# State-Conditioned Visual Evidence Retrieval for Fine-Grained Perception in Document Vision-Language Models

**Anonymous ACL submission**


## Abstract

Compared with typical vision–language tasks, document parsing places stronger demands on fine-grained visual perception. Existing vision–language model (VLM)–based parsing approaches rely on globally compressed visual tokens, where fine-grained details are entangled within a single representation and repeatedly accessed during decoding. However, we observe that the visual evidence for each prediction is typically localized and conditioned on the current decoding state, whereas such representations must be accessed in full at every decoding step, resulting in inefficient computation. To address this mismatch, we formulate perception as state-conditioned visual evidence retrieval (SCVER) during autoregressive decoding. The model operates on a compact global representation for coarse structure and retrieves a small set of relevant high-resolution regions conditioned on the current token state. This coarse-to-fine design enables on-demand access to fine-grained visual cues, relieving globally shared representations from encoding all fine-grained details. We further find that learning such state-conditioned retrieval in VLMs is challenging and unstable. To stabilize this process, we introduce a Spatially-Guided Learning Objective (SGLO) to guide the retrieval process. Experiments on document parsing benchmarks show that SCVER improves robustness under reduced input resolution and achieves a better accuracy–efficiency trade-off, demonstrating the effectiveness of on-demand visual evidence retrieval for fine-grained perception.


## 1 Introduction

Document parsing aims to convert document images into machine-readable formats (Cui et al., 2025a). Unlike typical vision–language (VL) tasks that primarily focus on cross-modal alignment and semantic reasoning, document parsing relies more on precise visual perception and fine-grained feature modeling (Wei et al., 2024).

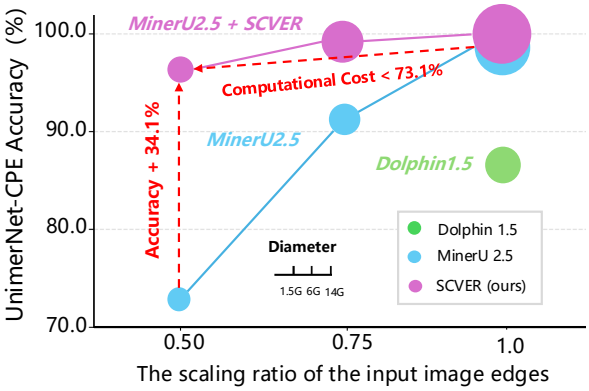
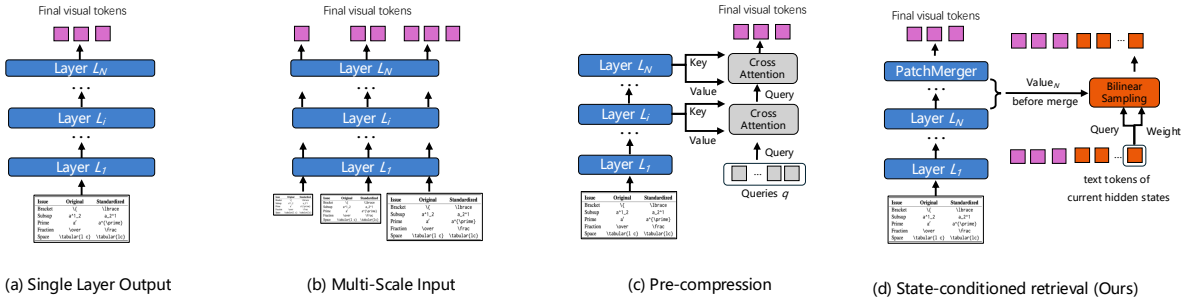



Figure 1: **Accuracy vs. Input Size on UniMERNet-CPE** (Wang et al., 2024a). Circle size indicates the average GFLOPs per generated token. SCVER reduces the reliance on high-resolution inputs, achieving a more favorable accuracy–efficiency trade-off compared to the dynamic-resolution MinerU2.5 (Niu et al., 2025) and the fixed-resolution Dolphin1.5 (Feng et al., 2025).

Despite recent advances in VLM-based document parsing, most methods encode the entire image into a fixed set of visual tokens and rely on them throughout decoding (Niu et al., 2025; Li et al., 2025). This forces a single global representation to capture both coarse semantics and fine-grained details, while requiring full interactions with all visual tokens at every decoding step. However, we observe that the visual evidence required for prediction is typically dynamic and spatially localized, with each step depending on only a small subset of regions. This creates a fundamental mismatch between localized, step-dependent evidence and global, fully accessed representations, leading to redundant computation. The problem further worsens as input resolution increases, since the number of visual tokens grows, making repeated full interactions increasingly expensive under quadratic-cost attention (Zhang et al., 2025a).

Based on this observation, we reformulate visual perception as state-conditioned visual evidence re-




(a) Single Layer Output (b) Multi-Scale Input (c) Pre-compression (d) State-conditioned retrieval (Ours)

Figure 2: (a) Most VLMs only use final layer output. (b) Llama-adapter(Zhang et al., 2024) resizes the image to different scales, and features are computed independently on each scale, which is inefficient. (c) VIVL(Cao et al., 2024) uses a small set of fixed queries to gather multi-level features, but these sparse visual samples stay constant throughout decoding, offering limited information. (d)Taking MinerU2.5(Niu et al., 2025) as an example, SCVER dynamically supplies pre-downsampling visual information required at each decoding step.

trieval (SCVER) during decoding, enabling on-demand access to fine-grained information. Instead of requiring a single global representation to encode all visual details upfront, the model first operates on compressed visual tokens to capture coarse structure and establish global text–vision alignment, and then dynamically retrieves a small set of relevant high-resolution regions conditioned on the current decoding state (Fig. 2(d)). As a result, global visual tokens are no longer required to preserve all fine-grained details, in contrast to conventional VLMs and multi-scale approaches that rely on pre-aggregated representations (Fig. 2(a–c)). This design reduces the dependence on high-resolution inputs while maintaining efficient decoding.

However, we find that naively applying SCVER leads to unstable optimization and limited gains. The difficulty stems from the absence of explicit spatial supervision in state-conditioned retrieval. While standard attention grounds relevance through query–key similarity, SCVER must predict retrieval locations directly, introducing latent spatial variables that are only indirectly supervised by token-level losses. This makes the spatial selection process underconstrained and prone to degenerate behaviors. To address this issue, we introduce a Spatially-Guided Learning Objective (SGLO), which provides explicit spatial guidance by leveraging cross-modal attention from a pretrained VLM as a weak prior over relevant regions, thereby stabilizing the retrieval process.

Our experiments show that SCVER consistently improves both efficiency and robustness, especially under reduced input resolution. When integrated with MinerU2.5 (Niu et al., 2025), it reduces computation by over 70% with negligible accuracy loss, leading to an improved accuracy–efficiency trade-off (Fig. 1). Moreover, SCVER significantly mitigates performance degradation under constrained resolutions, for example improving PubTabNet (Zhong et al., 2020) accuracy from 86.9 to 91.2 on MonkeyOCR (Li et al., 2025). The main contributions are summarized as follows:

- We identify a mismatch between localized, state-dependent visual evidence and global representations in document VLMs.
- We propose SCVER, a decoding-time mechanism that enables on-demand, token-specific access to fine-grained visual evidence.
- We introduce the SGLO to stabilize retrieval learning without additional annotations.

## 2 Related Works

### 2.1 Document Parsing Models

Document parsing involves layout detection and various recognition tasks. Early approaches integrated multiple specialized models(Zhao et al., 2024; Wang et al., 2024a; Da et al., 2023) into modular pipeline systems(Wang et al., 2024c). To reduce system complexity, research gradually shifted toward end-to-end approaches that directly map document images to structured text(Blecher et al., 2023; Wei et al., 2024). In recent years, the document parsing paradigm has evolved into a lightweight pipeline that first detects and then recognizes in parallel within a multi-task VLM (Feng et al., 2025; Cui et al., 2025a). However, these works mainly focus on the overall parsing paradigm, with limited attention to fine-grained visual information utilization during decoding.

