DODO: Discrete OCR Diffusion Models

Sean Man[*1] **Gilad Deutch**[*2] **Roy Ganz**[2] **Roi Ronen**[2] **Shahar Tsiper**[2] **Shai Mazor**[2] **Niv Nayman**[2]

Abstract

Optical Character Recognition (OCR) is a fundamental task for digitizing information, serving as a critical bridge between visual data and textual understanding. While modern Vision-Language Models (VLM) have achieved high accuracy in this domain, they predominantly rely on autoregressive decoding, which becomes computationally expensive and slow for long documents as it requires a sequential forward pass for every generated token. We identify a key opportunity to overcome this bottleneck: unlike open-ended generation, OCR is a highly deterministic task where the visual input strictly dictates a unique output sequence, theoretically enabling efficient, parallel decoding via diffusion models. However, we show that existing masked diffusion models fail to harness this potential; those introduce structural instabilities that are benign in flexible tasks, like captioning, but catastrophic for the rigid, exact-match requirements of OCR. To bridge this gap, we introduce **DODO**, the first VLM to utilize block discrete diffusion and unlock its speedup potential for OCR. By decomposing generation into blocks, DODO mitigates the synchronization errors of global diffusion. Empirically, our method achieves near state-of-the-art accuracy while enabling up to $5\times$ faster inference compared to autoregressive baselines.

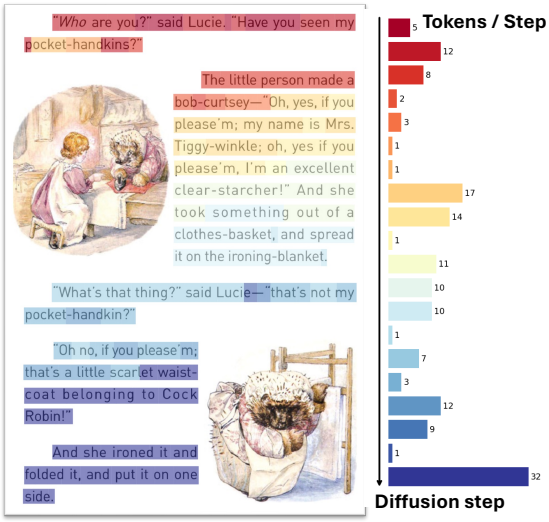

Figure 1. **DODO: High-throughput parallel generation.** Unlike autoregressive models constrained to a strict left-to-right sequence, DODO generates text across the entire canvas simultaneously (with same color) based on visual confidence. In this example, it resolves 148 tokens in just 20 forward passes ($\approx$ 7 tokens/step on average). Simultaneously generated tokens are adaptively distributed across steps.

1. Introduction

Optical character recognition (OCR) is a core component of modern document understanding systems, enabling the extraction of structured text from images such as scanned documents, forms, and natural scenes. Vision–language models are increasingly used for large-scale document parsing and multimodal reasoning (Alayrac et al., 2022; Li et al., 2022; 2023; Ganz et al., 2023; Dai et al., 2023; Wu et al., 2024; Li et al., 2024; Ganz et al., 2024; Chen et al., 2025; Bai et al., 2025b;a; Liu et al., 2025). However, the high computational cost and latency of these architectures have re-established OCR transcription as a critical bottleneck where both accuracy and inference efficiency are essential (Blecher et al., 2023; Wei et al., 2024; Abramovich et al., 2024; Nacson et al., 2025; Wei et al., 2025b).

Crucially, OCR differs fundamentally from semantically flexible tasks like image captioning (Sidorov et al., 2020; Chen et al., 2023; Lin et al., 2014) or visual question answering (VQA) (Antol et al., 2015; Singh et al., 2019; Mathew et al., 2021; Yue et al., 2024) as it is semantically rigid. Conditioned on the image, the posterior distribution is effectively unimodal, meaning the visual input strictly dictates a single valid sequence. This determinism exposes a critical inefficiency in standard Autoregressive (AR) models: they generate text sequentially, creating a significant latency bottleneck for long document sequences. Conversely, this characteristic makes OCR uniquely suited for Masked Dif-

[*]These authors contributed equally to this work. Work conducted during an internship at Amazon

[1]Technion - Israel Institute of Technology, Haifa, Israel. [2]Amazon Web Services. Correspondence to: Niv Nayman <nivay@amazon.com>.

fusion Models (MDMs) (Sahoo et al., 2024b). Because the output allows for little ambiguity, OCR satisfies the MDM assumption of *conditional independence*—the premise that tokens can be predicted independently given the input (Azangulov et al., 2025). Theoretically, this allows the model to resolve large spans of text simultaneously, similar to how traditional OCR pipelines (Wang et al., 2021; Ronen et al., 2022; Aberdam et al., 2023) recognize isolated regions in parallel without the risk of incoherence.

However, realizing this potential in practice reveals a structural paradox: the same rigidity that enables parallelization also makes OCR particularly sensitive to the instabilities of global decoding. While standard MDMs (Yu et al., 2025; Li et al., 2025a; You et al., 2025) can generate tokens in parallel, they introduce non-causal structural uncertainties, specifically regarding sequence length and absolute positional alignment. In flexible tasks like captioning, such errors are recoverable: the model can navigate a wide space of valid outputs to resolve a misalignment dynamically. OCR allows no such flexibility. Because the target is a single, immutable sequence, structural errors become irrecoverable; the model cannot "rewrite" the text to compensate for incorrect length estimates or token placement. Consequently, these rigidities force the model to either truncate valid text or hallucinate padding, leading to fractured, colliding outputs that fundamentally undermine the efficacy of standard masked diffusion for transcription.

To resolve this paradox, we propose DODO (Discrete OCR Diffusion Models), the first Vision–Language Model to adapt block discrete diffusion for document transcription. Unlike standard global diffusion, DODO decomposes the monolithic generation task into a sequence of causally anchored blocks. This structural change directly addresses the rigidities of OCR: by bounding the inference horizon and conditioning on a committed prefix, we eliminate the risk of long-range alignment drift and enable dynamic length adaptation without requiring a perfect global estimate. Crucially, we leverage the high-confidence nature of OCR combined with block-causal attention and KV-caching (Li et al., 2024) to achieve both causal consistency and fast inference. Moreover, unlike autoregressive models, the diffusion framework uniquely enables bidirectional attention across blocks, which we show pushes the utilizable block size from 32 to 256 tokens, further improving accuracy.

Empirically, DODO achieves transcription accuracy competitive with state-of-the-art autoregressive models while outperforming the equivalent autoregressive baseline in throughput. These results validate our hypothesis: OCR is indeed a regime where the conditional independence assumption holds, but it requires the structural safety rails of block diffusion to be realized in practice. Appropriately structured, DODO recovers the correctness of AR models while unlocking the efficiency benefits that motivate diffusion-based decoding in the first place.

Our contributions are summarized as follows:

- We identify a structural incompatibility between standard masked diffusion and the rigid requirements of OCR, explaining why positional and length errors that are benign in flexible tasks prove catastrophic for OCR.
- We introduce **DODO**, the first VLM to utilize block discrete diffusion. By decomposing generation into sequentially conditioned blocks, DODO enforces local alignment and enables dynamic length adaptation, resolving the rigidities of global diffusion.
- We demonstrate that DODO matches the accuracy of state-of-the-art autoregressive baselines while enabling up to $5\times$ faster inference, validating the potential of parallel decoding for dense text recognition.

