# Scroll-Net: Geometry-Aware Constrained Decoding for Multi-Page Document Hierarchy Parsing

**Anonymous ACL submission**


## Abstract

Hierarchical Document Structure Analysis (HDSA) aims to recover the logical tree and reading order of a document, providing a structural foundation for downstream tasks such as document question answering, information extraction, and long-form document understanding. Extending HDSA to multi-page documents remains challenging, as pagination introduces discontinuities between page-local coordinate systems, making it difficult to link structurally related elements across pages. We present Scroll-Net, a geometry-aware constrained decoding framework for multi-page document hierarchy parsing. Scroll-Net maps page-local coordinates into a unified scroll-space coordinate frame, enabling direct comparison of cross-page elements. It employs a dual-stream encoder to disentangle geometric and semantic cues, injecting geometric inductive bias before multimodal fusion. Under the pre-order convention used by the evaluated benchmarks, Scroll-Net constrains autoregressive parent decoding with a hard reading-order causal mask. This backward-pointer constraint makes decoded parent graphs acyclic under the imposed order while retaining $O(K)$ per-step decoding, reducing inference latency by 36.5% and peak memory by 46.5%. Across four benchmarks, Scroll-Net remains competitive on simple layouts while achieving state-of-the-art results on structurally complex documents, improving TEDS by 4.8 points on DocHieNet.


## 1 Introduction

Hierarchical Document Structure Analysis (HDSA) recovers the logical tree of a document together with a consistent reading order. Both serve as prerequisites for downstream document QA, information extraction, and long-form reading; recent multimodal work shows that explicit document-level structure substantially improves long-document understanding and agentic search (Fu et al., 2026; Li et al., 2026). Existing methods differ in how they represent layout and propagate dependencies. Page-wise pipelines such as DSG (Rausch et al., 2023) process pages independently and stitch predictions post hoc, severing cross-page geometric continuity. UniHDSA (Wang et al., 2025) adds a document-level grouping stage on top of unified relation prediction, but its per-page encoder still operates in page-centric coordinates. Sequence-based formulations (Huang et al., 2022; Beltagy et al., 2020) linearize 2D layouts into 1D streams: they scale well with Transformers, yet they obscure the spatial structure on which long-range linking depends. Multi-stage frameworks (Wang et al., 2024b) jointly infer ordering and hierarchy to avoid cascading errors, but rely on dense pairwise scoring with quadratic cost and soft constraints that admit invalid outputs.

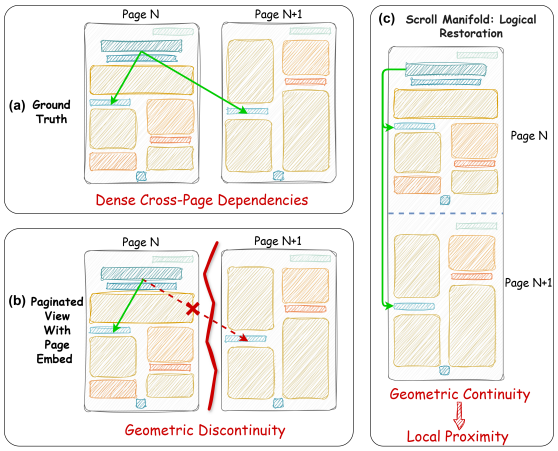


Figure 1: **Scroll-space coordinate mapping.** **(a)** Parent–child edges can span pages. **(b)** Page-local coordinates disrupt cross-page geometry. **(c)** Scroll-Net maps blocks into a unified scroll space $\mathcal{M}$.

Three obstacles remain in the multi-page setting. The most basic is geometric discontinuity (Fig. 1a-b): cross-page neighbors reside in disjoint coordinate frames, so distance-based locality does not extend across page boundaries. A second obstacle is the entanglement of geometry with semantics in dense layouts; spatial proximity becomes mis-

leading in multi-column pages or repeated patterns, and models often fall back on semantic matching to bridge the gap even when geometry would suffice. The third obstacle is decoding cost: dense pairwise parent scoring over all block pairs incurs $O(N^2)$ work and, without a hard structural prior, may produce invalid outputs such as cycles.

Scroll-Net addresses each obstacle directly. It maps page-local coordinates into a single scroll-space frame (Fig. 1c), so candidate pruning, attention, and pairwise scoring all operate in a unified geometric space. A dual-stream encoder keeps geometric and semantic representations separate until late fusion, which prevents spatial proximity from being conflated with semantic affinity. The third design choice rests on a structural observation: under the pre-order convention used by the evaluated benchmarks, a parent is expected to precede its child in the reading order. Scroll-Net uses the predicted reading order as a hard causal mask over parent candidates, yielding acyclic outputs under the imposed order and admitting efficient candidate pruning.

Our main contributions are: (1) We introduce a scroll-space coordinate mapping coupled with geometric inference, providing a shared coordinate frame for cross-page relation scoring and candidate pruning. (2) A causal tree decoding strategy restricts candidate parents to each node's reading-order prefix; under the imposed order, it yields acyclic decoded structures and removes the need for post-hoc cycle removal. (3) Top-$K$ geometric candidate pruning cuts sequential decoding complexity from $O(N^2)$ to $O(NK)$, reducing inference latency by 36.5% and peak memory by 46.5% while remaining competitive on simple layouts and improving TEDS by 4.8 points on the structurally complex DocHieNet.

## 2 Related Work

**Page-Wise Pipelines and Stitching.** Early visual document understanding systems treated layout analysis as a single-page task, driven by large-scale datasets (Zhong et al., 2019; Li et al., 2020b) and toolkits (Shen et al., 2021). A common paradigm processes pages independently and stitches predictions via post-processing (Rausch et al., 2019, 2023): DSG, for instance, processes documents as per-page renderings and applies a grammar-based module for structural consistency. Uni-HDSA (Wang et al., 2025) advances this with a unified relation prediction head and a dedicated document-level stage for cross-page grouping; however, its per-page encoding remains page-centric, leaving cross-page geometric comparability less explicit. Hi-VT5 (Tito et al., 2022) similarly aggregates page-level representations but retains page-centric coordinates.

**Sequence-Based Formulations and Linearization.** Transformer-based document models often linearize 2D layouts into 1D sequences, including LayoutLM-style encoders (Xu et al., 2020; Huang et al., 2022), BROS (Hong et al., 2022), LiLT (Wang et al., 2022), OCR-free parsers (Kim et al., 2022; Blecher et al., 2024; Lee et al., 2023), and long-context architectures (Beltagy et al., 2020; Zaheer et al., 2020). Such models scale naturally with sequence processing, but linearization obscures 2D locality and entangles geometry with semantics, making dense multi-column layouts and cross-page links difficult to model geometrically.

**Unified Joint Frameworks.** To avoid cascading errors, recent unified frameworks attempt to jointly infer ordering and hierarchy. Generative paradigms like Seg2Act (Li et al., 2024b) and DSG (Rausch et al., 2023), as well as end-to-end graph parsers like DRGG (Chen et al., 2025), achieve joint prediction but are limited to single-page documents. For multi-page understanding, multi-stage frameworks like Detect-Order-Construct (Wang et al., 2024b) jointly infer structure across documents but still face scalability and validity issues, often relying on dense pairwise prediction with $O(N^2)$ complexity. Autoregressive reading-order decoders can also become a bottleneck on rich-text documents, as observed in early end-to-end text reorganization models that couple graph encoders with pointer-style decoding (Li et al., 2020a). To scale, some adopt windowed/chunked training, which fundamentally restricts the global receptive field and obscures long-range dependencies. Without hard structural constraints, long-document inference can produce invalid structures (e.g., cycles) in practice. Reading order has been shown to be critical for downstream information extraction tasks (Zhang et al., 2023), and has been explicitly modeled with transformer-based RO predictors and dedicated resources such as LayoutReader/ReadingBank (Wang et al., 2021). RO-focused training (e.g., FocalOrder (Liu et al., 2026)) is complementary: it improves RO learning, whereas we use predicted RO as a hard causal prior at inference for constrained decoding.