## 2.2 Visual Information Utilization in VLMs

Dynamic-resolution encoders, such as NaViT, process images with flexible patch resolutions to handle aspect ratio variations (Dehghani et al., 2023), but do not explicitly enhance fine-grained perception. Multi-scale approaches aggregate features before decoding (Zhang et al., 2024; Cao et al., 2024; Park et al., 2024), improving representation capacity at the cost of efficiency. Recent work (e.g., V-Reflection (Zhou et al., 2026)) makes visual access adaptive to intermediate states, but remains within the standard attention framework, where visual evidence is implicitly aggregated rather than explicitly retrieved in a state-dependent manner. Some methods also incorporate external OCR modules (Nacson et al., 2024). Overall, these approaches do not explicitly enable state-dependent, selective access to fine-grained visual evidence during decoding.

## 2.3 Sparse Attention Mechanisms

In vision tasks, deformable attention samples a small set of spatial locations for each query to enable efficient multi-scale feature aggregation (Zhu et al., 2021). To reduce the quadratic cost of dense attention, prior work explores sparse attention by restricting interactions to subsets of tokens, including predefined patterns and similarity-based selection (Kitaev et al., 2020; Roy et al., 2020; Beltagy et al., 2020; Zaheer et al., 2021). Recent work further improves efficiency through token pruning, such as SparseVLM (Zhang et al., 2025b), which reduces redundant visual tokens. In contrast, our method enables state-dependent visual access by explicitly retrieving relevant visual evidence conditioned on the current decoding state.

# 3 Method

## 3.1 State-Conditioned Visual Evidence Retrieval

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the visual encoder produces high-resolution features $\tilde{F} \in \mathbb{R}^{\tilde{n}_v\times \tilde{d}}$. To reduce cost, a Patch Merger compresses them into $F \in \mathbb{R}^{n_v\times d}$ with $n_v < \tilde{n}_v$. The compressed visual tokens are concatenated with textual embeddings $T \in \mathbb{R}^{n_t\times d}$ to form $X_0 = [F; T]$, which is processed by a Transformer decoder.

Most existing approaches rely on this compressed representation throughout decoding, requiring it to encode both global semantics and fine-grained details. This leads to a trade-off: increasing $n_v$ preserves more detail but incurs quadratic attention cost $\mathcal{O}((n_v + n_t)^2)$, while reducing $n_v$ improves efficiency at the expense of fine-grained information.

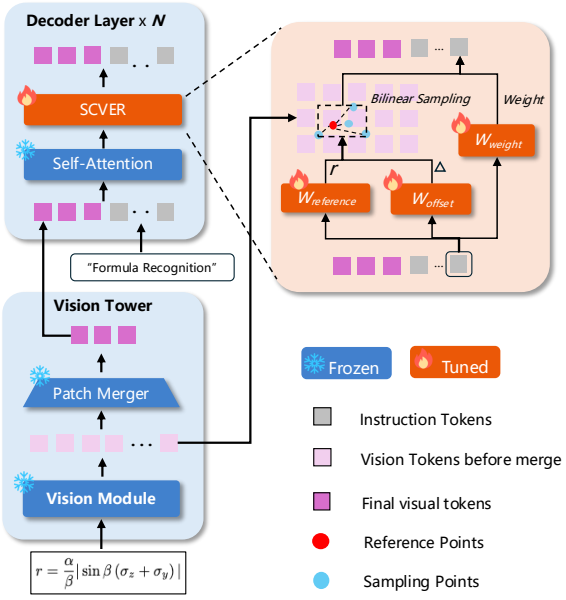



Figure 3: **Illustration of State-Conditioned Visual Evidence Retrieval.** SCVER efficiently retrieves fine-grained visual information from pre-downsampling features conditioned on the current decoding state.

However, in document parsing, the visual evidence required for each prediction is typically localized and varies across decoding steps. This suggests that not all visual information needs to be pre-aggregated into a fixed representation.

Based on this observation, we formulate visual perception as a state-conditioned visual evidence retrieval problem. Instead of relying solely on $F$, the decoder dynamically retrieves relevant visual cues from $\tilde{F}$ conditioned on its current hidden state.

Formally, at each decoder block $i$, we augment the Transformer as:

$$\begin{aligned} H_i &= X_{i-1} + \mathrm{Attn}(X_{i-1}), \\ \tilde{H}_i &= \mathrm{SCVER}(H_i, \tilde{F}), \\ X_i &= \tilde{H}_i + \mathrm{FFN}(\tilde{H}_i), \end{aligned} \tag{1}$$

SCVER enables the model to access token-specific, high-resolution visual evidence on demand, rather than requiring all details to be encoded in the compressed tokens. As a result, the decoder can maintain a compact global representation while retrieving fine-grained information only when needed, improving both efficiency and information fidelity.

### 3.2 Implementation of SCVER

To instantiate SCVER, we design a lightweight retrieval operator that selectively extracts localized visual evidence from the high-resolution feature map $\tilde{F}$. The retrieval is conditioned on the token representation, allowing different decoding states to access different spatial regions.

Our implementation is inspired by deformable attention (Zhu et al., 2021), but differs in that retrieval locations are directly predicted from the current token representation, enabling state-conditioned evidence access during decoding (Fig.3 right).

Given a token $t \in \mathbb{R}^d$, we predict a reference point, offsets, and weights:

$$r = \sigma(tW_{ref}), \quad \Delta = tW_{off}, \quad w = tW_{weight}, \tag{2}$$

where $W_{ref} \in \mathbb{R}^{d\times 2}$ predicts a normalized reference point, $W_{off} \in \mathbb{R}^{d\times(H\times P\times 2)}$ predicts offsets, and $W_{weight} \in \mathbb{R}^{d\times(H\times P)}$ predicts sampling weights. Here $H$ is the number of heads and $P$ the number of sampling points per head.

The offsets are reshaped to $\mathbb{R}^{H\times P\times 2}$ and normalized by $(W_f, H_f)$:

$$\ell_{h,p} = r + \frac{\Delta_{h,p}}{(W_f, H_f)}. \tag{3}$$

For each head $h$, we bilinearly sample $\tilde{F}$ at $\ell_{h,p}$ and aggregate:

$$t'_h = \sum_{p=1}^{P} w_{h,p}\,\mathrm{Bilinear}(\tilde{F}, \ell_{h,p}), \tag{4}$$

The outputs are concatenated across heads as $t' = \mathrm{Concat}(t'_1, \ldots, t'_H)$ and combined with the original token via a residual connection, enabling token-specific access to fine-grained visual evidence with minimal overhead while remaining compatible with KV caching.

### 3.3 Spatially-Guided Learning Objective

We find that naively applying SCVER leads to unstable training and slow convergence. This difficulty arises from a fundamental limitation of state-conditioned visual retrieval: the retrieval locations are predicted directly from token representations, introducing latent spatial variables that determine where visual evidence is accessed. Unlike standard attention, where relevance is explicitly grounded through query–key similarity ($QK^\top$), these spatial variables are only weakly constrained under token-level supervision, making the retrieval process prone to diffuse or degenerate behaviors.

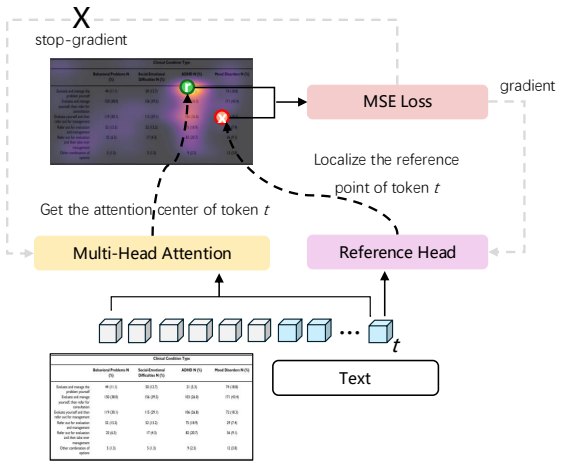



Figure 4: Illustration of the Computation Pipeline for the Spatially-Guided Learning Objective.

