HIERARCHICAL AND MULTIMODAL REPRESENTATION LEARNING FOR IRREGULAR AND LONG ESG REPORTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The principles of ESG (Environmental, Social, and Governance) are increasingly transforming the architecture of global financial governance. However, ESG reports, the main source for evaluating corporate ESG performance, are difficult to parse at scale because of (1) non-linear reading orders from fragmented, slide-like layouts and (2) hidden hierarchies within lengthy, weakly structured text. We introduce **Compass-ESG**, a unified framework that transforms ESG reports into structured representations through three core innovations: (1) *reading order modeling*, which integrates a page-level layout framework with block-level sequence ordering to recover coherent global-to-local flows; (2) *ToC-guided hierarchical reconstruction*, where ToC-RAP parses visually complex tables of contents and ToC-ALIGN anchors entries to body content, enabling accurate recovery of explicit and implicit hierarchies; and (3) *context-aware visual-to-text representation*, which integrates visual and structural cues under hierarchical guidance to transform images into grounded natural language. Extensive experiments on annotated benchmarks show that Compass-ESG significantly outperforms both specialized document parsers and general-purpose multimodal models. In addition, we release **ATLAS-ESG**, the first large-scale ESG dataset with multi-level annotations from China, Hong Kong, and the U.S., providing a landmark resource for structured ESG analysis and future research.

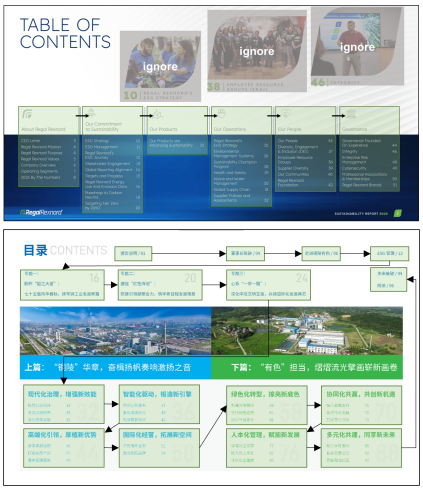
1 INTRODUCTION

With the rise of sustainable finance, ESG principles are becoming integral to capital allocation and regulatory oversight. As regulators move disclosures from voluntary to mandatory, ESG reporting now acts as a structural bridge among firms, investors, and regulators. However, ESG reports, the main medium for sustainability disclosures, are usually released as long, visually dense PDFs that pose two technical challenges. First, their layouts are highly varied, mixing text, tables, and charts in complex, slide-like formats, which complicates layout parsing and reading-order inference, even in structured sections such as directories (Figure 1a). Second, their content hierarchy is implicit: reports often exceed 40 pages and lack standardized cues like numbered headings or consistent formatting, making it difficult to recover structure (Figure 1b).

As a consequence of these challenges, financial research has largely fallen back on indirect ESG proxies, including disclosure indicators (Ni & Zhang, 2019), limited case studies (Florian Berg, 2022), and third-party ratings (Dane M. Christensen, 2022), thereby neglecting the semantic richness of the reports themselves. Meanwhile, parsers developed for structured domains such as academic articles, contracts, or forms (Palm et al., 2019; Xu et al., 2020; Huang et al., 2022) perform poorly on the irregular and implicit layouts typical of ESG reports.

To address these challenges, we present **Compass-ESG**, a unified framework for hierarchical and multimodal representation learning on complex ESG reports. As illustrated in the upper part of Figure 2, it consists of three core components that together provide a foundation for quantitative ESG analysis: ***(a) Reading order modeling:*** integrates deterministic page-level layout heuristics with data-driven block-level sequence ordering to reconstruct a coherent global-to-local reading flow. ***(b) ToC-guided hierarchical structure reconstruction:*** leverages table-of-contents (ToC) entries via region-aware prompting and anchor-based alignment to reconstruct explicit and implicit hierarchies in ultra-long ESG reports. ***(c) Context-aware visual-to-text representation:*** aligns images with broader document semantics via hierarchy-guided aggregation and context-guided reasoning, enabling faithful text representations in ESG reports.

The system produces a unified, hierarchical, and multimodal representation of ESG reports, in which text, tables, and images are integrated along the reconstructed reading order and enriched with structural and semantic annotations. Evaluations on expert-annotated benchmarks show that Compass-ESG achieves superior performance to both specialized parsers (e.g., MinerU (Huang et al., 2022), Marker (Tito et al., 2023), Docling (Kim et al., 2023)) and general-purpose multimodal models (e.g., ChatGPT5, Gemini 2.5 Pro, Doubao) across multiple tasks.



(a) Layout Diversity and Reading Order

(b) Implicit hierarchy and structure recovery

Figure 1: Representative challenges in ESG report parsing.

In addition, to advance ESG research at scale and to promote the development of multimodal long-document analysis, we release **ATLAS-ESG** (**A**nnotated **T**extual **L**ayout **A**nd **S**tructure for **ESG**), the first and largest public ESG report dataset, built on Compass-ESG outputs. ATLAS-ESG spans over 26K reports from China, Hong Kong, and the U.S., comprising more than 8 million content blocks, and establishes a landmark resource for future advances in ESG analytics and representation learning. The lower part of Figure 2 illustrates the ATLAS-ESG JSON schema, which organizes raw ESG reports into hierarchical elements at the document, page, and content block levels, with representative fields and examples.

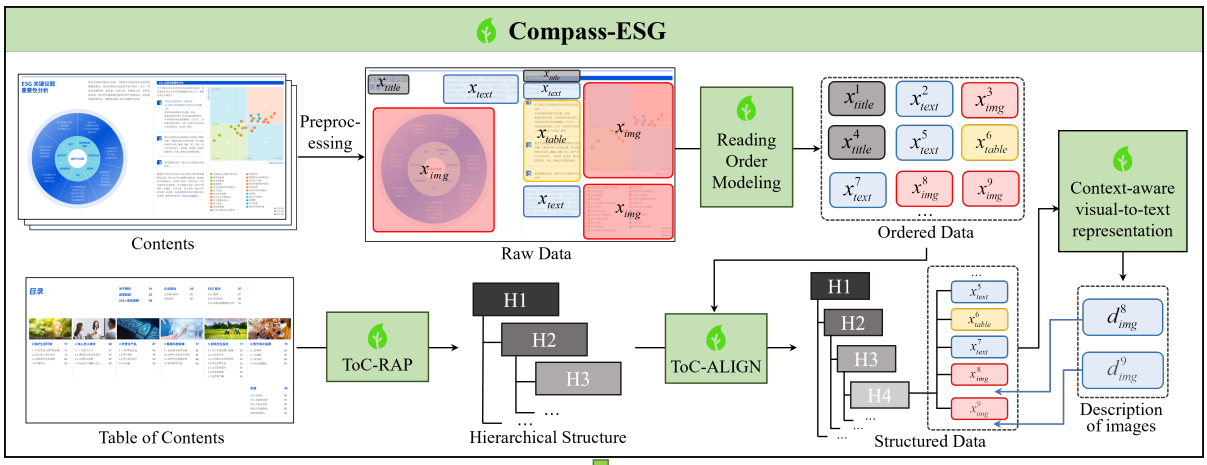
Contributions **(1)** We introduce Compass-ESG, the first framework capable of delivering a coherent interpretation of ultra-long ESG reports with highly irregular layouts and implicit hierarchies, thereby enabling an accurate and structured representation of ESG knowledge. **(2)** We extensively evaluate Compass-ESG in markets as distinct as the U.S. and Hong Kong, which differ in language, report format, and regulation. Across these challenging settings, Compass-ESG consistently proves robust to multilingual disclosures, irregular layouts, and diverse governance regimes. **(3)** We release ATLAS-ESG, the first large-scale dataset derived from ESG reports, designed to support research on parsing long, complex documents and to enable studies in green finance. **(4)** We open-source our code, models, and a subset of the ATLAS-ESG dataset: https://anonymous.4open.science/r/ICLR_2026-8358/. Given the dataset's scale, access to the full version is provided by the authors on request for academic purposes.

2 Related Work

Progress in document understanding has led to strong results on standard document benchmarks (Xu et al., 2020; Huang et al., 2022). In contrast, ESG reports remain difficult to process: slide-style layouts fragment the reading order, while verbose and loosely structured narratives obscure content hierarchy.

Specialized Parsing Systems and General-purpose Multimodal Models Specialized systems, such as Docling (Livathinos et al., 2025), MinerU (Wang et al., 2024), Marker (Datalab, 2025), and TextIn (intsig textin, 2023), typically rely on advanced layout analysis models, including EfficientViT (Liu et al., 2023), RT-DETR (Zhao et al., 2024), and LayoutLMv3 (Huang et al., 2022). These approaches achieve strong results on datasets with regular layouts, such as PubLayNet (Zhong et al., 2019), by converting documents into structured representations. However, their performance declines notably on structurally diverse benchmarks like DocLayNet (Pfitzmann et al., 2022), revealing limited generalization to ESG reports. Meanwhile, general-purpose multimodal models—including GPT (OpenAI), Gemini (Google), and DeepSeek (DeepSeek)—provide an end-to-end alternative by producing text directly from page images. Nevertheless, these models are prone to hallucinations and incur high computational costs when parsing hierarchies in long ESG reports with implicit structures.

Structural Modeling Challenges in ESG Reports Existing approaches often treat each page as an independent unit, making it difficult to exploit global context and identify hierarchical relations across pages (Huang et al., 2022; Xia et al., 2022). These methods also rely on layout signals such as numbering or indentation, which become unreliable in complex or inconsistent designs (Wang et al., 2023; Zhong et al., 2019; Xu et al., 2020). Recent studies introduced



ATLAS-ESG

Level	Category	Field	Description	Example
Document Level	Document Identifiers	stock_code	Stock code	300XXX.SZ
		company_name	Company name	FuXiang Pharma
		report_year	Reporting year	2024
		report_title	Title of the ESG report	ESG Report 2024
	Source Metadata	report_type	Report type	ESG Report
		market	Market the report belongs to	China
		original_filename	Original JSON file name	300XXX.SZ-...json
Page Level	Indexing	page_idx	Page index (starting from 1)	1, 2, 3...
	Rendering Links	page_markdown_url	Markdown image link	
		page_file_url / page_relative_path	Local image file path	/mnt/data/.../1.jpg
		page_http_url	Online accessible HTTP path	http://.../1.jpg
Content Block Level	Hierarchical depth	h1–h4	Structural depth of section headings	Environmental Management
	Content type and payload	data_type	Block type	text / table / image
		data	Block content	"...company 2024 energy..."
	Visual Resource Links	markdown_url	Markdown link (table/image)	
		file_url / relative_path	Local high-res path	/mnt/data/...jpg
		http_url	Online access path	http://.../12.jpg
	Ordering	reading_order	In-page reading order	0, 1, 2...
	Semantic tag	esg_category_label	ESG category label	E / S / G / N
		gri_label	One of the 32 GRI labels	Energy
		sentiment_label	Sentiment polarity	Positive / Neutral / Negative

Figure 2: Architecture of the Compass-ESG Pipeline and the Hierarchical JSON Schema of ATLAS-ESG.

Transformer-based frameworks for hierarchy reconstruction, exemplified by HRDoc (Marulanda et al., 2023) and DINO (Zhang et al., 2023). Nevertheless, attention mechanisms degrade on very long documents (e.g., over 20 pages), leading to recognition errors (Beltagy et al., 2020; Huang et al., 2022). These limitations are amplified in ESG reports, which are lengthy, provide weak or inconsistent structural cues, and show substantial variation across companies.

Challenges in Semantic Alignment for ESG Reports Transformer-based architectures, including DocFormer (Appalaraju et al., 2021), LayoutLMv3 (Huang et al., 2022), StrucTexT (Appalaraju et al., 2021), UDoc (Kim et al., 2023), and SelfDoc (Li et al., 2021), use cross-modal attention to link text with figures, tables, and other visual components. These models work well on documents with regular layouts. However, their effectiveness declines on ESG reports, where semantically related content is frequently scattered across distant positions, often spanning multiple pages.

ESG-Related Datasets To our knowledge, four recent datasets contribute to ESG research. Morio & Manning (2023) released more than 10,000 corporate climate policy documents. DynamicESG (Tseng et al., 2023) consists of 2,220 Taiwan-based ESG news articles. ESG-FTSE (Wang & Casey, 2024) contains 3,913 news articles on the UK's top 10 firms. A3CG (Ong et al., 2025) compiles ESG reports from 1,679 Singapore-listed companies. While valuable, existing datasets are limited by (i) the lack of complete long-form ESG reports, (ii) the omission of multimodal features such as tables and figures, and (iii) the absence of fine-grained annotations for downstream financial tasks.

3 METHODOLOGY

This section outlines the core innovations of Compass-ESG, with its overall architecture illustrated in Figure 2.

3.1 READING ORDER MODELING

Due to irregular layouts, multimodal elements, and fragmented blocks, ESG reports often resemble slide decks rather than traditional documents. To reconstruct a faithful reading sequence, Compass-ESG employs a hybrid framework where spatial heuristics capture page-level structure and relational modeling resolves block-level ordering, ultimately integrating macro-level layout with micro-level sequencing. (see Appendix A.2 for details) [1].

Page-level Layout Framework To determine the page's primary structural orientation, we estimate both Vertical–Manhattan partitions ($D_p \approx D_{\text{left}} \cup D_{\text{right}}$) and Horizontal–Manhattan partitions ($D_p \approx D_{\text{top}} \cup D_{\text{bottom}}$). The orientation is determined by comparing the resulting number of vertical columns (N_v) and horizontal bands (N_h), formalized as:

$$\text{RootCut}(p) = \begin{cases} \text{Horizontal–Manhattan} & N_h > N_v, \\ \text{Vertical–Manhattan} & N_v > N_h, \\ \text{Canonical XYCut} & \text{otherwise.} \end{cases} \tag{1}$$

where Canonical XYCut applies both a horizontal and a vertical split at the root. The selected orientation determines the first-level partition of D_p. We then apply an X–Y cut (Ha et al., 1995) to obtain a coarse region tree $T_p^{(1)} = \text{XYCut}(D_p)$, with nodes representing hierarchical subregions. Partitioning proceeds recursively: each subregion D_{sub} is re-evaluated for its dominant orientation and further split if meaningful, until no additional divisions are possible. This process yields the final hierarchical region tree $T_p^{(\text{final})}$, capturing both global and local layout organization.

Block-level Relational Sequence Modeling While $T_p^{(\text{final})}$ captures global structure, it does not determine fine-grained sibling order. We thus model child reordering as data-driven pairwise relation prediction in three steps:

(a) Pairwise Relational Modeling: For any pair of sibling blocks (D_i, D_j), we construct a rich multimodal feature vector ϕ_{ij} that encapsulates their spatial and semantic relationship:

$$\phi_{ij} = \big[E(w_i), E(w_j), \Delta x_{ij}, \Delta y_{ij}, \text{IoU}(b_i, b_j), d_{ij}, E(c_i), E(c_j)\big] \tag{2}$$

This vector comprises content embeddings $(E(w_i), E(w_j)$ for blocks i and $j)$; geometric features (normalized offsets $(\Delta x_{ij}, \Delta y_{ij})$, intersection over union $\text{IoU}(b_i, b_j)$, and normalized distance d_{ij}); and type features (embeddings of the semantic block types $E(c_i), E(c_j)$).

(b) Precedence Prediction: The feature vector ϕ_{ij} is fed into a Relation-Aware Transformer, which is trained to estimate the precedence probability s_{ij}—the likelihood that block i should be read before block j:

$$s_{ij} = \Pr(D_i \prec D_j) = \sigma\big(W \cdot \text{Transformer}(\phi_{ij}) + b\big) \tag{3}$$

(c) Directed Graph Decoding and Topological Sorting: Using the estimated precedence probabilities s_{ij}, we construct a directed graph G, where each block is a node. We add a directed edge $i \to j$ only if the precedence score s_{ij} exceeds a confidence threshold τ and further surpasses the reverse score s_{ji} by a margin [2]. To ensure acyclicity, potential cycles are resolved by removing the edge with the lowest confidence score. The resulting directed acyclic graph is then subjected to topological sorting to yield the definitive linear reading order R_k^{micro} for the sibling set.

This framework integrates macro-level layout cues with micro-level sequential signals, producing a consistent and faithful reading order for complex ESG report pages.

3.2 TOC-GUIDED HIERARCHICAL STRUCTURE RECONSTRUCTION

Existing methods for hierarchical modeling rely heavily on local visual cues such as numbering, indentation, and font size, which break down in ESG reports due to extreme length, weak layout signals, and heterogeneous formats. To overcome these issues, a top-down paradigm is introduced: *ToC-RAP (Region-aware Prompting)* robustly extracts hierarchical structure from the ToC, while *ToC-ALIGN (Anchor-based Linguistic Indexing for Granular Navigation)* anchors ToC entries to the corresponding body content (additional material is included in Appendix A.3). This innovative ToC-based global framework enables reconstruction of explicit and implicit hierarchies in ultra-long documents.

[1] Before Reading Order Modeling, we apply a structured preprocessing pipeline—including metadata extraction, layout analysis, content parsing, and noise reduction—to clean and structure multimodal elements (see Appendix A.1 for details).

[2] The threshold τ and margin were determined by grid search on a validation set, optimizing for the highest mean ROKT score.

