L^2)$ with respect to sequence length L , making it computationally expensive for long structured documents with thousands of tokens. Moreover, full-attention diffusion operates globally, which introduces positional instability as early denoising errors can

propagate across the sequence. Unlike autoregressive decoding [1], global diffusion lacks structural anchoring and is prone to long-range drift. Additionally, document structures exhibit strong locality, with high intra-region consistency and weak long-range dependencies. Full-attention unnecessarily couples independent regions, conflicting with the spatially constrained posterior structure of document OCR.

These observations suggest that purely full-attention diffusion is not structurally aligned with OCR. To address these limitations, we introduce MinerU-Diffusion, a block-attention [2,4] dVLM that incorporates structural locality into posterior refinement, as illustrated in Fig. 2. The output sequence is partitioned into B contiguous blocks:

$$y = (y^{(1)}, \dots, y^{(B)}), \quad y^{(b)} \in \mathcal{V}^{L'}, \quad L = BL'. \quad (4)$$

Rather than modeling the entire sequence as a monolithic denoising problem, we factorize the conditional posterior:

$$p_\theta(y \mid x) = \prod_{b=1}^B p_\theta(y^{(b)} \mid y^{(<b)}, x), \quad (5)$$

where $y^{(<b)}$ denotes all preceding blocks. Within each block, diffusion operates locally:

$$p_\theta(y_{t-1}^{(b)} \mid y_t^{(b)}, y^{(<b)}, x). \quad (6)$$

This hybrid factorization introduces coarse-grained autoregressive structure across blocks and parallel diffusion refinement within blocks. Block boundaries serve as structural anchors, preventing long-range alignment drift, while preserving parallel efficiency inside each block.

At each denoising step, a structured attention mask is applied. Tokens can attend fully to tokens within the same block and causally to tokens in preceding blocks, but cannot attend to tokens in future blocks. Formally, the attention mask $\mathcal{M}_{ij}$ is defined as:

$$\mathcal{M}_{ij} = \begin{cases} 1, & \text{if } b(i) = b(j), \\ 1, & \text{if } b(j) < b(i), \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

where $b(i)$ denotes the block index of token i . This structured masking reduces unnecessary global coupling, stabilizes positional alignment, and ensures that decoding errors remain locally bounded.

MinerU-Diffusion conditions the diffusion process on native-scale visual features [8,29,30,44], ensuring that posterior refinement remains grounded in visual evidence. Compared to full-attention diffusion, block-attention reduces complexity from $O(L^2)$ to $O(BL'^2)$. The causal structure across blocks enables efficient KV-caching during inference, while maintaining parallel decoding within each block. Overall, MinerU-Diffusion provides a structurally grounded and computationally scalable diffusion architecture tailored for document OCR.

3.3 Two-Stage Curriculum Learning with Uncertainty-Driven Refinement

To fully leverage large-scale heterogeneous data and alleviate the performance bottleneck caused by noisy labels, we propose a two-stage curriculum learning framework to train MinerU-Diffusion. Compared with autoregressive (AR) models that generate tokens in a fixed order, diffusion models can decode tokens in any order, introducing more intricate inter-token dependencies that make training less stable and more sensitive to noise, often requiring larger datasets and more carefully tuned training strategies [27, 37]. Furthermore, random masking typically achieves lower data utilization efficiency than AR models at the same data scale. This occurs because AR models make predictions based on complete prefix information at each step, while random masking disperses the supervisory signal, reducing the density of effective conditional information. To address this, we divide the dataset $\mathcal{D}$ into two subsets: $\mathcal{D}_{\text{base}}$, which is easier to train, and $\mathcal{D}_{\text{hard}}$, which is more challenging. We first train on $\mathcal{D}_{\text{base}}$ to establish foundational structural understanding, then fine-tune on $\mathcal{D}_{\text{hard}}$ to improve the model’s robustness to noisy labels and boundary precision. This framework mitigates the optimization complexity and data-inefficiency issues faced by diffusion models in any-order modeling.

Stage I: Diversity-Driven Foundational Learning In the first stage, our goal is to establish robust foundational representations and general parsing abilities across multiple document understanding tasks. To this end, we construct a large-scale, diverse, and balanced dataset $\mathcal{D}_{\text{base}}$ through data curation and automated annotation refinement.

This dataset satisfies:

$$\mathcal{D}_{\text{base}} \sim p_{\text{div}}(x), \quad (8)$$

where $p_{\text{div}}(x)$ denotes a high-entropy data distribution covering diverse layouts, languages, document types, and visual styles.

This stage emphasizes broad visual-semantic alignment, stable feature learning, and robust cross-domain generalization. Although $\mathcal{D}_{\text{base}}$ contains moderate annotation noise, its large scale and diversity enable effective representation learning. From an optimization perspective, training on $\mathcal{D}_{\text{base}}$ yields a relatively smooth loss landscape, facilitating stable convergence.