To address this issue, we introduce a Spatially-Guided Learning Objective (SGLO) that provides explicit spatial guidance for retrieval. We leverage the cross-modal attention learned by a pretrained VLM as a weak spatial prior, as such attention maps typically highlight regions that are informative for predicting each token.

During training, we utilize the causal attention over the sequence $X_i = [F; T]$ (Fig. 4). For a textual token $t$, let $\alpha_t \in \mathbb{R}^{H\times(n_v+n_t)}$ denote its multi-head attention weights. We extract the weights corresponding to visual tokens and average over heads:

$$\bar{\alpha}_t = \frac{1}{H}\sum_{h=1}^{H} \alpha_t^{(h)}[1:n_v], \tag{5}$$

which is reshaped into a spatial attention map $A_t \in \mathbb{R}^{H_f\times W_f}$.

We compute its spatial centroid:

$$c_t = \sum_{i,j} A_t(i,j)\,(x_{ij}, y_{ij}), \tag{6}$$

where $(x_{ij}, y_{ij}) \in [0,1]^2$ are normalized coordinates.

The centroid serves as a soft target to regularize the predicted reference point $r = \sigma(tW_{ref})$ via:

$$\mathcal{L}_{ref} = (r_x - c_{t,x})^2 + (r_y - c_{t,y})^2, \tag{7}$$

encouraging the retrieved regions to align with attention-induced visual focus.

The final objective combines the token-level cross-entropy loss with the alignment term:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{ref}. \tag{8}$$

| Method | R | Param | Formula$^{\text{CDM}\uparrow}$ | | | | Table$^{\text{TEDS}\uparrow}$ | | Text$^{1-\text{Edit}\uparrow}$ | | Page-Level |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | SPE | SCE | CPE | HWE | PubTabNet | FinTabNet | DocLaynet | B-MOD | OmniDoc |
| *General VLMs* | | | | | | | | | | | |
| GPT-4o | 1.0 | - | 96.1 | 88.2 | 83.3 | 85.9 | 76.6 | 83.9 | 91.7 | 0.318 | 75.0 |
| Qwen3-VL-235B | 1.0 | 235B | 98.4 | 96.2 | 97.4 | 94.2 | 85.4 | 84.9 | 92.3 | 0.321 | 89.7 |
| *Specialized VLMs* | | | | | | | | | | | |
| DeepSeek-OCR-2 | 1.0 | 3B | 95.5 | 77.1 | 92.0 | 81.6 | - | - | 92.1 | 81.2 | 90.2 |
| GLM-OCR | 1.0 | 0.9B | 98.4 | 97.7 | 96.7 | 95.1 | 85.2 | 92.1 | 93.3 | 84.5 | 95.2 |
| PaddleOCR-VL-1.5 | 1.0 | 0.9B | 99.2 | 94.9 | 98.8 | 92.3 | 84.6 | 94.7 | 93.6 | 85.1 | 94.3 |
| Dolphin-1.5 | 1.0 | 0.3B | 97.8 | 95.1 | 87.7 | - | 85.9 | 87.4 | 91.4 | 82.7 | 86.5 |
| Dolphin-1.5+SCVER | 1.0 | 0.3B | 98.4 | 97.6 | 90.2 | - | 91.9 | 92.4 | 92.6 | 94.0 | 87.4 |
| MonkeyOCR-Pro | 1.0 | 3B | 97.6 | 94.9 | 91.4 | 92.2 | 87.4 | 86.4 | 92.0 | 83.3 | 88.9 |
| MonkeyOCR-Pro | 0.5 | 3B | 75.6 | 64.3 | 74.2 | 72.6 | 58.3 | 52.9 | 77.3 | 62.4 | 71.3 |
| MonkeyOCR-Pro+SCVER | 0.5 | 3B | 98.1 | 94.5 | 91.0 | 96.0 | 92.8 | 89.7 | 91.3 | 91.3 | 89.1 |
| MonkeyOCR-Pro | 0.7 | 3B | 94.5 | 92.1 | 88.6 | 85.4 | 79.6 | 78.1 | 86.3 | 75.9 | 83.0 |
| MonkeyOCR-Pro+SCVER | 0.7 | 3B | 99.2 | **98.7** | 95.1 | **97.4** | **95.2** | 91.4 | 92.5 | **95.1** | 90.7 |
| MinerU2.5-Pro | 1.0 | 1.2B | 99.4 | 97.0 | 98.9 | 95.3 | 90.1 | 95.1 | 92.9 | 85.6 | 95.7 |
| MinerU2.5-Pro | 0.5 | 1.2B | 72.7 | 69.3 | 74.5 | 65.7 | 66.5 | 73.6 | 74.5 | 66.4 | 78.3 |
| MinerU2.5-Pro+SCVER | 0.5 | 1.2B | 99.1 | 97.5 | 98.9 | 96.1 | 90.4 | 95.7 | 92.3 | 92.4 | 95.5 |
| MinerU2.5-Pro | 0.7 | 1.2B | 95.4 | 94.3 | 93.7 | 90.4 | 82.0 | 89.1 | 86.8 | 73.3 | 89.2 |
| MinerU2.5-Pro+SCVER | 0.7 | 1.2B | **99.5** | 98.3 | **99.1** | 97.2 | 95.0 | **96.9** | **93.7** | 94.7 | **95.9** |

Table 1: Performance of different methods under varying image resolutions. R denotes the down-scaling ratio applied to each image dimension (width and height), corresponding to $R^2$ of the original image area. Values are approximate, as some models require padding or rounding to match implementation constraints. Bold numbers indicate the best results in each column.

| Model | Params | Encoder Type | Resolution |
|---|---|---|---|
| Dolphin1.5(2025) | 394M | Swin | Fixed |
| MinerU2.5(2025) | 1.2B | ViT | Dynamic |
| MonkeyOCR-Pro(2025) | 3B | ViT | Dynamic |

Table 2: Base models used in our experiments.

# 4 Experiments

## 4.1 Implementation Details

We evaluate VLMs with diverse encoder architectures and parameter scales (Table 2). During training, we freeze all pretrained backbone parameters and optimize only the components introduced by the state-aware attention module. The sampling module is applied every four layers, with further analysis in Sec. 4.7.2. Training data details are provided in Appendix A. All efficiency measurements are conducted on NVIDIA H100 GPUs with a batch size of 1.

## 4.2 Benchmarks and Evaluation Metrics

We consider three representative categories: (1) text recognition: B-MOD (Kišš et al., 2019) and the OCR-block subset of DocLayNet (Pfitzmann et al., 2022); (2) formula recognition (LaTeX): UniMER-1M (Wang et al., 2024a), including Simple Printed Expressions (SPE), Complex Printed Expressions (CPE), Screen-Captured Expressions (SCE), and Handwritten Expressions (HWE); (3) table structure recognition (HTML): PubTabNet (Zhong et al., 2020) and FinTabNet (Zheng et al., 2020). To assess overall document parsing capability, we further report end-to-end page-level evaluation on OmniDocBench v1.6 (Ouyang et al., 2025). Since most document parsing systems do not rely on VLMs for layout detection, we defer layout detection experiments to Appendix B.3. We use Edit Distance (Edit) for text recognition, Character Detection Matching (CDM) (Wang et al., 2024b) for formula recognition and Tree-Edit-Distance (TEDS) (Zhong et al., 2020) for table recognition.

## 4.3 Main Results

Table 1 compares SCVER with a broad range of methods on diverse document parsing benchmarks under both original and reduced input resolutions. We include specialized document recognition models such as DeepSeek-OCR-2 (Wei et al., 2026), GLM-OCR (Duan et al., 2026), and PaddleOCR-VL-1.5 (Cui et al., 2025a), as well as representative general-purpose VLMs (OpenAI et al., 2024; Bai et al., 2025).