2. Related Work

Specialized OCR and Document Understanding. Modern OCR systems leverage vision-language models for end-to-end document understanding. MonkeyOCR (Li et al., 2025c) introduces a multi-stage pipeline with detection, recognition, and reading order prediction. MinerU (Wang et al., 2024) and commercial systems like dots.ocr (Li et al., 2025b), DeepSeek-OCR (Wei et al., 2025a), Mistral-OCR (mis, 2025) achieve strong performance through careful engineering and large-scale training. These methods universally employ autoregressive decoding. This is the first successful attempt to achieve competitive OCR performance by MDMs.

Discrete Diffusion Models. Discrete diffusion models learn to reverse a corruption process over discrete tokens. D3PM (Austin et al., 2021) introduced structured transition matrices, while MDLM (Sahoo et al., 2024a) simplified training through masked diffusion with a tighter evidence lower bounds (ELBO). Recent work has scaled these models to language modeling (Zhou et al., 2024), though a gap remains compared to autoregressive models on perplexity benchmarks. This work narrows down the performance gap of MDMs with their plain autoregressive counterparts for the studied task.

Block Diffusion. Block Diffusion (BD3-LM) (Arriola et al., 2025) bridges autoregressive and diffusion models by generating blocks of tokens autoregressively during inference only, with each block decoded via masked diffusion. This enables KV-caching across blocks while maintaining parallel decoding within blocks. Prior work uses small block sizes (4–32 tokens) (Wu et al., 2025b) to minimize the

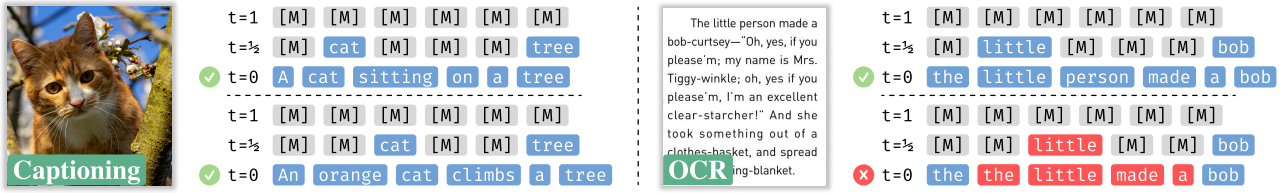

Figure 2. **Semantically flexible *vs*. semantically rigid vision–language tasks.** *Left:* Image captioning admits multiple, semantically equivalent descriptions of the same image. Different decoding trajectories can converge to distinct but equally valid captions, and lexical or structural variations are naturally absorbed. *Right:* OCR requires a single, exact transcription determined by the image. Even minimal local deviations, such as an incorrect token choice or boundary, render the output incorrect. As a result, conditioned on the image, OCR exhibits extremely low output variability, which makes it a natural candidate for parallel decoding, but also a demanding setting in which errors cannot be compensated by alternative phrasings or later corrections.

performance gap with autoregressive models on language modeling. (Wu et al., 2025a) further aligns the attention masks to be block causal during training. We implement this approach for VLMs.

Multimodal Diffusion Models. Dimple (Yu et al., 2025) extends discrete diffusion to vision-language tasks, training on the LLaVA recipe. However, it shows limited gains over autoregressive baselines and does not evaluate on OCR benchmarks. LaViDa (Li et al., 2025a) and LLaDA-V (You et al., 2025) explore diffusion for VQA but struggle with OCR tasks requiring precise text reproduction. This work is the first to successfully apply discrete diffusion for OCR.

3. Preliminaries

Notation. Let $\mathcal{V}$ be a vocabulary of size V, and let $\texttt{[M]} \notin \mathcal{V}$ denote a dedicated `MASK` token. We write $\tilde{\mathcal{V}} = \mathcal{V} \cup \{\texttt{[M]}\}$, and represent tokens either as categorical indices $v \in \tilde{\mathcal{V}}$ or as one-hot vectors $\mathbf{e}_v \in \{0,1\}^{|\tilde{\mathcal{V}}|}$. We use $\mathrm{Cat}(\cdot;\pi)$ for a categorical distribution with probabilities π.

3.1. OCR as Conditional Sequence Modeling

We formulate OCR as a conditional generation task where the goal is to map a document image I to a target sequence of discrete tokens $x^{1:L}$. This target is defined as $x^{1:L} = \tau(s(I))$, where $s(\cdot)$ represents a fixed serialization scheme (e.g., plain text, LaTeX, HTML) and $\tau(\cdot)$ is a tokenizer that maps the resulting string to vocabulary indices. A vision–language model (VLM) estimates the conditional distribution $p_\theta(x^{1:L} \mid I, c)$ given the image I and optional text context c. Autoregressive (AR) VLM decoding admits standard left-to-right factorization

$$\log p_\theta(x^{1:L} \mid I, c) = \sum_{\ell=1}^{L} \log p_\theta(x^\ell \mid x^{<\ell}, I, c), \quad (1)$$

with $x^{<l}$ the prefix tokens at step l out of L sequential steps.

3.2. Masked Diffusion Models (MDMs)

Forward (Masking) Process. MDMs define a coordinate-independent corruption process that replaces tokens with [M] according to a noise level $t \in [0,1]$. Writing x_0 for clean data and x_t for its noisy version, a common conditional probability distribution for the token masking is

$$q_{t|0}(x_t \mid x_0) = \prod_{i=1}^{L} \mathrm{Cat}\Big(x_t^i;\ \alpha_t\, \mathbf{e}_{x_0^i} + (1-\alpha_t)\, \mathbf{e}_{\texttt{[M]}}\Big), \quad (2)$$

where α_t is strictly decreasing with $\alpha_0 = 1$ and $\alpha_1 = 0$.

Training. MDMs train a denoiser to predict the original input tokens from partially masked ones. In continuous time, an ELBO-derived objective can be written (Sahoo et al., 2024a; Shi et al., 2024) as a weighted masked cross-entropy

$$\mathcal{L} = \mathbb{E}_{t,x_0,x_t|x_0}\left[\frac{\alpha_t'}{1-\alpha_t} \sum_{i:x_t^i=\texttt{[M]}} -\log p_\theta(x_0^i|x_t,t)\right], \quad (3)$$

where $\alpha_t' = \frac{d\alpha_t}{dt}$, and $p_\theta(\cdot \mid x_t, t)$ is frequently implemented without an explicit time embedding since x_t reveals t through its mask rate.

Sampling. MDM sampling starts from the fully masked sequence $x_1 = (\texttt{[M]}, \ldots, \texttt{[M]})$ and iterates noise levels $1 = t_K > \cdots > t_0 = 0$. Given an estimate of the marginal distribution of each token from the denoiser, a common and convenient decomposition (Sahoo et al., 2024a) of a single reverse step $t_{k+1} \to t_k$ proceeds by *first* choosing which masked positions to reveal via a selection rule (e.g., randomly (Sahoo et al., 2024a), top-k (Zheng et al., 2023), confidence-thresholding (Yu et al., 2025), or deterministically (Luxembourg et al., 2025)), and *second*, sampling token values for those positions from the denoiser's predicted distribution $x_{t_k}^i \sim p_\theta(x_{t_k}^i \mid x_{t_{k+1}})$.

This decomposition has two consequences described next.