Stage II: Uncertainty-Driven Boundary Refinement

Hard Case Mining via Inference Consistency After Stage I convergence, the model acquires strong general capabilities. However, performance is constrained by noisy supervision and limited exposure to complex edge cases. To overcome this limitation, we introduce an uncertainty-driven curriculum refinement stage.

For each unlabeled or weakly labeled sample x , we perform T stochastic inference passes:

$$\{\hat{y}^{(t)}\}_{t=1}^T = \{f_\theta(x; \xi_t)\}_{t=1}^T, \quad (9)$$

where ξ_t represents stochastic factors such as sampling temperature or dropout.

We define a task-specific consistency metric $S(\cdot, \cdot)$: (1) PageIoU for layout, (2) CDM for formula, (3) TEDS for tables.

The mean consistency score is computed as:

$$C(x) = \frac{2}{T(T-1)} \sum_{i < j} S(\hat{y}^{(i)}, \hat{y}^{(j)}). \quad (10)$$

Low values of $C(x)$ indicate high prediction uncertainty.

We select hard samples as:

$$\mathcal{D}_{hard} = \{x \mid C(x) < \tau_{train}\}, \quad (11)$$

where τ_{train} is a task-dependent curriculum threshold used only for hard-sample mining during Stage II training.

Expert Refinement and Dataset Construction Samples in $\mathcal{D}_{hard}$ are processed through an AI-assisted human annotation pipeline to produce high-precision labels:

$$\tilde{\mathcal{D}}_{hard} = \{x_i, \tilde{y}_i\}_{i=1}^M. \quad (12)$$

The final fine-tuning dataset is constructed as:

$$\mathcal{D}_{SFT} = \tilde{\mathcal{D}}_{hard} \cup \alpha \mathcal{D}_{rand}, \quad (13)$$

where $\mathcal{D}_{rand}$ is a randomly sampled subset of $\mathcal{D}_{base}$, and α controls the regularization ratio.

The Stage II optimization objective is defined as:

$$\mathcal{L}_{hard}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}_{SFT}} [w(x) \ell(f_\theta(x), y)], \quad (14)$$

where the sample weight is:

$$w(x) = 1 + \beta(1 - C(x)), \quad (15)$$

with β controlling the emphasis on hard samples.

This adaptive weighting further encourages the model to focus on decision-boundary regions. The progressive curriculum mitigates optimization instability and performance ceilings caused by diffusion’s any-order modeling by organizing data from broad to difficult, enabling MinerU-Diffusion to overcome annotation noise and long-tail complexity for superior real-world document OCR performance.

4 Experiments

In this section, we present a comprehensive quantitative evaluation of MinerU-Diffusion to demonstrate its effectiveness in document OCR tasks. More ablation studies and qualitative examples are included in the supplementary material.

4.1 Experimental Setups

Data All meta training data are derived from the MinerU2.5 dataset [29], with a total volume of approximately 7.5M samples. The dataset primarily focuses on Chinese and English document parsing tasks. Therefore, no dedicated evaluation was conducted for low-resource languages.

Models and Optimization Our experiments adopt a block-wise attention dVLM architecture. Specifically, we employ SDAR-1.7B-Chat-b32 [4] with a block size of 32. The remaining components follow the MinerU2.5 architecture, except that MRoPE [38, 44] is removed. We first fine-tune MinerU-Diffusion on the LLaVA-NeXT dataset [20] for visual question answering (VQA) tasks. Based on this initialization, we further conduct specialized training for optical character recognition (OCR). Additional optimization details are provided in the supplementary material.

Evaluation All experiments are conducted with a block size of 32 and a dynamic decoding strategy. The inference-time decoding threshold is set to $\tau_{\text{decode}} = 0.95$, with $\text{top-}k = 0$, temperature = 1.0, and $\text{top-}p = 1.0$.

For full document parsing and layout analysis, we evaluate our models on OmniDocBench v1.5 [32]. Table recognition is assessed using CC-OCR [50] and OCRBench v2 [9], while formula recognition is evaluated on UniMER-Test [42]. More detailed evaluation settings and prompt templates are provided in the supplementary material.

4.2 Full-Document Parsing Task Results

As shown in Tab. 1, under the fully automatic setting without GT Layout, MinerU-Diffusion achieves an Overall score of 88.94, outperforming most AR-based models and demonstrating strong end-to-end parsing capability without relying on oracle layout information. When evaluated with GT Layout, MinerU-Diffusion further improves to 93.37 Overall, reaching a score that is close to top-tier AR-based systems and indicating high competitiveness in overall parsing performance. Meanwhile, the sizable gap between the two settings suggests that layout understanding remains a key bottleneck: MinerU-Diffusion can leverage accurate layout signals effectively, and its remaining weaknesses are primarily attributed to layout analysis, leaving clear room for improvement in layout prediction to further close the gap in fully automatic parsing.