**Vision-Language Models and Benchmarks.** Recent document-oriented VLMs and LLM extensions improve visual-document reasoning through layout-aware generation, multi-image inputs, or structural maps (Wang et al., 2024a; Li et al., 2024a; Fu et al., 2026; Li et al., 2026; Jin et al., 2025), but they do not provide specialized HDSA decoding guarantees. Benchmarks such as HRDoc (Ma et al., 2023) and DocHieNet (Xing et al., 2024) further motivate models for cross-page structural reasoning. Overall, prior work leaves three gaps (page-local geometry, entangled cues, quadratic inference), which Scroll-Net addresses via scroll-space mapping (Section 3.2), dual-stream fusion (Section 3.3), and sequentially constrained decoding (Section 3.4).

# 3 Method

## 3.1 Task Formulation and Overview

### 3.1.1 Problem Statement

Let $\mathcal{D} = \{b_i\}_{i=1}^{N}$ denote a multi-page document comprising $N$ layout blocks distributed across $P$ pages. Each block $b_i$ is characterized by a bounding box $\mathbf{x}_i = (c_i^x, c_i^y, w_i, h_i) \in \mathbb{R}^4$ expressed in page-local coordinates, a page index $p_i \in \{1, \ldots, P\}$, and a semantic feature vector $\mathbf{s}_i \in \mathbb{R}^{d_s}$ extracted from a LayoutLMv3 encoder (Huang et al., 2022).

The goal of Hierarchical Document Structure Analysis (HDSA) is to recover the document's logical structure, represented as a *rooted ordered tree*:

$$\mathcal{T} = (V,\ E,\ r,\ \pi), \tag{1}$$

where $V = \{b_i\}_{i=1}^{N}$ is the node set (layout blocks), $E = \{(v, \mathrm{pa}(v)) : v \in V\} \subset V \times (V \cup \{r\})$ is the directed child→parent edge set, $r \notin V$ is a *virtual document root* representing the document itself, and $\pi : \{1, \ldots, N\} \to V$ is a bijection representing the *reading order*. The reading order provides a total ordering over $V$ that matches the document's intended sequential consumption. By convention we set $\pi^{-1}(r) = 0$, so the root precedes all blocks.

### 3.1.2 Reading Order as a Causal Prefix

Logical hierarchy (containment) and reading order (sequential traversal) are tightly coupled under the depth-first pre-order convention used to linearize document trees in our evaluated benchmarks: a parent is expected to precede its child, so reading order provides a causal prefix for parent selection in most cases.

**DFS Pre-Order Linearization.** Document tree annotations in our evaluated benchmarks largely follow depth-first pre-order traversal, placing a parent node (e.g., a section header) before its descendants (e.g., the section's paragraphs).

**Sequential Consistency.** This traversal protocol motivates a temporal constraint on the reading sequence: the parent of a node should appear earlier in the sequence than the node itself. We formalize this sequential-consistency constraint as:

$$\forall\, v \in V : \quad \pi^{-1}\big(\mathrm{pa}(v)\big) \ < \ \pi^{-1}(v), \tag{2}$$

where $\pi^{-1}(v)$ denotes the integer position of node $v$ in the reading order, and we set the virtual root's position as $\pi^{-1}(r)$=0.

**Causal Prefix Constraint.** We enforce Eq. 2 as a hard prior: each node selects its parent only from the reading-order prefix. This yields acyclic outputs under the imposed order, and we quantify the exclusion rate and net benefit on Comp-HRDoc in Section 4.5.

### 3.1.3 Architecture Overview

**Model Pipeline.** Fig. 2 provides an overview of Scroll-Net. We map paginated geometry into scroll space (Section 3.2), encode blocks via a dual-stream encoder (Section 3.3), predict reading order via Sinkhorn (Mena et al., 2018), and autoregressively decode parent pointers under a hard causal mask (Section 3.4), yielding acyclic decoded outputs with $O(NK)$ decoding complexity.

**Training and Runtime.** The framework is trained end-to-end via a composite multi-task loss $\mathcal{L}$. During training, the decoder receives the ground-truth reading order to isolate structure prediction from ordering errors. At inference, the predicted order $\hat{\pi}$ drives autoregressive decoding. While the encoding stage requires $O(N^2)$ complexity (shared with standard Transformers), the decoding stage runs in $O(NK)$ time due to geometric pruning ($K \ll N$), yielding substantial savings on long documents.

## 3.2 Scroll-Space Coordinate Mapping

The paginated coordinate space $\mathcal{P}$ induces incomparable cross-page distances under page-local Euclidean geometry. We define a scroll-space mapping $\Phi$ that projects all blocks into a single global frame, enabling the downstream geometric inference pipeline (candidate construction, 2D-RPE

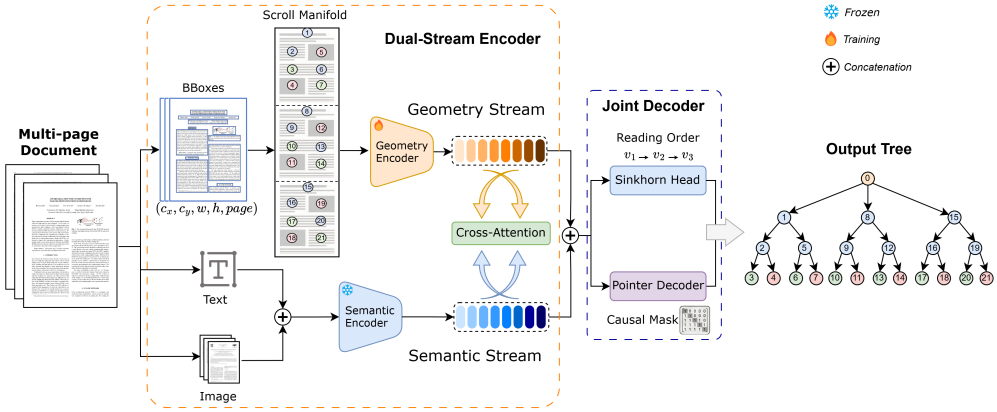


Figure 2: **Overview of Scroll-Net.** Scroll-Net maps paginated coordinates into scroll space, fuses geometric and semantic cues, and decodes parent pointers under a hard causal mask.

bias, and structured decoding) to operate uniformly over the full document.

#### 3.2.1 Scroll-Space Coordinate Frame

We address this discontinuity by mapping all blocks onto a scroll space $\mathcal{M} = [0, 1] \times \mathbb{R}_{\geq 0}$, which models the document as a single vertical strip. For a page-local box $\mathbf{c} = (c^x, c^y, w, h)$, the mapping $\Phi : \mathcal{P} \to \mathcal{M}$ outputs $\tilde{\mathbf{c}} = \Phi(\mathbf{c}, p)$ via a page-dependent vertical translation:

$$\begin{aligned} \tilde{c}^x &= c^x, & \tilde{c}^y &= c^y + p - 1, \\ \tilde{w} &= w, & \tilde{h} &= h. \end{aligned} \tag{3}$$

This maps page $p$ to the interval $[p-1, p)$ on the global vertical axis. This mapping assumes unit-normalized page heights, which is a reasonable approximation for digitally typeset documents (e.g., LaTeX-compiled papers in HRDoc); perturbation experiments in Appendix E show robustness under synthetic page-height perturbations. After applying $\Phi$, $\ell_2$ distance $d_{\mathcal{M}}$ is well-defined for all block pairs, and $\mathcal{M}$ serves as a globally consistent frame for downstream geometry-driven modules.

#### 3.2.2 Pairwise Geometric Features on $\mathcal{M}$

The mapped coordinates enable uniform computation of pairwise geometric features for all block pairs, regardless of whether they reside on the same page. For each candidate pair $(b_i, b_j)$, we compute a $D_g$-dimensional geometric descriptor $\mathbf{f}_{ij}$ on $\mathcal{M}$, including relative offsets, scroll distance, scale ratios, and alignment cues. An MLP scorer $f_\theta : \mathbb{R}^{D_g} \to \mathbb{R}$ maps $\mathbf{f}_{ij}$ to an additive scalar geometric score for the pointer decoder (Section 3.4).