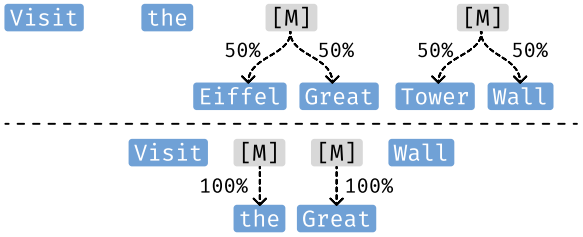

Figure 3. **Conditional independence assumption.** *(Top)* In open-ended tasks, ambiguity between valid options risks sampling incoherent mixtures. *(Bottom)* In OCR, the strong visual signal resolves this ambiguity, enabling conflict-free parallel decoding.

Conditional Independence Assumption. At each sampling step, the decoded tokens are sampled independently from one another. This may lead to incorrect results if the decoded tokens in fact are conditionally dependent given the context, as illustrated in Figure 3. On the other hand, when this assumption holds, there is a large parallelization potential to leverage.

Carry-Over Unmasking. We follow the common practice (Wu et al., 2025b;a; You et al., 2025; Li et al., 2025a; Yu et al., 2025) and only allow the sampling of masked tokens at each step. While this formulation is shown by (Sahoo et al., 2024a) to be the key for the derivation of the discrete ELBO equivalent loss in Equation (1), this means one cannot revise previously decoded tokens. While this has a lesser effect on generative tasks where multiple responses are acceptable, it might be detrimental for tasks where only one response is correct, as illustrated in Figure 2.

4. Method

We analyze OCR through the lens of MDM training and sampling (Section 3.2), focusing on the two sampling assumptions made explicit there: *conditional independence* of tokens sampled within a step, and *carry-over unmasking*, where revealed tokens are not revised.

We first argue that OCR is especially compatible with the conditional-independence assumption. We then show why this potential is difficult to realize with vanilla MDM inference, as early mistakes persist and this requires caution when decoding many tokens in parallel. Finally, we argue that block discrete diffusion mitigates these failure modes, while retaining high parallelism and enjoying KV-Caching.

4.1. Parallel Decoding Potential

OCR typically yields long sequences, making standard autoregressive decoding a significant latency bottleneck, as it requires L sequential forward passes to decode L tokens. Masked diffusion models offer a compelling alternative by enabling parallel token generation; however, their effectiveness hinges on the validity of the *conditional independence assumption*.

We argue that OCR is uniquely suited for this parallel paradigm. Unlike semantically flexible vision-language tasks, the posterior distribution $p(x^{1:L}|I, c)$ for document transcription is highly peaked, often approaching a Dirac delta function around a single ground-truth sequence. In this low-entropy regime, the strong conditioning on the visual input effectively decouples token predictions, allowing the joint probability of masked tokens to be factorized:

$$p(x_{t_k}^{1:L} \mid x_{t_{k+1}}^{1:L}, I, c) \approx \prod_{\ell=1}^{L} p(x_{t_k}^{\ell} \mid x_{t_{k+1}}^{1:L}, I, c). \quad (4)$$

This independence allows the decoding of large spans of tokens simultaneously, as shown in Figure 1.

4.2. Brittleness of Parallel Decoding in OCR

Standard masked diffusion models operate on a fixed-length canvas and rely on *carry-over unmasking* (Section 3.2), treating tokens revealed in early steps as immutable context. While this constraint is manageable for semantically flexible tasks like captioning, where the model can paraphrase content or alter semantics to fit the available space, it presents a fundamental challenge for OCR. Document transcription is a rigid, zero-tolerance task: the ground-truth text consists of a specific set of characters in an immutable order. This lack of flexibility exposes two critical failure modes for parallel decoding:

Length Mismatch. Because the true sequence length is unknown at inference time, the decoding canvas size is effectively an estimate. In generative tasks like captioning, this is rarely fatal; the model can simply generate a valid shorter or longer description to match the canvas. In OCR, however, this mismatch creates a structural vulnerability. If the initial L is incorrect, or the model predicts an end-of-sequence `[EOS]` token prematurely, the model is forced to either truncate valid text (if the effective length is too short) or hallucinate (if too long) to satisfy the imposed constraint.

Positional Anchoring. Even with a valid length estimate, parallel decoding binds content to absolute positional indices in a non-causal manner. This introduces a critical synchronization risk: the model may predict a segment at an incorrect offset, such as placing a table header 50 tokens too early or too late. Because *carry-over unmasking* prevents the revision of revealed tokens, this error is locked in place. Unlike autoregressive decoding, which inherently align content to its history without looking ahead, diffusion parallel decoding is bound to errors made ahead. Consequently, the subsequent text cannot shift to accommodate the offset. Due to the unimodal nature of OCR, the model cannot simply

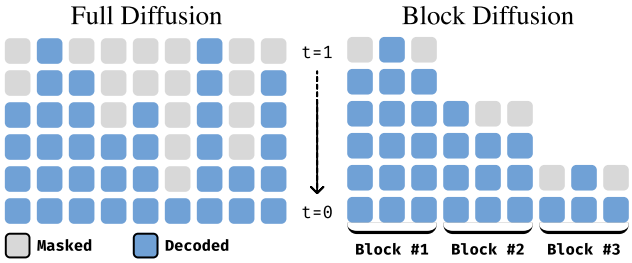

Figure 4. **Full *vs*. block diffusion**. In standard full diffusion (*left*), MDM sampling is applied globally to the entire sequence. In contrast, block diffusion (*right*) restricts parallel sampling to discrete windows, processing blocks sequentially from left to right.

paraphrase or "tweak" the surrounding text to bridge this misalignment, leading to fractured outputs where disjoint segments collide – a fundamental challenge that limits the efficacy of purely parallel OCR.

4.3. Block Diffusion as a Structural Remedy

Block discrete diffusion can mitigate these failure modes by replacing a single length-L denoising problem with a sequence of bounded-span problems, conditioned of a prefix composed on previous blocks decoded sequentially.

Block discrete diffusion models (Arriola et al., 2025) combine AR structure at a coarse granularity with diffusion within blocks. Partition the sequence into B contiguous blocks of length L' (so $L = BL'$) and write $x^{(b)}$ for block b and $x^{(<b)}$ for the prefix blocks. The model factorizes as

$$p_\theta(x^{1:L}|I,c) = \prod_{b=1}^{B} p_\theta\big(x^{(b)} \mid x^{(<b)}, I, c\big), \qquad (5)$$

where each conditional distribution $p_\theta(x^{(b)} \mid x^{(<b)}, I, c)$ is implemented via a masked diffusion sampling over the L' tokens of the block, conditioned on prefix states. This formulation narrows the performance gap between MDMs and AR models and enables significant computational efficiency by allowing the KV-cache of committed prefix blocks $x^{(<b)}$ to be reused rather than recomputed. An illustrative example of the difference between full and block diffusion models sampling is given in Figure 4. This factorization anchors indices and conventions at block boundaries, reduces length sensitivity, and retains parallel token updates within each block while enabling variable-length generation via block-level stopping.

Previous unimodal MDMs (Wu et al., 2025b;a) use small block sizes (4–32 tokens) to minimize the performance gap with autoregressive models on text-only tasks. By utilizing the properties of the OCR task and equipping the model with bidirectional attention across blocks, we are able to scale the block size from 32 to 256 tokens. Finally, this work is the first to apply block-causal masking both during training and inference in the multimodal VLM settings.