Type	Methods	Params.	Overall $\uparrow$	Text $\downarrow$	Formula $\uparrow$	Table _{TEDS} $\uparrow$	Table _{TEDS_s} $\uparrow$	Reading Order $\downarrow$	GT Layout
Pipeline	MinerU2-pipeline [43]	-	75.51	0.209	76.55	70.90	79.11	0.225	$\times$
	PP-StructureV3 [7]	-	86.73	0.073	85.79	81.68	89.48	0.073	$\times$
AR	Qwen2.5-VL-72B [40]	72B	87.02	0.094	88.27	82.15	86.22	0.102	$\times$
	Gemini-2.5 Pro [6]	-	88.03	0.075	85.82	85.71	90.29	0.097	$\times$
	MinerU2-VLM [43]	0.9B	85.56	0.078	80.95	83.54	87.66	0.086	$\times$
	Nanonets-OCR-s [26]	3B	85.59	0.093	85.90	80.14	85.57	0.108	$\times$
	MonkeyOCR-pro-1.2B [18]	1.9B	86.96	0.084	85.02	84.24	89.02	0.130	$\times$
	MonkeyOCR-pro-3B [18]	3.7B	88.85	0.075	87.25	86.78	90.63	0.128	$\times$
	dots.ocr [17]	3B	88.41	0.048	83.22	86.78	90.62	0.053	$\times$
	MinerU2.5 [29]	1.2B	90.67	0.047	88.46	88.22	92.38	0.044	$\times$
	PaddleOCR-VL [7]	0.9B	92.56	0.035	91.43	89.76	93.52	0.043	$\times$
dLM	MinerU-Diffusion	2.5B	88.94	0.061	86.41	86.50	90.29	0.059	$\times$
AR	MinerU2.5 [29]	1.2B	93.44	0.025	91.98	90.84	95.10	-	$\checkmark$
	PaddleOCR-VL [7]	0.9B	93.91	0.021	92.13	91.70	95.42	-	$\checkmark$
dLM	MinerU-Diffusion	2.5B	93.37	0.028	91.92	91.00	94.86	-	$\checkmark$

Table 1: Comprehensive evaluation of document parsing on OmniDocBench v1.5. $\uparrow$ denotes higher is better, $\downarrow$ denotes lower is better.

Type	Method	CC-OCR		OCRBench v2		UniMER-Test			
		TEDS $\uparrow$	TEDS-S $\uparrow$	TEDS $\uparrow$	TEDS-S $\uparrow$	CPE $\uparrow$	HWE $\uparrow$	SCE $\uparrow$	SPE $\uparrow$
AR	InternVL3.5-241B [45]	62.87	69.52	79.50	85.81	91.7	93.2	95.1	97.8
	Qwen2.5-VL-72B [40]	81.22	86.48	81.33	86.58	88.9	91.8	95.5	96.2
	GPT-4o [13]	66.98	79.04	70.51	79.55	82.7	85.9	87.8	96.7
	dots.ocr [17]	75.42	81.65	82.04	86.27	86.8	90.5	94.7	97.5
	MinerU2.5 [29]	79.76	85.16	87.13	90.62	96.6	94.4	96.4	98.4
dLM	MinerU-Diffusion	73.77	82.06	81.18	88.66	91.6	91.6	92.0	96.8

Table 2: Comprehensive recognition results on CC-OCR [50], OCRBench v2 [9], and UniMER-Test [42]. $\uparrow$ denotes higher is better.

4.3 Element-Specific Parsing Task Results

Table Recognition We evaluate MinerU-Diffusion’s table recognition performance on the table subsets of CC-OCR [50] and OCRBench v2 [9], as shown in Tab. 2. MinerU-Diffusion is competitive with other AR-based models on both table subsets. It achieves 81.18/88.66 (TEDS/TEDS-S) on OCRBench v2 and 73.77/82.06 on CC-OCR, matching or surpassing several strong AR baselines in the same evaluation setting and remaining close to the top group. In particular, the strong TEDS-S scores indicate that MinerU-Diffusion can reliably preserve table structure during diffusive decoding, demonstrating solid end-to-end table parsing capability and overall competitiveness among AR approaches.