#### 3.2.3 Scroll-Space Candidate Construction

The scroll distance $d_{\mathcal{M}}$ also serves as the basis for *geometric candidate pruning* in the pointer decoder. We precompute $\mathcal{C}_i = \text{top-K}_{j \neq i}\, s_\Phi(b_i, b_j)$ using a lightweight scorer on the pairwise geometry. The pointer decoder then restricts each step to $\mathcal{C}_i$ (Section 3.4), reducing sequential decoding to $O(NK)$.

### 3.3 Geometry-Guided Dual-Stream Encoder

Geometric proximity and logical affiliation are distinct properties. Spatially adjacent blocks may belong to different columns, while a section header and its last paragraph may span multiple pages. To separate these signals, we use a dual-stream encoder $\mathcal{E}$. It independently encodes geometry and semantics, exchanges information via bidirectional cross-attention, and then adaptively fuses the two streams for each node.

#### 3.3.1 Stream-Specific Encoding

The *geometry stream* takes each block's mapped coordinates $\tilde{\mathbf{x}}_i$ on $\mathcal{M}$, concatenates a learnable page embedding (dimension $d_p$), and projects the result to dimension $d$ via a two-layer MLP with GELU. The *semantic stream* independently projects each block's LayoutLMv3 feature (dimension $d_s$) to the same dimension $d$. Both projected sequences are then encoded by separate Pre-LN Transformer encoders ($L_e$ layers), ensuring geometric and semantic representations occupy distinct subspaces before fusion.

#### 3.3.2 Bidirectional Cross-Attention Fusion

We exchange complementary signals between the two streams through $L_f$ layers of standard bidirectional cross-attention: in each layer, the geometry stream attends to the semantic stream and

vice versa, each followed by a feed-forward sublayer with Pre-LN residual connections. We update semantics conditioned on the geometry-updated stream, injecting a geometry-first inductive bias; ablations show gains (Section 4.3).

#### 3.3.3 Adaptive Gating

After cross-attention, we fuse the two streams via a node-dependent gating mechanism. A two-layer MLP $\psi$ maps the concatenated stream outputs to softmax weights $(\alpha_i^{\text{geo}}, \alpha_i^{\text{sem}})$, yielding $\mathbf{h}_i = \alpha_i^{\text{geo}}\,\mathbf{h}_i^{\text{geo}} + \alpha_i^{\text{sem}}\,\mathbf{h}_i^{\text{sem}}$. A final linear projection with layer normalisation produces the unified node representations $\mathbf{H} \in \mathbb{R}^{N\times d}$ consumed by all downstream heads.

#### 3.3.4 Geometric Contrastive Alignment

To strengthen the structural discriminability of $\mathbf{H}$, we introduce an auxiliary contrastive alignment loss during training. For each node whose ground-truth parent is not the virtual root, we form a positive pair using the true parent's representation $\mathbf{h}_{\text{pa}^*(b_i)}$ and draw hard negatives $\mathcal{N}_i^-$ from the top-scoring candidate parents (by current model logits) excluding $\text{pa}^*(b_i)$. Let $\mathcal{S}_{\text{nr}} = \{i : \text{pa}^*(b_i) \neq r\}$, $p_i^* = \text{pa}^*(b_i)$, $z_i^+ = \text{sim}(\mathbf{h}_i, \mathbf{h}_{p_i^*})$, and $z_{ik}^- = \text{sim}(\mathbf{h}_i, \mathbf{h}_k)$. We then apply an InfoNCE objective (van den Oord et al., 2018) with temperature $\tau_c$:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{|\mathcal{S}_{\text{nr}}|}\sum_{i\in\mathcal{S}_{\text{nr}}} \log \frac{\exp(z_i^+/\tau_c)}{Z_i}. \quad (4)$$

where $Z_i = \exp(z_i^+/\tau_c) + \sum_{k\in\mathcal{N}_i^-}\exp(z_{ik}^-/\tau_c)$, and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. By mining hard negatives from the model's own candidate set, this loss specifically separates confusable parent–non-parent pairs in embedding space, complementing the cross-entropy signal from the pointer decoder.

### 3.4 Sequentially Constrained Decoding

Under the sequential consistency constraint (Eq. 2), inferring a node's parent becomes a left-to-right, autoregressive process over the reading-order prefix. This formulation resembles incremental parsing schemes that restrict decisions to the already-processed context (Ma et al., 2018). We adopt a pointer-network style decoder $\mathcal{P}$ (Vinyals et al., 2015) to implement this mechanism. It steps through the document nodes according to the reading order, selecting the parent from the valid causal prefix using a combination of semantic attention and geometric bias.

#### 3.4.1 Decoding Formulation

Let $\pi$ denote the reading order (ground-truth during training; predicted $\hat{\pi}$ at inference). At step $t$, the current node is $i = \pi(t)$, and the causal prefix $\mathcal{A}_t = \{0\} \cup \{\pi(1), \ldots, \pi(t-1)\}$ contains the virtual root (indexed by 0) and all nodes that precede $i$ in the imposed reading order. When Eq. 2 holds under the ground-truth reading order, the true parent lies in $\mathcal{A}_t$; if the annotated order violates this convention or the predicted order misorders a parent–child pair, the mask can exclude the true parent, which we quantify in Section 4.5. The decoder selects $\widehat{\text{pa}}(\pi(t)) = \arg\max_{j\in\mathcal{A}_t} \ell_t[j]$, where the logit $\ell_t[j]$ fuses three complementary signals:

#### 3.4.2 Geometry-Aware Pointer Attention

A learnable root token $\mathbf{r}_0$ is prepended to the reading-order-sorted node representations to form the decoder memory bank. At each step, an autoregressive hidden state $\mathbf{g}_t$ (updated via a GRU cell) is projected into a query, and the memory bank provides keys and values via standard learned projections. The attention logit combines three terms:

$$\ell_t[j] = \frac{\mathbf{q}_t^\top \mathbf{k}_j}{\sqrt{d}} + \beta^{\text{geo}}(t,j) + \gamma f_\theta(\mathbf{f}_{tj}). \quad (5)$$

The first term is the semantic attention score. The **2D-RPE bias** $\beta^{\text{geo}}(t,j)$ is a learned 2D relative positional bias on $\mathcal{M}$. The **explicit geometry score** projects the $D_g$-dimensional pairwise geometric feature vector $\mathbf{f}_{tj}$ (Section 3.2.2) through a two-layer MLP $f_\theta$, scaled by a learnable factor $\gamma$.

**Hard Causal Constraint.** To impose the sequential consistency constraint (Eq. 2), we apply an additive causal mask to the attention logits:

$$m_t[j] = \begin{cases} 0, & j \in \mathcal{A}_t, \\ -\infty, & \text{otherwise}. \end{cases} \quad (6)$$

The mask assigns zero probability to the current node and all nodes outside $\mathcal{A}_t$. Thus each non-root decoded edge points to an earlier node in the imposed order, preventing self-loops, forward pointers, and directed cycles among document nodes.