Reducing input resolution is commonly used to control computational cost, but often leads to severe performance degradation. To evalu-

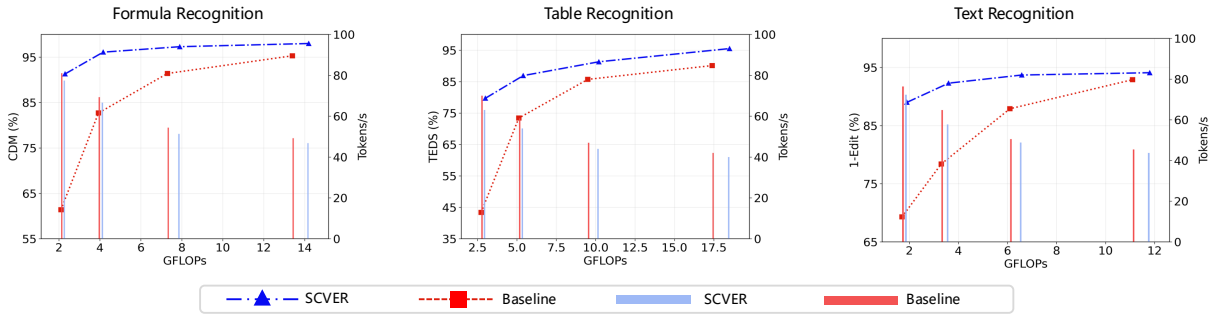



Figure 5: Accuracy and efficiency analysis. The x-axis shows input image size (0.4, 0.6, 0.8, 1.0) corresponding to per-token GFLOPs. Lines on the left y-axis indicate model performance and bars on the right y-axis indicate tokens per second. Integrating SCVER achieves a better accuracy-efficiency trade-off compared to the baseline (MinerU2.5).

ate robustness under such constraints, we integrate SCVER with the three backbone models in Sec. 4.1. For Dolphin-1.5, results are reported at the original resolution, while for MinerU2.5-Pro and MonkeyOCR-Pro, we additionally evaluate reduced resolutions.

SCVER consistently improves performance across both standard and constrained settings. On Dolphin-1.5, it improves all benchmarks, e.g., PubTabNet from 85.9 to 91.9 and B-MOD from 82.7 to 94.0. More importantly, under reduced resolutions, SCVER significantly mitigates performance degradation. For example, at 0.5 resolution (about 25% of the original visual tokens), MonkeyOCR-Pro drops from 87.4 to 58.3 on PubTabNet and from 83.3 to 62.4 on B-MOD, while SCVER recovers them to 92.8 and 91.3. A similar trend is observed on MinerU2.5-Pro, where SCVER improves PubTabNet from 82.0 to 95.0 and B-MOD from 73.3 to 94.7 at 0.7 resolution (about 49% tokens). These results demonstrate that SCVER reduces the reliance of VLMs on the initial input resolution.

### 4.4 Accuracy–efficiency trade-off

Figure 5 analyzes the accuracy–efficiency trade-off under varying input resolutions. As input size decreases, the baseline suffers a rapid performance drop across all three tasks, indicating strong reliance on high-resolution inputs for preserving fine-grained visual details.

In contrast, SCVER consistently achieves higher accuracy at comparable or lower computation, with the largest gains appearing in low-compute regimes. At roughly 2 GFLOPs, SCVER improves formula recognition from about 61% to 91%, table recognition from about 43% to 80%, and text recognition from about 69% to 89%. The advantage remains substantial at moderate computation levels, where SCVER continues to outperform the baseline by around 10–13 points across tasks.

Moreover, SCVER often reaches the baseline's high-cost performance at much lower computation. For example, in text recognition, SCVER achieves about 89% at roughly 1.8 GFLOPs, while the baseline requires more than 6 GFLOPs to approach a similar level. Although SCVER introduces additional retrieval operations, the throughput remains comparable to the baseline, with only a small drop in tokens per second (as indicated by the bars), indicating that the practical efficiency is largely preserved.

| | HWE | PubTabNet | B-MOD | FLOPs/Token |
|---|---|---|---|---|
| MinerU2.5-Pro | 95.3 | 90.1 | 85.6 | 1.15e+10 |
| +VIVL | 94.1(-1.2) | 90.7(+0.7) | 91.7(+6.1) | 1.19e+10 |
| +HVFA | 96.2(+0.9) | 91.4(+1.3) | 92.1(+6.5) | 1.34e+10 |
| +SCVER (Ours) | 97.4(+2.1) | 95.2(+5.1) | 95.1(+9.5) | 1.21e+10 |

Table 3: Comparison of different methods.

### 4.5 Comparison with Existing Methods

We compare our method with VIVL (Cao et al., 2024) and HVFA (Park et al., 2024), both of which enhance visual representations before decoding, whereas our method dynamically retrieves fine-grained visual evidence during decoding. As shown in Table 3, we conduct experiments using MinerU2.5 as the backbone model. VIVL introduces the smallest computational overhead but provides limited performance gains and even slightly degrades text recognition accuracy, possibly because it mainly emphasizes semantic-level feature fusion rather than improving access to fine-grained visual details. HVFA improves performance across datasets but incurs the largest computational cost.

| Method | Tok | DocVQA | TextVQA | InfoVQA |
|---|---|---|---|---|
| Qwen2.5-VL | 576 | 92.7 | 80.9 | 71.6 |
| DocVLM | 320 | 92.3(-0.4) | 79.7(-1.2) | 67.3(-4.3) |
| SCVER (Ours) | 320 | 90.4(-2.3) | 78.0(-2.9) | 69.4(-2.2) |
| DocVLM | 576 | 94.2(+1.5) | 83.8(+2.9) | 72.8(+1.2) |
| SCVER (Ours) | 576 | 93.5(+0.8) | 82.9(+2.0) | 73.1(+1.5) |

Table 4: Results on document VQA benchmarks under different visual token budgets. **Tok** denotes the maximum number of visual tokens after image resizing.

In comparison, our method achieves the best overall performance with only a small increase in computation, demonstrating the effectiveness of dynamic visual evidence retrieval for enhancing fine-grained perception.

### 4.6 Generalization to Text-Rich VQA

Text-rich VQA also requires models to identify fine-grained, answer-relevant visual evidence from complex images. Since attention cost grows quadratically with the number of visual tokens, reducing image resolution is a common way to improve efficiency, but it may also remove local details needed for answering questions.

As shown in Table 4, we further evaluate SCVER on DocVQA(Mathew et al., 2021b), TextVQA(Biten et al., 2019), and InfoVQA(Mathew et al., 2021a) using Qwen2.5-VL-7B as the backbone. We compare with DocVLM (Nacson et al., 2024), which introduces an additional OCR module to compensate for information loss under low-resolution inputs. Under a limited token budget of 320, SCVER remains competitive with DocVLM on DocVQA and TextVQA, and performs better on InfoVQA. Under 576 tokens, both methods improve over the backbone, with comparable overall performance.

Although SCVER does not consistently outperform the OCR-augmented baseline, it achieves competitive results without relying on an external OCR system. In contrast, DocVLM is specifically designed for OCR-assisted document question answering, while SCVER operates as a model-internal, decoding-time retrieval mechanism that can be directly applied across different vision-language settings. This suggests that state-conditioned visual evidence retrieval is not limited to document parsing, and can also benefit broader text-rich grounded understanding tasks.

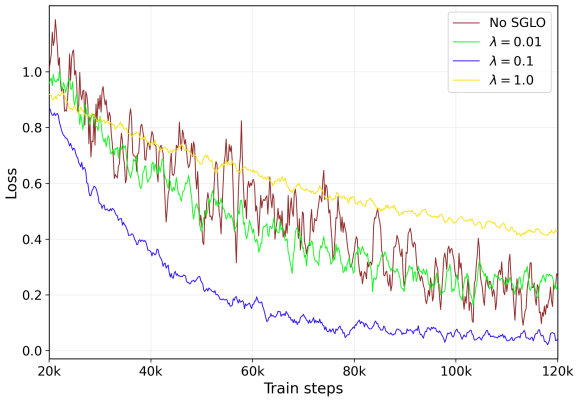



Figure 6: Training loss curves with and without the Spatially-Guided Learning Objective (SGLO). $\lambda$ denotes the weight of the SGLO term.