5. Experiments

5.1. DODO

We instantiate our proposed framework as **DODO** (Discrete OCR Diffusion Models), built upon the Qwen2.5-VL-3B architecture (Bai et al., 2025c). We train DODO on `olmOCR-mix-1025` (Poznanski et al., 2024), a large-scale OCR dataset comprising approximately 270K document-text pairs derived from PDFs. The dataset covers diverse document types, including academic papers, books, reports, and web pages, with text extracted using a combination of PDF parsing and OCR pipelines. While block-based diffusion has previously been explored for text-only models (Arriola et al., 2025; Wu et al., 2025a), we are the first to adapt this paradigm to the multimodal domain.

DODO employs a *block-causal* attention mask: tokens within the active block $x^{(b)}$ attend bidirectionally to one another and to all previous blocks $x^{(<b)}$, but attention from previous blocks to the current one is masked. This strict causality ensures that the representations of the committed prefix remain fixed, enabling the use of an exact Key-Value (KV) cache. Consequently, only the active block needs to be computed at each step, resulting in significant speedups. Furthermore, unlike autoregressive models, the diffusion framework also permits full bidirectional attention across blocks, which we show in Section 6.1 can further improve accuracy by enabling larger effective block sizes.

5.2. Implementation Details

We fine-tune the model with a maximum sequence length of 8192 tokens to accommodate dense document texts, retaining the default image preprocessing pipeline. A visualization of the attention structure is provided in Appendix A.

Training Configuration. We train DODO for a total of $200,000$ steps on a node of $8\times$NVIDIA A100 (40GB) GPUs, using a global batch size of 8. Optimization is performed using AdamW with a peak learning rate of 5×10^{-6} and a weight decay of 0.01. We employ a Warmup-Steady-Decay (WSD) learning rate scheduler, consisting of a linear warmup for $5,000$ steps, a constant steady phase, and a linear cooldown to zero over the final $20,000$ steps. To ensure training stability and efficiency, we utilize `bfloat16` precision throughout the process.

Diffusion Setup. Following recent advances in discrete diffusion (Wu et al., 2025a; Li et al., 2025a), we utilize complementary masking during training. For timestep sampling, we employ stratified uniform scheduling to ensure balanced coverage of the noise levels during training.

5.3. Evaluation Setup

Benchmarks. We evaluate our models on two distinct benchmarks. First, we use `OmniDocBench` (Ouyang et al., 2025), a comprehensive testbed for layout-sensitive transcription containing 290 English documents across 9 diverse types (e.g., academic papers, financial reports), annotated with structured ground truth for text, tables, and formulas. Second, we evaluate on `Fox-Page-EN` (Liu et al., 2024), a dataset of 112 document pages focused exclusively on pure text without figures or tables. This combination allows us to assess performance on both complex, multimodal document layouts and standard dense text transcription.

Metrics. We assess performance using two metrics. First, we report accuracy using the Normalized Edit Distance (NED) between the predicted and ground-truth text, with lower scores indicating higher fidelity. Second, to quantify inference efficiency, we measure throughput as the number of generated Tokens Per Second (TPS) for each model.

Baselines. We compare DODO against three model categories: (1) specialized OCR models, including dots.ocr, DeepSeek-OCR, MinerU 2.0 VLM, MonkeyOCR, MistralOCR, olmOCR, Nanonets-OCR-s, and SmolDocling (Li et al., 2025b; Wei et al., 2025a; Wang et al., 2024; Li et al., 2025c; mis, 2025; Poznanski et al., 2025; Mandal et al., 2025; Nassar et al., 2025); (2) general-purpose autoregressive VLMs, specifically the Qwen2.5-VL family (Bai et al., 2025c), which serves as our backbone; and (3) diffusion VLMs, represented by Dimple, LaViDa, and LLaDA-V (Yu et al., 2025; Li et al., 2025a; You et al., 2025).

5.4. Main Results

Table 1 presents the OCR performance on OmniDocBench (breakdown by document length and type is provided in Appendix D) and Fox-Pages. Compared to prior diffusion-based VLMs, DODO demonstrates a substantial performance leap. Standard diffusion models such as Dimple, LaViDa, and LLaDA-V struggle significantly with dense document transcription, incurring high error rates (> 0.5 NED on OmniDocBench). In contrast, DODO achieves an NED of 0.069. While DODO benefits from specialized OCR training, we demonstrate in our ablations (Section 6.1) that data distribution alone does not explain this gap; even when trained on identical OCR data, standard full-sequence diffusion fails to converge on dense text. This confirms that the improvement is primarily structural: DODO's block-wise constraints effectively mitigate the alignment failures that plague global diffusion models.

Second, DODO proves highly competitive against strong autoregressive and specialized baselines. On the layout-intensive OmniDocBench, it surpasses its own autoregressive backbone, the Qwen2.5-VL family, across all model scales. Furthermore, DODO outperforms various specialized models and achieves near-parity with robust engines such as MonkeyOCR (Li et al., 2025c) and Mistral OCR (mis, 2025). These results establish that discrete diffusion is a viable, high-performance alternative to the dominant autoregressive paradigm in the challenging domain of dense text recognition.

Table 1. **DODO results.** OCR performance on the English subset of OmniDocBench and Fox-Pages. Bold numbers indicate the best value across each model type (MDM and AR).

Method	Size	Edit Distance ↓ OmniDoc	Fox
Specialized OCR			
dots.ocr	3B	**0.032**	0.034
DeepSeek-OCR	3.4B	0.049	0.100
MinerU 2.0 VLM	0.9B	0.045	-
MonkeyOCR-pro	3B	0.058	0.084
Mistral OCR	-	0.072	**0.013**
olmOCR	7B	0.097	0.023
Nanonets-OCR-s	3B	0.134	-
SmolDocling	256M	0.262	0.022
Autoregressive VLMs			
Qwen 2.5 VL	72B	0.092	0.039
Qwen 2.5 VL	7B	0.135	0.025
Qwen 2.5 VL	3B	0.184	0.051
Diffusion VLMs			
Dimple	7B	0.856	0.932
LaViDa-L	8B	0.994	-
LLaDA-V	7B	0.524	0.336
DODO	3B	**0.069**	**0.038**

5.5. Throughput Analysis

Figure 5 illustrates the inference throughput across different model architectures. For qualitative visualizations of the parallel decoding process on dense documents, we refer

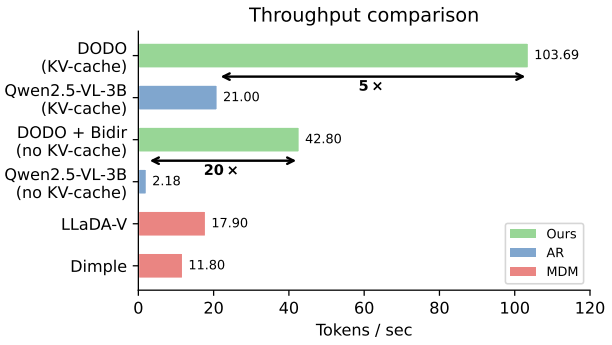

Figure 5. **Inference throughput comparison.** DODO leverages parallel decoding, block-causal attention and KV-caching to achieve ≈ 104 tokens/sec, a $5\times$ speedup over the autoregressive baseline.

Table 2. **Impact of block structure.** Vanilla MDM fails even with Oracle length; block-wise *training* is essential.