#### 3.4.3 Geometric Candidate Pruning

Although the causal mask restricts attention to prefix $\mathcal{A}_t$, the prefix grows linearly with $t$, yielding $O(|\mathcal{A}_t|)$ cost per step and $O(N^2)$ total. We reduce

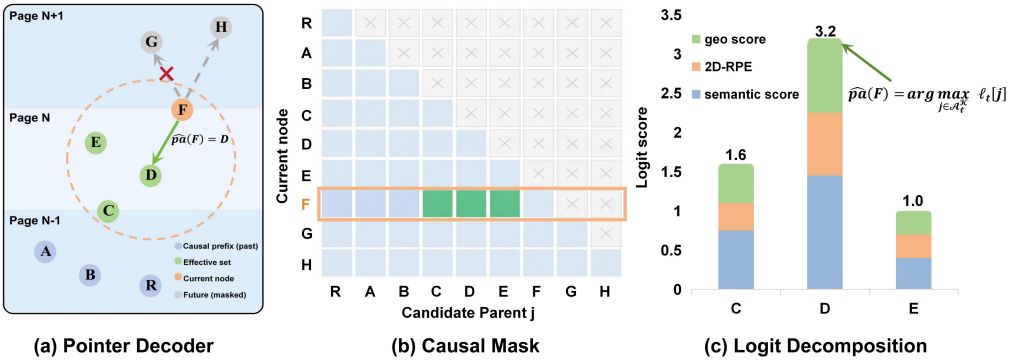


Figure 3: **Constrained parent decoding.** The pointer decoder selects a parent from the causal prefix; geometric pruning narrows candidates and the score combines semantic, 2D-RPE, and geometry terms.

this to $O(K)$ per step via geometric candidate pruning.

For the current node $i = \pi(t)$, we retrieve the candidate set $\mathcal{C}_i$ pre-computed by the auxiliary edge scorer: the top-$K$ nodes ranked by global geometric compatibility on the scroll space $\mathcal{M}$. The effective attention set is then:

$$\mathcal{A}_t^K = \mathcal{A}_t \cap \big(\{0\} \cup \mathcal{C}_i\big), \tag{7}$$

and all positions $j \notin \mathcal{A}_t^K$ receive mask value $-\infty$. Since $|\mathcal{C}_i| = K$ and $K \ll N$, we have $|\mathcal{A}_t^K| \leq K + 1$.

**Complexity.** Under geometric candidate pruning with budget $K$, the autoregressive decoding complexity reduces from $O(N^2)$ to $O(NK)$. At each step, the decoder computes attention and geometric biases only over the pruned candidate set $\mathcal{A}_t^K$ (at most $K+1$ elements). The auxiliary edge scorer retains $O(N^2)$ complexity but runs as a batched parallel operation outside the sequential loop. Empirically, end-to-end latency reduces by 36.5% under a matched protocol (Fig. 4).

**Decoder State and Root Classification.** After each decoding step, a GRU cell (Cho et al., 2014) updates the hidden state $\mathbf{g}_t$ by consuming the concatenation of the current node representation and the attention-weighted parent context, with a learnable residual connection for stability. An independent root classifier produces a calibrated logit that is additively injected into the pointer score at the root column, decoupling the root decision from the relative ranking of non-root candidates. Detailed formulations are provided in Appendix A.

### 3.5 Training and Inference

#### 3.5.1 Training Objective

We train Scroll-Net with a multi-task objective:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{root}} &+ \lambda_{\text{parent}}\mathcal{L}_{\text{parent}} \\ &+ \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}} \\ &+ \lambda_{\text{RO}}\mathcal{L}_{\text{RO}} + \lambda_{\text{contrast}}\mathcal{L}_{\text{contrast}}.\end{aligned} \tag{8}$$

The terms supervise root prediction, parent selection, edge classification, path consistency, reading-order prediction, and contrastive alignment. Full definitions and weights are in Appendix A.

#### 3.5.2 Hierarchical Structure Loss

Predicting whether a node is a root and selecting its parent are distinct tasks. Merging them into a single $(N+1)$-way softmax allows the $N$ non-root alternatives to dominate the root class. We therefore supervise root classification and parent selection separately with $\mathcal{L}_{\text{root}}$ and $\mathcal{L}_{\text{parent}}$ in Eq. 8.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** HRDoc-Simple and HRDoc-Hard (Ma et al., 2023) comprise academic papers with simple and complex layouts, respectively. E-Periodica (Rausch et al., 2023) contains historical periodicals with challenging visual characteristics. DocHieNet (Xing et al., 2024) provides diverse real-world documents with fine-grained block-level hierarchies. For fine-grained cross-page and component analyses (Table 3), we additionally adopt Comp-HRDoc (Wang et al., 2024b), which extends HRDoc-Hard with multi-page annotations supporting hierarchical structure reconstruction. All results are reported on the standard held-out test partition. We use all existing artifacts only for research

Table 1: **Main results across four HDSA benchmarks.** Link F1 is micro-F1 over directed child→parent edges; TEDS measures normalized tree-edit similarity. Bold and underlined figures indicate the best and second-best results, respectively. "–" indicates the metric was not reported by the original authors under this protocol.

| Method | HRDS | HRDH | E-Periodica | DocHieNet |
|---|---|---|---|---|
| | Link F1↑ / TEDS↑ | Link F1↑ / TEDS↑ | Link F1↑ / TEDS↑ | Link F1↑ / TEDS↑ |
| DocParser (Rausch et al., 2019) | 56.8 / 28.7 | 47.4 / 22.4 | 35.2 / 18.7 | 23.3 / 6.8 |
| DSG (Rausch et al., 2023) | 84.8 / 83.2 | 74.0 / 64.3 | 67.2 / 60.1 | 53.5 / 33.9 |
| DOC (Wang et al., 2024b) | – / 95.1 | – / 85.5 | – / – | – / – |
| DHFormer (Xing et al., 2024) | **99.3 / 98.7** | 93.4 / 89.1 | **92.5** / 84.9 | 77.8 / 57.6 |
| Scroll-Net-B (ours) | 95.3 / 91.2 | 90.5 / 87.4 | 87.9 / 83.5 | 78.3 / 61.6 |
| **Scroll-Net-L (ours)** | 98.2 / 96.5 | **94.2 / 89.8** | 91.5 / **87.4** | **80.1 / 62.4** |

Table 2: **Comparison on Comp-HRDoc under UniHDSA's STEDS protocol.** Results for DOC and UniHDSA are from their original papers.

| Method | Micro↑ | Macro↑ |
|---|---|---|
| DOC-R18 (Wang et al., 2024b) | 83.7 | 83.7 |
| UniHDSA-R18 (Wang et al., 2025) | 88.0 | 87.8 |
| UniHDSA-R50 (Wang et al., 2025) | 88.9 | 88.6 |
| Scroll-Net-B (ours) | 87.5 | 87.3 |
| **Scroll-Net-L (ours)** | **89.4** | **89.1** |

evaluation under their released benchmark protocols and do not redistribute derivative data.

**Evaluation Metrics.** We evaluate hierarchy quality using Link F1, computed as micro-F1 over directed child→parent edges, and Root F1 for identifying roots under the virtual root $r$. This Link F1 is directly comparable to the relation-triple F1 reported in prior work (Rausch et al., 2023) because HDSA uses a single relation type (`parent-of`). For reading order, we report Kendall's $\tau$ between the predicted permutation and the benchmark order under a deterministic linearization (sorting by $(p_i, \mathrm{ro}_i, i)$ to break ties). Finally, TEDS (Gao et al., 2019) measures normalized tree-edit similarity between predicted and ground-truth hierarchies.

**Implementation Details.** Scroll-Net uses a dual-stream encoder ($d$=128, $L_f$=4), an 8-head pointer decoder, and geometric pruning $K$=64. Semantic features are extracted via LayoutLMv3 (Huang et al., 2022) (BASE and LARGE variants) with the top-2 Transformer layers unfrozen. We train for 40 epochs on two NVIDIA GeForce RTX 3090 GPUs using AdamW (lr=$2\times10^{-4}$). Full training details are in Appendix A for reproducibility.

**Input Blocks.** Layout blocks (bounding boxes and reading order) are obtained from dataset-provided annotations; we do not train or evaluate the upstream detection stage. Block-level semantic features are extracted by LayoutLMv3 from the annotated bounding boxes and associated text.

**Experimental Protocol.** All variants share the same backbone and splits; we change only the component under study and compute Comp-HRDoc metrics offline from exported predictions. Our quantitative comparisons focus on methods with compatible HDSA or STEDS evaluation protocols, while Section 2 discusses broader multi-page document models with different output spaces or evaluation settings. We report UniHDSA separately under its released STEDS protocol in Table 2.