ToC Structure Parsing ToC-RAP is proposed as a visual-context modeling strategy for identifying and reconstructing hierarchical structures in visually complex ToCs [3]. Unlike entry-centric approaches that treat each heading line in isolation, a region-centric paradigm is adopted, jointly modeling structural elements with their surrounding visual context. ToC-RAP integrates four mechanisms. *(a) Merging multi-line regions.* Long headings split across lines or blocks are restored by detecting linguistic continuity (e.g., syntactic dependency) and spatial proximity (e.g., minimal gaps), preserving structural integrity. *(b) Aggregating fragmented entries.* Semantically linked entries scattered across columns, shaded areas, or partitions are merged to avoid false segmentation. *(c) Expanding context-aware labels.* Ambiguous headings (e.g., "Part", "Section") are enriched with adjacent descriptors (e.g., titles, numbers, themes), improving alignment with body content. *(d) Inferring region-based hierarchy.* Hierarchical depth is predicted when conventional cues (e.g., font size, indentation) are weak, by exploiting layout signals such as spacing, alignment, and grouping. Finally, a structured ToC set is obtained:

$$H = \{\, h_k = (t_k, l_k, m_k) \,\}_{k=1}^{K}, \tag{4}$$

where t_k is the text, l_k the level, and m_k the metadata. This underpins ToC-ALIGN, anchoring each h_k to the body.

ToC-driven Anchor-based Fine-grained Alignment This step is to align H with the body text, a task complicated by the "semantic gap" between ToC entries and body headings: entries are often abbreviated, paraphrased, or may correspond to implicit thematic paragraphs. To address this, we introduce ToC-ALIGN, a two-stage hybrid paradigm that combines deterministic matching with anchor-based context engineering. Formally, the body is represented as $B = \{b_i = (x_i, p_i, \text{bbox}_i)\}_{i=1}^{N}$, where x_i is the block text, p_i the page, and bbox_i its bounding box.

(a) deterministic matching, ToC titles t_k and candidates x_i are normalized by $\text{canon}(\cdot)$ (case folding and cleaning). An exact match yields $(h_k \to b_i)$ if $\text{canon}(t_k) = \text{canon}(x_i)$; otherwise, fuzzy similarity is computed:

$$\text{sim}_{\text{fuzzy}}(t_k, x_i) = \lambda\, \text{Lev}(t_k, x_i) + (1-\lambda)\, \mathbf{1}[t_k \subseteq x_i], \tag{5}$$

where $\text{Lev}(\cdot)$ is the normalized Levenshtein similarity (Mehlhorn & Näher, 1990) and $\mathbf{1}[\cdot]$ is an indicator function that equals 1 when t_k is a substring of x_i. Alignment is accepted if the maximum fuzzy similarity exceeds τ_{fuzzy}.

(b) anchor-based context engineering, unaligned entries h_k are linked via anchor sets $A(h_k) = \{a_j\}$ derived from nearby landmarks (e.g., parent start, sibling end, page boundaries). For each anchor a_j, a context window $W(a_j) = [\text{pre}_L, a_j, \text{post}_L]$ is constructed and stored in a knowledge base

$$K(a_j) = \langle W(a_j), \pi(h_k), \textit{numbering}, p(a_j) \rangle, \tag{6}$$

where $W(a_j)$ provides local text, $\pi(h_k)$ encodes the parent path, *numbering* captures style cues, and $p(a_j)$ specifies page constraints. The LLM then consumes $(t_k, K(a_j))$ in a structured prompt and outputs

$$\text{out} = \{\text{decision} \in \{\text{ANCHOR}, \text{INSERT\_AFTER}\},\ \text{target\_anchor},\ \text{confidence}\}, \tag{7}$$

where ANCHOR is direct alignment and INSERT_AFTER a new start. To avoid overconfidence, the final score is:

$$s_k = \alpha \cdot s_{\text{LLM}} + (1-\alpha) \cdot \text{sim}_{\text{topic}}(t_k, \text{summary}(K(\text{target\_anchor}))), \tag{8}$$

balancing LLM confidence with semantic similarity. Finally, the mapping $M = (h_k \to a_k, s_k)$ is obtained, which preserves sibling order and, when necessary, inserts anchors to ensure structural and semantic coherence.

3.3 CONTEXT-AWARE VISUAL-TO-TEXT REPRESENTATION

Methods such as LayoutLMv3 (Huang et al., 2022) and Donut (Kim et al., 2023) associate adjacent text and visual content, but struggle on ESG reports with heterogeneous layouts and extreme length, often producing fragmented or inaccurate image descriptions. This arises because key contextual cues (e.g., section headings or captions) may be distant from the corresponding images or even span multiple pages, thereby hindering the capture of global structure. To address these limitations, we propose a context-centric global paradigm with two innovations: *hierarchy-guided multimodal aggregation*, which leverages reconstructed document hierarchy to group related text and visuals into coherent clusters, and *context-guided multimodal reasoning*, which operates on these clusters to align images with broader semantics, thereby enabling globally consistent and semantically faithful image-to-text representations(additional material is included in Appendix A.4).

Hierarchy-guided Multimodal Aggregation Unlike prior methods that rely on spatial adjacency, we construct a semantic context space defined by the reconstructed directory hierarchy. Within this space, the target image is aggregated with sibling nodes under the same parent chapter, forming a coherent multimodal cluster. In parallel, absolute reading order and page numbers are encoded as spatial position features and retained alongside the content. This design enables semantically correlated but physically distant elements to be processed together, fundamentally resolving cross-page dependencies that hinder traditional methods.

[3] As most ESG reports exceed 40 pages, they typically include a ToC. The few very short reports (fewer than 20 pages) without a ToC are excluded, as they lack the structural and layout challenges targeted by our framework.

Context-guided Multimodal Reasoning Building on the aggregated clusters, we introduce a context-engineering strategy that transforms an image into a structured form within information-rich contexts. For any target image, sibling elements (text, tables, or figures) and the hierarchical path embedding h_i (supplying a macro-level semantic frame) are jointly encoded into a structured instance:

$$I_i = (h_i, \{x_1, \ldots, x_{\text{target}}, \ldots, x_k\}, q_i) \tag{9}$$

Here, the ordered set $x_1, \ldots, x_{\text{target}}, \ldots, x_k$ includes the target image and all sibling elements arranged in natural reading order as contextual evidence, and q_i specifies the task directive. Conditioning on I_i integrates local visual signals, mitigating fragmentation and yielding coherent visual-to-text representations for long ESG reports.

4 EXPERIMENT

In this section, we conduct some experiments to evaluate the proposed framework, Compass-ESG.

4.1 DATA

To evaluate the parsing capabilities of Compass-ESG, we focus on Chinese ESG reports, which are generally longer and more heterogeneous than those from other markets. They feature design-driven formatting such as cross-page tables, dense text-figure mixtures, and decorative elements, making them particularly challenging to parse and thus a rigorous testbed for our framework. We collected 50 Chinese reports (2,383 pages) from Wind for evaluation, along with 20 reports each from Hong Kong and the U.S. (1743 pages in total) to assess cross-market generalization across diverse formats and languages [4]. For the construction of the ATLAS-ESG dataset, we ultimately processed 26,006 reports spanning all three markets.

4.2 EXPERIMENTAL DESIGN

We evaluate Compass-ESG on the 70 ESG reports described above, using expert annotations at the document, page, and block levels from three independent domain experts. After consolidation, these annotations were standardized into JSON format as gold-standard references, forming the basis for our baselines and task-specific evaluation metrics.

Baselines To evaluate the effectiveness of Compass-ESG, we conduct comparisons with two groups of systems. The first group includes *specialized document parsers*, namely MinerU Huang et al. (2022), Marker Tito et al. (2023), Docling Kim et al. (2023), and Textin Appalaraju et al. (2021), which are assessed on reading order modeling and hierarchical alignment. The second group consists of *general-purpose multimodal models*, including ChatGPT5, Gemini 2.5 Pro, Doubao, DeepSeek-V3, DeepSeek-R1, and Qwen 3, evaluated on reading order modeling, ToC structure parsing, and ToC-driven fine-grained alignment.

Evaluation Metrics Task-specific metrics are defined as follows: (i) reading order prediction, measured by Reading Order Kendall's Tau (ROKT); (ii) ToC structure extraction, assessed by Content Completeness (CC), Region Order Consistency (RC), and Hierarchical Consistency (HC); (iii) hierarchy alignment, measured by ToC-Body Title Alignment (TBTA); and (iv) comprehensive ESG report analysis, reflecting end-to-end structured transformation, measured by Precision, Recall, and Macro-F1. Detailed metric definitions are provided in Appendix B.1.

4.3 RESULTS

Leveraging their ability to handle long documents, Compass-ESG and specialized parsers are evaluated in batch mode on 50 full Chinese reports, while general-purpose multimodal models, constrained by context length, are assessed on partitioned 5-page segments. Key findings are presented below, with extended results provided in Appendix B.2.

4.3.1 MODULE-WISE PERFORMANCE BREAKDOWN

Table 1 reports module-wise results on reading order, ToC extraction, and ToC-body alignment.

Reading Order Modeling Specialized parsers achieve ROKT scores around 0.8, with higher ROKT reflecting stronger structural parsing, while Marker falls far behind at 0.34. General-purpose multimodal models are less consistent (mostly below 0.7), constrained by context windows and prone to hallucinations, with only ChatGPT5 and Gemini 2.5 Pro reaching 0.75. In contrast, Compass-ESG attains 0.92, effectively capturing long-range dependencies and maintaining sequence alignment in complex layouts.

ToC Extraction With general-purpose multimodal models, ToC-RAP consistently surpasses GP, yielding notable improvements in content completeness (CC), region order consistency (RC), and hierarchical consistency (HC). Models such as Qwen3, Doubao, and ChatGPT5 even reach perfect (100%) scores in certain dimensions, highlighting ToC-RAP's robustness across model families. To lower deployment costs and mitigate API-related privacy risks,

[4] In total, 70 reports were selected through a stratified sampling strategy, ensuring diversity across industry sectors, company sizes, reporting standards, disclosure practices, and temporal coverage.

Table 1: Performance on reading order prediction, ToC extraction, and hierarchy alignment. Since specialized document parsers lack modules for ToC identification or hierarchical parsing, no ToC extraction results are reported.

Category	Method	Reading Order	ToC Extraction						Hierarchy Alignment
		ROKT	CC (%)		RC (%)		HC (%)		TBTA (%)
			GP	ToC-RAP	GP	ToC-RAP	GP	ToC-RAP	
Specialized Document Parsers	Marker	0.34	-	-	-	-	-	-	3.79
	Docling	0.79	-	-	-	-	-	-	16.43
	MinerU	0.82	-	-	-	-	-	-	6.94
	Textin	0.80	-	-	-	-	-	-	9.68
General-purpose Multimodal Models	DeepSeek-V3	0.48	64.35	84.16	56.43	89.11	69.31	92.08	35.29
	Qwen3	0.67	94.06	**100**	**96.04**	98.00	**92.08**	97.02	56.41
	DeepSeek-R1	0.56	82.18	93.00	87.13	94.06	77.23	98.02	55.88
	Doubao	0.45	**95.04**	**100**	81.19	97.00	89.11	97.00	42.50
	ChatGPT5	0.75	85.15	99.00	88.20	**100**	71.29	**99.00**	43.55
	Gemini 2.5 Pro	0.75	93.07	97.02	93.07	96.03	91.09	94.06	64.30
Ours	**Compass-ESG**	**0.92**	81.64	93.19	79.68	97.52	68.19	93.81	**92.46**

Table 2: End-to-end performance on structured ESG report analysis. The reported metrics (Prec., Rec., Macro-F1) are composite measures that integrate ToC extraction and hierarchy alignment, offering a holistic view of system dependencies. The same composite evaluation protocol is also applied in Tables 3 and 4.

Category	Method	End-to-End Performance		
		Prec.	Rec.	Macro-F1
Specialized Document Parsers	Marker	45.77	35.33	39.88
	Docling	74.12	75.31	74.71
	MinerU	75.16	78.69	76.89
	Textin	89.65	76.50	82.55
General-purpose Multimodal Models	DeepSeek-V3	42.11	59.65	36.04
	Qwen3	47.83	55.00	51.16
	DeepSeek-R1	67.44	60.42	63.74
	Doubao	66.22	64.47	65.33
	ChatGPT5	75.90	74.44	75.17
	Gemini 2.5 Pro	86.15	88.89	87.50
Ours	**Compass-ESG**	**92.23**	**95.00**	**92.04**

Compass-ESG integrates a locally deployed Qwen2.5-VL-7B-Instruct, which maintains competitive performance under constrained settings.

ToC-Body Title Hierarchy Alignment Specialized parsers perform poorly on this task (TBTA $< 20\%$), as they lack explicit ToC–body alignment modules and instead rely on shallow heuristics (e.g., font size, indentation), which collapse under the diverse and implicit layouts of ESG reports. General-purpose multimodal models achieve better but still limited performance (TBTA remaining below 65%), as context length constraints and input fragmentation hinder consistent hierarchy prediction. In contrast, Compass-ESG reaches 92.46% TBTA by leveraging ToC structures as alignment anchors, demonstrating its advantage in reconstructing implicit hierarchical structures.

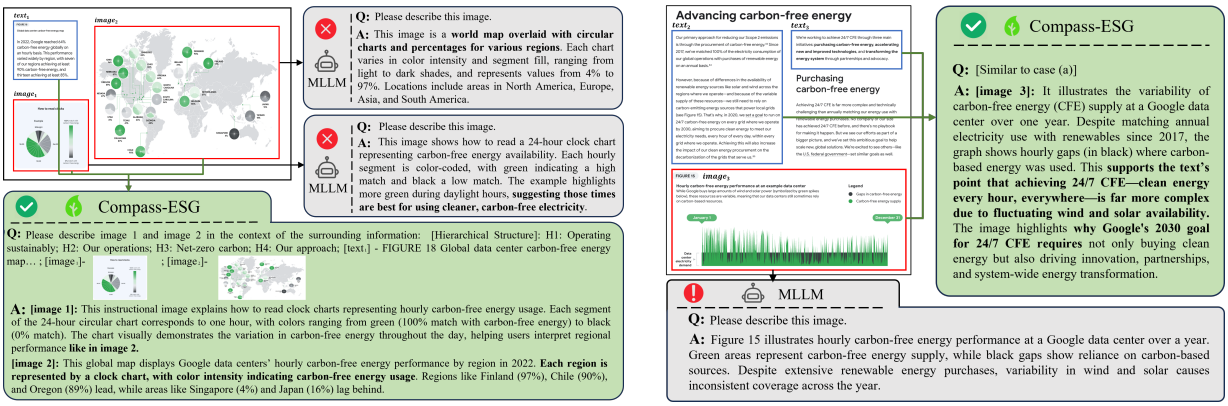
4.3.2 End-to-End Performance on ESG Report Analysis

Table 2 summarizes the end-to-end performance of all systems on structured ESG report analysis. Specialized parsers yield mixed results: Textin attains relatively high precision (89.65) but low recall (76.50), resulting in a moderate Macro-F1 of 82.55. General-purpose multimodal models are less stable, with some (e.g., DeepSeek-V3, Qwen3) underperforming on both precision and recall, while others (e.g., Gemini 2.5 Pro) achieve a better balance (F1 = 87.50). Compass-ESG outperforms all baselines, achieving 92.23 precision, 95.00 recall, and a Macro-F1 of 92.04, underscoring its robustness for comprehensive ESG report analysis.

4.4 Case Studies of Multimodal Reasoning and Contextual Grounding in Compass-ESG

To illustrate the effectiveness of context-aware visual-to-text representation, Figure 3 presents two case studies where Compass-ESG leverages contextual grounding and multimodal reasoning to produce coherent and semantically faithful descriptions.

Compositional Reasoning over Multiple Visual Elements Figure 3a shows a page containing two images and one text block. When interpreted independently, these visual elements lose their intended meaning, as reflected in the



(a) Compositional multimodal reasoning (b) Contextual text-guided grounding

Figure 3: Case studies demonstrating Compass-ESG's ability in multimodal reasoning and contextual text grounding.

Table 3: Ablation study of Compass-ESG's key components.

Config ID	Reading Order Modeling	ToC Structure Parsing	ToC-Body Alignment	End-to-End Performance		
				Prec.	Rec.	Macro-F1
1	✗	✗	✗	75.10	78.90	76.95
2	✗	*GP*	✗	78.63	79.20	78.79
3	✗	*ToC-RAP*	✗	84.43	82.76	83.57
4	✓	*ToC-RAP*	✗	86.01	89.40	88.14
5	✓	*ToC-RAP*	DM	89.12	92.85	90.05
6	✓	***ToC-RAP***	✓	**92.33**	**95.00**	**92.04**

erroneous outputs of multimodal large language models (MLLMs) shown in the figure. In contrast, Compass-ESG models their interdependence: it identifies $image_2$ as the legend for $image_3$ based on hierarchical and reading-order cues, and incorporates $text_1$ to enrich temporal and statistical understanding, yielding an accurate interpretation of the carbon-free energy distribution.

Contextual Grounding of Visual Information through Text Figure 3b demonstrates Compass-ESG's ability to ground visual content in surrounding discourse. Rather than treating the chart ($image_3$) as a standalone visual object, the model links it to contextual cues such as wind and solar intermittency, thereby aligning visual evidence with the broader textual argument concerning the challenges of 24/7 carbon-free energy. This results in coherent, domain-aware reasoning that simple captioning approaches cannot achieve.