Train Config	Max Len	Inf. Block	Edit Dist.	TPS
Vanilla	Oracle	-	0.100	8.77
Vanilla	8192	-	0.834	3.19
Vanilla	8192	32	0.951	7.47
Block	8192	32	0.067	29.5
Blk Causal	8192	32	**0.069**	**103.7**

readers to Figure 1 and Section E. Two key trends emerge from this comparison.

First, DODO achieves $\approx$ 105 tokens per second, a $5\times$ speedup over the autoregressive Qwen 2.5 VL baseline. For analysis purposes alone, we isolate the contribution of parallel decoding to the speedup, by running both the AR baseline and DODO+Bidir without KV-caching. The speedup of parallel decoding alone is evaluated as $20\times$ (42.8 vs. 2.2 TPS), demonstrating that a significant part of the throughput gain stems from generating multiple tokens per step. By enabling the use of an exact KV-cache for completed blocks, DODO eliminates the redundant re-computation of prefix representations, maximizing deployment efficiency. While AR generation requires L sequential steps, DODO decouples latency from sequence length, allowing parallel predictions of multiple tokens to amortize the cost of heavier forward passes.

Second, DODO achieves significantly higher throughput than competing diffusion VLMs. Prior methods rely on full-sequence attention from the outset, incurring the maximal computational cost proportional to the total length L at every denoising step. In contrast, DODO combines parallel decoding within blocks with constant cost per block by caching the prefix, allowing it to process generation much faster than standard diffusion baselines, which are immediately burdened by the full sequence complexity. For more comparisons see Appendix C.

6. Ablation and Empirical Analysis

We validate the structural design of DODO by isolating the impact of block-based training and analyzing the interaction between block size and caching strategies. For an ablation of sampling schedules, we refer readers to Section B.1.

6.1. Vanilla *vs*. Block Training

Table 2 serves as the empirical validation of the theoretical challenges outlined in Section 4. We compare our block-based approach against a "Vanilla" baseline: a standard masked diffusion model trained on global sequences (up to 8192 tokens) without block decomposition.

Table 3. **Block size and caching.** Approx. KV-Cache collapses; block-causal training enables exact caching with $5\times$ speedup.

KV-Cache	Train	Test	Block	Edit Dist.	TPS
No KV-Cache					
	Bidir	Bidir	32	0.067	29.5
	Bidir	Bidir	128	0.068	43.1
	Bidir	Bidir	256	**0.057**	42.8
	Bidir	Bidir	512	0.111	35.4
	Bidir	Bidir	1024	0.192	26.5
Approx. KV-Cache					
	Bidir	Blk-Causal	32	0.805	134.0
	Bidir	Blk-Causal	128	0.731	167.4
	Bidir	Blk-Causal	256	0.978	142.3
Exact KV-Cache					
	Blk-Causal	Blk-Causal	32	**0.069**	**103.7**
	Blk-Causal	Blk-Causal	128	0.089	123.1
	Blk-Causal	Blk-Causal	256	0.177	153.4

The baseline model exhibits high error rates compared to DODO. Crucially, this performance gap persists even when the model is provided with the *oracle* sequence length. This suggests that the limitation is not solely due to length estimation, but also stems from the positional anchoring inherent to parallel decoding. Attempting to resolve thousands of tokens simultaneously on a fixed canvas creates synchronization risks; because the model cannot adjust the global offset of disjoint text segments, the output becomes fractured.

We further investigate if this issue can be mitigated solely at inference time by applying block decoding to the baseline model. The results show that restricting the decoding window without a corresponding training objective leads to poor performance. This contrasts with findings in fast-dLLM (Wu et al., 2025b), where inference-time blocking remained effective for math and coding benchmarks like GSM8K (Cobbe et al., 2021) and HumanEval (Chen et al., 2021). We hypothesize that this divergence stems from the nature of the tasks, as unlike OCR, they allow some semantic or syntactic flexibility. This indicates that block diffusion serves as a necessary structural prior for OCR, conditioning the model to treat the prefix $x^{(<b)}$ as a stable anchor.

Finally, the baseline exhibits significantly lower throughput. This is due to the computational cost of the global canvas: the model must compute attention over the full sequence length (up to 8192 tokens) at every denoising step. In contrast, DODO decomposes the workload, ensuring a more efficient distribution of computational cost.

6.2. Block Size and Caching Strategies

Table 3 investigates the impact of block size and KV-caching strategies on performance. For the standard bidirectional model (No KV-Cache), we observe a non-monotonic trend as the block size increases. Initially, enlarging the block

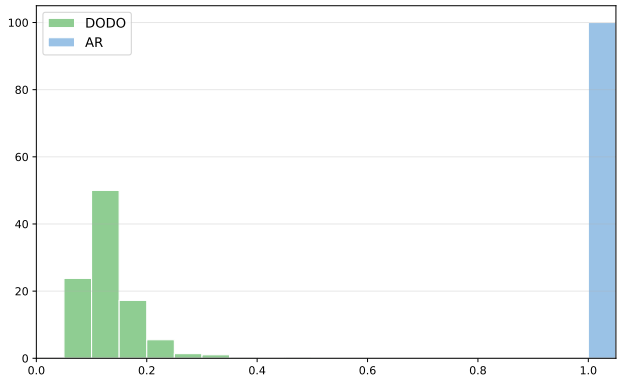

Figure 6. **Decoding Efficiency.** DODO requires fewer than 0.1 steps per token, compressing inference by an order of magnitude vs. autoregressive models.

size improves both throughput and accuracy, peaking at an intermediate size ($B = 256$). This is because generating more tokens in parallel reduces the total number of sequential inference steps, amortizing the cost of the forward pass. However, further increasing the block size ($B = 512, 1024$) leads to diminishing returns in throughput and a notable degradation in accuracy. We attribute this to the recurrence of positional anchoring issues: as the block becomes sufficiently large, it begins to suffer from the same internal synchronization failures that plague the global canvas.

The "Approx. KV-Cache" configurations investigate the feasibility of reusing computed keys and values in a bidirectional model without retraining. Unlike prior findings which observed minimal degradation when freezing history in standard diffusion models (Wu et al., 2025b), our experiments show a sharp accuracy collapse. We attribute this discrepancy to the rigid nature of the OCR task.

DODO resolves this incompatibility by explicitly training with block-causal masks, enabling exact KV-caching. However, the diffusion framework offers an additional capability unavailable to autoregressive models: bidirectional attention across blocks. Equipping it with bidirectional attention allows the context to dynamically adapt to the current active block by recomputing prefix representations at every step. This reduces the risk of representation drift when generating large blocks, effectively pushing the utilizable block size from 32 to 256 tokens and thus further improve accuracy. We observe that while DODO performs best with smaller blocks ($B = 32$), equipping it with bidirectional attention uniquely scales to larger blocks ($B = 256$), demonstrating that this is a powerful mechanism for enhancing parallel decoding in diffusion models.

6.3. Inference Efficiency Analysis

To disentangle the source of DODO's throughput advantage, we analyze the number of model forward passes required for generation in Figure 6. Autoregressive models are structurally bound to a $1:1$ ratio (one step per token). In contrast, DODO leverages parallel decoding to compress the inference process, typically requiring fewer than 0.1 steps per token. This order-of-magnitude reduction in the number of sequential steps explains the throughput results shown in Figure 5. Although DODO's bidirectional forward pass (without caching) is computationally heavier than a cached autoregressive step, the sheer reduction in the number of calls, generating 10 to 20 tokens per AR step, overcomes the per-step cost. This allows DODO to outperform optimized autoregressive baselines even without KV caching.