Formula Recognition We evaluate formula recognition on UniMER-Test [42], with results summarized in Tab. 2. MinerU-Diffusion achieves 91.6/91.6/92.0/96.8 on CPE/HWE/SCE/SPE, consistently demonstrating strong performance across complex, handwritten, and printed settings. Compared with other autoregressive (AR) baselines, MinerU-Diffusion is clearly competitive, substantially outperforming general-purpose models such as GPT-4o [13] and reaching similar ranges to strong AR systems on multiple subsets. The remaining gap to the best specialized pipeline (e.g., MinerU2.5 [29]) is most visible on harder categories like

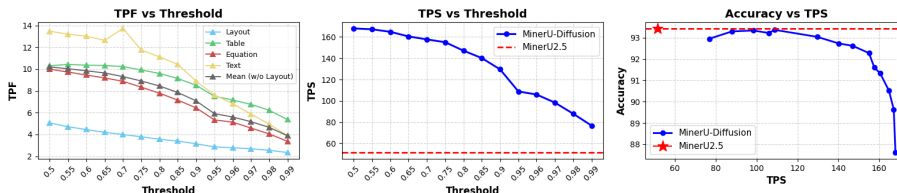


Fig. 3: Threshold sensitivity analysis of TPF, TPS, and accuracy. TPF denotes tokens per forward, and TPS refers to throughput measured on an NVIDIA H200 GPU with a batch size of 1.

CPE, suggesting future improvements should focus on more precise symbol-level modeling and structure-aware decoding for complex printed expressions.

4.4 Ablation Study

Decoding Confidence Threshold vs. Decoding Parallelism We systematically analyze the impact of the inference-time decoding threshold τ_{decode} on decoding parallelism and inference efficiency. This threshold is distinct from the training-time hard-sample mining threshold τ_{train} introduced in Sec. 3.3. As shown in Fig. 3, as τ_{decode} increases from 0.5 to 0.99, both Tokens Per Forward (TPF) and Throughput Per Second (TPS) decrease monotonically: lower thresholds relax confirmation constraints, enabling greater parallel decoding and higher throughput, whereas thresholds above 0.95 cause decoding to approach near token-by-token generation, leading to a marked efficiency drop. The decoding threshold thus serves as a controllable system-level knob that enables a continuous and predictable trade-off between efficiency and conservativeness. Compared with MinerU2.5 at matched accuracy (fixed at ~ 52 TPS), our method achieves 108.9 TPS at 93%+ accuracy ($\tau_{\text{decode}} = 0.95$), corresponding to a $2.1\times$ speedup, and reaches 164.8 TPS while maintaining over 90% accuracy ($\tau_{\text{decode}} = 0.6$), yielding a peak acceleration of approximately $3.2\times$, thereby unlocking substantial parallelism without sacrificing quality.

Decoding Parallelism vs Accuracy This experiment analyzes the impact of decoding parallelism on model accuracy by varying τ_{decode} from 0.5 to 0.99 under the *w/ GT Layout* setting. As shown in Fig. 4, as τ_{decode} increases, performance first improves, then saturates, and slightly declines: the Overall score rises from 87.61 ($\tau_{\text{decode}} = 0.5$) to 93.37 ($\tau_{\text{decode}} = 0.95$), indicating that moderately reducing parallelism mitigates low-confidence error propagation and enhances accuracy. However, further increasing τ_{decode} to 0.99 leads to a slight drop, suggesting that overly conservative decoding weakens the benefits of parallel information interaction. At the task level, higher decoding thresholds consistently improve structural consistency, benefiting text, table, and formula recognition (e.g., Table TEDS improves from 81.10 to 90.99), as reduced parallelism alleviates structural error accumulation. Combined with throughput analysis in

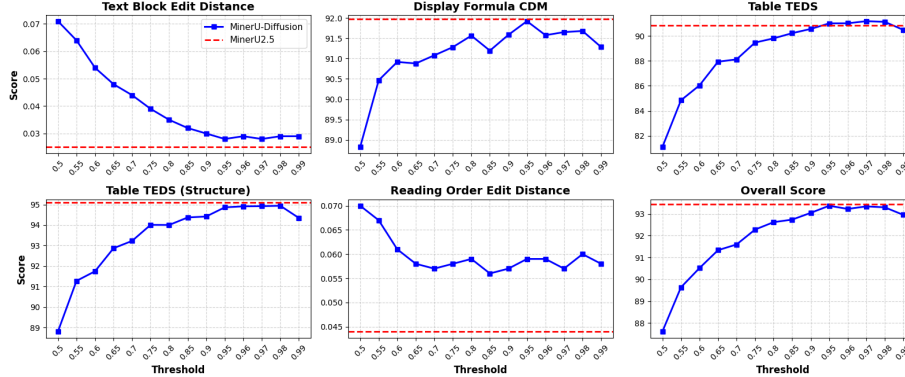


Fig. 4: Visualizing the accuracy-throughput trade-off of different models across different OCR tasks under the *w/ GT Layout* setting.