4.5 Ablation Study

We performed ablation studies (Table 3) by incrementally enabling three core components—reading order modeling, ToC structure parsing, and ToC-body alignment—to assess their contributions to Compass-ESG's overall performance.

Starting from the baseline (**Config 1**), where all modules are disabled, the system achieves an F1 score of 76.95, comparable to general-purpose document parsers. Introducing the GP-based ToC parser (**Config 2**) improves precision from 75.10 to 78.63 by incorporating ToC information, though it captures only the simplest hierarchical structures and struggles with irregular section labels. Replacing GP with the ToC-RAP strategy (**Config 3**) significantly improves precision to 84.43 by leveraging region-centric visual-context modeling for ToCs, enabling robust parsing of visually complex structures. ToC-RAP serves as the foundation for subsequent modules, and as the cornerstone of the entire system. Adding the reading order module (**Config 4**) lifts the F1 score to 88.14 by reconstructing coherent sequences across irregular, slide-like layouts, using page-level layout frameworks and block-level relational ordering to capture long-range dependencies. **Config 5** introduces deterministic matching (DM), the first stage of ToC-body alignment, where exact and fuzzy checks capture straightforward correspondences, yielding a 3.45-point recall improvement. Finally, the complete system (**Config 6**) with the full ToC-ALIGN module achieves the highest performance (F1 = 92.04), as anchor-based context reasoning resolves semantic gaps and ensures globally consistent alignment, demonstrating the complementary benefits of all three components.

Table 4: Cross-market generalization performance of Compass-ESG on ESG reports. ToC extraction metrics are omitted, as Hong Kong and U.S. directories are simple and yield 100% performance.

Market	Reading Order	Hierarchy Alignment	End-to-end performance		
	ROKT	TBTA (%)	Prec.	Rec.	Macro-F1
China stock	0.92	92.46	92.23	95.00	92.04
HK stock	0.88	89.50	85.42	93.00	89.05
US stock	0.94	93.13	93.61	95.50	94.30

4.6 Cross-Market Generalization of Compass-ESG

Compass-ESG is evaluated on Hong Kong and U.S. reports to assess adaptability across languages, formats, and document structures. The model—trained exclusively on Chinese ESG reports—was kept unchanged; for cross-market evaluation, only prompt localization was applied (e.g., translating Chinese prompts into English). Table 4 presents both module-wise and end-to-end results (a detailed analysis, including full cross-market baseline results, is provided in Appendix B.3). On U.S. reports, Compass-ESG surpasses its performance on Chinese reports, benefiting from their more standardized layouts and clearer hierarchies. Its performance on Hong Kong reports, while slightly below the Chinese baseline, remains strong. This minor variance is likely attributable to the domain shift between the training corpus and evaluation corpus. Although both markets use Chinese, their ESG reports differ in terminology, conventions, and phrasing, posing generalization challenges for models trained on a single domain. The above conclusion demonstrates strong cross-lingual and cross-market generalization across diverse layouts and reporting styles.

5 ATLAS-ESG Dataset

On the basis of the structured data extracted by Compass-ESG, we further apply **MLPDH** (**M**ulti-**L**evel **P**rediction with **D**ocument **H**ierarchy)(see Appendix C for details), which assigns three-level labels tailored to financial analysis needs for each content block: ESG-N category → GRI indicator (Global Reporting Initiative, 2021)→ sentiment (see Appendix B.4 for details). This supports downstream multimodal parsing and financial integration, and results in the construction of the **ATLAS-ESG** dataset [5]—the largest structured ESG dataset to date. It comprises 3,376 preprocessed reports and 1,457K blocks from 2,257 China-listed companies (2021–2025, 41.86% disclosure rate); 11,139 preprocessed reports and 4,413K blocks from 2,631 Hong Kong-listed companies (2020–2025, full disclosure); and preprocessed 11,491 reports and 2,097K blocks from 3,769 US-listed companies (2020–2025, 69.99% disclosure rate). For details of ATLAS-ESG, see Appendix D.

Hierarchical Composition and Content Each ESG report in ATLAS-ESG is structured into three levels: *document*, *page*, and *content block*, capturing metadata, page references, and fine-grained semantics, respectively (Figure 2).

Potential Applications (1) *Multimodal Reasoning over Long ESG Reports.* ATLAS-ESG delivers full-length documents that preserve reading order, hierarchical layout, and interleaved modalities, enabling the development and evaluation of long-context models on tasks that mirror the complexity of real-world ESG analysis. (2) *Cross-Document Analysis and Greenwashing Identification.* By mapping disclosure segments to standardized GRI indicators, the dataset makes possible fine-grained comparisons across firms and years, helping to uncover inconsistencies, omissions, and signals of potential greenwashing. (3) *Investor Sentiment and Decision-Making.* With dedicated sentiment labels tailored to ESG contexts, ATLAS-ESG supports research into how disclosure tone shapes investor perceptions, levels of trust, and sustainability-oriented decisions in behavioral finance studies. (4) *Comparative Benchmarking Across Markets.* Incorporating ESG reports from China, Hong Kong, and the United States, ATLAS-ESG establishes a unified resource for cross-market comparisons. It facilitates structured evaluation of disclosure style, completeness, and tone under differing regulatory frameworks.

6 Conclusion

This study introduce Compass-ESG, a framework for large-scale and structured analysis of ESG reports. The approach transforms fragmented semantics into structured representations by reconstructing disrupted reading flows and revealing implicit hierarchical structures within complex, heterogeneous long-document layouts. It further converts visual elements into text through a context-aware processing pipeline. Experimental results show that Compass-ESG consistently outperforms both specialized document parsers and general-purpose multimodal models. To facilitate future research, we release ATLAS-ESG, a comprehensive dataset with multi-level annotations covering corporate reports from China, Hong Kong, and the U.S., providing a foundation for both financial analysis and document parsing.

[5] ATLAS-ESG is constructed through an end-to-end automated pipeline, combining Compass-ESG for structural parsing and MLPDH for hierarchical labeling, thereby ensuring consistency and scalability.

Ethics Statement

This work utilizes publicly available ESG reports and related datasets. All data were collected from legal and compliant public sources without involving any private, sensitive, or personally identifiable information. The use of large language models was limited to language refinement, and the authors remain fully responsible for the accuracy, validity, and originality of the content. No human subjects, animal studies, or other sensitive domains were involved in this research.

Reproducibility Statement

To ensure reproducibility, we have provided detailed descriptions of our dataset construction, model configurations, and experimental protocols in the main paper and appendix. The source code, data processing scripts, and model checkpoints will be made available in an open repository upon publication. Hyperparameters, evaluation metrics, and training details are explicitly documented to facilitate replication and further research.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 993–1003, 2021.

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

Anywhere Sikochi Dane M. Christensen, George Serafeim. Why is corporate virtue in the eye of the beholder? the case of esg ratings. In *The Accounting Review, Volume 97 ,Issue 1*, pp. 147–175, 2022.

Datalab. Marker: Document parsing toolkit. https://github.com/datalab-to/marker, 2025. Accessed: 2025-07-22.

Roberto Rigobon Florian Berg, Julian F Kölbel. Aggregate confusion: The divergence of esg ratings. In *Review of Finance, Volume 26, Issue 6, November 2022*, pp. 1315–1344, 2022.

Global Reporting Initiative. GRI Standards. https://www.globalreporting.org/standards/, 2021. Accessed: 2025-09-24.

Jaekyu Ha, R.M. Haralick, and I.T. Phillips. Recursive x-y cut using bounding boxes of connected components. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pp. 952–955 vol.2, 1995. doi: 10.1109/ICDAR.1995.602059.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4083–4091, 2022.

intsig textin. Textin xparse frontend. https://github.com/intsig-textin/xparse-frontend, 2023. Accessed: 2025-07-22.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun park. Donut: Document understanding transformer without ocr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16579–16589, 2023.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I. Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5652–5660, 2021.

Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. Efficientvit: Memory efficient vision transformer with cascaded group attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14420–14430, 2023.

Nikos Livathinos, Christoph Auer, Maxim Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, et al. Docling: An efficient open-source toolkit for ai-driven document conversion. In *AAAI Conference on Artificial Intelligence*, 2025.

Esteban Marulanda, Alejandro Restrepo, and Johans Restrepo. Correspondence between the energy equipartition theorem in classical mechanics and its phase-space formulation in quantum mechanics. In *Entropy,volume:25,number:6*, pp. 939, 2023.

Kurt Mehlhorn and Stefan Näher. Dynamic fractional cascading. In *Algorithmica,Volume 5*, 1990.

Gaku Morio and Christopher D Manning. An nlp benchmark dataset for assessing corporate climate policy engagement. *Advances in Neural Information Processing Systems*, 36:39678–39702, 2023.

Xiaoran Ni and Huilin Zhang. Mandatory corporate social responsibility disclosure and dividend payouts: evidence from a quasi-natural experiment. In *accounting and finance,volume 58,issue 5*, 2019.

Keane Ong, Rui Mao, Deeksha Varshney, Erik Cambria, and Gianmarco Mengaldo. Towards robust esg analysis against greenwashing risks: Aspect-action analysis with cross-category generalization. *arXiv preprint arXiv:2502.15821*, 2025.

Rasmus Berg Palm, Florian Laws, and Ole Winther. Attend, copy, parse end-to-end information extraction from documents. In *2019 International Conference on Document Analysis and Recognition (ICDAR)l Conference on Multimedia*, pp. 329–336, 2019.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3743–3751, 2022.

Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pp. 629–633. IEEE, 2007.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing Hea, Fei Wu, and Jiwei Li. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. In *arXiv preprint arXiv:2106*, 2021.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage docvqa. In *Pattern Recognition 144,Volume 144*, 2023.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 5412–5416, 2023.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*, 2024.

Mariya Pavlova Miaosen Wang and Bernard Casey. Esg-ftse: A corpus of news articles with esg relevance labels and use cases. *LREC-COLING 2024*, 137, 2024.

Xinyu Wang, Qingqing Cao, Xiaojian Ma, Yujia Xie, Junyang Lin, Lidong Bing, and Zhiyuan Liu. Docparser: End-to-end ocr-free information extraction from visually rich documents. In *Document Analysis and Recognition - ICDAR 2023*, pp. 155–172, 2023.

Jnatas Wehrmann, Ricardo Cerri, and Rodrigo C Barros. Hierarchical multi-label classification networks. In *International Conference on Machine Learning*, 2019.

Mengzhou Xia, Zexuan Zhong, and Danqi Chen. Structured pruning learns compact and accurate models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1192–1200, 2020.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1174. URL https://aclanthology.org/N16-1174/.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2023.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974, 2024.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pp. 1015–1022. IEEE, 2019.

A IMPLEMENTATION DETAILS

A.1 PREPROCESSING PIPELINE

Table 5: Key hyperparameters of the Compass-ESG preprocessing pipeline. This table provides a comprehensive overview of the parameters for each pipeline component. For heuristic-based stages like ToC Identification and PDF-to-Image conversion, we detail the key operational parameters. For model-based stages, we specify the settings for the MinerU framework's built-in YOLOv10 and PaddleOCR engines. Based on the unique characteristics of ESG reports, we have established distinct detection thresholds for general text versus those for structured table areas. The table also delineates the rationale behind various IoU threshold settings for refinement rules. For the final image data cleaning in post-processing, we opted for the more efficient Tesseract OCR (Smith, 2007) engine.

Component	Hyperparameter	Setting	Description
ToC Identification	`max_pages_to_check`	5	Max number of pages to scan for a ToC.
	`ocr_render_dpi`	200	DPI for rendering pages for ToC feature detection.
	`toc_density_threshold`	0.5	Minimum density of ToC indicators on a page.
	`min_indicator_count`	5	Minimum number of ToC indicators required.
PDF-to-Image	`render_dpi`	144	DPI for rendering pages for content analysis.
	`max_width_or_height`	2560	Max dimension for rendered images.
Layout Analysis (YOLOv10)	`imgsz`	1280	Input image size for the YOLOv10 model.
	`conf`	0.1	Confidence threshold for object detection.
	`iou`	0.45	IoU threshold for Non-Maximum Suppression (NMS).
Text Recognition (PaddleOCR)	`rec_algorithm`	CRNN	The specific algorithm for text line recognition.
	`drop_score`	0.5	Recognition results with scores below this are discarded.
	`det_db_box_thresh`	0.3	Binarization threshold for general text detection.
	`det_db_unclip_ratio`	1.8	Box expansion ratio for general text detection.
Table Recognition	`det_db_box_thresh`	0.5	A dedicated threshold for text detection within tables.
	`det_db_unclip_ratio`	1.6	A dedicated expansion ratio for text boxes in tables.
Refinement Rules	`text_title_overlap_iou`	0.8	IoU threshold to merge text blocks with title blocks.
	`table_merge_iou`	0.7	IoU threshold for merging adjacent table cells.
	`min_ocr_confidence`	0.5	Minimum confidence to keep a recognized character.
Post-processing & Cleaning	`ocr_engine`	Tesseract	Engine used for filtering decorative images.
	`confidence_threshold`	0.7	Min OCR confidence for an image to be kept.
	`min_text_length`	5	Min number of characters for an image to be kept.

To transform raw ESG report PDF documents, obtained from public channels in various formats, into a clean and structured corpus suitable for our models, we designed and implemented an end-to-end preprocessing pipeline. This pipeline first performs a deep structural analysis of the raw PDFs, followed by refinement and cleaning of the extracted data. The hyperparameter configurations for the key components in this process are detailed in Table 5.

Stage 1: Structural Analysis and Raw Data Extraction This stage converts raw PDFs into structured JSON.

(1) ToC Identification: We scan the first few pages (defined by `max_pages_to_check`) to locate ToC pages using heuristic rules based on keywords and layout features (e.g., `toc_density_threshold`).

(2) Deep Layout Analysis and Content Extraction: We then perform a comprehensive deep parse on all pages using the `MinerU`[6] framework, which coordinates a `YOLOv10`[7]-based layout analysis model and the `PaddleOCR`[8] engine. We configure dedicated, stricter detection thresholds for text recognition within table areas (see Table 5) to ensure accuracy.

(3) Result Refinement: Refinement rules merge integral elements such as text-title blocks and table cells based on IoU thresholds.

Stage 2: Post-Processing and Data Cleaning This stage refines and validates the JSON data to ensure the quality of the final corpus.

[6] https://github.com/opendatalab/MinerU/tree/master
[7] https://github.com/THU-MIG/yolov10
[8] https://github.com/PaddlePaddle/PaddleOCR

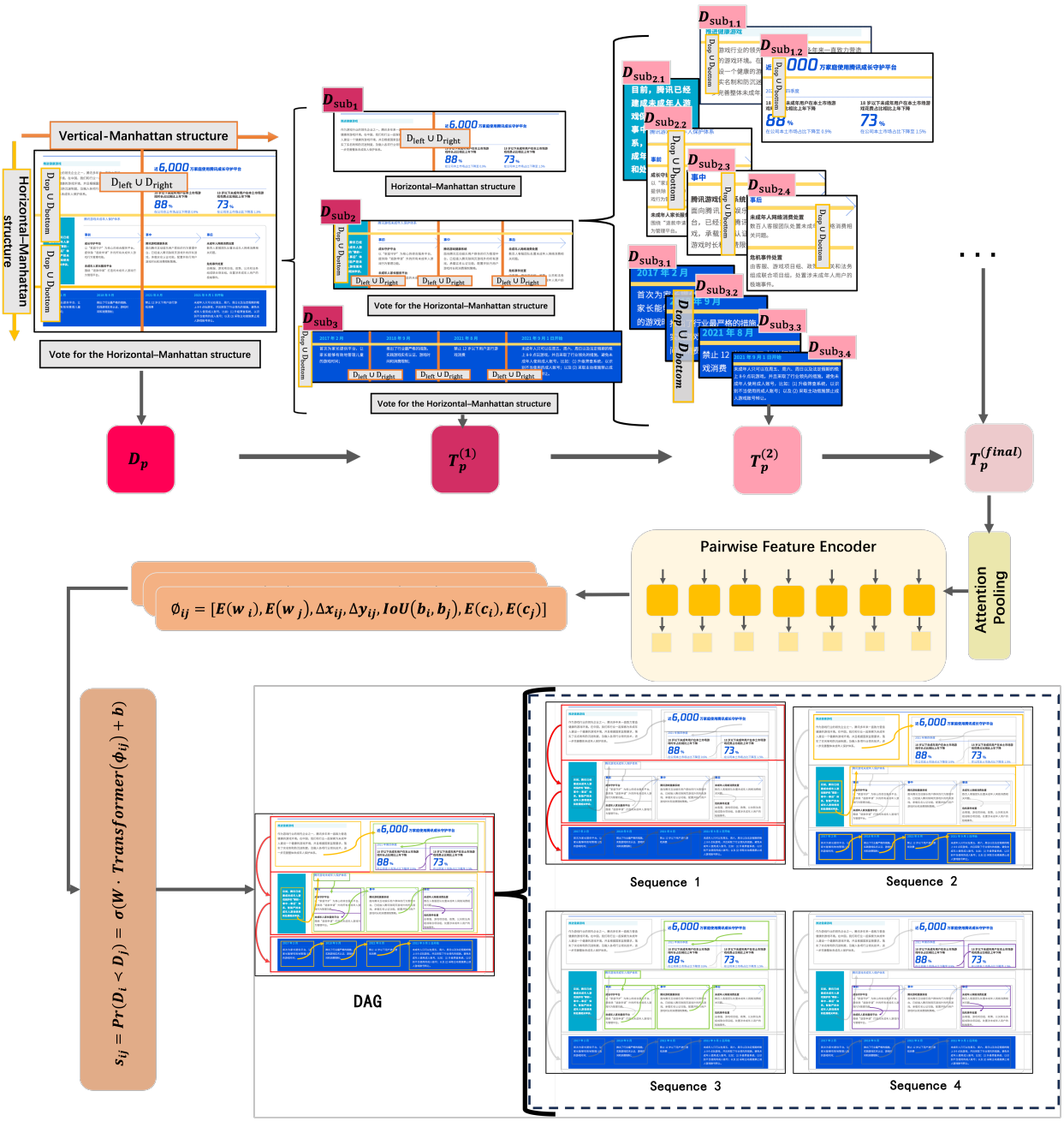


Figure 4: Overview of two-phase hybrid reading order model. *Phase 1 (Top): Page-level Layout Framework.* First, a dominant layout type (e.g., Horizontal-Manhattan) is inferred by detecting and voting on all possible structures within the page's content boxes (D_p). Guided by this, a recursive XY-cut algorithm decomposes the page into a deep, hierarchical region tree ($T_p^{(1)} \rightarrow \cdots \rightarrow T_p^{(\text{final})}$). *Phase 2 (Bottom): Block-level Relational Sequence Modeling.* A Pairwise Feature Encoder constructs a multimodal feature vector ϕ_{ij} using semantic embeddings (refined via Attention Pooling) and geometric features. This vector is used to predict precedence scores (s_{ij}), which are then decoded into a final, coherent reading order via a Directed Acyclic Graph (DAG) and topological sorting. The final linearized sequence is visualized on the page, demonstrating the model's effectiveness.