7. Conclusion

In this work, we introduce DODO, a framework that unlocks the potential of masked diffusion models to accelerate OCR via parallel decoding. Our analysis reveals that standard parallel decoding is fundamentally limited by the brittleness of the monolithic canvas, which leads to catastrophic synchronization failures due to length mismatches and positional anchoring. By decomposing this task into semi-autoregressive blocks, DODO provides a structural remedy that reconciles the stability of causal anchoring with the efficiency of parallel generation. Empirically, DODO sets a new standard for non-autoregressive OCR VLMs. It surpasses prior diffusion-based VLMs by an order of magnitude and achieves performance competitive with state-of-the-art specialized and autoregressive systems. Through block-causal attention and exact KV-caching, DODO achieves a $5\times$ inference speedup over the autoregressive baseline, establishing discrete diffusion not just as a theoretical capability, but as a practical, high-performance alternative for latency-critical applications. Furthermore, we show that the diffusion framework uniquely enables bidirectional attention across blocks, which further improves accuracy by pushing the utilizable block size from 32 to 256 tokens.

Limitations. While DODO significantly accelerates inference, its reliance on exact KV-caching enforces a static history, limiting accuracy at larger block sizes compared to the bidirectional variant. Future work will focus on bridging this gap and exploring diffusion samplers tailored to OCR.

References

Mistral ocr. https://mistral.ai/news/mistral-ocr, 2025. Mistral AI Optical Character Recognition model.

Aberdam, A., Bensaïd, D., Golts, A., Ganz, R., Nuriel, O., Tichauer, R., Mazor, S., and Litman, R. Clipter: Looking at the bigger picture in scene text recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21706–21717, 2023.

Abramovich, O., Nayman, N., Fogel, S., Lavi, I., Litman, R., Tsiper, S., Tichauer, R., Appalaraju, S., Mazor, S., and Manmatha, R. Visfocus: Prompt-guided vision encoders for ocr-free dense document understanding. In *European Conference on Computer Vision*, pp. 241–259. Springer, 2024.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J. L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M. a., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Arriola, M., Gokaslan, A., Chiu, J. T., Yang, Z., Qi, Z., Han, J., Sahoo, S. S., and Kuleshov, V. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*, 2025.

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 2021.

Azangulov, I., Pandeva, T., Prasad, N., Zazo, J., and Karmalkar, S. Parallel sampling from masked diffusion models via conditional independence testing, 2025. URL https://arxiv.org/abs/2510.21961.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025c.

Blecher, L., Cucurull, G., Scialom, T., and Stojnic, R. Nougat: Neural optical understanding for academic documents, 2023. URL https://arxiv.org/abs/2308.13418.

Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., and Lin, D. Sharegpt4v: Improving large multimodal models with better captions, 2023. URL https://arxiv.org/abs/2311.12793.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code. 2021.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Zhang, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., and Wang, W. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dai, W., Li, J., LI, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P. N., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

Ganz, R., Nuriel, O., Aberdam, A., Kittenplon, Y., Mazor, S., and Litman, R. Towards models that can see and read. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21718–21728, October 2023.

Ganz, R., Kittenplon, Y., Aberdam, A., Avraham, E. B., Nuriel, O., Mazor, S., and Litman, R. Question aware vision transformer for multimodal reasoning, 2024. URL https://arxiv.org/abs/2402.05472.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Li, J., Li, D., Xiong, C., and Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Li, S., Kallidromitis, K., Bansal, H., Gokul, A., Kato, Y., Kozuka, K., Kuen, J., Lin, Z., Chang, K.-W., and Grover, A. Lavida: A large diffusion model for vision-language understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Li, Y., Yang, G., Liu, H., Wang, B., and Zhang, C. dots.ocr: Multilingual document layout parsing in a single vision-language model, 2025b. URL https://arxiv.org/abs/2512.02498.

Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., and Bai, X. Monkeyocr: A unified ocr system with multi-stage pipelines. *arXiv preprint arXiv:2506.05218*, 2025c.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, C. et al. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*, 2024.

Liu, Z., Zhu, L., Shi, B., Zhang, Z., Lou, Y., Yang, S., Xi, H., Cao, S., Gu, Y., Li, D., Li, X., Fang, Y., Chen, Y., Hsieh, C.-Y., Huang, D.-A., Cheng, A.-C., Nath, V., Hu, J., Liu, S., Krishna, R., Xu, D., Wang, X., Molchanov, P., Kautz, J., Yin, H., Han, S., and Lu, Y. Nvila: Efficient frontier visual language models, 2025. URL https://arxiv.org/abs/2412.04468.

Luxembourg, O., Permuter, H., and Nachmani, E. Plan for speed–dilated scheduling for masked diffusion language models. *arXiv preprint arXiv:2506.19037*, 2025.

Mandal, S., Talewar, A., Ahuja, P., and Juvatkar, P. Nanonets-ocr-s: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging, 2025.

Mathew, M., Karatzas, D., and Jawahar, C. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Nacson, M. S., Aberdam, A., Ganz, R., Ben Avraham, E., Golts, A., Kittenplon, Y., Mazor, S., and Litman, R. Docvlm: Make your vlm an efficient reader. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29005–29015, 2025.

Nassar, A., Marafioti, A., Omenetti, M., Lysak, M., Livathinos, N., Auer, C., Morin, L., de Lima, R. T., Kim, Y., Gurbuz, A. S., Dolfi, M., Farré, M., and Staar, P. W. J. Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion, 2025. URL https://arxiv.org/abs/2503.11576.

Ouyang, L. et al. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Poznanski, A. et al. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *Hugging Face Datasets*, 2024. https://huggingface.co/datasets/allenai/olmOCR-mix-1025.

Poznanski, J., Rangapur, A., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Wilhelm, C., Lo, K., and Soldaini, L. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*, 2025.

Ronen, R., Tsiper, S., Anschel, O., Lavi, I., Markovitz, A., and Manmatha, R. Glass: Global to local attention for scene-text spotting. In *European Conference on Computer Vision*, pp. 249–266. Springer, 2022.

Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., and Kuleshov, V. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024a.

Sahoo, S. S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J. T., Rush, A., and Kuleshov, V. Simple and effective masked diffusion language models, 2024b. URL https://arxiv.org/abs/2406.07524.

Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. Simplified and Generalized Masked Diffusion for Discrete Data. *Advances in Neural Information Processing Systems*, 37:103131–103167, December 2024.

Sidorov, O., Hu, R., Rohrbach, M., and Singh, A. Textcaps: a dataset for image captioning with reading comprehension, 2020. URL https://arxiv.org/abs/2003.12462.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read, 2019. URL https://arxiv.org/abs/1904.08920.

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.

Wang, B. et al. Mineru-diffusion: A diffusion-based multi-stage document content extraction model. *arXiv preprint arXiv:2603.22458*, 2026.

Wang, H., Pan, C., Guo, X., Ji, C., and Deng, K. From object detection to text detection and recognition: A brief evolution history of optical character recognition. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13 (5):e1547, 2021.

Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., Han, C., and Zhang, X. General ocr theory: Towards ocr-2.0 via a unified end-to-end model, 2024. URL https://arxiv.org/abs/2409.01704.