Sec. 4.4, these results reveal a clear accuracy–efficiency trade-off: lower τ_{decode} favors speed, while higher τ_{decode} improves stability at the cost of efficiency. Notably, at $\tau_{\text{decode}} = 0.95$, MinerU-Diffusion achieves an Overall score of 93.37, comparable to MinerU2.5 while retaining superior decoding efficiency, making it an effective balance point between accuracy and efficiency.

Type Method	Inference		Score (w/ GT Layout)					Score (w/o GT Layout)	
	TPF	TPS	Overall↑	Text↓	Formula↑	TableTEDS↑	TableTEDS-S↑	Overall↑	Reading Order↓
dLM	Static (step=6)	5.33 91.56	88.31	0.064	85.52	85.81	91.20	82.73	0.077
	Dynamic ($\tau_{\text{decode}} = 0.97$)	5.18 98.32	93.34	0.028	91.65	91.17	94.92	89.04	0.057
	Static (step=32)	1.00 21.86	93.02	0.030	91.10	90.95	94.73	89.76	0.060
AR	MinerU2.5 [29]	1.00 51.46	93.44	0.025	91.98	90.84	95.10	90.67	0.044
	PaddleOCR-VL [7]	1.00 40.77	93.91	0.021	92.13	91.70	95.42	92.79	0.043

Table 3: Comparison of different inference strategies and model types in terms of efficiency (TPF, TPS) and document parsing performance.

Decoding Strategy Comparisons We compare static decoding with dynamic scheduling on OmniDocBench v1.5 [32], as summarized in Tab. 3. Increasing the total decoding steps from 6 to 32 improves overall performance from 88.31 to 93.02, indicating that decoding fewer tokens per step reduces error accumulation, albeit at the cost of requiring more rounds of inference and thus lower throughput (21.86 TPS). In contrast, dynamic scheduling outperforms static step=6 in both overall and sub-task metrics (93.34 vs. 88.31) while achieving higher throughput (98.32 vs. 91.56 TPS), indicating that confidence-based adaptive token selection effectively reduces premature low-confidence commitments and balances efficiency with accuracy. As further shown in Tab. 3, compared with strong AR baselines such as MinerU2.5 [29] and PaddleOCR-VL [7], our approach achieves competitive parsing performance while maintaining high throughput, highlighting the potential of parallel generation strategies for complex document understanding.

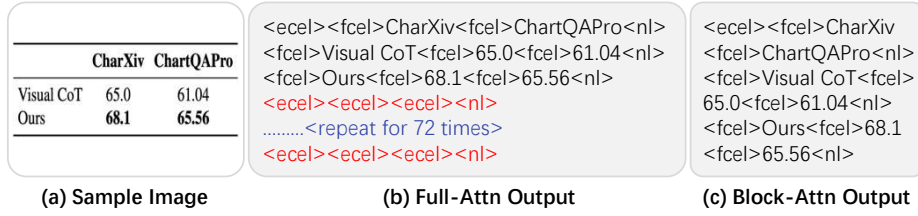


Fig. 5: Full-Attn vs. Block-Attn outputs on a table OCR example.

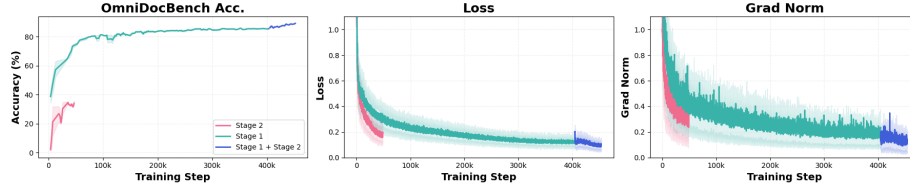
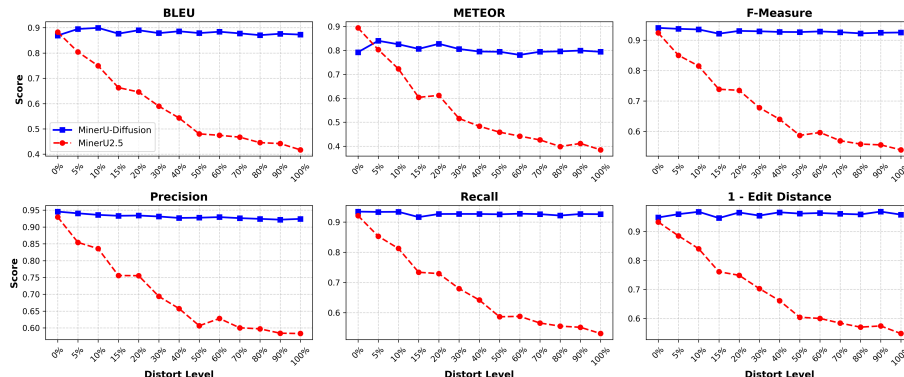


Fig. 6: Comparison of training dynamics across different curriculum strategies. The two-stage framework achieves smoother optimization and higher final accuracy compared to single-stage baselines.