### 4.7 Ablation Study

In the main paper, we analyze the scaling behavior of state-conditioned visual evidence retrieval and the effect of the Spatially-Guided Learning Objective. Further analyses of other design choices and whether the observed performance gains arise from data scaling effects are provided in Appendix B.

#### 4.7.1 Effect of Training Strategy

Directly applying SCVER to VLMs yields only limited performance improvements (Table 5, first two rows), as learning state-conditioned visual retrieval is challenging and often leads to unstable optimization. This difficulty arises because the retrieval process must predict where to access visual evidence without explicit spatial supervision. To address this issue, we introduce the Spatially-Guided Learning Objective (SGLO), which leverages cross-modal attention as a weak spatial prior to guide the retrieval process.

With SGLO, the advantage of SCVER becomes significantly more pronounced. For example, on PubTabNet, the performance gain increases from 2.5% to 6.9%. Training dynamics further show that SGLO stabilizes optimization (Fig. 6). Without SGLO, training exhibits noticeable oscillations. A small weight ($\lambda = 0.01$) provides limited stabilization, while a large weight ($\lambda = 1.0$) overly constrains optimization and slows convergence. Empirically, $\lambda = 0.1$ achieves a better balance between stability and performance.

Overall, SGLO effectively regularizes the retrieval process, improving both training stability and downstream performance.

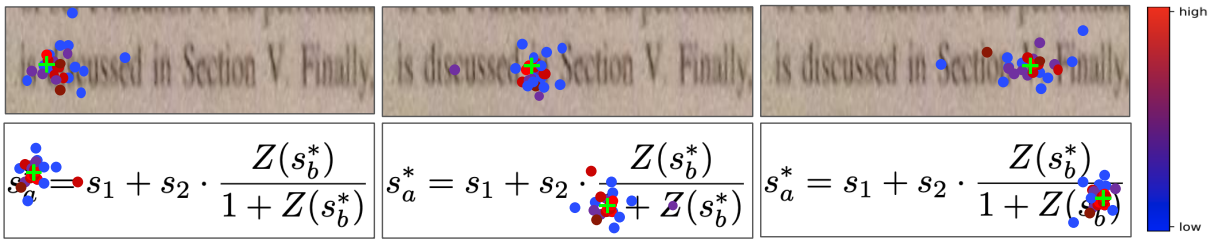



Figure 7: **Visualization of state-conditioned retrieval points.** Multi-head retrieval points (after applying offsets to the reference point) and their attention weights, with and without the Spatially-Guided Learning Objective (SGLO). Each circle denotes a retrieval point, and color indicates its attention weight.

| *SCVER* | *SGLO* | SCE | PubTabNet | B-Mod |
|---|---|---|---|---|
| | | 95.1 | 85.9 | 82.7 |
| ✓ | | 95.4 | 88.1 | 85.9 |
| ✓ | ✓ | **97.6** | **91.9** | **94.0** |

Table 5: Ablation study on the effect of SGLO.

### 4.7.2 Effect of SCVER Application Frequency

We further analyze the effect of the SCVER module frequency on both accuracy and computational cost. Experiments are conducted using MinerU2.5 (Niu et al., 2025) as the backbone model on the PubTabNet (Zhong et al., 2020).

As shown in Fig. 8, increasing the interval between SCVER applications significantly improves decoding speed, while the accuracy decreases gradually. This indicates that applying SCVER too frequently yields diminishing returns in performance.

Based on this trade-off between efficiency and performance, we apply SCVER every four layers. Compared with applying it at every decoder layer, this configuration results in only a 0.7% drop in accuracy while reducing computational cost by approximately 10%. This setting achieves a better balance between efficiency and performance.

### 4.8 Case study.

We analyze the behavior of SCVER using a representative example in Fig. 7, which visualizes retrieved locations and their weights for different tokens in text and formula regions. The retrieved points are highly localized and vary across tokens: tokens requiring fine-grained visual details, such as mathematical symbols or specific characters, focus on corresponding local regions, while different tokens attend to distinct spatial areas. This demonstrates that SCVER performs token-specific, state-dependent visual evidence retrieval, enabling dynamic access to relevant fine-grained information during decoding.

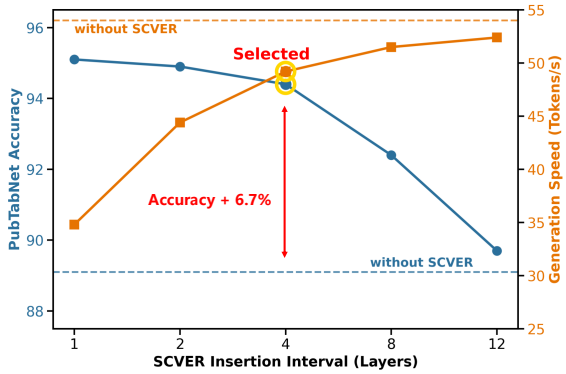



Figure 8: Effect of retrieval module application frequency on accuracy. The x-axis indicates the interval of decoder layers at which sampling module is inserted.

## 5 Conclusion

In this paper, we introduce State-Conditioned Visual Evidence Retrieval (SCVER), a new paradigm for fine-grained visual perception in vision–language models (VLMs). Instead of relying on static visual representations that must encode both coarse structure and fine-grained details, SCVER formulates perception as the retrieval of token-specific visual evidence conditioned on the current decoding state. To realize this paradigm, we combine a coarse-to-fine retrieval design with a Spatially-Guided Learning Objective (SGLO), enabling efficient access to fine-grained visual evidence during autoregressive decoding. Extensive experiments show that SCVER consistently improves robustness and achieves a better accuracy–efficiency trade-off on document parsing benchmarks, especially under reduced input resolution. Overall, these results suggest that decoding-time visual evidence retrieval provides a practical and effective direction for enhancing fine-grained perception in future VLMs.

## Limitations

The proposed SGLO training strategy relies on cross-modal attention maps returned by the backbone model to provide spatial guidance. As a result, it is not directly compatible with certain optimized attention implementations, such as FlashAttention or SDPA, during training. While this limitation does not affect inference after training, it may introduce additional computational overhead during the training stage.

## References


Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. *Preprint*, arXiv:1905.13648.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *Preprint*, arXiv:2308.13418.

Yun-Hao Cao, Kaixiang Ji, Ziyuan Huang, Chuanyang Zheng, Jiajia Liu, Jian Wang, Jingdong Chen, and Ming Yang. 2024. Towards better vision-inspired vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13537–13547.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025a. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model. *Preprint*, arXiv:2510.14528.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025b. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. 2023. Vision grid transformer for document layout analysis. *Preprint*, arXiv:2308.14978.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey Gritsenko, Mario Lučić, and Neil Houlsby. 2023. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. *Preprint*, arXiv:2307.06304.

Shuaiqi Duan, Yadong Xue, Weihan Wang, Zhe Su, Huan Liu, Sheng Yang, Guobing Gan, Guo Wang, Zihan Wang, Shengdong Yan, Dexin Jin, Yuxuan Zhang, Guohong Wen, Yanfeng Wang, Yutao Zhang, Xiaohan Zhang, Wenyi Hong, Yukuo Cen, Da Yin, and 4 others. 2026. Glm-ocr technical report. *Preprint*, arXiv:2603.10910.

Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, Jingqun Tang, Hao Liu, and Can Huang. 2025. Dolphin: Document image parsing via heterogeneous anchor prompting. *Preprint*, arXiv:2505.14059.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *Preprint*, arXiv:2001.04451.