(1) JSON to Markdown Conversion: For ease of manual inspection and debugging, the structured JSON files are converted into a human-readable Markdown format.

(2) Image Link Conversion: All image reference paths within the documents are normalized to a consistent relative path format.

(3) Image Text Detection: This is a critical noise filtering step. We use a lightweight OCR engine (specified by `ocr_engine`) to detect and filter out purely decorative images that lack meaningful text, controlled by the `min_text_length` and `confidence_threshold` parameters.

After all the above processing steps, the final output of this pipeline is a high-quality, structurally-rich, and clean multimodal JSON dataset.

A.2 Reading Order Modeling

This section provides a detailed, visual walkthrough of our two-phase hybrid reading order model. Figure 4 illustrates the entire pipeline, from initial page decomposition to the final sequence generation, using a sample ESG report page with a complex layout.

Page-level Layout Framework As illustrated in the top half of Figure 4, the process begins by establishing a structural hierarchy. The system detects both Vertical- and Horizontal-Manhattan partitioning candidates and selects the dominant orientation via a voting mechanism. In this example, the page is classified as primarily Horizontal-Manhattan. Guided by this global orientation, a recursive X–Y cut algorithm generates a coarse region tree ($T_p^{(1)}$) with high-level nodes such as $D_{\text{sub1}}, D_{\text{sub2}}, \dots$. This process is repeated within each sub-region, producing progressively finer trees ($T_p^{(2)}, \dots$) until the final hierarchical tree $T_p^{(\text{final})}$ is obtained.

Block-level Relational Sequence Modeling The second phase, depicted in the bottom half of the Figure 4, refines the reading order for sibling nodes in $T_p^{(\text{final})}$ using a data-driven approach. A Pairwise Feature Encoder integrates (1) semantic embeddings for each block (e.g., $E(w_i), E(c_i)$), obtained via Attention Pooling, and (2) geometric relations (e.g., Δx_{ij}, IoU). The resulting vector ϕ_{ij} is processed by a Transformer-based Relation Prediction Head to estimate precedence probability s_{ij}. These pairwise scores define a directed graph, where an edge $i \to j$ indicates i precedes j; cycles are removed to obtain a DAG. A topological sort then yields the definitive reading order for each sibling group.

Final Output Visualization The panels at the bottom right of Figure 4 show how the model's predictions are re-assembled into the final reading order. Each sequence corresponds to a local ordering of sibling blocks obtained from the DAG. These locally consistent sequences are then concatenated following the page's hierarchical tree, ensuring that content from different columns and regions is merged in the correct global order. In "Sequence 1–4," the numbered overlays directly mark the order in which blocks should be read on the page. This explicit reconstruction demonstrates that the system not only resolves fine-grained ambiguities within sibling groups but also preserves the macro-level flow across complex, multi-column layouts.

A.3 TOC-GUIDED HIERARCHICAL STRUCTURE RECONSTRUCTION

This section provides a detailed breakdown of the core components of our TOC-guided hierarchical structure reconstruction framework: Region-aware Prompting (ToC-RAP) and the Anchor-based Linguistic Indexing for Granular Navigation (ToC-ALIGN) module. To clearly illustrate their inner workings, we will use a case-study format to deconstruct the specific reasoning process of each component.

A.3.1 ToC-RAP Demonstration

Our ToC-RAP method is designed to transform the complex task of ToC parsing into a structured, sequential reasoning process for the model. This is achieved through a carefully crafted prompting framework, which consists of the two key components detailed below. First, the *System Message* establishes the model's role and high-level objective (as shown in Box A.1). Second, a series of four explicit *Directive Messages* are provided, which the model must execute sequentially to handle common layout challenges like multi-line titles and hierarchical inference (as shown in Box A.2). To further illustrate the method, the Step-by-Step Reasoning Workflow demonstrates how the model applies the four directives to a real-world ToCs, highlighting both effectiveness and interpretability (see Figure 5).

Box A.1: System Message in ToC-RAP

```
As a document directory extraction assistant, identify and extract the
hierarchical headings from the provided image and output the corresponding list.
Before producing the result, sequentially execute the reasoning for four
directives and provide justification for each step.
```

Box A.2: Directive Message in ToC-RAP

```
Directive 1: Multi-line Region Merging
  Objective: To merge a single logical entry that has been split across multiple
  lines.
  Rule: If adjacent lines exhibit linguistic continuity and close vertical
  proximity (e.g., "Compliance Governance" on line N and "Escorting Development"
  on line N+1), you must combine their text into a single string joined by a
  space.

Directive 2: Cross-Region Entry Aggregation
  Objective: To aggregate semantically linked but spatially fragmented elements
  into a single entry.
  Rule: When you identify a module where a numeric index (e.g., "01"), a short
  tag (e.g., "huan jing pian"), and a long title (e.g., "Cultivating Green
  Shoots") are spatially proximate, you must aggregate them into a single string
  in the format "Tag Title". The required output in this example is "huan jing
  pian Cultivating Green Shoots".

Directive 3: Context-Aware Abbreviated Region Expansion
  Objective: To resolve vague, abbreviated labels by replacing them with
  specific, contextually relevant titles.
  Procedure: You must execute the following steps sequentially.

  (3.1) Find Candidate (X): Identify any line X if its length is 3 characters or
  less AND it ends with a generic tag ("pian", "zhang", "bu"). Store its stem as
  X'.
  (3.2) Scan Vicinity for Replacement (Y): For each X, search the visual vicinity
  (10 lines below, 40% page width to the right) for a visually prominent line Y.
  (3.3) Validate Y: A line Y is a valid replacement only if it meets two
  conditions. First, its text length must be 3 characters or more. Second, it
  must satisfy at least one of the following semantic checks:
        (i) Jaccard similarity to X' is 0.5 or greater;
        (ii) The normalized edit distance to X' is 0.5 or less;
        (iii) It shares at least two characters with X'.
  (3.4) Execute Resolution: If a valid Y is found, replace X with Y. If no valid
  Y is found, search for a line Z (length 6 chars or more) within 3 lines below
  X; if found, replace X with the merged string "X'". If neither Y nor Z is
  found, discard X.
  (3.5) Post-Process: After processing all candidates, remove any duplicate Y
  entries, retaining only the first. Apply a final filter to exclude any
  remaining short, generic tags from the final output.

Directive 4: Region-Based Hierarchy Inference
  Objective: To determine the hierarchical level of each entry.
  Inference Priority: You must evaluate hierarchy cues in the following strict
  order.

  (4.1) Aligned Numbering: Entries with independent, aligned numbering are of the
  same, non-nested level.
  (4.2) Typographic Cues: If numbering is absent, determine hierarchy based on
  the sequence: Font Size/Weight -> Color -> Indentation.
  (4.3) Spatial Cues: If typographic cues are inconclusive, a significantly
  larger vertical space above an entry designates it as a parent-level heading.
  (4.4) Sub-level Cues: Within a parent's scope, a bolder/darker entry is
  Level-2. Subsequent, less prominent entries are Level-3.
```

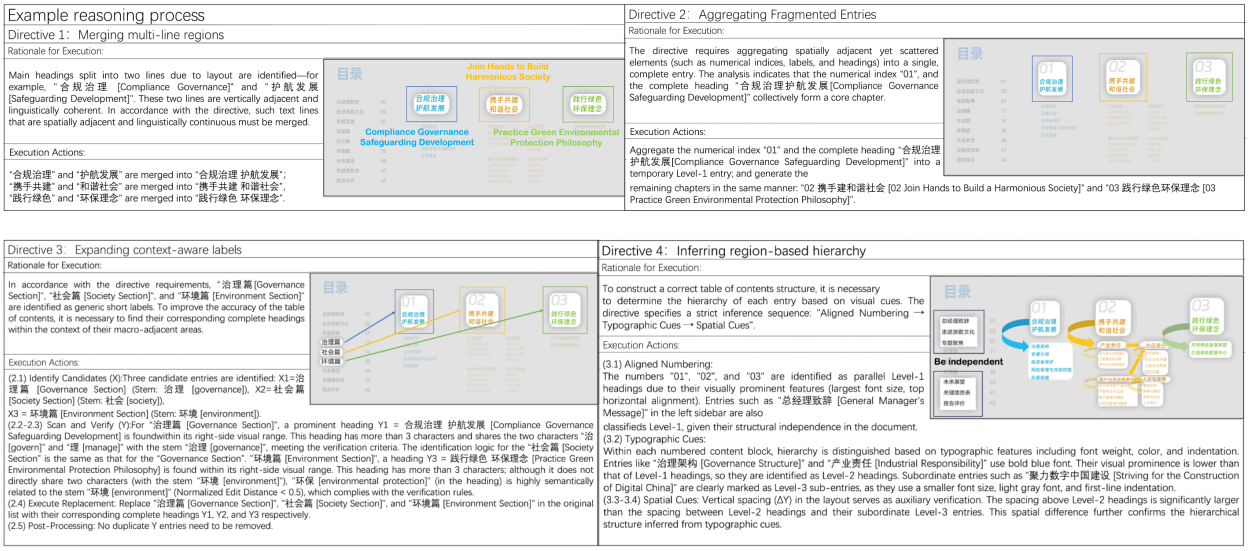


Figure 5: Step-by-step reasoning workflow of ToC-RAP applied to a real-world ESG report's ToCs. The workflow follows four sequential directives: (1) merge multi-line text into coherent entries; (2) aggregate spatially fragmented elements such as indices and titles; (3) expand abbreviated labels with contextually relevant names; and (4) infer the hierarchical structure using numbering, typographic, and spatial cues.

A.3.2 ToC-ALIGN Demonstration

The ToC-ALIGN module employs a two-stage hybrid paradigm that combines *Deterministic Matching* with *Contextual Engineering Insertion* to align ToC entries with the main document content at high precision. The following describes the detailed implementation of each stage.

Stage 1: Deterministic Matching and Candidate Identification This stage resolves the majority of straightforward alignment cases through efficient rule-based methods, while leaving unresolved or ambiguous entries for the more advanced procedures in Stage 2. The process applies a multi-step matching strategy—exact, fuzzy, and containment—to identify correspondences between ToC titles and candidate headings in the document body [9]. The key hyperparameters for this stage are summarized in Table 6.

Table 6: Key hyperparameters of the Deterministic Matching stage in ToC-ALIGN.

Hyperparameter	Setting	Description
Fuzzy Match Threshold	0.75	Minimum similarity score required for a fuzzy match.
Containment Match Similarity	0.9	Similarity threshold for a containment match.
Unaligned Title Level	4	Default level for Markdown titles not found in the ToC.

Stage 2: LLM-based Contextual Alignment This stage resolves the most challenging alignment cases for ToC entries that remain unlinked after the initial matching phase. ToC-ALIGN leverages a Large Language Model (LLM) guided by a structured context-engineering and prompting schema to determine the correct placement of titles based on deep semantic understanding. The process consists of three main steps.

(1) Anchor-based Context Construction: For each unaligned ToC entry, we avoid a costly and inefficient global search. Instead, we use its correctly aligned preceding and succeeding titles from the ToC as structural anchors. These anchors specify a bounded, contextually relevant search region in the document; an example is provided in Box A.5 (lines 121–149). The content within this region is then extracted and annotated with global line numbers and content tags (e.g., `[Content]`) to create a machine-readable context for the LLM.

[9] A critical step in this stage is the handling of document titles that do not match any ToC entry. These titles are not deleted but are programmatically demoted to a specific low level (e.g., Markdown's `####`), marking them as "unaligned." This serves as a crucial structural cue for the next stage. The output of this stage is a partially aligned document and a list of ToC entries that remain unaligned.

(2) Structured Reasoning Prompt Schema: The annotated context is then embedded into a sophisticated, multi-part prompt designed (as shown in Box A.6) to elicit a structured reasoning process from the LLM. This prompt schema includes:

(i) A System Message that assigns the LLM the role of a "structured document analysis expert" (as shown in A.3).

(ii) The specific Target Title (e.g., `"ESG Key Performance Indicators"`) and its intended hierarchical level (an illustrative example is shown in Box A.4).

(iii) The contextual Anchor Titles to bound the model's focus.

(iv) A set of explicit Inference Instructions, including special logic (like replacing `####`-marked titles).

(v) A strictly enforced Required Output Format (e.g., `Insert global line number:` ⟨`line number`⟩) to ensure the response is unambiguous and easily parsable.

(3) Multi-Phase Reasoning and Decision Execution: The LLM is instructed to follow a multi-phase chain of thought before providing its final answer. As demonstrated in the example reasoning process (Box A.7), this involves

(i) a Semantic Flow Analysis to understand the topical progression,

(ii) a Thematic Boundary Identification to locate the most logical insertion point, and

(iii) a Final Decision Logic to select the action.

The final output from the LLM is not a natural language sentence, but the simple, structured response we require (e.g., `Insert global line number: 136`). Our system then parses this instruction and executes the final insertion or replacement action, completing the document's structural alignment with high precision.

Box A.3: System Message in ToC-ALIGN

```
You are a structured document analysis expert, specializing in determining the
insertion points for titles in unstructured text. Your task is to judge the most
natural and logical insertion position for a target title based on its semantic,
structural, and thematic relationship with the document content...
```

Box A.4: Example of Target Title

```
{
  "title_info": {
    "title": "ESG Key Performance Indicators",
    "level": 2
  }
}
```

Box A.5: Example of Contextual Anchors

```
[content_for_llm]:{
121. [title_info['prev_title']] Sustainable Development Governance
122. [Content] The company has established a board-led...
...
135. [Content] Training on ESG risk identification was delivered to 1,200
employees...
136. [Content] In terms of environmental performance, this year's
GHG emissions...
137.[Content]  Energy intensity fell by 9%, and 62% of electricity came from
renewable sources...
...
148.[title_info['next_title']] Stakeholder Communication
149.[Content]  We actively engage with stakeholders through various channels...
}
```

Box A.6: Example of Prompt in ToC-ALIGN

```
Take a deep breath and work on this step by step.
You must determine the best position to insert a missing title into a document.
Target Title: {{"title_info" }}
Intended Level: {{"level"}}
Contextual Information:
- Preceding Title in ToCs: {"title_info['prev_title']"}
- Succeeding Title in ToCs: {"title_info['next_title']"}
The Annotated Document Context:{content_for_llm}
Inference Instructions:
- Analyze the semantic flow of the annotated context.
- Identify the most logical insertion point for the Target Title.
- Special Replacement Logic: Pay close attention to any title marked with ####.
If you find one that is semantically equivalent to the Target Title, recommend
replacing it.
Required Output Format (Choose ONE):
Insert global line number: <line number>
OR
Replace global line number: <line number>
```

Box A.7: Example of Reasoning Process in TOC-ALIGN

```
The model's inference process for determining the insertion point follows a
structured, multi-step analysis:

Phase1: Semantic Flow Analysis:
The model first analyzes the sequence of topics within the provided context
between the two anchors.It identifies a clear thematic progression:
- Starts with general governance ("Sustainable Development Governance").
- Transitions to a specific social metric (employee training on line 135).
- Concludes with a series of quantitative environmental metrics (GHG emissions,
energy intensity on lines 136-137).

Phase2: Thematic Boundary and Header Identification:
- A distinct thematic boundary is detected between the social metric (line 135)
and the environmental metrics (line 136).
- The sequence of quantitative data starting on line 136 is recognized as a
cohesive block of Key Performance Indicators (KPIs).
- The analysis concludes that this emergent block lacks a unifying header, and
the target title, "ESG Key Performance Indicators", is a semantically fitting
match.

Phase3: Final Decision Logic:
- A check is performed for any ####-marked titles that would trigger the special
replacement logic; none are found in this context.
- Based on the identified thematic boundary and the semantic match, the model
determines that the optimal action is to insert the title.
- Conclusion: The insertion point is set immediately preceding the start of the
KPI block on line 136.
```

A.4 Context-Aware Visual-to-Text Representation

This section provides a concrete case study to illustrate the two-stage process of our context-aware visual-to-text representation framework, as described in the main text. We use a common and challenging example from an ESG report—a Materiality Assessment Matrix—to demonstrate how our method overcomes the limitations of traditional, proximity-based approaches. The corresponding System Message for Visual-to-Text Representation is provided in Box A.8.