Full-Attn vs Block-Attn To analyze the impact of attention mechanisms on efficiency and stability, we conducted ablation experiments on Full-Attn and Block-Attn Diffusion using LLaDA-MoE-7B-A1B-Instruct [55], trained on an OCR task. The main difference lies in the amount of attention computed per forward step. Full-Attn computes roughly $L \times L$ attention, leading to quadratic growth in memory and computational costs with longer sequences, which slows inference and can induce repetitive decoding in OCR, as illustrated in Fig. 5 where the Full-Attn output shows obvious repeated patterns on a table sample. In contrast, Block-Attn computes roughly $B \times L$ attention per forward step, using a “self-regressive across blocks, diffusion within blocks” design, achieving near-linear scalability while maintaining bidirectional context modeling. Experimental results show Block-Attn reduces memory and latency significantly and mitigates repetition in OCR generation, as shown by the cleaner Block-Attn output in Fig. 5, with better scalability and stability. A more detailed discussion of these ablation results is deferred to the supplementary material.

Two-Stage Curriculum Learning To validate the proposed two-stage curriculum learning framework, we conduct systematic ablations on OmniDocBench v1.5, comparing Stage 1 only, Stage 2 only, and the full two-stage strategy. As shown in Fig. 6, training dynamics (Accuracy, Loss, and Grad Norm) reveal that Stage 2 alone exhibits rapid early loss reduction but low accuracy and significant gradient oscillation, indicating high-variance optimization caused by emphasizing high-uncertainty samples without stable representation initialization. In contrast, Stage 1 achieves smooth and stable convergence with accuracy improving to around 85%, yet performance saturates due to limited boundary refinement. The full two-stage strategy integrates both advantages: Stage 1 establishes low-variance global representations under diverse data, while Stage 2

Setting	Score (w/ GT Layout)					Score (w/o GT Layout)	
	Overall↑	Text↓	Formula↑	TableTEDS ↑	TableTEDS-S ↑	Overall↑	Reading Order↓
Stage 1	92.89	0.032	91.58	90.28	94.93	86.13	0.056
Stage 2	89.33	0.038	88.50	83.29	88.93	35.71	0.472
Stage 1 + Stage 2	93.37	0.028	91.92	91.00	94.86	88.94	0.059

Table 4: Comparison of different training stages under GT and non-GT layout settings.**Fig. 7:** Semantic Shuffle benchmark results across distortion levels.

performs uncertainty-driven boundary refinement, enabling continued improvement beyond the Stage 1 plateau and achieving over 89% accuracy. As shown in Tab. 4, under the challenging *w/o GT Layout* setting, Stage 2 alone largely fails (35.712), whereas the full framework reaches 88.937, confirming that the progressive curriculum effectively mitigates optimization instability and performance ceilings induced by diffusion’s any-order modeling.

4.5 Semantic Shuffle Analysis

Most modern VLM-based OCR systems [7, 17, 29, 47] follow a “vision encoder + autoregressive (AR) language decoder” pipeline: the image is converted into visual tokens, and an LM-style decoder [39] generates the transcription token by token. Although this paradigm achieves strong OCR scores, it is often unclear whether these gains come from faithful visual reading or from the decoder’s ability to “fill in” plausible text using linguistic priors and global context.

To better isolate true visual decoding ability, we introduce **Semantic Shuffle**, a benchmark that removes semantic coherence while keeping the visual presentation comparable. Starting from 112 English document images in the FOX dataset [19], we shuffle a controlled proportion of words and re-render the resulting text into new document images with consistent formatting. Pseudocode for the construction pipeline is provided in the supplementary material.

As shown in Fig. 7, experiments reveal a clear contrast between decoding paradigms. As the distortion level increases, the performance of AR decoders degrades sharply and consistently, indicating a heavy reliance on linguistic plausibility. In contrast, a diffusion-style decoder remains robust, maintaining nearly

constant performance across distortion levels. These results suggest that AR decoders tend to reconstruct text using global language regularities when semantics are disrupted, whereas diffusion-based decoding aligns more directly with the visual signal under semantically invalid inputs.

5 Conclusion

In this paper, we propose MinerU-Diffusion, a diffusion-based framework for document OCR that replaces autoregressive decoding with block-level parallel diffusion decoding and confidence-guided scheduling to improve efficiency and scalability. A two-stage curriculum learning strategy stabilizes training and enhances boundary precision and robustness. Experiments across document, table, formula, and Semantic Shuffle benchmarks show competitive performance against AR baselines and stronger resilience to disrupted semantics, highlighting diffusion-based parallel decoding as a promising alternative for document OCR. We hope our work inspires future research on efficient and reliable diffusion-based decoding for document OCR.