Martin Kišš, Michal Hradiš, and Oldřich Kodym. 2019. Brno mobile ocr dataset. *Preprint*, arXiv:1907.01307.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. 2025. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *Preprint*, arXiv:2506.05218.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. Infographicvqa. *Preprint*, arXiv:2104.12756.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021b. Docvqa: A dataset for vqa on document images. *Preprint*, arXiv:2007.00398.

Mor Shpigel Nacson, Aviad Aberdam, Roy Ganz, Elad Ben Avraham, Alona Golts, Yair Kittenplon, Shai Mazor, and Ron Litman. 2024. Docvlm: Make your vlm an efficient reader. *Preprint*, arXiv:2412.08746.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2025. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. *Preprint*, arXiv:2412.07626.

Jaeyoo Park, Jin Young Choi, Jeonghyung Park, and Bohyung Han. 2024. Hierarchical visual feature aggregation for ocr-free document understanding. *Preprint*, arXiv:2411.05254.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter Staar. 2022. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 3743–3751. ACM.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2020. Efficient content-based sparse attention with routing transformers. *Preprint*, arXiv:2003.05997.

Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. 2024a. Unimernet: A universal network for real-world mathematical expression recognition. *Preprint*, arXiv:2404.15254.

Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Bo Zhang, and Conghui He. 2024b. Cdm: A reliable metric for fair and accurate formula recognition evaluation. *Preprint*, arXiv:2409.03643.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024c. Mineru: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024d. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *Preprint*, arXiv:2409.01704.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2026. Deepseek-ocr 2: Visual causal flow. *Preprint*, arXiv:2601.20552.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences. *Preprint*, arXiv:2007.14062.

Qintong Zhang, Bin Wang, Victor Shea-Jay Huang, Junyuan Zhang, Zhengren Wang, Hao Liang, Conghui He, and Wentao Zhang. 2025a. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *Preprint*, arXiv:2410.21169.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2024. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *Preprint*, arXiv:2303.16199.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2025b. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *Preprint*, arXiv:2410.04417.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *Preprint*, arXiv:2410.12628.

Xinyi Zheng, Doug Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2020. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. *Preprint*, arXiv:2005.00589.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. *Preprint*, arXiv:1911.10683.

Jiazhou Zhou, Yucheng Chen, Hongyang Li, Qing Jiang, Hu Zhou, Ying-Cong Chen, and Lei Zhang. 2026. V-reflection: Transforming mllms from passive observers to active interrogators. *Preprint*, arXiv:2604.03307.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable detr: Deformable transformers for end-to-end object detection. *Preprint*, arXiv:2010.04159.

# A Training Data

Our training data are mainly derived from the training splits corresponding to the evaluation benchmarks described in the main text. These datasets provide diverse supervision signals for layout understanding, formula recognition, and text recognition.

However, most existing open-source document parsing datasets are primarily constructed from English documents. In contrast, the OmniDoc benchmark contains a large number of non-English document elements, such as Chinese formulas and tables, while it does not provide an official training set. This mismatch between training resources and evaluation data may limit the robustness of models in multilingual document scenarios.

To address this issue, we supplement additional multilingual training resources and further refine existing datasets. Specifically, we construct additional data for formula recognition and text recognition to better support multilingual document parsing tasks.

## A.1 Formula Recognition Dataset

Existing open-source formula datasets are typically constructed by extracting mathematical expressions from a large number of arXiv files and re-rendering them into image–LaTeX pairs. Although this pipeline largely avoids syntactic errors, variations in authors' writing styles often lead to multiple LaTeX representations of the same mathematical expression. Such diversity introduces semantic ambiguities and inconsistencies in the ground truth.

To mitigate this issue, we identify common divergent LaTeX representations and standardize them to ensure semantic and structural consistency. After normalization, the standardized LaTeX code is re-rendered to verify correctness, producing a cleaner and more coherent dataset whose input images and output LaTeX sequences are strictly aligned.

In addition, to better support multilingual document parsing scenarios such as OmniDoc, we incorporate an additional 0.5M multilingual formula samples from the open-source project oleehyo_latex_formulas_80M, which provides formula expressions in multiple languages.

## A.2 Text Recognition Dataset

Our text recognition data mainly come from two sources. The first is DocLayNet (Pfitzmann et al., 2022), which provides fine-grained annotations for layout elements across a wide variety of document domains. While DocLayNet offers high-quality layout segmentation annotations, it does not explicitly account for inline formulas in scientific documents, which may introduce ambiguity during training.

| Issue | Original | Standardized |
|---|---|---|
| Bracket | `\{` | `\lbrace` |
| Subsup | `a^1_2` | `a_2^1` |
| Prime | $a'$ | `a^{\prime}` |
| Fraction | `\over` | `\frac` |
| Space | `\array{l c}` | `\array{lc}` |

Table 6: Examples of LaTeX standardization for various symbols and expressions.

To complement this limitation, we construct a scientific-document-oriented paragraph recognition dataset based on the Text_Completion_arXiv corpus, a page-level annotated dataset of arXiv papers. Building upon these annotations, we generate large-scale block-level image–text pairs using a combination of detection and recognition models supported by heuristic rules.

Specifically, we first employ a layout detection model to segment page-level images into individual layout elements and crop the corresponding regions using detected bounding boxes. Each cropped region is then processed by a state-of-the-art text recognition model to obtain textual predictions. Meanwhile, paragraph-level ground truth is extracted from the original page-level text using `\n\n` as separators. Finally, we align the recognized text with extracted paragraphs using edit-distance filtering and order-aware matching with strict thresholds to ensure correctness.

Through this pipeline, we construct a large-scale, high-quality element-level dataset tailored for robust scientific document text recognition.

# B Supplementary experiments

## B.1 Training Details

We conducted all training and inference speed evaluations using the PyTorch framework on NVIDIA H100 GPUs. During training, we used an initial learning rate of $1e{-}5$ and employed a cosine learning rate scheduler to progressively adjust the model parameters.

## B.2 Page-level Evaluation Metrics

We conduct page-level evaluation on the OmniDocBench benchmark. To eliminate the influence

| Method | Textual | Image | Table | Equation | Page Margins | Full Page |
|---|---|---|---|---|---|---|
| D4LA | | | | | | |
| DocLayout-YOLO (Zhao et al., 2024) | 90.8 | 62.6 | 89.8 | 91.1 | - | 88.5 |
| PP-StructureV3 (Cui et al., 2025b) | 90.0 | 67.9 | 89.7 | 92.1 | 79.1 | 88.3 |
| MinerU2.5 (Niu et al., 2025) | 94.6 | 72.8 | 91.4 | 91.0 | 84.2 | 91.4 |
| MinerU2.5+SCVER | 95.8 | 74.9 | 93.2 | 92.9 | 88.5 | 93.2 |
| DocLaynet | | | | | | |
| DocLayout-YOLO (Zhao et al., 2024) | 91.2 | 91.3 | 94.8 | 82.8 | - | 92.0 |
| PP-StructureV3 (Cui et al., 2025b) | 93.8 | 94.2 | 96.7 | 92.1 | 77.4 | 94.0 |
| MinerU2.5 (Niu et al., 2025) | 94.8 | 95.9 | 97.1 | 93.5 | 86.3 | 95.2 |
| MinerU2.5+SCVER | 95.2 | 96.7 | 97.6 | 94.7 | 89.7 | 95.5 |

Table 7: Comparison of layout analysis performance across different methods and content types on multiple benchmark datasets.

```
<|box_start|>100 200 300 400<|box_end|><|ref_start|>title<|ref_end|><|rotate_up|>
<|box_start|>400 500 600 700<|box_end|><|ref_start|>text<|ref_end|><|rotate_up|>
```

Figure 9: Output format of layout detection.

of layout detection errors on downstream evaluation, the page-level assessment focuses on three major document layout elements: text, tables, and formulas.