Stage 1: Hierarchy-guided Multimodal Aggregation This stage extends beyond spatial adjacency to form semantically coherent multimodal clusters. A representative example of an aggregated multimodal context bundle is provided in Box A.9.

(1) Target Identification: The system identifies the target visual element, which in this case is the materiality matrix chart (`image 1`).

(2) Hierarchical Context: Based on the previously reconstructed document hierarchy, the system recognizes that the image's parent is the heading "h4 Materiality Assessment of ESG Issues".

(3) Sibling Node Aggregation: Instead of just looking at nearby text, our method aggregates all sibling nodes that share the same parent heading. This forms a rich multimodal cluster containing not only the target image, but also all surrounding explanatory text blocks (`text 1-5`) and the crucial legend (`image 2`). This step is critical, as it ensures that semantically related but physically distant information (like a caption on the next page) is included in the context.

Stage 2: Context-guided Multimodal Reasoning In this stage, the aggregated cluster is encoded into a structured input instance, $I_i = (h_i, x_1, \ldots, x_{\text{target}}, \ldots, x_k, q_i)$, which guides the LLM's reasoning process. A representative example of this reasoning in image representation is provided in Box A.10.

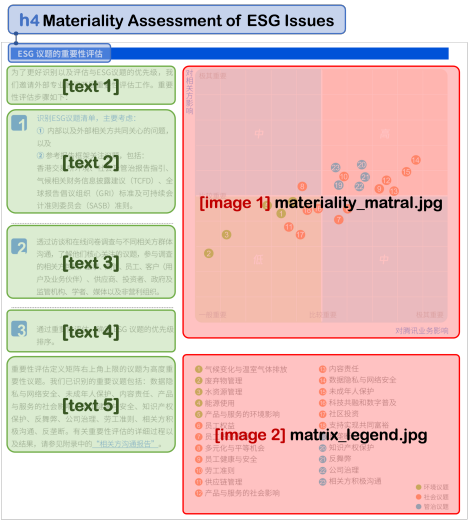
Box A.8: System Message in Visual-to-Text Representation

```
Role: Expert Multimodal Document Analyst---specializing in data visualization and
structured report interpretation.

Task: Interpret the target visual element with semantic depth and contextual
accuracy, using all structured input materials.
Principles:
1. Beyond Description: Synthesize across visual, textual, and structural cues to
uncover deeper meaning.
2. Evidence-Driven: Ground reasoning in the provided context (e.g., surrounding
text, legends).
3. Concise Insight: Deliver a clear, well-articulated summary of key insights.
```

Box A.9: Example of an Aggregated Multimodal Context Bundle

```
Hierarchical Context:
h1: ESG Governance
h2: Analysis of Material ESG Issues
h3: null
h4: Materiality Assessment of ESG Issues
Content Components:
[text 1] To better identify and prioritize
ESG issues, we invite...
                .
                .
                .
[text 5] The issues at the upper limit of
the upper right corner of the materiality
assessment definition matrix are highly
material issues. The material issues we
have identified include: ...
[image 1] ![](.../materiality_matrix.jpg)
[image 2] ![](.../matrix_legend.jpg)
Positional Indices:
data_indices: [(15, 7),...,(15, 8), (15,
↪  9),(15,10)]
```



Example content in the original PDF

The tuples (page_idx, reading_order) in this index list correspond one-to-one with each item in the "specific content within the cluster".

Box A.10: Example of Reasoning Process in Image Representation

```
The model's inference process for determining the insertion point follows a
structured, multi-step analysis:
- Semantic Flow Analysis: The model first analyzes the sequence of topics within
the provided context. It identifies a clear thematic progression starting from
general governance, transitioning to specific social metrics (employee training),
and concluding with a series of quantitative environmental metrics (GHG
emissions, energy intensity).

- Thematic Boundary and Header Identification: A distinct thematic boundary is
detected between the social metrics and the environmental metrics. The sequence
of quantitative data is recognized as a cohesive block of Key Performance
Indicators (KPIs). The analysis concludes that this emergent block lacks a
unifying header, and the target title is a semantically fitting match.

- Final Decision Logic: Based on the identified thematic boundary and the
semantic match, the model determines that the optimal action is to insert the
title immediately preceding the start of the KPI block.

- Conclusion:The materiality matrix clearly indicates that for the company
represented (Tencent), its most critical ESG risks and opportunities lie in the
responsible operation of its digital platforms, protecting user rights
(especially minors), maintaining data security, and ensuring a high level of
corporate governance and business ethics. In contrast, while traditional
environmental issues are noted, their urgency and importance are ranked
relatively lower.
```

B Detailed Experimental Setup and Analysis

This appendix provides a detailed extension, offering a deeper description of the evaluation framework and a more fine-grained analysis of the results.

B.1 Evaluation Metrics and Calculation Methodology

To comprehensively evaluate the performance of Compass-ESG, we design a multi-level evaluation framework. At the global level, we measure the *end-to-end performance* of the entire pipeline, assessing how well the system produces accurate and coherent ESG report analyses. At the local level, we introduce *content extraction metrics* that focus on the model's ability to capture document layout, hierarchical organization, and fine-grained structural elements.

B.1.1 End-to-End Performance of Comprehensive ESG Report Analysis

Our evaluation framework for this task is based on the standard metrics of Precision, Recall, and Macro-F1 score. The Macro-F1, as the harmonic mean of Precision and Recall, offers a balanced measure of a model's performance. In document analysis, this is particularly crucial as it simultaneously penalizes False Positives (incorrectly identified elements) and False Negatives (missed elements), providing a robust assessment framework superior to simple accuracy.

Our choice of a **macro-averaging** evaluation strategy is a direct response to the severe class imbalance inherent in document analysis. Unlike a micro-average, which would be skewed by the high frequency of `Paragraph` elements and thus mask failures on rare classes, the macro-average provides a more robust assessment. It computes the metric independently for each semantic category—from `Paragraph` to structurally critical `Heading` elements—and then calculates the unweighted mean. This ensures every class contributes equally to the final score, offering a faithful evaluation of the model's ability to comprehend the entire document structure.

To implement this, we define the parsing task as a classification problem at the "semantic unit" level. The set of semantic units (C) used in our study is defined in Table 7. The establishment of this classification scheme is vital: it not only provides the transparency required for reproducible research but also defines the precise scope and granularity of the "comprehensive analysis" task. Furthermore, it enables a detailed analysis of a model's specific failure modes, which would be obscured by a single, un-decomposed Macro-F1.

For each semantic category c in the set of all categories C, we determine the counts of True Positives (TP_c), False Positives (FP_c), and False Negatives (FN_c):

Table 7: Definition of Semantic Unit Categories and Associated Metadata.

ID / Metadata	Description	Example
Semantic Unit Categories		
`Heading`	Hierarchical section headings(L1-L4).	"3.1. Greenhouse Gas Emissions"
`Paragraph`	A continuous block of prose text.	"Our commitment to sustainability..."
`List_Item`	A single item in a bulleted or numbered list.	"- Reduce water consumption by 15%."
`Table`	A block of structured, tabular data.	(A complete table element)
`Figure_URL`	A link or reference to an image or chart.	`![](image_1.jpg)`
Associated Metadata		
`Position_Index`	Positional metadata for each unit above.	`[page:n,reading_order:m]`

(1) True Positive (TP_c): A semantic unit that exists in the ground truth as category c is correctly identified and classified as category c by the model.

(2) False Positive (FP_c): A semantic unit is incorrectly classified as category c. This includes both cases where a unit of a different category is misclassified as c, and where the model hallucinates a unit of category c that does not exist in the ground truth.

(3) False Negative (FN_c): A semantic unit that exists in the ground truth as category c is either missed entirely by the model or misclassified as a different category.

Based on these counts, the Precision, Recall, and Macro-F1 are first computed for each individual class c:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{10}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{11}$$

$$\text{Macro-F1}_c = 2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{12}$$

The final macro-averaged scores, which are reported in our experimental results, are the unweighted arithmetic mean of these per-class scores over all $|C|$ classes:

$$\text{Macro-Precision} = \frac{1}{|C|} \sum_{c \in C} \text{Precision}_c \tag{13}$$

$$\text{Macro-Recall} = \frac{1}{|C|} \sum_{c \in C} \text{Recall}_c \tag{14}$$

$$\text{Macro-F1} = \frac{1}{|C|} \sum_{c \in C} \text{Macro-F1}_c \tag{15}$$

B.1.2 Content Extraction Metrics

For this task, we employ specialized metrics to assess the model's understanding of document layout and global hierarchy.

ROKT (Reading Order Kendall's Tau) To quantitatively measure the correctness of the reading order prediction, we use Kendall's Tau correlation coefficient (τ), which we term ROKT. Kendall's Tau is a non-parametric statistic that measures the ordinal association between two sequences. In our context, it compares the model-predicted sequence of semantic units on a page with the ground-truth sequence. For a set of n elements, it is calculated based on the number of concordant pairs (pairs ordered the same way in both sequences) and discordant pairs (pairs ordered differently). A ROKT score of +1 indicates a perfect match, 0 indicates no correlation, and -1 indicates a perfect inversion. The formula is:

$$\text{ROKT} = \tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)} \tag{16}$$

where N_c is the number of concordant pairs, N_d is the number of discordant pairs, and n is the number of semantic units being sequenced.

ToC Extraction To specifically evaluate the performance of our RAP module on the ToCs page, we utilize three dedicated metrics. These metrics assess the quality of the parsed ToC, which is a critical prerequisite for the final hierarchical alignment.

(1) CC (Content Completeness): This metric measures the percentage of titles within the ToC that are extracted with their text perfectly matching the ground truth.

$$\text{CC} = \frac{|T_{\text{correct}}|}{|T_{\text{total}}|} \times 100\% \tag{17}$$

where T_{correct} is the set of correctly extracted ToC title texts, and T_{total} is the total set of ToC titles in the ground truth.

(2) RC (Region order Consistency): This measures the percentage of ToC entry pairs that maintain the same relative order as their corresponding titles appear in the document body. It assesses if the ToC sequence is consistent with the document's narrative flow.

$$\text{RC} = \frac{|\text{Pairs with consistent order}|}{|\text{Total sequential pairs}|} \times 100\% \tag{18}$$

(3) HC (Hierarchical Consistency): This measures the percentage of parent-child relationships (e.g., a section and its sub-section) within the ToC that are correctly identified by the model.

$$\text{HC} = \frac{|R_{\text{correct}}|}{|R_{\text{total}}|} \times 100\% \tag{19}$$

where R_{correct} is the set of correctly identified hierarchical relations, and R_{total} is the total set of such relations in the ground truth.

TBTA (ToC-Body Title Alignment) We introduce TBTA as a novel and crucial metric to evaluate the model's ability to perform global, cross-page structural reasoning. It measures the percentage of ToC entries that are correctly *linked* to their corresponding title blocks in the main body of the document. A successful link requires not only that the ToC entry is parsed correctly, but also that its corresponding title in the body is accurately identified and associated. TBTA, therefore, directly assesses the most critical aspect of hierarchical document understanding.

$$\text{TBTA} = \frac{|L_{\text{correct}}|}{|L_{\text{total}}|} \times 100\% \tag{20}$$

where L_{correct} is the set of correctly established links between ToC entries and body titles, and L_{total} is the total number of ToC entries that should be linked.

B.2 Extended Analysis of Compass-ESG Results

This section extends the experimental analysis (Tables 1 and 2) to address two core questions: Why do existing paradigms falter on complex ESG reports, and how does Compass-ESG achieve its breakthrough performance? We analyze the performance of different technical paradigms, highlighting their inherent strengths and limitations to elucidate the mechanisms behind Compass-ESG's success. In addition, we provide extended results on its cross-market generalization to diverse layouts and reporting styles.

B.2.1 Specialized Document Parsers: Performance Analysis and Limitations

The first group of baselines comprises document parsers—Textin, MinerU, Docling, and Marker—targeting sub-tasks like text block segmentation and table recognition. Their methodologies and limitations, along with experimental comparisons, are summarized in Table 8. Performance analysis further shows a strong link between architectural design and empirical outcomes.

Marker yielded the weakest performance, with a Macro-F1 of only 39.88%. This can be attributed to two primary factors. First, its optimization for books and scientific papers in English-like languages is a poor fit for our core dataset of Chinese ESG reports, likely causing significant errors at the basic text extraction level. Second, its pipeline's heavy reliance on heuristic rules for block cleaning and sorting is ineffective against the highly heterogeneous and non-standard layouts found in these reports.

Docling delivered a robust Macro-F1 of 74.71%. Its strength lies in its modular design philosophy, integrating best-in-class specialized models such as DocLayNet for layout analysis and TableFormer for table recognition. By leveraging state-of-the-art tools for each sub-task, Docling ensures high-quality parsing of well-defined elements like tables and lists, contributing to its solid overall performance.

Table 8: Comparison of Architectures and Methodologies of Specialized Document Parsers.

Model	Core Architecture / Methodology	Key Advantages	Key Limitations (Based on Our Experiments)
Marker	A pipeline based on deep learning models, including a vision transformer (ViT) and a DONUT model, but with extensive use of heuristic rules for sorting and post-processing.	Optimized for scientific papers and books, achieving high speed on specific formats.	When processing complex Chinese ESG reports, Marker exhibits recognition errors, incorrect layout interpretation, and inaccurate attribution of broken lines/columns, leading to a significant degradation in precision.
Docling	Modular pipeline architecture, integrating state-of-the-art specialized models, such as DocLayNet for layout analysis and TableFormer for table recognition.	Each module is powerful and independent, ensuring high-quality analysis for specific element types like tables and lists.	The modular design is weighted towards local tasks; lacks an integrated semantic inference mechanism, leading to failure in hierarchical alignment.
MinerU	Utilizes a modular pipeline (OCR + Layout Detection + Table Recognition + Post-processing) to extract document content and produce structured output.	Ensures high recall of layout elements by performing a comprehensive, decoupled layout analysis.	Heavily relies on visual heuristic rules for structure determination, unable to handle global hierarchical alignment, leading to extremely low TBTA.
Textin	Proprietary model; the specific architecture is not publicly disclosed.	High precision for fine-grained extraction of local elements; effectively utilizes visual and spatial cues.	Performs poorly on modeling complex and implicit hierarchical structures, resulting in extremely low TBTA. It also exhibits occasional errors in reading order prediction for complex layouts.

MinerU achieved the highest recall in this category at 78.69%, a result consistent with its distinct two-stage pipeline architecture. By first performing a comprehensive page segmentation to identify all potential content regions before recognition, its "segment-then-recognize" strategy effectively minimizes the number of missed elements, thereby maximizing recall.

Textin emerged as the top performer among all specialized parsers, attaining the highest precision (89.65%) and F1 score (82.55%). This superior and stable performance, likely attributable to its proprietary internal algorithms, confirms its strength in core content parsing and reading order modeling, showing no major flaws in these specific areas.

Moreover, our analysis shows a strong correlation between Reading Order Kendall's Tau (ROKT) and Macro-F1: accurate reading order is essential for structural parsing. Misordered blocks—such as swapped columns—prevent correct heading–paragraph or figure–caption associations. Consequently, systems like MinerU and Docling, which include dedicated reading order algorithms, achieve higher ROKT and thus stronger downstream performance in semantic grouping and classification.

Despite strengths in local extraction, all parsers fail catastrophically on the ToC-Body Title Alignment (TBTA) task, with scores below 17%. This exposes the *Semantic–Visual Gap*: a reliance on low-level cues (font size, indentation) without grasping high-level semantics (e.g., true heading levels). Such heuristics may work on simple layouts but collapse in ESG reports where visual prominence often contradicts semantic role (e.g., subheadings appearing visually larger than their parent headings; see Figure 6). Lacking semantic grounding, parsers misplace entries and cannot build stable hierarchies. Crucially, this failure reflects not an implementation bug but an *architectural limitation*. Existing parsers are *page-constrained*, treating each page in isolation. Yet ToC alignment is inherently *document-global*, requiring links across distant pages (e.g., ToC entry on page 3 vs. heading on page 78). By design, these systems cannot capture such global context, making poor TBTA performance an inevitable outcome when applied to complex ESG reports.

Figure 6: A typical example of an implicit hierarchical relationship in an ESG report. The ground-truth hierarchy, overlaid as H2, H3, and H4 labels, must be inferred from a complex combination of visual cues including font size, iconography, and spatial containment (e.g., the H4 block nested within the H3 block). This non-linear, design-driven layout presents a significant challenge for automated document parsing systems.