### 4.2 Main Results

Table 1 presents comparisons across four benchmarks. Scroll-Net achieves the best results on HRDoc-Hard (94.2 Link F1, 89.8 TEDS) and DocHieNet (80.1 Link F1, 62.4 TEDS). On E-Periodica, Scroll-Net-L surpasses DHFormer in TEDS (87.4 vs. 84.9) despite trailing in Link F1 (91.5 vs. 92.5). On HRDoc-Simple, DHFormer retains a lead (99.3 vs. 98.2), as near-deterministic layouts are well-captured by dense attention. Table 2 provides a protocol-matched Comp-HRDoc comparison.

### 4.3 Ablation Study

**Cross-Page Analysis.** Table 3(a) shows that scroll-space mapping substantially improves cross-page edge prediction compared to page-local baselines.

**Component Ablations.** Table 3(b) shows that the scroll-space frame yields the largest gain, identifying cross-page geometric comparability as the primary source of improvement. Geometric bias, sequential consistency, and contrastive alignment (Eq. 4) each add complementary improvements, with the latter contributing +0.8 Link F1 and +1.0 TEDS.

Table 3: **Ablation results on Comp-HRDoc.** (a) Cross-page performance, (b) cumulative components under oracle RO, and (c) mask robustness.

**(a) Cross-page performance**

| Model | Intra | CPE | Overall | Pred Cross |
|---|---|---|---|---|
| Single-Page | 88.0 | — | 56.8 | — |
| Page-Emb | 88.5 | 81.5 | 84.8 | 55.8 |
| **Scroll-Net** | **92.8** | **88.4** | **90.5** | **52.5** |

**(b) Cumulative component ablation**

| Component added | Link | CPE | TEDS |
|---|---|---|---|
| Single-page baseline | 56.8 | — | 43.2 |
| + Scroll-space frame | 82.4 | 78.6 | 76.5 |
| + Geometric bias | 86.9 | 84.3 | 82.8 |
| + Sequential consistency | 89.7 | 87.5 | 86.4 |
| + Contrastive alignment | **90.5** | **88.4** | **87.4** |

**(c) Mask robustness**

| Model | Mask | RO | Link | CPE |
|---|---|---|---|---|
| Page-Emb | None | Oracle | 84.8 | 81.5 |
| Scroll-Net | Soft | Predicted | 62.0 | 55.3 |
| Scroll-Net | None | Predicted | 84.2 | 80.8 |
| Scroll-Net | Hard | Predicted | 87.4 | 83.9 |
| **Scroll-Net** | **Hard** | **Oracle** | **90.5** | **88.4** |

**Robustness to Predicted Reading Order.** In Table 3(c), the hard mask is consistently stronger than soft-mask and no-mask settings under predicted RO. Moving from oracle to predicted reading order yields only a moderate drop ($-3.1$ Link F1). Although noisy predicted RO can exclude a true parent, the hard mask remains beneficial overall by trading a bounded recall ceiling for structural validity and more stable decoding.

**Additional Ablations.** Appendix C reports leave-one-out ablations: removing scroll-space mapping causes the largest single-component drop ($-7.7$ Link F1), followed by sequential consistency ($-3.6$) and geometric bias ($-3.1$). Removing the contrastive alignment loss (Eq. 4) decreases Link F1 by 1.4 points (Appendix C, Table 5), supporting its complementary role.

### 4.4 Efficiency Analysis

**Efficiency.** With $K$=64, pruning reduces latency by 36.5% and peak memory by 46.5% without measurable loss in Link F1/CPE-F1 (Fig. 4 and Table 4). Geometric candidate recall@64 is 95.6% and rises to 98.9% at $K$=256 (Appendix E, Table 9).

Representative qualitative examples are provided in Appendix F.

### 4.5 Sequential Consistency Analysis

We report this fine-grained analysis on Comp-HRDoc. The ground-truth DFS violation rate is below 1%, and under predicted RO the hard mask yields +3.2 Link F1 over unconstrained decoding (Table 3c).

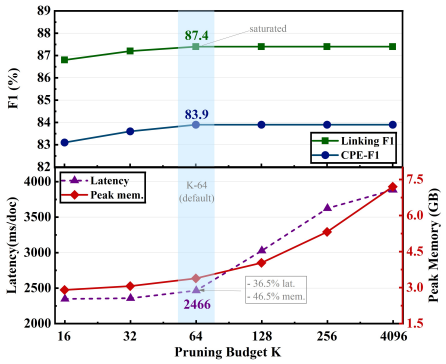


Figure 4: **Accuracy–efficiency trade-off on Comp-HRDoc.** Accuracy saturates at $K$=64, while latency and peak memory increase with $K$.

Table 4: **Efficiency metrics at different pruning budgets.** Memory (MB), latency (ms), and hierarchy scores; bold numbers indicate lowest cost.

| Setting | Cand. | Mem. | Lat. | Link | CPE |
|---|---|---|---|---|---|
| $K$=64 | 69.8 | **3384** | **2466** | 87.4 | 83.9 |
| $K$=256 | 161.9 | 5318 | 3624 | 87.4 | 83.9 |
| No Pruning | 167.5 | 6324 | 3884 | 87.4 | 83.9 |

### 4.6 Discussion

**Role of Scroll-Space Geometry.** The leave-one-out ablation (Appendix C) confirms that scroll-space mapping is the single most impactful component. These results suggest that cross-page geometric comparability helps causal constraints and geometric scoring add complementary gains.

## 5 Conclusion

We present Scroll-Net, a multi-page HDSA framework that maps page-local layouts into a unified scroll-space frame and combines geometry-aware encoding with efficient $O(NK)$ parent decoding. The scroll-space mapping restores geometric continuity across page boundaries, while geometric candidate pruning and a hard causal mask jointly guarantee acyclic tree structures with negligible accuracy loss. Across four benchmarks, Scroll-Net remains competitive on simple layouts and achieves state-of-the-art results on structurally complex documents. These results demonstrate that restoring geometric comparability is essential for robust multi-page document understanding.

## Limitations

Although Scroll-Net performs well across multiple HDSA benchmarks, multi-page document structure analysis remains constrained by the diversity of available evaluation settings. Existing datasets cover important document families, but they cannot fully reflect the long tail of documents encountered in real-world applications, including heterogeneous templates, multilingual reports, heavily annotated pages, degraded scans, and layouts with multiple plausible reading paths. Practical systems also depend on upstream OCR and layout parsing pipelines, whose quality can vary substantially across domains. Future work could expand evaluation to more diverse and challenging document collections and study robustness under different upstream parsing conditions, so hierarchy parsers can generalize more reliably to unseen document types.

## References


Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *CoRR*, abs/2004.05150.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural Optical Understanding for Academic Documents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yufan Chen, Ruiping Liu, Junwei Zheng, Di Wen, Kunyu Peng, Jiaming Zhang, and Rainer Stiefelhagen. 2025. Graph-based Document Structure Analysis. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL.

ShunLiang Fu, Yanxin Zhang, Yixin Xiang, Xiaoyu Du, and Jinhui Tang. 2026. DMAP: Human-Aligned Structural Document Map for Multimodal Document Understanding. *Preprint*, arXiv:2601.18203.

Liangcai Gao, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Maria Lang. 2019. ICDAR 2019 Competition on Table Detection and Recognition (cTDaR). In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1510–1515. IEEE.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10767–10775. AAAI Press.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4083–4091. ACM.

Yiqiao Jin, Rachneet Kaur, Zhen Zeng, Sumitra Ganesh, and Srijan Kumar. 2025. SlideAgent: Hierarchical Agentic Framework for Multi-Page Visual Document Understanding. *Preprint*, arXiv:2510.26615.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-Free Document Understanding Transformer. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 18893–18912. PMLR.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *Preprint*, arXiv:2407.07895.