The overall page-level score is computed by averaging the performance on these three elements. Specifically, the evaluation uses Edit Distance for text, TEDS for tables, and CDM for formulas. The overall score is calculated as follows:

$$\text{Overall} = \frac{(1-\text{Text}^{\text{Edit}})\times 100+\text{Table}^{\text{TEDS}}+\text{Formula}^{\text{CDM}}}{3}$$

### B.3 Layout Detection

In the main paper, we observe that incorporating the proposed State-Conditioned Visual Evidence Retrieval (SCVER) paradigm into VLMs leads to only marginal improvements on page-level document parsing benchmarks. A key reason is that many document parsing systems rely on additional layout detection modules, where the final parsing results are largely determined by the outputs of these detectors. Since layout detection itself requires fine-grained visual perception, improvements brought by SCVER may not be fully reflected in page-level evaluations.

To further evaluate the effectiveness of SCVER in layout detection tasks, we adapt SCVER to the MinerU2.5 model and conduct experiments on the publicly available DocLayNet (Pfitzmann et al., 2022) and D$^4$LA (Da et al., 2023) datasets. We compare our approach with representative layout detection methods, including the YOLO-based DocLayout-YOLO (Zhao et al., 2024) and the Detection Transformer-based PP-StructureV3 (Cui et al., 2025b). Since the two datasets adopt different label taxonomies, we follow the label mapping strategy proposed in the MinerU2.5 report (Niu et al., 2025) to convert the annotations into a unified label space.

Table 7 presents detailed results across different layout element categories. The results show that the vision–language model MinerU2.5 significantly outperforms traditional detection models such as YOLO, as well as attention-based detection architectures like Detection Transformer. We attribute this advantage to the joint training of MinerU2.5 on both layout detection and layout element recognition tasks. This joint learning allows the model to implicitly leverage textual modality information during layout detection, enabling not only more accurate bounding box localization but also substantially improved category prediction.

Building upon this strong baseline, the proposed SCVER further enhances the model's ability to utilize fine-grained visual information during decoding. As a result, the generative vision–language model can better capture detailed visual structures, leading to additional improvements in layout detection performance.

| Method | Formula$^{\text{CDM}\uparrow}$ | | | | Table$^{\text{TEDS}\uparrow}$ | | Text$^{\text{Edit}\downarrow}$ | |
|---|---|---|---|---|---|---|---|---|
| | SPE | SCE | CPE | HWE | PubTabNet | FinTabNet | DocLaynet | B-MOD |
| Dolphin-1.5 (Feng et al., 2025) | 97.8 | 95.1 | 87.7 | - | 85.9 | 87.4 | 0.017 | 0.304 |
| Dolphin-1.5* | 97.9 | 95.3 | 87.4 | - | 86.1 | 87.0 | 0.014 | 0.244 |
| Dolphin-1.5 + SCVER | 98.4 | 97.6 | 90.2 | - | 91.9 | 92.4 | 0.011 | 0.068 |
| MonkeyOCR-Pro (Li et al., 2025) | 97.6 | 94.9 | 91.4 | 92.2 | 87.4 | 86.4 | 0.020 | 0.204 |
| MonkeyOCR-Pro* | 97.4 | 95.3 | 91.7 | 92.0 | 88.0 | 86.2 | 0.018 | 0.154 |
| MonkeyOCR-Pro + SCVER | **99.2** | 97.2 | 93.9 | 97.2 | **95.2** | 91.4 | 0.014 | **0.028** |
| MinerU2.5 (Niu et al., 2025) | 98.4 | 96.4 | 96.6 | 94.4 | 89.1 | 95.6 | 0.013 | 0.218 |
| MinerU2.5* | 98.6 | 96.1 | 96.9 | 93.9 | 89.2 | 95.8 | 0.011 | 0.158 |
| MinerU2.5 + SCVER | **99.2** | **98.7** | **97.1** | **97.4** | 95.1 | **95.9** | **0.010** | 0.052 |

Table 8: Analysis of potential data leakage to verify whether the observed performance gains are caused by training data overlap. The symbol * indicates that the model is further fine-tuned on the same data used to train SCVER.

| Configuration | | | Performance | | | Efficiency |
|---|---|---|---|---|---|---|
| H | P | F | Formula | Table | Text | FLOPs |
| 8 | 2 | 1 | 95.6 | 87.6 | 0.028 | 9.58e+08 |
| 16 | 4 | 1 | 97.3 | 91.9 | 0.011 | 9.90e+08 |
| 24 | 8 | 2 | 97.6 | 92.4 | 0.011 | 1.27e+09 |

Table 9: Performance and efficiency under different configurations. $H$, $P$, and $F$ correspond to the number of heads, sampling points, and high-scale feature maps used in the SCVER computation.

| $\lambda$ | Formula | Table | Text |
|---|---|---|---|
| 0 | 96.4 | 95.6 | 0.218 |
| 0.01 | 96.7 | 95.3 | 0.166 |
| 0.1 (selected) | **98.7** | **95.9** | 0.028 |
| 1.0 | 98.2 | 95.7 | **0.017** |

Table 10: Performance under different $\lambda$ values on SGLO.

### B.4 Internal Hyperparameters of SCVER

In the main paper, we primarily investigate the impact of the application frequency of SCVER on model accuracy and inference efficiency. Here, we further explore several internal hyperparameters of SCVER, including the number of high-resolution feature maps used, the number of attention heads, and the number of sampling points per head.

In our previous experiments, MinerU2.5 is built upon the Qwen2VL architecture (Wang et al., 2024d), which employs a NaViT-based visual encoder. This encoder only provides a single high-resolution feature map before downsampling, limiting our ability to study the influence of multi-scale visual features. To conduct a more comprehensive analysis, we instead adopt Dolphin1.5, which utilizes a hierarchical visual encoder and thus allows us to better explore these hyperparameters.

As shown in Table 9, $F$ denotes the number of high-resolution feature maps used by SCVER. When $F = 1$, the model uses a $2 \times 2$ high-resolution feature map, while $F = 2$ indicates the use of both $2 \times 2$ and $4 \times 4$ feature maps. The results show that when $F = 1$, increasing the number of sampling points per head is already sufficient to approach the peak performance. Incorporating additional high-resolution feature maps brings only marginal accuracy gains while significantly increasing computational cost. This observation suggests that the key benefit of SCVER mainly comes from dynamic, state-conditioned sampling rather than simply aggregating more high-resolution visual features.

### B.5 Impact of Weight Coefficients on SGLO

As described earlier, we introduce the State-Conditioned Visual Evidence Retrieval (SCVER) module to enable state-conditioned dynamic fine-grained perception in vision–language models. However, when integrating SCVER into VLMs, the conventional token-level loss cannot provide direct supervision for the sampling locations, which limits the performance gains. Visualization results further reveal that the predicted sampling points tend to collapse into nearly uniform global sampling patterns.

To address this issue, we introduce the Spatially-Guided Learning Objective (SGLO) as an additional loss term to provide explicit guidance for the sampling process. We further conduct an ablation study on the weight of this loss term, as shown in

Table 10. When the weight of SGLO is too small, the performance gain remains marginal, suggesting that the supervisory signal is insufficient to effectively guide the sampling behavior. In contrast, assigning an excessively large weight to this loss term may dominate the training objective, potentially constraining the model's ability to learn semantic representations. Consequently, a balanced weighting between the two objectives is necessary to achieve the best performance.

### B.6 Data Leakage Ablation Study

Since the training data of the backbone model are not publicly available, it is difficult to determine whether the performance improvements observed in our experiments originate from the proposed design or from potential distribution differences between the backbone model's training data and the evaluation datasets. To investigate this factor, we conduct additional experiments to analyze the impact of such data-related effects.

As shown in Table 8, fine-tuning the backbone model with these datasets results in only limited improvements. Except for a moderate gain on the B-MOD dataset, most benchmarks show negligible changes, and some even exhibit slight performance degradation. This suggests that simply increasing the amount of training data does not consistently lead to performance improvements.

Furthermore, even on B-MOD, where visual quality is relatively degraded, the performance gains obtained from additional training data remain smaller than those achieved by introducing our perception mechanism. These observations indicate that the improvements observed in our experiments are unlikely to be explained solely by data scaling. Instead, they highlight the importance of explicitly modeling fine-grained visual perception in generative vision–language models.