B.2.2 General-purpose Multimodal Models: Performance Analysis and Limitations

As shown in Tables 1 and 2, general-purpose multimodal models (GMMs) often outperform specialized parsers, likely due to pre-training on large, diverse datasets that support broad semantic and cross-modal reasoning. Their advantage is most apparent in the semantically demanding TBTA task, where performance reaches 64.30%. At the same time, GMMs exhibit limitations when processing complex ESG reports. To control context length and mitigate hallucinations, all models were restricted to 5-page segments—a constraint that exposes a bottleneck in current GMMs for long-document understanding. The performance of open-source GMMs reflects design trade-offs in their respective architectures. Specifically,

DeepSeek-R1 emphasizes Chain-of-Thought reasoning but shows weaker visual grounding, evidenced by a ROKT score of 0.56. On pages with complex layouts (e.g., multi-column or dense text–figure mixtures), inaccurate estimation of the physical reading flow can yield a mis-ordered input sequence, which in turn reduces downstream language reasoning quality and yields an F1-score of 63.74

Qwen3 demonstrates a stronger visual foundation with a higher ROKT of 0.67, yet its overall F1-score is 51.16%. This suggests that while Qwen3 can order content blocks more reliably, it still struggles with fine-grained semantic classification (e.g., distinguishing a level-two heading from a table caption) without task-specific fine-tuning.

Gemini 2.5 Pro attains a Macro-F1 of 87.50%. A remaining limitation is the inconsistency of its hierarchical understanding. Without an explicit schema as reference, its TBTA score is 64.30%. The model readily identifies visually salient headings (e.g., large, centered fonts) but is less reliable for subtle or implicit hierarchical cues common in ESG reports.

These observations point to a central trade-off in document analysis: *Implicit Structure Inference vs. Explicit Structure Guidance*. GMMs are trained to infer implicit structures from raw data, learning probabilistic patterns (e.g., larger text is more likely to be a heading). In contrast, a document's ToC provides a deterministic, author-defined map of the logical structure. Our results indicate that relying solely on implicit inference, even for a strong native multimodal model like Gemini 2.5 Pro, is not sufficient for high-fidelity Hierarchical Document Structure Analysis (HDSA). In addition, performance is further affected by context window limits. Processing a long report in 5-page chunks gives

Figure 7: A sample ToC with a complex, multi-column layout. This example was used for the qualitative analysis to demonstrate the failure modes of the GP method.

the model only a local view. Lacking a global coordinate system, the model's judgment of heading levels can drift across segments, creating a bottleneck for TBTA. The 5-page setting is an engineering compromise that reflects the $O(n^2)$ complexity of Transformer self-attention, which increases the cost of single-pass long-document processing.

In summary, GMMs perform well on ToC extraction, indicating substantial capacity to capture document structure. However, architectural and computational constraints currently limit their extension from local cues to document-level consistency. This creates a practical tension: strong local performance on isolated tasks versus incomplete global consistency across the full document.

B.2.3 Qualitative Analysis and Failure Modes of the General Prompt

The experimental results in the "ToC Extraction" section of Table 1 show that ToC-RAP consistently and significantly outperforms the GP across all tested GMMs and all metrics. This performance increase is particularly prominent for metrics that evaluate structural understanding, namely RC and HC. Key performance gains include: For *Qwen3*, the RC score surged from 56.43% with GP to 96.04% with ToC-RAP, while the HC score increased from 69.31% to 92.08%. *ChatGPT5* saw its RC score improve from 88.20% to a perfect 100%, and its HC score rose from 71.29% to 99.00%. *Gemini 2.5 Pro* also demonstrated significant progress, with its RC score increasing from 93.07% to 96.03%. These results quantitatively confirm that ToC-RAP is a far more effective method than the General Prompt for accurately parsing the complex structure of ToC pages in ESG reports.

A qualitative analysis reveals the root cause of this performance gap. Box B.2, B.3, B.4, and B.5. showcases the parsing results from four leading GMMs on a complex, multi-column ToC (as shown in Figure 7) when guided by the GP (see Box B.1). The GP is a well-engineered but generic prompt for document structural analysis. It lacks the specific heuristics needed to navigate the unique layout challenges of a ToC page, such as multi-column formats and implicit hierarchies. As a result, all models faced a unifying problem: an inability to correctly handle the relationship between semantic main titles and presentational, summary, or symbolic labels. For example, in the ToC, "Social Section" is merely a symbolic label for presentational purposes, summarizing the true main title "Jointly Building a Harmonious Society" and its subordinate level-2 and level-3 titles. The GP cannot guide the model to understand

this complex, non-hierarchical, attributive relationship, leading to various parsing errors. The specific failure modes varied:

Qwen3-Max_Preview: This model incorrectly replaced the true main title with the symbolic label. Furthermore, it failed to recognize the connection between the numerical indices (e.g., "01", "02") and their corresponding titles, treating them as separate entities.

Doubao-Deepthinking: Its most critical error was treating the left and right sides of the sample ToC as two independent regions without connecting their relationship. This approach fundamentally breaks the document's intended sequence, which is a classic error measured by the RC metric.

ChatGPT5 and Gemini 2.5 Pro: These two models shared the same flaw, misidentifying the relationship between the main title and the summary label as a primary-secondary (parent-child) hierarchical relationship. This initial, fundamental misjudgment caused a cascading failure, leading to a collapse in the recognition of all subsequent level-2 and level-3 titles. This type of misjudgment of parent-child relationships is precisely the core issue that the HC metric is designed to capture.

In contrast, ToC-RAP (Appendix A.3.1), with its specialized multi-step directive framework, provides explicit guidance that enables the model to handle complex layouts, thereby avoiding the above failure modes and achieving precise ToC structure parsing.

Box B.1: General Prompt for ToC Extraction

```
## Role Definition
You are a high-precision, multi-source, heterogeneous PDF structural analysis
system. Your core capability is to convert PDF content with chaotic layouts and
diverse data formats (text blocks, tables, images) into structured Markdown text
with a clear hierarchy and a well-defined reading order, ensuring no loss of
original information and no deviation in format.

## Core Mission
- Full Content Parsing: Completely extract all elements (text blocks, tables,
images) from the PDF without omitting any content.
- Structural Conversion: Deconstruct content into a "Page  Content Block"
hierarchy, assigning a unique reading-order identifier to each content block.
- Format Standardization: Standardize different content types into a specified
Markdown format, preserving the original data structure and logical
relationships.

## Execution Guidelines
1. Table Parsing: Strictly replicate the row/column count and cell merge states
of the PDF table into raw Markdown code format. Clearly identify the header (if
present). The reading order tag must immediately follow the end of the table.
2. Image Parsing: Use `![](image_{n}_{m}.jpg)` (where n = page number, m =
reading order number) as the placeholder path. If a caption exists, add "Image
Caption: [Caption Content]". The reading order tag must immediately follow the
caption or the path.
3. All content blocks must be appended with a reading order tag
`[page:n,reading_order:m]` with no extra spaces. Here, 'n' is the accurate page
number of the content block in the original PDF (e.g., page 3 is tagged as
`page:3`), and 'm' is the sequential number of the content block within page 'n',
incrementing from 1 according to the natural human reading logic (e.g., the third
content block on the second page is tagged as `reading_order:3`).

## Output Requirements
1. Output only raw Markdown syntax code without rendering any visual effects.
2. Enclose the output in a ```markdown code block. Use `##` for headings and `-`
for lists.
3. Retain all syntax symbols so the output can be directly copied and pasted into
an editor.
```

Box B.2: Comparison of GMM Parsing Outputs - Qwen3-Max-Preview

Chinese

```
# 总经理致辞
01
# 走进浙数文化
02
# 专题聚焦
07

# 治理篇
11
### 治理架构
### 党建引领
### 投资者保护
### 风险管理与内部控制
### 反腐倡廉

# 社会篇
16
## 产业责任
### 助力数字中国建设
### 正能量信息传播
### 参与行业标准建设

## 社区责任
### 民生服务
### 社区文化传播
### 公益与志愿者
### 园区共建

## 客户与供应商责任
### 自主研发和产品创新
### 产品安全与合规
### 客户管理与服务
### 供应商管理与服务

## 人才发展
### 雇佣关系
### 薪酬与福利
### 发展与培训
### 人文关怀

# 环境篇
35
## 共筑绿色智慧家园
## 打造绿色数据中心

# 未来展望
40
# 关键绩效表
41
# 报告评价
43
```

English

```
# President's Address
01
# Entering Zhejiang Digital Culture
02
# Special Focus
07

# Governance Section
11
## Governance Framework
## Party Leadership
## Investor Protection
## Risk Management and Internal Control
## Anti-Corruption and Integrity

# Social Section
16
## Industrial Responsibility
### Contributing to Digital China Construction
### Disseminating Positive Information
### Participating in Industry Standard Setting

## Community Responsibility
### Public Services
### Community Cultural Communication
### Public Welfare and Volunteering
### Community Co-Building

## Customer and Supplier Responsibility
### Independent R&D and Product Innovation
### Product Safety and Compliance
### Customer Management and Services
### Supplier Management and Services

## Talent Development
### Employment Relations
### Compensation and Benefits
### Development and Training
### Humanistic Care

# Environmental Section
35
## Building a Green Smart Park
## Creating a Green Data Center

# Future Outlook
40
# Key Performance Indicators
41
# Report Evaluation
43
```

Box B.3: Comparison of GMM Parsing Outputs - Doubao-Deepthinking

Chinese

```
# 总经理致辞 01
# 走进浙数文化 02
# 专题聚焦 07
# 治理篇 11
# 社会篇 16
# 环境篇 35
# 未来展望 40
# 关键绩效表 41
# 报告评价 43

## 01 合规治理 护航发展
## 治理架构
## 党建引领
## 投资者保护
## 风险管理与内部控制
## 反腐倡廉

# 02 携手共建 和谐社会
## 产业责任
### 助力数字中国建设
### 正能量信息传播
### 参与行业标准建设

## 客户与供应商责任
### 自主研发和产品创新
### 产品安全与合规
### 客户管理与服务
### 供应商管理与服务

## 社区责任
### 民生服务
### 社区文化传播
### 公益与志愿者
### 园区共建

## 人才发展
### 雇佣关系
### 薪酬与福利
### 发展与培训
### 人文关怀

# 03 践行绿色 环保理念
## 共筑绿色智慧园区
## 打造绿色数据中心
```

English

```
# President's Address 01
# Entering Zhejiang Digital Culture 02
# Special Focus 07
# Governance Section 11
# Social Section 16
# Environmental Section 35
# Future Outlook 40
# Key Performance Indicators 41
# Report Evaluation 43

# 01 Compliance Governance Safeguarding
Development
## Governance Framework
## Party Leadership
## Investor Protection
## Risk Management and Internal Control
## Anti-Corruption and Integrity

# 02 Jointly Building a Harmonious Society
## Industrial Responsibility
### Contributing to Digital China Construction
### Disseminating Positive Information
### Participating in Industry Standard Setting

## Customer and Supplier Responsibility
### Independent R&D and Product Innovation
### Product Safety and Compliance
### Customer Management and Services
### Supplier Management and Services

## Community Responsibility
### Public Services
### Community Cultural Communication
### Public Welfare and Volunteering
### Community Co-Building

## Talent Development
### Employment Relations
### Compensation and Benefits
### Development and Training
### Humanistic Care

# 03 Practicing Green Environmental Protection
Concept
## Building a Green Smart Park
## Creating a Green Data Center
```

Box B.4: Comparison of GMM Parsing Outputs - ChatGPT5

Chinese

```
# 总经理致辞
# 走进浙数文化
# 专题聚焦

# 治理篇
## 01 合规治理 护航发展
### 治理架构
### 党建引领
### 投资者保护
### 风险管理与内部控制
### 反腐倡廉

# 社会篇
## 02 携手共建 和谐社会
### 产业责任
· 助力数字中国建设
· 正能量信息传播
· 参与行业标准建设

### 社区责任
· 民生保障
· 社区文化传播
· 公益与志愿者
· 国家共青

### 客户与供应商责任
· 自主研发和产品创新
· 产品安全与合规
· 客户管理与服务
· 供应商管理与服务

### 人才发展
· 雇佣关系
· 薪酬与福利
· 发展与培训
· 人文关怀

# 环境篇
## 03 践行绿色 环保理念
### 共筑绿色智慧园区
### 打造绿色数据中心

# 未来展望
# 关键绩效表
# 报告评价
```

English

```
# President's Address
# Entering Zhejiang Digital culture
# Special Focus

# Governance Section
## 01  Compliance Governance Safeguarding
Development
### Governance Framework
### Party Leadership
### Investor Protection
### Risk Management and Internal Control
### Anti-Corruption and Integrity

# Social Section
## 02 Jointly Building a Harmonious Society
### Industrial Responsibility
· Contributing to Digital China Construction
· Disseminating Positive Information
· Participating in Industry Standard Setting

### Community Responsibility
· Public Services
· Community Cultural Communication
· Public Welfare and Volunteering
· Community Co-Building

### Customer and Supplier Responsibility
· Independent R&D and Product Innovation
· Product Safety and Compliance
· Customer Management and Services
· Supplier Management and Services

### Talent Development
· Employment Relations
· Compensation and Benefits
· Development and Training
· Humanistic Care

# Environmental Section
## 03 Practicing Green Environmental
Protection Concept
### Building a Green Smart Park
### Creating a Green Data Center

# Future Outlook
# Key Performance Indicators
# Report Evaluation
```

Box B.5: Comparison of GMM Parsing Outputs: Gemini 2.5 Pro

Chinese

```
# 总经理致辞
# 走进浙数文化
# 专题聚焦

# 治理篇
## 01 合规治理 护航发展
### 治理架构
### 党建引领
### 投资者保护
### 风险管理与内部控制
### 反腐倡廉

# 社会篇
## 02 携手共建 和谐社会
### 产业责任
· 助力数字中国建设
· 正能量信息传播
· 参与行业标准建设

### 客户与供应商责任
· 自主研发和产品创新
· 产品安全与合规
· 客户管理与服务
· 供应商管理与服务

### 社区责任
· 民生服务
· 社区文化共建
· 公益与志愿者
· 园区共建

### 人才发展
· 雇佣关系
· 薪酬与福利
· 发展与培训
· 人文关怀

# 环境篇
## 03 践行绿色 环保理念
### 共筑绿色智慧园区
### 打造绿色数据中心

# 未来展望
# 关键绩效
```

English

```
# President's Address
# Entering Zhejiang Digital Culture
# Special Focus

# Governance Section
## 01 Compliance Governance Safeguarding
Development
### Governance Framework
### Party Leadership
### Investor Protection
### Risk Management and Internal Control
### Anti-Corruption and Integrity

# Social Section
## 02 Jointly Building a Harmonious Society
### Industrial Responsibility
· Contributing to Digital China Construction
· Disseminating Positive Information
· Participating in Industry Standard Setting

### Customer and Supplier Responsibility
· Independent R&D and Product Innovation
· Product Safety and Compliance
· Customer Management and Services
· Supplier Management and Services

### Community Responsibility
· Public Services
· Community Cultural Communication
· Public Welfare and Volunteering
· Community Co-Building

### Talent Development
· Employment Relations
· Compensation and Benefits
· Development and Training
· Humanistic Care

# Environmental Section
## 03 Practicing Green Environmental
Protection Concept
### Building a Green Smart Park
### Creating a Green Data Center

# Future Outlook
# Key Performance Indicators
# Report Evaluation
```

B.3 Cross-Market Generalization Analysis of Compass-ESG

To comprehensively evaluate the robustness of Compass-ESG and its adaptability across different languages, formats, and document structures, we further conducted cross-market generalization tests on ESG reports from the Hong Kong and U.S. markets. As described in Section 4.6, our model was trained exclusively on Chinese (A-share) ESG reports. For the cross-market evaluation, the model itself and all hyperparameters were kept unchanged, with only prompt localization applied (e.g., translation into Traditional Chinese or English).

B.3.1 Hong Kong Market Experimental Analysis

This section aims to evaluate the generalization capability of each model on Hong Kong ESG reports after being trained on the Mainland China (A-share) market. This cross-market test utilized 20 ESG reports from Hong Kong-listed companies, with the objective of verifying the models' robustness when facing a new data distribution, rather than pursuing absolute precision, as the models' accuracy has already been validated on the more complex A-share dataset. While reports from the Hong Kong market share thematic similarities with A-share reports, they also exhibit a significant domain shift in financial terminology, reporting conventions, and stylistic phrasing, as mentioned in the main text, posing a true test for the models' generalization capabilities. Detailed experimental data are presented in Table 9.

Table 9: Comparison of specialized document parsers, multimodal models, and Compress-ESG (HK) on the HK ESG benchmark.

Category	Method	Reading Order	Hierarchy Alignment	End-to-end performance		
		ROKT	TBTA (%)	Prec.	Rec.	Macro-F1
Specialized Document Parsers	Marker	0.65	15.50	52.27	50.30	51.27
	Docling	0.78	14.00	68.22	76.30	72.03
	MinerU	0.80	11.20	73.80	71.32	72.54
	Textin	0.87	15.00	78.94	80.32	79.62
General-purpose Multimodal Models	DeepSeek-V3	0.48	30.43	51.17	56.56	53.72
	DeepSeek-R1	0.51	51.33	47.43	55.53	51.16
	Qwen3	0.57	52.00	57.10	46.04	51.00
	Doubao	0.66	35.84	67.92	65.16	66.51
	ChatGPT5	0.67	31.33	67.26	62.36	64.71
	Gemini 2.5 Pro	0.82	55.33	83.39	79.82	81.57
Ours	**Compress-ESG (HK)**	**0.88**	**89.50**	**85.42**	**93.00**	**89.05**

Baseline Model Performance Analysis On the Hong Kong market dataset, the performance of the two baseline categories diverged in an interesting and insightful manner, further exposing the inherent limitations of their respective methodologies.