Liangcheng Li, Feiyu Gao, Jiajun Bu, Yongpan Wang, Zhi Yu, and Qi Zheng. 2020a. An End-to-End OCR Text Re-organization Sequence Learning for Rich-Text Detail Image Comprehension. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV*, volume 12370 of *Lecture Notes in Computer Science*, pages 85–100. Springer.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020b. DocBank:

A Benchmark Dataset for Document Layout Analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 949–960, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhanli Li, Huiwen Tian, Lvzhou Luo, Yixuan Cao, and Ping Luo. 2026. DeepRead: Document Structure-Aware Reasoning to Enhance Agentic Search. *Preprint*, arXiv:2602.05014.

Zichao Li, Shaojie He, Meng Liao, Xuanang Chen, Yaojie Lu, Hongyu Lin, Yanxiong Lu, Xianpei Han, and Le Sun. 2024b. Seg2Act: Global Context-aware Action Generation for Document Logical Structuring. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18077–18088. Association for Computational Linguistics.

Fuyuan Liu, Dianyu Yu, He Ren, Nayu Liu, Xiaomian Kang, Delai Qiu, Fa Zhang, Genpeng Zhen, Shengping Liu, Jiaen Liang, Wei Huang, Yining Wang, and Junnan Zhu. 2026. FocalOrder: Focal Preference Optimization for Reading Order Detection. *CoRR*, abs/2601.07483.

Jiefeng Ma, Jun Du, Pengfei Hu, Zhenrong Zhang, Jianshu Zhang, Huihui Zhu, and Cong Liu. 2023. HRDoc: Dataset and Baseline Method toward Hierarchical Reconstruction of Document Structures. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 1870–1877. AAAI Press.

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard H. Hovy. 2018. Stack-Pointer Networks for Dependency Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1403–1414. Association for Computational Linguistics.

Gonzalo E. Mena, David Belanger, Scott W. Linderman, and Jasper Snoek. 2018. Learning Latent Permutations with Gumbel-Sinkhorn Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Johannes Rausch, Octavio Martinez, Fabian Bissig, Ce Zhang, and Stefan Feuerriegel. 2019. DocParser: Hierarchical Structure Parsing of Document Renderings. *CoRR*, abs/1911.01702.

Johannes Rausch, Gentiana Rashiti, Maxim Gusev, Ce Zhang, and Stefan Feuerriegel. 2023. DSG: An End-to-End Document Structure Generator. In *IEEE International Conference on Data Mining, ICDM 2023, Shanghai, China, December 1-4, 2023*, pages 518–527. IEEE.

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. *CoRR*, abs/2103.15348.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2022. Hierarchical multimodal transformers for Multi-Page DocVQA. *CoRR*, abs/2212.05935.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2692–2700.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024a. DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8529–8548, Bangkok, Thailand. Association for Computational Linguistics.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022. LiLT: A Simple yet Effective Language-Independent Layout Transformer for Structured Document Understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7747–7757. Association for Computational Linguistics.

Jiawei Wang, Kai Hu, and Qiang Huo. 2025. UniHDSA: A unified relation prediction approach for hierarchical document structure analysis. *Pattern Recognit.*, 165:111617.

Jiawei Wang, Kai Hu, Zhuoyao Zhong, Lei Sun, and Qiang Huo. 2024b. Detect-order-construct: A tree construction based approach for hierarchical document structure analysis. *Pattern Recognition*, 156:110836.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. LayoutReader: Pre-training of Text and Layout for Reading Order Detection. *CoRR*, abs/2108.11591.

Hangdi Xing, Changxu Cheng, Feiyu Gao, Zirui Shao, Zhi Yu, Jiajun Bu, Qi Zheng, and Cong Yao. 2024. DocHieNet: A Large and Diverse Dataset for Document Hierarchy Parsing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA,*

*November 12-16, 2024*, pages 1129–1142. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of Text and Layout for Document Image Understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1192–1200. ACM.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Chong Zhang, Ya Guo, Yi Tu, Huan Chen, Jinyang Tang, Huijia Zhu, Qi Zhang, and Tao Gui. 2023. Reading Order Matters: Information Extraction from Visually-rich Documents by Token Path Prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13716–13730, Singapore. Association for Computational Linguistics.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. PubLayNet: Largest Dataset Ever for Document Layout Analysis. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1015–1022. IEEE.

# A Implementation Details

**Model and Optimization.** Scroll-Net uses a dual-stream encoder with hidden dimension $d = 128$, $L_f = 4$ cross-attention fusion layers, and a pointer decoder with 8 attention heads. We set the geometric pruning budget to $K = 64$ and use a log-ratio geometric bias. Semantic features are extracted via LayoutLMv3 (BASE and LARGE variants) with the top-2 Transformer layers unfrozen. We train Scroll-Net with AdamW using a learning rate of $2 \times 10^{-4}$ and cosine decay for 40 epochs, with batch size 8 and gradient clipping at 1.0 on two NVIDIA GeForce RTX 3090 GPUs. The loss weights are $\lambda_{\text{parent}} = 1.2$, $\lambda_{\text{edge}} = 0.3$, $\lambda_{\text{path}} = 1.2$, $\lambda_{\text{RO}} = 1.0$, and $\lambda_{\text{contrast}} = 0.5$.

**Root Classification ($\mathcal{L}_{\text{root}}$).** The independent root classifier $\rho_t = \text{MLP}_{\text{root}}(\mathbf{h}_{\pi(t)})$ uses weighted binary cross-entropy on the calibrated logit $\tilde{\rho}_t = \alpha_r \rho_t + \beta_r$:

$$\mathcal{L}_{\text{root}} = -\frac{1}{N}\sum_{t=1}^{N}\Big[w_+ y_t^r \log\sigma(\tilde{\rho}_t) + (1-y_t^r)\log\big(1-\sigma(\tilde{\rho}_t)\big)\Big], \tag{9}$$

where $y_t^r \in \{0,1\}$ is the ground-truth root label, $\sigma$ is the sigmoid function, and the positive weight $w_+ = (1-\hat{\rho})/\hat{\rho}$ adapts to the batch root ratio $\hat{\rho}$. At inference, the same $\tilde{\rho}_t$ is injected into the pointer logits at the root column via $\ell_t[0] \leftarrow \tilde{\rho}_t + \max_{j>0} \ell_t[j]$. During training, $\alpha_r$ and $\beta_r$ are learned through $\mathcal{L}_{\text{root}}$, while $\mathcal{L}_{\text{parent}}$ is computed only over non-root candidates.

**Parent Selection ($\mathcal{L}_{\text{parent}}$).** For non-root nodes $\mathcal{S} = \{t : y_t^r = 0\}$, we apply cross-entropy over the pointer logits:

$$\mathcal{L}_{\text{parent}} = -\frac{1}{|\mathcal{S}|}\sum_{t\in\mathcal{S}}\log\frac{\exp\big(\ell_t[y_t^*]\big)}{Z_t}, \qquad Z_t = \sum_{j\in\mathcal{A}_t^K\setminus\{0\}}\exp\big(\ell_t[j]\big). \tag{10}$$

where $y_t^*$ is the ground-truth parent index in the reading-order sorted vocabulary, and the normalizer sums over non-root candidates in the masked set $\mathcal{A}_t^K$ (excluding the virtual root 0), since root decisions are supervised by $\mathcal{L}_{\text{root}}$ and $y_t^* \neq 0$ for all $t \in \mathcal{S}$.

## A.1 Auxiliary Losses

**Edge Classifier Loss ($\mathcal{L}_{\text{edge}}$).** An auxiliary pairwise edge classifier (operating on original node order) provides dense gradient signal:

$$\mathcal{L}_{\text{edge}} = -\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big[y_{ij}\log\sigma(e_{ij}) + (1-y_{ij})\log\big(1-\sigma(e_{ij})\big)\Big], \tag{11}$$

where $y_{ij} = \mathbf{1}[\text{pa}^*(b_i) = b_j]$ and $e_{ij}$ is the edge logit.