Wei, H., Sun, Y., and Li, Y. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025a.

Wei, H., Sun, Y., and Li, Y. Deepseek-ocr: Contexts optical compression, 2025b. URL https://arxiv.org/abs/2510.18234.

Wu, C., Zhang, H., Xue, S., Diao, S., Fu, Y., Liu, Z., Molchanov, P., Luo, P., Han, S., and Xie, E. Fast-dllm v2: Efficient block-diffusion llm. *arXiv preprint arXiv:2509.26328*, 2025a.

Wu, C., Zhang, H., Xue, S., Liu, Z., Diao, S., Zhu, L., Luo, P., Han, S., and Xie, E. Fast-dllm: Training-free acceleration of diffusion language models. *arXiv preprint arXiv:2505.22618*, 2025b.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., Xie, Z., Wu, Y., Hu, K., Wang, J., Sun, Y., Li, Y., Piao, Y., Guan, K., Liu, A., Xie, X., You, Y., Dong, K., Yu, X., Zhang, H., Zhao, L., Wang, Y., and Ruan, C. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

You, Z., Nie, S., Zhang, X., Hu, J., Zhou, J., Lu, Z., Wen, J.-R., and Li, C. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025.

Yu, R., Ma, X., and Wang, X. Dimple: Discrete diffusion multimodal large language model. *arXiv preprint arXiv:2505.16990*, 2025.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., and Chen, W. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Zheng, L., Yuan, J., Yu, L., and Kong, L. A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*, 2023.

Zhou, J., Ding, T., Chen, T., Jiang, J., Zharkov, I., Zhu, Z., and Liang, L. Dream: Diffusion rectification and estimation-adaptive models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8342–8351, 2024.

A. Attention Structure Visualization

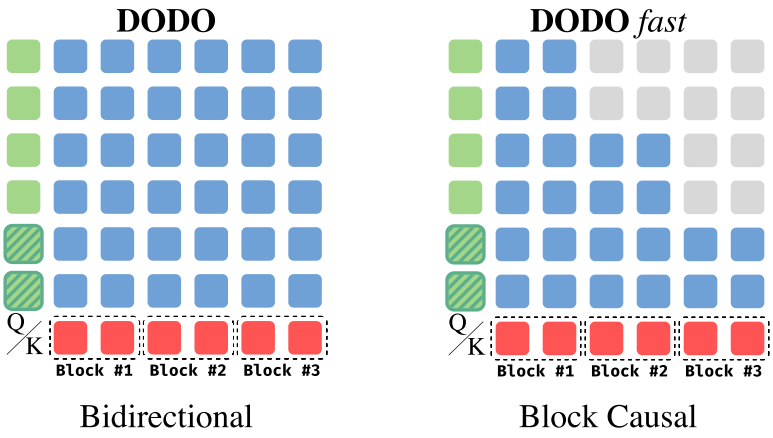

Figure A1. **Visualization of the attention structure.** Full bidirectional attention allows prior blocks to attend to the current block (green hatched), meaning their internal representations dynamically adapt during the forward pass. Block-causal masking prevents prior blocks from attending to the current block. This ensures the history representations remain invariant, enabling exact KV-caching for faster inference.

B. Additional Ablations

B.1. Sampling Strategies

To optimize the inference process, we investigate the impact of the decoding schedule. Given the intolerance of OCR tasks to transcription errors, our primary objective is not merely to maximize speed, but to identify the fastest strategy that maintains high fidelity. We compares two distinct sampling strategies:

- **Confidence Thresholding (Yu et al., 2025):** A dynamic strategy that unmasks all tokens whose prediction probability exceeds a fixed threshold p. This allows for adaptive step sizes: the model accelerates through clear text and slows down for ambiguous regions. This is the default strategy used in our main experiments (with $p = 0.99$).
- **Confidence Top-K (Zheng et al., 2023):** A fixed-rate strategy that unmasks exactly K tokens per step, guaranteeing a steady generation pace regardless of model confidence.

Figure B2 illustrates the performance trade-offs. In the high-accuracy regime required for OCR, Confidence Thresholding emerges as the superior strategy. By strictly enforcing a high confidence floor ($p = 0.99$), it ensures that the model only commits to tokens when certainty is high. While other strategies (such as aggressive Top-K) may achieve higher raw throughput, they do so at the cost of unacceptable error rates. Thus, confidence thresholding provides the optimal balance, maximizing speed strictly within the bounds of usable accuracy.

C. Comparison with MinerU-Diffusion

We additionally compare DODO against MinerU-Diffusion (Wang et al., 2026), a concurrent work that applies discrete diffusion to document parsing. Unlike DODO, which operates on full pages end-to-end, MinerU-Diffusion is designed to work at the crop level, employing a two-stage pipeline: a layout detection stage that identifies document blocks, followed by a content extraction stage that parses each detected crop independently. Notably, MinerU-Diffusion was trained on a proprietary closed OCR dataset, while DODO is trained exclusively on the publicly available olmOCR-mix-1025. For fair comparison, we run their official code on the same hardware and OmniDocBench dataset version used for all other methods. We process crops sequentially; however, a key advantage of MinerU-Diffusion's crop-level design is the ability to trade memory for speed by parallelizing multiple crops within the same batch, significantly increasing throughput. Similarly, DODO can trade memory for speed by parallelizing across pages.

On OmniDocBench, MinerU-Diffusion achieves an edit distance of 0.114 at an effective throughput of 45.25 tokens/sec. Breaking down the pipeline: the layout detection stage operates at an effective rate of 163 tokens/sec (total output tokens /

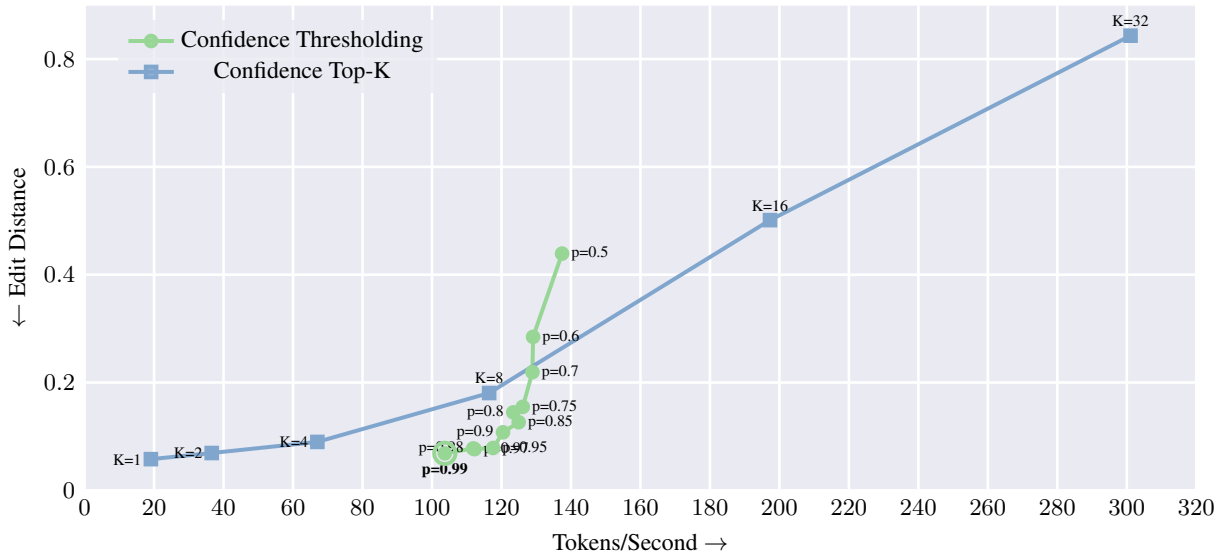

Figure B2. Edit Distance vs Speed for different sampling strategies. **Confidence Thresholding (Green)** achieves the optimal balance at $p = 0.99$, maintaining high accuracy while offering adaptive speedups.