Specialized Document Parsers: Similar to their performance on the A-share market, these models performed poorly on the TBTA metric in the Hong Kong market, with all scores failing to exceed 16%. Although some models (e.g., Textin's TBTA improved from 9.68% to 15.00%) saw slight improvements due to certain layout features of Hong Kong reports, their extremely high failure rate remained unchanged. This once again confirms that their underlying method, which relies on visual heuristics, lacks the adaptive capability to face new reporting styles, and their 'page-constrained' architecture cannot support cross-domain hierarchical structure understanding. In terms of Reading Order (ROKT), the performance of specialized parsers was relatively stable (scores between 0.65–0.87), indicating their capability to handle conventional page flows is adequate. However, the stark contrast between the high ROKT and low TBTA scores precisely highlights their core 'Semantic-Visual Gap': being able to correctly order blocks is a fundamentally different and less complex problem than truly understanding their hierarchical relationships.

General-purpose Multimodal Models: GMMs demonstrated stronger generalization capabilities than specialized parsers, achieving significantly higher TBTA scores (e.g., Qwen3 at 52.00%, Gemini 2.5 Pro at 55.33%). However, compared to their performance in the A-share training domain, their performance showed a marked decline. For instance, Gemini 2.5 Pro's TBTA score dropped from 64.30% to 55.33%. This decline validates the "domain shift" hypothesis: the powerful implicit structure reasoning ability of GMMs is inherently based on the general patterns and statistical regularities learned from their massive pre-training data. When faced with the different terminology and layout paradigms of Hong Kong reports, the effectiveness of these learned patterns diminishes, causing the models' inference accuracy to be affected and unable to maintain the peak levels seen in a more familiar

data domain. Furthermore, the inherent architectural constraints of GMMs remain a significant factor. Due to context window limitations, long documents must be processed in fragmented 5-page chunks, which prevents a truly holistic, end-to-end analysis in a single pass. This fragmented processing, coupled with potential generative hallucination risks, collectively contributes to the volatility we observe in their end-to-end performance metrics like Precision and Recall. In terms of Reading Order (ROKT), the performance of GMMs showed significant variance (from 0.48 for DeepSeek-V3 to 0.82 for Gemini 2.5 Pro). This indicates that the visual grounding capabilities of different GMMs are inconsistent. For models with weaker visual positioning abilities, like the DeepSeek series, an incorrect reading order severely undermines subsequent semantic understanding, thereby lowering their overall F1 scores and reaffirming the foundational role of reading order in end-to-end parsing tasks.

Compress-ESG's Powerful Generalization In stark contrast to the performance degradation or instability commonly observed in the baseline models, Compress-ESG demonstrated powerful cross-domain generalization and performance robustness in the Hong Kong market test.

On End-to-end Performance and Hierarchy Alignment: As shown in Table 9, Compass-ESG achieved a Macro-F1 score of 89.05% and a high TBTA score of 89.50%. The most critical finding lies not in the absolute numbers, but in the performance stability and the vast gap compared to the baselines. While the TBTA performance of all baseline models was limited by the domain shift, peaking at just 55.33%, Compass-ESG still maintained a state-of-the-art level close to 90%. This decisively proves that our proposed methodology of using the ToCs as a "Structural Scaffold" is key to achieving robust generalization. By relying on the explicit, language- and style-agnostic structural information provided by the document itself (the ToC), rather than on fragile implicit structure inference from the content, our model can significantly reduce its sensitivity to specific report styles and terminologies.

On Reading Order Modeling: Compass-ESG also demonstrated the superiority of its architecture in reading order modeling, achieving a high ROKT score of 0.88, surpassing all baseline models. Although this score is slightly lower than its performance in the A-share market (0.92), it again validates the strong generalization capability of our two-stage hybrid modeling approach. Its strategy of combining "top-down recursive decomposition" with "bottom-up relational modeling" enables it to effectively adapt to the different layout styles of Hong Kong reports, thereby maintaining a state-of-the-art level of sequence understanding accuracy in cross-domain testing.

B.3.2 U.S. Market Experimental Analysis

The U.S. market ESG reports serve as the final test for the model's cross-lingual and cross-style generalization capabilities. Compared to the Mainland China and Hong Kong markets, U.S. reports are typically more uniform and standardized in format, with the clearest hierarchical structures, providing a relatively "friendly" testing environment for baseline models that rely on rules and general patterns. This section aims to analyze how different technical paradigms perform under these idealized conditions and to compare their results with those from more complex markets. Detailed experimental data are presented in Table 10.

Table 10: Comparison of specialized document parsers, multimodal models, and Compress-ESG (US) on the HK ESG benchmark.

Category	Method	Reading Order	Hierarchy Alignment	End-to-end performance		
		ROKT	TBTA (%)	Prec.	Rec.	Macro-F1
Specialized Document Parsers	MinerU	0.81	32.88	87.65	86.18	86.89
	Marker	0.82	37.62	71.61	75.32	73.42
	Docling	0.83	36.44	64.65	65.19	64.89
	Textin	0.87	39.68	92.26	83.60	87.72
General-purpose Multimodal Models	DeepSeek-R1	0.74	51.67	72.41	60.93	66.18
	Qwen3	0.75	43.00	66.43	56.42	61.02
	DeepSeek-V3	0.79	44.38	65.47	60.93	63.12
	ChatGPT5	0.80	57.00	76.70	69.95	73.17
	Doubao	0.80	45.60	69.86	56.20	62.29
	Gemini 2.5 Pro	0.87	73.67	89.78	85.59	87.64
Ours	**Compress-ESG (US)**	**0.94**	**93.13**	**93.61**	**95.50**	**94.30**

Baseline Model Performance Analysis On the relatively more standardized reports of the U.S. market, both categories of baseline models achieved significant performance improvements, yet this also made their respective performance ceilings and limitations clearer.

Specialized Document Parsers: The performance of these models, particularly on hierarchical alignment (TBTA), was substantially improved. For example, Textin's TBTA score leaped from 9.68% on the A-share market to 39.68%. This decisively proves that when document layouts are relatively regular and visual cues (fonts, indentation) are reliable and consistent, the heuristic rules they rely on can achieve better results. However, it is noteworthy that even under these more ideal conditions, the highest TBTA score in this category is still less than 40%, meaning over 60% of hierarchical relationships remain incorrectly aligned. This not only confirms the high dependency of these models on layout regularity but also, conversely, proves that even the "simpler" U.S. ESG reports possess a level of hierarchical complexity that fundamentally exceeds the capabilities of purely visual heuristic-based methods. In terms of Reading Order (ROKT), all specialized parsers achieved high scores (0.81–0.87). This is attributable to the more traditional single- or double-column layouts common in U.S. reports, which makes the intra-page reading flow very clear and easy to determine with geometric rules.

General-purpose Multimodal Models: In contrast to their poor performance in the Hong Kong market, the performance of GMMs also improved across the board in the U.S. market. Gemini 2.5 Pro's TBTA score reached 73.67%, far exceeding its performance in the Hong Kong market and surpassing its baseline in the A-share market. This indicates that the clear and consistent structural patterns in U.S. reports align well with the general patterns learned by GMMs during their massive (primarily English-language) pre-training. When the basis for "implicit inference" becomes reliable, the performance of GMMs is correspondingly enhanced. However, despite their excellent performance on TBTA, GMMs remain constrained by several systemic architectural limitations. First, context window limits require us to process long reports in small chunks, which prevents true batch processing and leads to lower efficiency. Second, GMMs still suffer from hallucination issues when processing long texts. These factors combined mean that their Precision (Prec.) and Recall (Rec.) scores, while high for some categories, cannot be stably maintained at a high level across all semantic categories, showing a degree of volatility. In terms of Reading Order (ROKT), leading GMMs (like Gemini 2.5 Pro) also achieved high scores (0.87), comparable to the specialized parsers. This again shows that for clearly laid-out pages, determining the reading order is no longer a problem. However, it also highlights a key point: a high ROKT score does not directly translate to perfect hierarchical understanding, proving that hierarchical alignment is a far more complex challenge than sequence determination.

Compress-ESG's Powerful Generalization The performance recovery of the baseline models in the U.S. market provides the perfect backdrop to verify the superiority of Compress-ESG. As shown in Table 10, Compass-ESG achieved a Macro-F1 score of 94.30% and a high TBTA score of 93.13%.

On End-to-end Performance and Hierarchy Alignment: As stated in the main text, Compass-ESG's performance in the U.S. market even surpassed its baseline performance on the more complex A-share dataset (F1 score increased from 92.04% to 94.30%, TBTA from 92.46% to 93.13%), which demonstrates that its framework can effectively leverage the clear structure of standardized reports. The most telling evidence of its superiority is the comparison on the TBTA metric: even as the strongest baseline model (Gemini 2.5 Pro) reached its performance peak of 73.67% in this favorable environment, Compass-ESG (93.13%) still maintained a massive leading margin of nearly 20 percentage points. This result specifically demonstrates that the methodology of using the ToCs as a "Structural Scaffold" is not just a strategy to "win from behind" on complex documents, but a more fundamental and robust paradigm under all conditions, capable of achieving near-perfect structural alignment in scenarios where other methods still exhibit significant failure rates.

On Reading Order Modeling: Compass-ESG achieved an extremely high ROKT score of 0.94, also surpassing the best performance of all baseline models (0.87). This result validates the strong generalization capability of our two-stage hybrid modeling approach. The standardized and more linear layouts of U.S. reports allow the first-stage "top-down recursive decomposition" to establish the macro-structure with extreme reliability; subsequently, the second-stage "bottom-up relational modeling" can accurately handle any remaining local ambiguities. The entire framework demonstrated strong adaptability when faced with a new domain that is entirely different in both language (English) and layout style, proving that the reading order model itself also possesses strong cross-domain universality.

In summary, our methodology of using the ToCs as a structural scaffold demonstrates clear advantages, whether applied to documents with complex, variable layouts or to those with uniform formats and well-defined structures. It represents a fundamental paradigm that delivers consistent and robust performance across diverse document types.

B.4 Extended Results on Multi-Level Label Prediction in Financial Markets

Based on structured ESG report data, this section evaluates the MLPDH module within Compress-ESG, which maps content blocks to a three-level label structure: ESG-N category (E for Environmental, S for Social, G for Governance, N for Non-target) → GRI indicator (Global Reporting Initiative indicators correspond to regulatory framework-

Table 11: Multi-level Label Prediction Performance.

Method	Multi-level F1-score			Macro-F1	HLA
	ESG-N	GRI	Sentiment		
SVM+TF-IDF	72.14	61.59	68.31	67.35	-
XGBoost	75.33	65.21	71.18	70.57	-
RoBERTa	80.21	72.30	77.61	76.71	81.31
HAN	81.51	74.11	78.93	78.18	82.12
HMCN	82.70	76.86	79.07	79.54	88.15
MLPDH	**85.62**	**84.23**	**89.11**	**86.32**	**94.78**

specified items) → sentiment (Characterizes enterprises' attitudinal tendencies toward the corresponding indicators). The evaluation is conducted on 15,213 expert-annotated blocks (10,213 train / 1,500 validation / 3,000 test) collected from 50 ESG reports.

As shown in Table 11, MLPDH significantly outperforms all baselines. It achieves a macro-F1 score of 86.32, surpassing the strongest baseline HMCN (Wehrmann et al., 2019) by 6.78%, with the most notable improvements observed at the sentiment level. Most baseline models obtain multi-level F1 scores below 80, with SVM+TF-IDF and XGBoost performing around 70 due to their limited semantic modeling capabilities.

In HLA (Hierarchical Label Accuracy), MLPDH scores highest at 94.78. RoBERTa drops to 81.31 due to parent-child inconsistencies from missing cross-level constraints. HAN (Yang et al., 2016), though using hierarchical attention, trails MLPDH by 12.66%. HMCN adds hierarchical prediction but lacks triplet embedding and cross-level attention, causing a 6.63% drop.

C Multi-Level Prediction with Document Hierarchy

To support fine-grained analysis in financial applications, we propose **MLPDH** (**M**ulti-**L**evel **P**rediction with **D**ocument **H**ierarchy), a hierarchical classification framework for multilayer label prediction in ESG disclosures. Each content block is annotated with: (1) a ESG-N category; (2) a GRI indicator; (3) a sentiment label. MLPDH follows a three-stage pipeline: *ternary embedding → hierarchical attention → hierarchy-aware prediction*.

Ternary Embedding Each content block is represented by a composite embedding that integrates textual semantics, hierarchical context, and global reading order:

$$\mathbf{e}_{\text{blk}} = \mathbf{E}_{\text{text}} + \mathbf{E}_{\text{lvl}} + \mathbf{E}_{\text{pos}}, \tag{21}$$

where $\mathbf{E}_{\text{text}}$ is obtained from the `[CLS]` token of Chinese-RoBERTa-wwm-ext Sun et al. (2021); $\mathbf{E}_{\text{lvl}}$ encodes the heading path $\{h_1, h_2, h_3, h_4\}$ via a GRU:

$$\mathbf{E}_{\text{lvl}} = \mathbf{W}_{\text{lvl}} \cdot \texttt{GRU}([\text{Emb}(h_1), \ldots, \text{Emb}(h_4)]); \tag{22}$$

and $\mathbf{E}_{\text{pos}}$ captures the block's global reading order.

Hierarchical Attention To extract level semantics, we apply stacked attention to propagate hierarchical signals. For each level h, the semantic vector is:

$$\mathbf{v}_{\text{blk}}^{(h)} = \text{softmax}\left(\frac{(\mathbf{W}_q \mathbf{e}_{\text{blk}})^\top (\mathbf{W}_k \mathbf{v}_{\text{ref}}^{(h-1)})}{\sqrt{d}}\right) \cdot \mathbf{W}_v \mathbf{v}_{\text{ref}}^{(h-1)} \tag{23}$$

with $\mathbf{v}_{\text{ref}}^{(0)} = \mathbf{E}_{\text{lvl}}$.

Hierarchy-aware Prediction Each level's label is predicted via sigmoid classification, with hierarchical consistency enforced by a parent-child constraint that penalizes violations of label dependencies:

$$\mathcal{L}_{\text{hier}} = \sum_{h=2}^{H} \sum_{c^h} \max\left(0, P(c^h) - P(\text{parent}(c^h))\right), \tag{24}$$

where $P(c^h)$ is the predicted probability for label c^h at level h. The final objective combines binary cross-entropy loss with the hierarchical constraint:

$$\mathcal{L}_{\text{total}} = \sum_{h=1}^{H} \text{BCE}(P^h, Y^h) + \lambda \cdot \mathcal{L}_{\text{hier}}. \tag{25}$$

During inference, labels with $P > \theta$ (default: 0.5) are selected to form a coherent multi-level label path (e.g., `E` → `e-gri30` → `negative`).

D ATLAS-ESG

There is currently no publicly available dataset that supports the study of deep hierarchical structure understanding in long, heterogeneous, design-driven documents such as ESG reports. To address this gap, we introduce ATLAS-ESG, a large-scale dataset constructed to provide a robust foundation for next-generation AI systems and financial research. In the following, we present its design philosophy, construction process, data schema, and potential value to the research community.

D.1 Comparative Analysis of the ATLAS-ESG Dataset with Existing Works

Although several large-scale document image analysis datasets already exist, they are mostly designed for general, single-page tasks (such as page classification or layout analysis) and have fundamental limitations when processing long, heterogeneous documents with deep semantic hierarchies, such as ESG reports. To clearly position the unique contributions of ATLAS-ESG, we conduct a systematic comparison with seven mainstream related datasets, including DocVQA, RVL-CDIP, PubLayNet, DocBank, DocLayNet, and HierText. The details are presented in Table 12.

Table 12: A systematic comparison of ATLAS-ESG with classic document understanding datasets. The table highlights ATLAS-ESG's unique features designed to overcome the limitations of existing single-page benchmarks for long-document analysis. Key differentiators include: (1) its document-level granularity, comprising complete, multi-page reports instead of isolated pages; (2) its deep hierarchical annotations, providing explicit heading levels (h1-h4) in contrast to the flat structural labels of datasets like PubLayNet and DocBank; and (3) its unique provision of global relational annotations and rich, domain-specific (ESG) semantic labels, which are absent in previous works.