**Path Consistency Penalty ($\mathcal{L}_{\text{path}}$).** While the hard causal mask zeroes out invalid edges during parent selection, the unmasked pointer logits may still assign large scores to future candidates. We therefore penalize raw logits at positions that violate the reading order:

$$\mathcal{L}_{\text{path}} = \frac{1}{|\mathcal{V}|}\sum_{(t,j)\in\mathcal{V}}\max\big(0,\ \ell_t[j]\big), \qquad \mathcal{V} = \big\{(t,j) : j \in \{1,\dots,N\},\ j \geq t\big\}. \tag{12}$$

Here $j$ indexes positions in the RO-sorted memory $\mathbf{M} = [\mathbf{r}_0; \mathbf{h}_{\pi(1)}; \dots; \mathbf{h}_{\pi(N)}]$, excluding the root at index 0. Memory position $j \in \{1,\dots,N\}$ corresponds to node $\pi(j)$; equivalently, $\pi^{-1}(\pi(j)) = j$ by construction. Therefore, 'self-or-future' positions for step $t$ are exactly those with $j \geq t$. This regularizer is applied to the unmasked logits and complements the hard causal mask used during decoding.

**Reading Order Loss ($\mathcal{L}_{\text{RO}}$).** The Sinkhorn head produces a doubly-stochastic matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$ via iterative row–column normalisation. We supervise it with a permutation cross-entropy:

$$\mathcal{L}_{\text{RO}} = -\frac{1}{N}\sum_{i=1}^{N}\log\mathbf{S}\big[i,\ \pi^*(i)\big], \tag{13}$$

where $\pi^*(i)$ is the ground-truth position of node $b_i$.

**Contrastive Alignment Loss ($\mathcal{L}_{\text{contrast}}$).** We use the contrastive alignment loss defined in the main text.

## A.2 Full Objective

The complete training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{root}} + \lambda_{\text{parent}}\mathcal{L}_{\text{parent}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}} + \lambda_{\text{RO}}\mathcal{L}_{\text{RO}} + \lambda_{\text{contrast}}\mathcal{L}_{\text{contrast}}. \tag{14}$$

**Algorithm 1** Scroll-Net Inference

**Require:** Node features $\{\mathbf{x}_i, \mathbf{s}_i, p_i\}_{i=1}^{N}$
**Ensure:** Tree $\hat{\mathcal{T}} = (V, \hat{E}, r, \hat{\pi})$
$\tilde{\mathbf{x}}_i \leftarrow \Phi(\mathbf{x}_i, p_i)$ for all $i$
$\mathbf{H} \leftarrow \mathcal{E}(\{\tilde{\mathbf{x}}_i\}, \{\mathbf{s}_i\})$
$\mathbf{S} \leftarrow \mathrm{Sinkhorn}(\mathbf{H})$
$\hat{\pi} \leftarrow \mathrm{Hungarian}(\mathbf{S})$
Precompute $\mathcal{C}_i$ for all nodes using the auxiliary scorer
$\mathbf{g}_1 \leftarrow \mathbf{r}_0$
**for** $t = 1, \ldots, N$ **do**
  $i \leftarrow \hat{\pi}(t)$
  $\mathcal{A}_t^K \leftarrow \{0\} \cup \big(\mathcal{C}_i \cap \{\hat{\pi}(1), \ldots, \hat{\pi}(t-1)\}\big)$
  Compute logits $\ell_t[j]$ for $j \in \mathcal{A}_t^K$ using the unified pointer score
  Inject calibrated root score into $\ell_t[0]$
  $\widehat{\mathrm{pa}}(i) \leftarrow \arg\max_{j \in \mathcal{A}_t^K} \ell_t[j]$
  Update $\mathbf{g}_{t+1} \leftarrow \mathrm{GRU}([\mathbf{h}_{\hat{\pi}(t)}; \mathbf{c}_t],\ \mathbf{g}_t)$
**end for**
**return** $\hat{\mathcal{T}}$

where the $\lambda$ values are trade-off weights balancing the multi-task objectives.

### A.3 Target Inclusion During Training

During training, geometric candidate pruning may exclude the ground-truth parent from $\mathcal{A}_t^K$. If the annotated parent falls outside the pruned candidate set, it is included for loss computation during training. This affects training only and does not alter the inference-time candidate set.

### A.4 Inference Algorithm

At test time, Scroll-Net performs joint inference in two sequential stages:
At inference, the parent decision is made over the unified candidate set $\mathcal{A}_t^K$ after injecting the calibrated root logit, ensuring consistency with the decoupled training objective. The representation stage (scroll-space mapping + encoder) has $O(N^2)$ complexity from self-attention. Candidate precomputation is $O(N^2)$ but fully parallel. The autoregressive pointer decoding therefore runs in $O(NK)$ time under geometric candidate pruning.

## B Additional Analyses

**Cross-Page Evaluation Details. CPE-F1 Definition.** We define the cross-page edge set as $E_{\text{cross}} = \{(c \to p) \mid \text{page}(c) \neq \text{page}(p),\ p \neq r\}$. CPE-F1 is the micro-F1 computed over $E_{\text{cross}}$; if a model predicts no cross-page edges while the ground truth contains cross-page edges, CPE-F1 is 0. **Configuration Details.** The Single-Page variant uses page-local normalized coordinates $[0, 1]^2$ and restricts candidate parents to the same page. The Page-Emb variant keeps page-local coordinates but adds learnable page embeddings and allows cross-page candidates. Scroll-Net applies scroll-space mapping to obtain continuous global coordinates and uses geometric pruning with log-ratio bias. **Observed Results.** In the cross-page comparison, Single-Page predicts no cross-page edges. Page-Emb obtains CPE-F1 81.5 with Cross% 55.8, while Scroll-Net reaches CPE-F1 88.4 with Cross% 52.5.

Table 5: **Leave-one-out ablation on Comp-HRDoc (evaluated under oracle RO).** Each row removes a single component from the full model. Bold figures indicate the best results.

| Variant | Link | CPE | Δ |
|---|---|---|---|
| Full Model | **90.5** | **88.4** | — |
| w/o Scroll-space mapping | 82.8 | 77.2 | −7.7 |
| w/o Sequential Consistency | 86.9 | 84.3 | −3.6 |
| w/o Geometric bias | 87.4 | 84.8 | −3.1 |
| w/o Dual-Stream Encoder | 87.8 | 85.5 | −2.7 |
| w/o Contrastive Alignment | 89.1 | 86.8 | −1.4 |
| Sem-first Fusion Order | 89.6 | 87.2 | −0.9 |

**Robustness to Predicted Reading Order.** Under predicted RO, the soft, no-mask, and hard-mask settings obtain Link F1/CPE-F1 of 62.0/55.3, 84.2/80.8, and 87.4/83.9, respectively. Under oracle RO, the hard mask reaches 90.5/88.4.

**Oracle and End-to-End Protocols.** The cross-page analysis in the main text isolates geometric contributions by feeding ground-truth reading order to the decoder (Oracle RO). Experiments marked Predicted or E2E use the Sinkhorn-predicted reading order (mean Kendall's $\tau \approx 0.91$ on Comp-HRDoc), whereas experiments marked Oracle RO use the ground-truth order. Under E2E inference, Scroll-Net achieves 87.4 Link F1 and 83.9 CPE-F1. Under Oracle RO, the corresponding numbers are 90.5 and 88.4. The Page-Emb baseline with oracle RO obtains 84.8 Link F1.

## C Additional Ablations

**Leave-One-Out Analysis.** To complement the cumulative ablation, Table 5 removes each component individually from the full model, isolating its marginal contribution.
Removing scroll-space mapping causes the largest drop ($-7.7$ Link F1), indicating that globally comparable coordinates are the most impactful single component among the tested components on Comp-HRDoc. Sequential consistency ($-3.6$) and geometric bias ($-3.1$) rank second and third. Removing the dual-stream encoder costs $-2.7$, contrastive alignment $-1.4$, and reversing the fusion

Table 6: **Dataset statistics.** Cross-page edge ratio is macro-averaged over documents. E-Periodica is a single-page benchmark and is included for completeness.

| Dataset | #Docs | #Pages | Max P. | Cross % | Span |
|---|---|---|---|---|---|
| HRDoc | 2,500 | 31,651 | 35 | 24.9 | 2.4 |
| E-Periodica | 542 | 542 | 1 | 0.0 | 0.0 |
| DocHieNet | 1,673 | 15,610 | 50 | 37.4 | 5.4 |
| Comp-HRDoc | 1,500 | 21,408 | 35 | 48.7 | 3.6 |

Table 7: **Reading-order stability across datasets.** Kendall's $\tau$ is computed between predicted and annotated reading order on each test split.

| Dataset | Kendall's $\tau\uparrow$ |
|---|---|
| HRDoc-Simple | 0.97 |
| Comp-HRDoc | 0.91 |
| E-Periodica | 0.90 |
| DocHieNet | 0.87 |

order $-0.9$.