layout time $= 1{,}546{,}107/9{,}480\text{s} = 163$), while the content extraction stage, which generates the actual text, operates at 66 tokens/sec (total output tokens / extraction time $= 1{,}546{,}107/23{,}360\text{s} = 66$), indicating that the crop-level extraction dominates the computational cost. Additionally, while MinerU-Diffusion was not trained for full-page inference, we evaluate it in a single-stage setting (bypassing layout detection and applying it directly on full pages), which yields an ED of 0.552 at 82.5 tokens/sec. In contrast, DODO achieves 0.069 edit distance at 103.7 tokens/sec using a single end-to-end model, representing both higher accuracy and $2.3\times$ higher throughput without requiring an explicit layout detection stage.

For reference, fine-tuning the same Qwen2.5-VL-3B backbone autoregressively on the same OCR data yields an ED of 0.044, comparable to DODO's 0.069, but at only 21 tokens/sec, a $5\times$ slower throughput.

D. Fine-Grained Evaluation Breakdown

We provide a detailed breakdown of OmniDocBench results by document length and document type.

Table 4 reveals that DODO's advantage over autoregressive VLMs grows with document length: at 4096+ tokens, DODO achieves 0.079 vs. Qwen 7B's 0.185, as longer sequences amplify the benefit of parallel decoding. Conversely, for very short documents (0–128 tokens), all methods perform similarly since sequential overhead is minimal.

Table 5 shows that DODO performs consistently across document types, with particularly strong results on PPT slides (0.014) and academic papers (0.050). The bidirectional variant further improves on structured layouts (Academic, Book) where long-range context aids alignment.

E. Document Parsing Examples

We present qualitative results in Figures E4 and E5. Selected from the OmniDocBench dataset, these examples demonstrate DODO's capability to transcribe dense text and preserve complex layout structures, while the accompanying heatmaps visualize the underlying parallel decoding process.

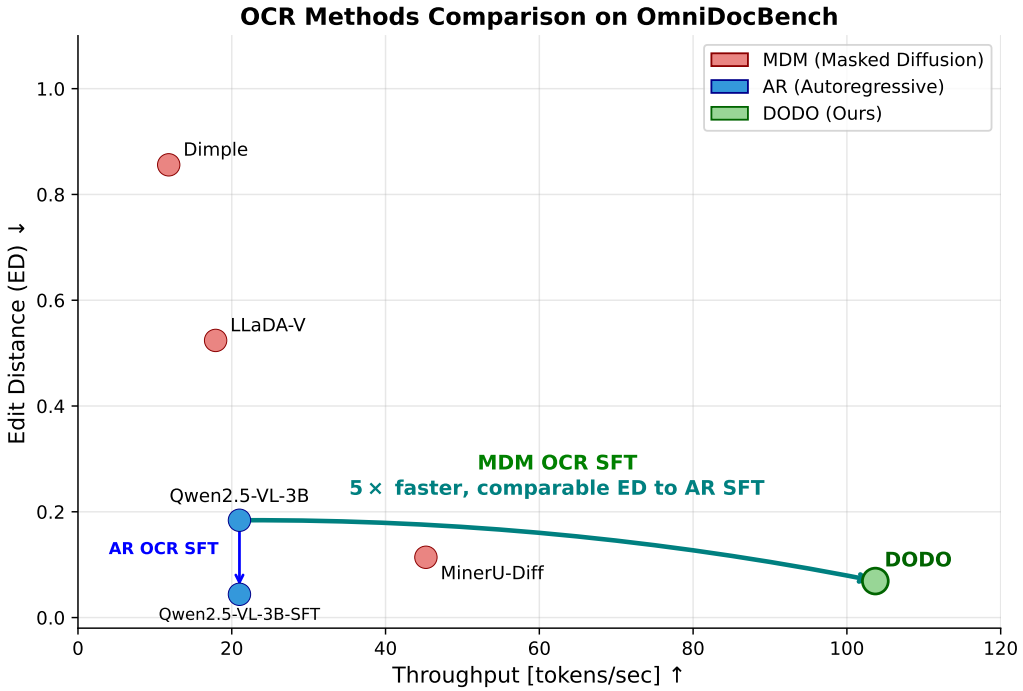

Figure C3. **OCR Methods Comparison on OmniDocBench.** Edit distance vs. throughput. DODO achieves both higher accuracy and higher throughput than all competing diffusion-based methods.

Table 4. **Edit distance by document length (tokens).** Lower is better.

Model	Pages	0–128	128–512	512–1024	1024–4096	4096+	Overall
Specialized OCR							
MinerU	275	0.328	0.166	0.027	0.032	0.030	0.041
Autoregressive VLMs							
Qwen 2.5 VL 7B	277	0.032	0.241	0.074	0.096	0.185	0.136
Qwen 2.5 VL 3B	276	0.032	0.156	0.062	0.159	0.363	0.237
Diffusion VLMs							
Dimple	269	0.343	0.697	0.750	0.868	0.900	0.857
LaViDa-L	259	0.743	0.995	0.998	1.000	1.000	0.995
LLaDA-V	212	0.485	0.306	0.235	0.471	0.782	0.524
DODO+Bidir	276	**0.015**	0.146	**0.025**	**0.051**	**0.065**	**0.057**
DODO	276	0.016	**0.133**	0.037	0.060	0.079	0.068

Table 5. **Edit distance by document type.** Lower is better.

Model	Academic	Book	News	Magazine	Textbook	Exam	PPT	Overall
Specialized OCR								
MinerU	0.015	0.049	0.047	0.043	0.052	0.045	0.261	0.041
Autoregressive VLMs								
Qwen 2.5 VL 7B	0.098	0.189	0.306	0.119	0.067	0.115	0.020	0.136
Qwen 2.5 VL 3B	0.180	0.216	0.629	0.189	0.157	0.138	0.019	0.237
Diffusion VLMs								
Dimple	0.900	0.872	0.880	0.836	0.805	0.871	0.404	0.857
LaViDa-L	1.000	1.000	1.000	0.999	1.000	1.000	0.865	0.995
LLaDA-V	0.681	0.350	0.895	0.404	0.290	0.313	0.301	0.524
DODO+Bidir	**0.031**	**0.048**	0.109	**0.074**	0.086	0.111	0.043	**0.057**
DODO	0.050	0.099	**0.093**	0.088	**0.061**	**0.084**	**0.014**	0.068

Figure E4. **Qualitative Results.** Each row displays a different document from OmniDocBench. **Left:** Input document image. **Center:** DODO's generated transcript rendered to PDF. **Right:** Visualization of the decoding process (heatmap of token commitment order). DODO successfully recovers complex layouts, including multi-column text, tables, and mathematical formulas, while maintaining high parallel efficiency.

Figure E5. **Additional Qualitative Results.**