Dataset	Document Composition	Structural Annotation Depth	Global Relation Annotation	Domain & Semantic Labels
ATLAS-ESG (This work)	Full ESG reports, cross-page long documents (26,006 reports, 8,657 listed companies, 7.96M content blocks)	Hierarchical Document Structure Analysis (HDSA), Information Extraction; paragraph, table, image URL, image caption blocks; explicit heading levels h1–h4; page index and reading order for block-level positioning	Provides block-level page index, reading order, and links to original document pages	ESG-specific semantics supporting downstream analysis
DocVQA	Mainly single-page; limited multi-page (12,767 images)	No hierarchical annotation; only Q–A pairs for VQA	No global relation annotation	General-purpose, manually annotated Q–A pairs, no domain semantics
RVL-CDIP	Single-page grayscale images, 400K docs, 16 classes (25K per class)	No structural depth; only document-level class labels	No global relation annotation	16 document types (ads, tables, handwriting, etc.)
PubLayNet	Single-page PDF pages from scientific articles (364,232 images)	Page-level layout annotation (text, title, list, figure, table); only flat "title" label, no nested hierarchy	No global relation annotation	General academic publishing domain; no semantic labels
DocBank	Single-page pages (500K, split 400K/50K/50K)	12 semantic units annotated at token-level (Abstract, Author, Caption, Equation, Figure, Footer, List, Paragraph, Reference, Section, Table, Title); still flat "section" class without depth	No global relation annotation	General academic papers; semantic labels cover common structures
DocLayNet	80,863 single-page documents, diverse sources	Manually annotated bounding boxes for 11 layout objects; page-level only	No global relation annotation	Sources include financial reports, scientific docs, patents, tenders; no domain semantic labels
HierText	11,639 natural scene + document images; avg. 100+ words per image	Word → line → paragraph hierarchy for geometric grouping; does not cover document heading hierarchy	No global relation annotation	No domain semantics; focused on text detection

D.1.1 Document Granularity and Data Scale

While mainstream existing datasets are vast in total page count—for example, RVL-CDIP and DocBank each contain nearly 500,000 pages, and PubLayNet contains 360,000—their common characteristic is that the fundamental processing unit is the independent, single-page image. This "page-level" granularity forces a focus on the recognition of local, intra-page elements, rendering these datasets incapable of capturing or evaluating a model's ability to understand the context of long documents that spans across page boundaries. In stark contrast, the fundamental unit of ATLAS-ESG is the complete ESG report, designed specifically for long-document research. Our dataset comprises a total of 26,006

reports covering 8,657 listed companies, which parse into 7.96 million content blocks. The average report length is well over 20 pages, with A-share market reports containing an average of 431 content blocks per document, Hong Kong reports 396, and U.S. reports 182. This "document-level" granularity, combined with its massive scale in terms of content blocks, provides a crucial foundation for training and evaluating models capable of true global structural reasoning and of breaking through the traditional "page-constrained" architectural paradigm.

D.1.2 Hierarchical Depth and Global Linkage

The structural annotations of current datasets are predominantly flat. For instance, RVL-CDIP only provides document-level category labels; DocVQA offers question-answer pairs but lacks an explicit hierarchy; PubLayNet, DocLayNet, and DocBank, while annotating layout elements like "title" within a page, treat them as a single category, lacking differentiation in the nesting depth of titles. Although HierText introduces a three-level structure of word-line-paragraph, its purpose is text-line clustering and it is similarly confined to single-page scenarios, not addressing the logical outline hierarchy of a document.

ATLAS-ESG achieves two major breakthroughs in this dimension:

Deep Hierarchical Annotation: At the content block level, it explicitly annotates each title with a hierarchical depth from h1 to h4. To the best of our research team's knowledge, this is the first time a direct, fine-grained supervisory signal has been provided for the complex task of Hierarchical Document Structure Analysis (HDSA).

Global Linkage Annotation: By integrating page numbers, reading order, and the correspondence between the ToC and the main body content, ATLAS-ESG is currently the only dataset that can support global linkage modeling (such as the TBTA task). These fine-grained hierarchical, sequential, and positional annotations collectively form a complete document structure representation, providing the necessary cues for a model to understand how elements are organized across dozens of pages.

D.1.3 Domain Specificity and Semantic Richness

Most classic datasets are oriented towards general-purpose documents (like letters, memos) or scientific papers and lack in-depth, domain-specific semantic information. For example, the text content in DocVQA and RVL-CDIP is domain-agnostic; PubLayNet and DocBank focus on the layout of general academic publications and do not contain specific domain labels.

ATLAS-ESG, in contrast, is deeply focused on the financial (ESG) domain and provides rich, domain-specific semantic labels for each content block (such as Environment, Social, Governance categories, GRI indicators, and sentiment polarity). This vastly expands the dataset's utility, making it not only a benchmark for structural analysis but also a high-value resource that can directly support downstream NLP tasks such as ESG issue classification, information extraction, and sentiment analysis. Furthermore, its design, which spans the three major markets of Mainland China, Hong Kong, and the U.S., provides an ideal testbed for cross-lingual and cross-market adaptability research.

D.1.4 Innovation, Core Value, and Limitations of ATLAS-ESG

Innovation and Core Value ATLAS-ESG is not a simple incremental dataset but a strategic resource for driving a paradigm shift in document intelligence research. It provides a high-quality platform for training and evaluating next-generation AI models capable of understanding complex long documents. With its three-level organization (document, page, content block), hierarchical annotations, global linkage, and ESG semantic labels, ATLAS-ESG fills critical gaps in existing resources. Its contributions span three dimensions: global context modeling, hierarchical supervision, and linkage-based navigation.

First, its composition of complete, long documents is key to pushing models beyond the "page-constrained" architecture to perform global contextual modeling.

Second, the deep logical hierarchy annotation (h1-h4) provides, for the first time, a direct supervisory signal for the HDSA task, enabling models to learn and reproduce the precise outline structure of a document.

Finally, the global linkage annotation (ToC-body links) opens up a new research direction, namely, the evaluation and optimization of a model's document-level navigation and information localization capabilities.

Limitations Despite its significant advantages, ATLAS-ESG also has certain limitations. For instance, the ESG disclosure rates vary significantly across different markets (approx. 41.86% for A-shares, full disclosure for Hong Kong, and approx. 69.99% for the U.S.); this imbalance could lead to model bias towards certain markets. Furthermore, the current dataset focuses on long, formal reports, and its applicability to other document types (such as short news articles or social media texts) remains to be explored.

D.2 Dataset Construction Process

Data Sourcing and Filtering The construction of the ATLAS-ESG dataset aims to provide a benchmark that spans key, representative global financial markets. Our data sources are as follows:

- ESG reports for Mainland China (A-share) and Hong Kong markets were sourced from the `Wind` database[10].
- ESG reports for the U.S. market were obtained via legal and compliant web scraping of the public data source `responsibilityreports.com`[11].

We initially collected a large corpus of reports and applied a strict filtering criterion: each report must be longer than 20 pages. Based on our observations, this threshold effectively ensures the presence of a ToC, which is crucial for our hierarchical and global linkage research. After preprocessing and filtering, the final dataset statistics are as follows:

- *Mainland China Market:* From 4,077 reports collected from 2,257 listed companies (2021–2025, 41.86% disclosure rate), 3,376 reports were ultimately processed, comprising 1.457 million content blocks.
- *Hong Kong Market:* From 13,057 reports collected from 2,631 listed companies (2020–2025, full disclosure), 11,139 reports were processed, comprising 4.413 million content blocks.
- *U.S. Market:* From 11,873 reports collected from 3,769 listed companies (2020–2025, 69.99% disclosure rate), 11,491 reports were processed, comprising 2.097 million content blocks.

Data Preprocessing and Quality Assurance Before the annotation phase, a multi-stage preprocessing pipeline was applied to ensure data quality and structural integrity. Key steps include:

1. *Layout Classification:* To ensure the robustness of the macro-level reading order, we first classify each page based on its dominant layout (e.g., Vertical-Manhattan or Horizontal-Manhattan). This allows for more adaptive downstream processing.
2. *Rigorous ToC Extraction:* A critical step is the meticulous extraction of the ToCs. We employ a multi-round filtering process to guarantee the successful extraction of all ToC pages. This process includes specialized logic to handle complex cases such as multi-page and partial-page ToCs, providing a reliable foundation for the subsequent ToC-RAP and ToC-ALIGN modules.
3. *Image URL Verification:* To ensure the integrity of the multimodal components, all extracted image URLs are verified for accessibility, guaranteeing their availability for the multimodal aggregation step.

Annotation Pipeline To achieve high-precision annotation at a large scale, we adopted a two-stage "human-in-the-loop," expert-led "Bootstrapping" annotation pipeline.

1. *Automated Pre-annotation.* All filtered PDF reports were first processed by our proposed Compass-ESG model for fully automated structural analysis. The model outputs initial structured JSON files containing hierarchy, reading order, and content block types.
2. *Manual Verification and Correction by Domain Experts.* The pre-annotated results were then reviewed by a team of experts with backgrounds in ESG and finance. The verification process strictly followed our established Annotation Guideline, ensuring the final quality of the annotations.

Quality Control We implemented rigorous quality control measures to ensure the consistency and accuracy of the annotated data.

- *Multi-round, Back-to-Back Verification:* The annotations for each document were independently verified by at least two domain experts. In cases of disagreement, a third senior expert was brought in for final arbitration.
- *Inter-Annotator Agreement (IAA):* As a gold standard for measuring dataset quality, we calculated the Inter-Annotator Agreement (IAA) score. This provided a quantitative basis for the dataset's reliability.

D.3 Dataset Schema in Detail

The ATLAS-ESG dataset adopts a three-level nested structure (document, page, and content block level) to simultaneously capture the macro-level metadata of reports and the micro-level, fine-grained structural and semantic information. This design provides comprehensive and flexible data support for complex document analysis tasks. The overall field structure of the dataset is shown in Table 13.

[10] https://www.wind.com.cn/mobile/WFT/zh.html
[11] https://www.responsibilityreports.com/Companies?exch=1

Table 13: Hierarchical Field Structure of Each Structured ESG Report in ATLAS-ESG

Level	Category	Field	Description	Example
Document Level	Document Identifiers	stock_code	Stock code	300XXX.SZ
		company_name	Company name	FuXiang Pharma
		report_year	Reporting year	2024
		report_title	Title of the ESG report	ESG Report 2024
	Source Metadata	report_type	Report type	ESG Report
		market	Market the report belongs to	China
		original_filename	Original JSON file name	300XXX.SZ-...json
Page Level	Indexing	page_idx	Page index (starting from 1)	1, 2, 3...
	Rendering Links	page_markdown_url	Markdown image link	
		page_file_url / page_relative_path	Local image file path	/mnt/data/.../1.jpg
		page_http_url	Online accessible HTTP path	http://.../1.jpg
Content Block Level	Hierarchical depth	h1–h4	Structural depth of section headings	Environmental Management
	Content type and payload	data_type	Block type	text / table / image
		data	Block content	"...company 2024 energy..."
	Visual Resource Links	markdown_url	Markdown link (table/image)	
		file_url / relative_path	Local high-res path	/mnt/data/...jpg
		http_url	Online access path	http://.../12.jpg
	Ordering	reading_order	In-page reading order	0, 1, 2...
	Semantic tag	esg_category_label	ESG category label	E / S / G / N
		gri_label	One of the 32 GRI labels	Energy
		sentiment_label	Sentiment polarity	Positive / Neutral / Negative

D.3.1 Detailed Description of Fields at Each Level

Document Level This level contains two core types of information. `Document Identifiers` provide basic identity information for the report, such as the stock code (`stock_code`) and company name (`company_name`). `Source Metadata` records the source information of the report, such as the market it belongs to (`market`) and the original filename (`original_filename`).

Page Level This level is responsible for indexing and localization. The `Indexing` field (`page_idx`) provides the accurate page number of the content in the original PDF. `Rendering Links` offers multiple paths to access the page image: `page_markdown_url` is used for referencing the image when converting JSON to Markdown, while `page_relative_path` (local relative path) and `page_http_url` (online path) facilitate direct access and rendering of the page when processing the JSON file.

Content Block Level This is the most information-dense level in our dataset, containing rich structural and semantic annotations:

Hierarchical depth (`h1-h4`): This is an explicit annotation of the structural depth of the block and one of the core features of our dataset. A key design choice is that `h1` to `h3` are used to annotate true hierarchical titles that explicitly exist in the ToCs, while `h4` is used to annotate "pseudo" hierarchical titles or important indicative texts that do not appear in the ToC but are presented with visual styles like bolding or color changes in the main body.

Content type and payload (`data_type` and `data`): `data_type` indicates the data type of the content block in the original document, primarily including text, table, and image (notably, ESG reports generally do not contain formulas common in scientific papers). The `data` field stores the standardized content processed by Compass-ESG.

Visual Resource Links: This field links each content block to its visual resources. The `markdown_url` is used for referencing tables/images in Markdown format, `file_url`/`relative_path` provide local high-resolution copies, and `http_url` enables online access, supporting accurate rendering and external reference.

Ordering (`reading_order`): This is an explicit annotation of the intra-page reading order. This field, combined with the page-level `page_idx`, can provide a globally unique and precise coordinate for every content block in the entire document.

Semantic tag: This field provides rich, domain-specific semantic labels. This allows ATLAS-ESG to be used not only for structural analysis but also to directly support downstream ESG content analysis tasks. The specific GRI indicators and sentiment labels are categorized as shown in Tables 14, 15, 16, and 17.

Table 14: Classification of Environmental (E) category labels in ATLAS-ESG. These labels provide fine-grained annotations for content blocks on environmental management, supporting tasks such as climate risk analysis and emissions tracking. Categories like E-EMISSION and E-CLIMATE align with global ESG standards (e.g., GRI) and are defined by scope and typical keywords to ensure consistency.

Code	Name	Typical Keywords/Expressions
E-ENERGY	Energy Management	Energy consumption, clean energy, conservation
E-WATER	Water & Wastewater Management	Water withdrawal, efficiency, treatment, discharge
E-BIODIV	Biodiversity Protection	Ecosystem, conservation, restoration
E-EMISSION	Emissions Management	Greenhouse gas (GHG), NOx, SOx
E-WASTE	Waste Management	Waste classification, recycling, disposal
E-ENVCOMP	Environmental Compliance	Fines, penalties, compliance
E-CLIMATE	Climate Change Adaptation & Mitigation	Carbon neutrality, emissions reduction, adaptation
E-GREENSC	Green Supply Chain & Procurement	Green procurement, environmental supply chain
E-SUPPENV	Supplier Environmental Assessment	Supplier audit, environmental standards

Table 15: Classification of Social (S) category labels in ATLAS-ESG. This schema annotates content blocks on labor practices, human rights, and community relations, central to social responsibility evaluation. Fine-grained labels such as S-LABORPRAC and S-HUMANRTS support downstream social impact analysis and are grounded in themes from major ESG reporting frameworks.

Code	Name	Typical Keywords/Expressions
S-EMPLOY	Employment & Labor Management	Recruitment, compensation, contract, benefits
S-LABORREL	Labor-Management Relations	Union, communication, satisfaction
S-OHS	Occupational Health & Safety	Workplace safety, work injury, training
S-TRAINING	Training & Education	Training, learning, development
S-DIVERSITY	Diversity & Equal Opportunity	Diversity, equality, anti-discrimination
S-HUMANRTS	Non-discrimination & Human Rights	Human rights, discrimination, supply chain
S-FREEDOM	Freedom of Association & Collective Bargaining	Union, collective bargaining, freedom
S-LABORPRAC	Child Labor & Forced Labor	Child labor, forced labor, audit
S-COMMUNITY	Community Involvement & Impact	Community, public welfare, investment
S-CUSTHS	Customer Health & Safety	Product safety, consumer, health
S-MARKETING	Marketing & Product Labeling	Marketing, advertising, transparency
S-CUSTPRIV	Customer Privacy Protection	Privacy, data protection, data breach
S-SUPPSOC	Supplier Social Assessment	Supplier, social responsibility, assessment

Table 16: Classification of Governance (G) category labels in ATLAS-ESG. These labels capture content on corporate governance, risk management, and business ethics. Tags such as G-ANTICORR and G-RISK enable detailed analysis of governance structures and policies, aligning with disclosure requirements in major ESG standards.

Code	Name	Typical Keywords/Expressions
G-GOV	Org. Governance & Structure	Board, committee, governance framework
G-STRATEGY	Strategy & Targets	Strategy, targets, execution
G-RISK	Risk Management	Risk identification, assessment, control
G-ANTICORR	Anti-corruption	Bribery, corruption, compliance
G-ANTICOMP	Anti-competitive Behavior	Monopoly, competition, compliance
G-TAX	Tax Compliance	Taxation, compliance, transparency
G-LEGAL	Legal & Policy Compliance	Law, regulation, compliance
G-DISCLOSE	Transparency & Information Disclosure	Disclosure, transparency, reporting
G-INVESTREL	Investor Relations Management	Investor, communication, shareholder
G-ESGMGMT	Overall ESG Management	ESG management, responsibility, performance

Table 17: Classification of fine-grained sentiment polarity labels in ATLAS-ESG. Beyond the standard Positive, Negative, and Neutral, two additional categories—Cautiously Optimistic and Concerned—capture nuanced rhetorical stances in ESG reports, such as acknowledging risks while projecting a positive outlook. This schema supports more accurate analysis of a document's underlying tone.

Code	Sentiment	Description
SENT-P	Positive	Showcasing positive results or receiving authoritative recognition.
SENT-N	Negative	Disclosing negative events such as violations, accidents, or losses.
SENT-NEU	Neutral	Objective, factual statements without significant emotional coloring.
SENT-CO	Cautiously Optimistic	Acknowledging challenges while emphasizing strategies and positive outlooks.
SENT-C	Concerned	Implying or explicitly stating potential risks, uncertainties, or future challenges.

E Statement on the Use of Large Language Models

Large language models were employed to assist with language refinement. The authors remain solely responsible for the accuracy, validity, and originality of all content, including any text generated with AI assistance.