## C Study Case

In the last few pages of the appendix, we present additional examples demonstrating the accuracy gains brought by the SCVER module through enhanced fine-grained visual perception. During the analysis of failure cases, we also observed an interesting counterexample. As shown in Fig. 10, we initially assumed that the base model made an error in text recognition. However, upon closer inspection, we found that the spelling in the original image was actually incorrect, while the base model produced a semantically plausible output. According to previous studies, this phenomenon may be attributed to the strong language priors accumulated by the decoder during large-scale pretraining. When encountering ambiguous or uncertain visual inputs, the model's predictions may be dominated by the decoder's language modeling capability, resulting in outputs that conform to linguistic statistics rather than strictly following the visual evidence.

In contrast, when the model is equipped with the SCVER module, this behavior appears to change. The model tends to produce outputs that more faithfully follow the visual information in the image rather than relying on language priors for semantic correction. This observation suggests that the SCVER module not only improves the fine-grained visual perception ability of vision–language models but may also potentially mitigate hallucination caused by language-prior dominance, thereby encouraging the model to better adhere to visual evidence.

Nevertheless, it should be noted that this issue is not the primary focus of our work, and due to space limitations we do not provide a systematic investigation of this phenomenon. Moreover, from the perspective of text recognition tasks, it remains an open question whether models should possess the ability to automatically correct spelling errors or instead strictly reproduce the text as it appears in the image. We therefore leave this interesting observation for future exploration.

*During the documentary shoot, positioned above the river that runs through downtown, the camera captures a wide angel of the city.*

**Output without SaSA :**

During the documentary shoot, positioned above the river that runs through downtown, the camera captures a wide angle of the city.

**Output with SaSA :**

During the documentary shoot, positioned above the river that runs through downtown, the camera captures a wide angel of the city.

Figure 10: Comparison between the base model and the model equipped with the SCVER module on text recognition. The base model tends to rely on the decoder's language prior, while the model with SCVER follows the visual evidence more faithfully.

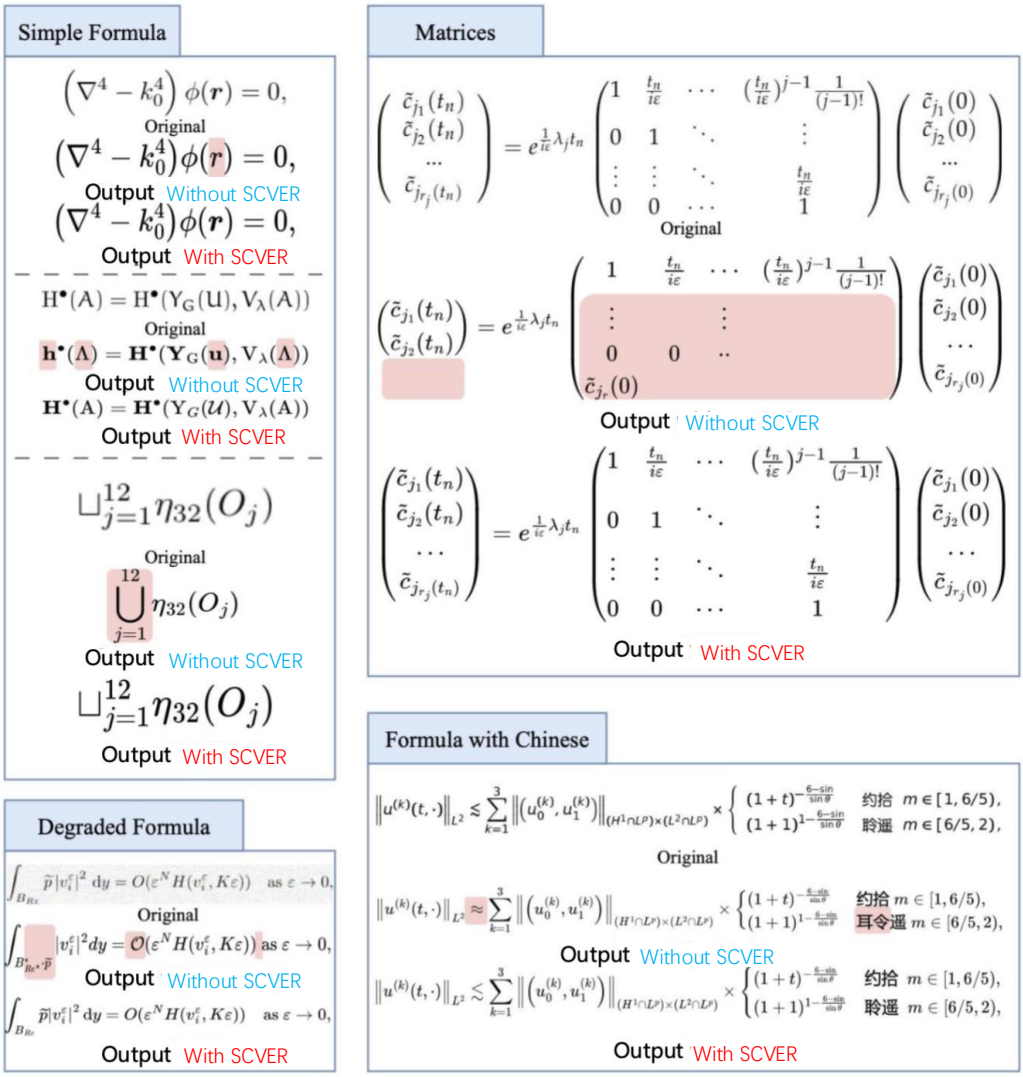



Figure 11: Impact of the SCVER strategy on formula recognition. This figure compares model outputs with and without the SCVER module across various types of mathematical expressions, including simple formulas, matrices, degraded formulas, and formulas containing Chi- nese text. The highlighted regions indicate the positions where the model without SCVER fails to preserve structural correctness or symbol fidelity. These comparisons demonstrate that SCVER effectively mitigates such errors by maintaining fine-grained structural details and improving symbol-level accuracy, resulting in outputs that more closely align with the original expressions.

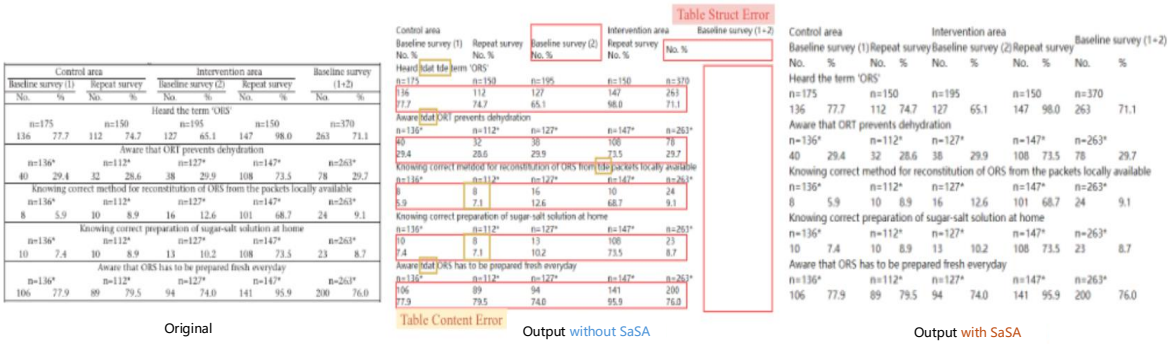



Figure 12: Impact of the SCVER strategy on table recognition.This figure compares model outputs with and without the SCVER module on a complex table containing multi-row, multi-column, and merged-cell structures. The highlighted regions indicate structural and content errors produced by the model without SCVER, including misaligned cell boundaries, incorrect row–column associations, and incorrectly recognized text. These comparisons demonstrate that SCVER effectively preserves table-structure integrity and improves cell-level content fidelity, enabling the model to generate outputs that more accurately reflect the original tabular layout and data.