Acknowledgements

This project was supported by Shanghai Artificial Intelligence Laboratory.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Arriola, M., Gokaslan, A., Chiu, J.T., Yang, Z., Qi, Z., Han, J., Sahoo, S.S., Kuleshov, V.: Block diffusion: Interpolating between autoregressive and diffusion language models. arXiv preprint arXiv:2503.09573 (2025)
3. Cheng, S., Bian, Y., Liu, D., Jiang, Y., Liu, Y., Zhang, L., Wang, W., Guo, Q., Chen, K., Qi*, B., Zhou, B.: Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation (2025), <https://github.com/JetAstra/SDAR>
4. Cheng, S., Bian, Y., Liu, D., Zhang, L., Yao, Q., Tian, Z., Wang, W., Guo, Q., Chen, K., Qi, B., et al.: SDAR: A synergistic diffusion-autoregression paradigm for scalable sequence generation. arXiv preprint arXiv:2510.06303 (2025)
5. Cheng, S., Jiang, Y., Zhou, Z., Liu, D., Tao, W., Zhang, L., Qi, B., Zhou, B.: Sdar-vl: Stable and efficient block-wise diffusion for vision-language understanding. arXiv preprint arXiv:2512.14068 (2025)
6. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
7. Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., et al.: PaddleOCR 3.0 technical report. arXiv preprint arXiv:2507.05595 (2025)

8. Dehghani, M., Mustafa, B., Djolonga, J., Heek, J., Minderer, M., Caron, M., Steiner, A., Puigcerver, J., Geirhos, R., Alabdulmohsin, I.M., et al.: Patch n’pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems* **36**, 2252–2274 (2023)
9. Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., et al.: Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321* (2024)
10. Google DeepMind: Gemini diffusion. <https://blog.google/technology/google-deepmind/gemini-diffusion/> (2025), accessed: 2025-05-20
11. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1.5-VL technical report. *arXiv preprint arXiv:2505.07062* (2025)
12. Hu, Z., Meng, J., Akhauri, Y., Abdelfattah, M.S., Seo, J.s., Zhang, Z., Gupta, U.: Accelerating diffusion language model inference via efficient kv caching and guided diffusion. *arXiv preprint arXiv:2505.21467* (2025)
13. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: GPT-4o system card. *arXiv preprint arXiv:2410.21276* (2024)
14. Labs, I., Khanna, S., Kharbanda, S., Li, S., Varma, H., Wang, E., Birnbaum, S., Luo, Z., Miraoui, Y., Palrecha, A., et al.: Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298* (2025)
15. Leviathan, Y., Kalman, M., Matias, Y.: Fast inference from transformers via speculative decoding. In: *International Conference on Machine Learning*. pp. 19274–19286 (2023)
16. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024)
17. Li, Y., Yang, G., Liu, H., Wang, B., Zhang, C.: dots.ocr: Multilingual document layout parsing in a single vision-language model. *arXiv preprint arXiv:2512.02498* (2025)
18. Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., Bai, X.: Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218* (2025)
19. Liu, C., Wei, H., Chen, J., Kong, L., Ge, Z., Zhu, Z., Zhao, L., Sun, J., Han, C., Zhang, X.: Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295* (2024)
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023)
21. Liu, Y., Hallyburton, S., Kim, J., Lin, Y., Li, Y., Wang, Q., Ye, H., Sun, J., Pajic, M., Chen, Y., et al.: Llavida: A large language vision driving assistant for explicit reasoning and enhanced trajectory planning. *arXiv preprint arXiv:2512.18211* (2025)
22. Liu, Z., Yang, Y., Zhang, Y., Chen, J., Zou, C., Wei, Q., Wang, S., Zhang, L.: dllm-cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295* (2025)
23. Livathinos, N., Auer, C., Lysak, M., Nassar, A., Dolfi, M., Vagenas, P., Ramis, C.B., Omenetti, M., Dinkla, K., Kim, Y., et al.: Docling: An efficient open-source toolkit for ai-driven document conversion. *arXiv preprint arXiv:2501.17887* (2025)
24. Ma, X., Yu, R., Fang, G., Wang, X.: dkv-cache: The cache for diffusion language models. *arXiv preprint arXiv:2505.15781* (2025)