### C.1 Empirical Verification of Sequential Consistency

**Violation Statistics.** We compute the fraction of parent–child edges where the parent does not precede the child in a deterministic document-level reading order, using the deterministic linearization defined in Section 4.1 of the main text. On Comp-HRDoc, the overall violation rate is $0.57\%$. We do not observe cross-page violations in the annotated train (1,000 docs, 66,817 cross-page edges) or test (500 docs, 31,094 cross-page edges) splits.

## D Dataset Statistics

Table 6 summarizes the scale and cross-page characteristics of the evaluated benchmarks. For HRDoc, E-Periodica, and DocHieNet, we use the statistics reported by DocHieNet (Xing et al., 2024). For Comp-HRDoc, we compute the statistics from the official released annotations (Wang et al., 2024b).

## E Robustness Analysis

**Page-Height Perturbation.** The scroll mapping (Eq. 3 in the main text) assumes unit-normalized page heights. To evaluate robustness to heterogeneous page geometries, we apply random page-height perturbation at inference: each page $p$ is assigned a height $h_p \sim \mathrm{Uniform}(1-\sigma,\ 1+\sigma)$, and the mapped vertical coordinate becomes $\tilde{c}^y = \sum_{j<p} h_j + c^y \cdot h_p$. Table 8 reports results on Comp-HRDoc under predicted RO.

Table 8: **Robustness to page-height perturbation on Comp-HRDoc (predicted RO).** $\sigma$ denotes the maximum relative deviation from unit height.

| Perturbation | $\tau\uparrow$ | Link F1$\uparrow$ | CPE-F1$\uparrow$ | $\Delta$F1 |
|---|---|---|---|---|
| None (unit height) | 0.91 | 87.4 | 83.9 | — |
| $\pm5\%$ | 0.908 | 87.2 | 83.6 | $-0.2$ |
| $\pm10\%$ | 0.905 | 86.9 | 83.2 | $-0.5$ |
| $\pm20\%$ | 0.895 | 86.1 | 82.2 | $-1.3$ |
| $\pm30\%$ | 0.882 | 85.1 | 80.8 | $-2.3$ |

Table 9: **Geometric candidate recall at different pruning budgets on Comp-HRDoc (predicted RO).**

| | $K$=32 | $K$=64 | $K$=256 |
|---|---|---|---|
| Candidate Recall | 91.2% | 95.6% | 98.9% |

At $\pm5\%$, Link F1 drops by only $0.2$. At $\pm20\%$, a substantial perturbation setting, the drop is $1.3$ and the result (86.1) remains above the unperturbed no-mask baseline (84.2 in Table 3(c) of the main text). At $\pm30\%$, Link F1 (85.1) still exceeds the unperturbed no-mask setting, suggesting gradual degradation under large geometric perturbations.

**Geometric Candidate Recall.** We report candidate recall@$K$ to quantify how often the ground-truth parent remains in the geometric candidate set $\mathcal{C}_i$ before reading-order masking. Table 9 reports results on Comp-HRDoc under end-to-end inference.

At the default budget $K$=64, 95.6% of ground-truth parents fall within the geometric candidate set before reading-order masking; increasing the budget to $K$=256 raises this coverage to 98.9%. These values isolate geometric candidate coverage and do not measure additional exclusions introduced by the predicted reading-order mask.

## F Qualitative Analysis

Figure 5 provides representative cross-page linking examples on dense multi-page documents.

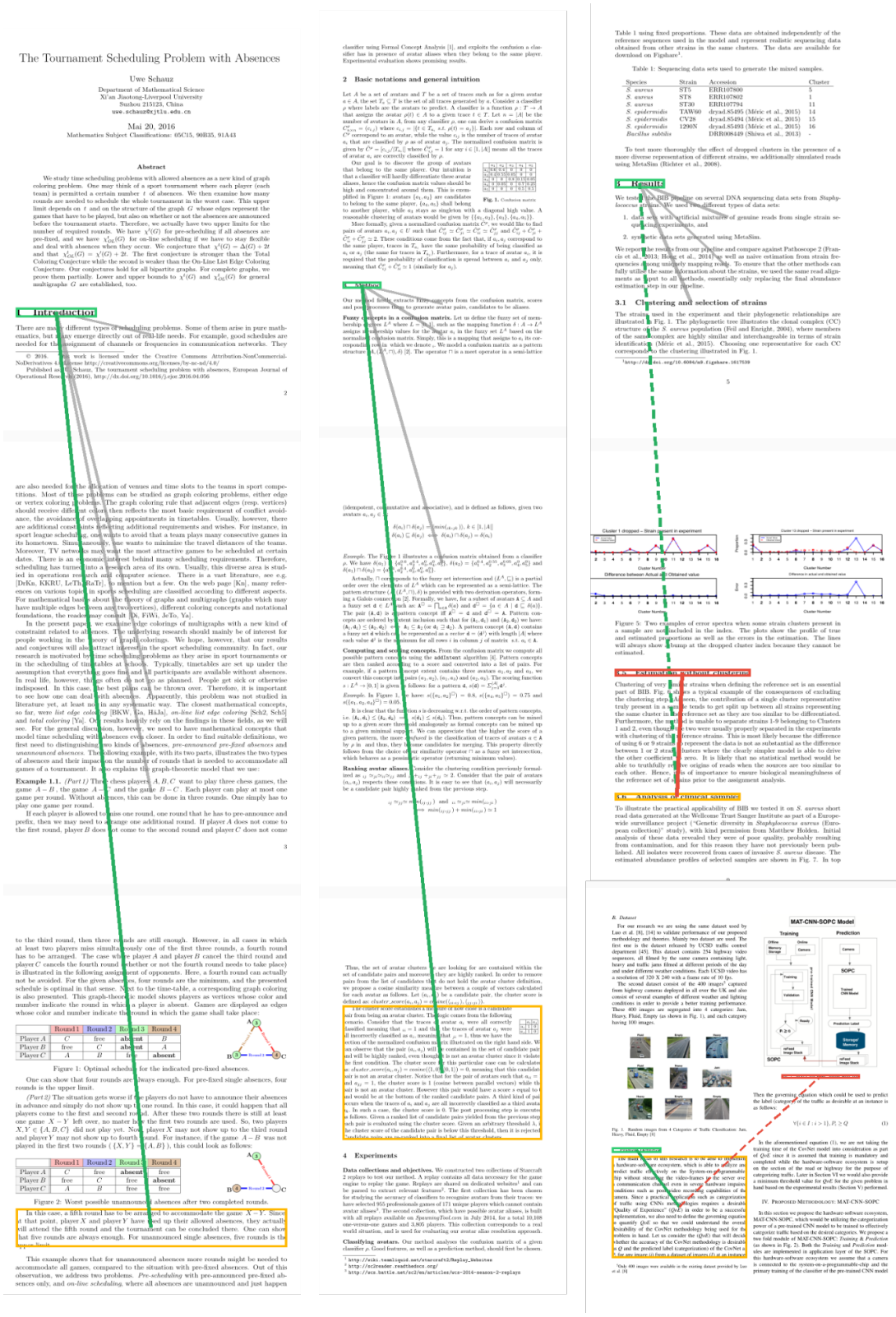


Figure 5: **Qualitative cross-page linking examples.** Green/red lines denote correct/incorrect predicted child→parent links.