25. Man, S., Ganz, R., Ronen, R., Tsiper, S., Mazor, S., Nayman, N.: DODO: Discrete OCR diffusion models. arXiv preprint arXiv:2602.16872 (2026)
26. Mandalm, S.: Nanonets-ocr-s. <https://nanonets.com/research/nanonets-ocr-s/> (2025), accessed:2025-09-25
27. Ni, J., Liu, Q., Dou, L., Du, C., Wang, Z., Yan, H., Pang, T., Shieh, M.Q.: Diffusion language models are super data learners. arXiv preprint arXiv:2511.03276 (2025)
28. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.R., Li, C.: Large language diffusion models. arXiv preprint arXiv:2502.09992 (2025)
29. Niu, J., Liu, Z., Gu, Z., Wang, B., Ouyang, L., Zhao, Z., Chu, T., He, T., Wu, F., Zhang, Q., et al.: MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. arXiv preprint arXiv:2509.22186 (2025)
30. Niu, J., Zheng, Y., Miao, Z., Dong, H., Ge, C., Liang, H., Lu, M., Zeng, B., Zheng, Q., He, C., et al.: Native visual understanding: Resolving resolution dilemmas in vision-language models. arXiv preprint arXiv:2506.12776 (2025)
31. Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., Li, C.: Your absorbing discrete diffusion secretly models the conditional distributions of clean data. arXiv preprint arXiv:2406.03736 (2024)
32. Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., et al.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24838–24848 (2025)
33. Paruchuri, V.: Marker. <https://github.com/datalab-to/marker> (2025), accessed:2025-09-25
34. Poznanski, J., Rangapur, A., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Wilhelm, C., Lo, K., Soldaini, L.: olmocr: Unlocking trillions of tokens in pdfs with vision language models. arXiv preprint arXiv:2502.18443 (2025)
35. Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., Kuleshov, V.: Simple and effective masked diffusion language models. Advances in Neural Information Processing Systems **37**, 130136–130184 (2024)
36. Shi, J., Han, K., Wang, Z., Doucet, A., Titsias, M.: Simplified and generalized masked diffusion for discrete data. Advances in neural information processing systems **37**, 103131–103167 (2024)
37. Song, Y., Zhang, Z., Luo, C., Gao, P., Xia, F., Luo, H., Li, Z., Yang, Y., Yu, H., Qu, X., et al.: Seed diffusion: A large-scale diffusion language model with high-speed inference. arXiv preprint arXiv:2508.02193 (2025)
38. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
39. Team, Q.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)
40. Team, Q.: Qwen2.5-vl (January 2025), <https://qwenlm.github.io/blog/qwen2.5-vl/>
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
42. Wang, B., Gu, Z., Liang, G., Xu, C., Zhang, B., Shi, B., He, C.: Unimernet: A universal network for real-world mathematical expression recognition. arXiv preprint arXiv:2404.15254 (2024)
43. Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et al.: MinerU: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839 (2024)

44. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
45. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
46. Wang, Y., Yang, L., Li, B., Tian, Y., Shen, K., Wang, M.: Revolutionizing reinforcement learning framework for diffusion large language models. arXiv preprint arXiv:2509.06949 (2025)
47. Wei, H., Sun, Y., Li, Y.: DeepSeek-OCR: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
48. Wu, C., Zhang, H., Xue, S., Liu, Z., Diao, S., Zhu, L., Luo, P., Han, S., Xie, E.: Fast-dLLM: Training-free acceleration of diffusion LLM by enabling KV cache and parallel decoding. arXiv preprint arXiv:2505.22618 (2025)
49. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models. arXiv preprint arXiv:2505.15809 (2025)
50. Yang, Z., Tang, J., Li, Z., Wang, P., Wan, J., Zhong, H., Liu, X., Yang, M., Wang, P., Bai, S., et al.: Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21744–21754 (2025)
51. Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., Kong, L.: Dream 7b (2025), <https://hkunlp.github.io/blog/2025/dream>
52. You, Z., Nie, S., Zhang, X., Hu, J., Zhou, J., Lu, Z., Wen, J.R., Li, C.: LLaDA-V: Large language diffusion models with visual instruction tuning. arXiv preprint arXiv:2505.16933 (2025)
53. Yu, R., Ma, X., Wang, X.: Dimple: Discrete diffusion multimodal large language model with parallel decoding. arXiv preprint arXiv:2505.16990 (2025)
54. Zheng, K., Chen, Y., Mao, H., Liu, M.Y., Zhu, J., Zhang, Q.: Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. arXiv preprint arXiv:2409.02908 (2024)
55. Zhu, F., You, Z., Xing, Y., Huang, Z., Liu, L., Zhuang, Y., Lu, G., Wang, K., Wang, X., Wei, L., Guo, H., Hu, J., Ye, W., Chen, T., Li, C., Tang, C., Feng, H., Hu, J., Zhou, J., Zhang, X., Lan, Z., Zhao, J., Zheng, D., Li, C., Li, J., Wen, J.R.: Lladamoe: A sparse moe diffusion language model. arXiv preprint arXiv:2509.24389 (2025)
56. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)