TETHER: Training-Free, Elastic, ToC-Anchored Hierarchy Extraction and Recovery in Design-Driven Long Documents

Anonymous ACL submission

Abstract

Converting complex documents into structured knowledge constitutes the critical infrastructure urgently required for deploying Large Language Models (LLMs) in downstream research. However, blind to the Table of Contents (ToC) as a macro-structure, existing bottom-up paradigms rely on localized visual cues, inducing heuristic fragility in high-entropy layouts where visual prominence decouples from logical depth; meanwhile, direct LLM extraction lacks spatial grounding, triggering hallucinations in long contexts. To address this, we introduce **TETHER**, a training-free framework that elevates the ToC from discarded noise to a vital structural prior, enabling robust adaptation across design-driven long documents. TETHER operates via a two-stage neuro-symbolic mechanism: first, it constructs a document skeleton from the raw ToC using a Large Multimodal Model (LMM) empowered by spatio-semantic constraints; second, it employs an anchor-based alignment strategy to tether content blocks to this skeleton, effectively resolving boundary ambiguities. Evaluated across a multilingual benchmark of ESG reports, product manuals, and annual reports, TETHER delivers highly precise, state-of-the-art (SOTA) zero-shot hierarchy recovery, outperforming deep hierarchical recovery models, dedicated parsers, and general-purpose LMMs. Moreover, stress tests highlight TETHER's resilience to sparse omissions and textual noise, demonstrating that its precise semantic alignment is fundamentally anchored in topological and depth fidelity. Crucially, ablations confirm the framework's architectural irreducibility, while achieving 90% token reduction and drastic speedups over end-to-end baselines, establishing a compelling equilibrium between structural precision and deployment scalability.

1 Introduction

Deploying Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) in real-world domains heavily relies on processing massive, complex documents. To unlock their reasoning potential, converting these unstructured layouts spanning hundreds of pages into rigorous hierarchical formats has become an urgent prerequisite, providing the essential infrastructure to navigate deep cross-page dependencies.

However, current paradigms struggle to maintain structural coherence in long-form documents. Bottom-up models (e.g., deep hierarchical recovery models (Hu et al., 2022; Wang et al., 2023) and dedicated parsers (Niu et al., 2025; Datalab, 2025)) rely on local visual cues to infer structure, leading to severe structural distortion whenever visual formatting contradicts logical hierarchy. Moreover, such specialized approaches are inherently prone to domain overfitting. Concurrently, direct LLM extraction lacks spatial grounding, inducing hallucinations across unanchored extended contexts (Perot et al., 2024; Liu et al., 2025).

To mitigate the challenges of extended lengths and high-entropy layouts in documents, a deterministic global map is imperative. Fortunately, virtually all ultra-long documents conventionally feature a native structural blueprint: the ToC. Surprisingly, this structural prior remains largely unexploited in prior work, as real-world ToCs are frequently structurally implicit, typographically heterogeneous, and severely corrupted by noise.

Therefore, we introduce **TETHER** (**T**raining-Free, **E**lastic, **T**oC-Guided **H**ierarchy **E**xtraction and **R**ecovery), an architecture tailored for long-form documents with complex hierarchical structures. Rather than treating the ToC as extraneous information (Hu et al., 2022; Wang et al., 2023), TETHER formulates it as a native structural prior, serving as an explicit reference schema to anchor cross-page content alignment. TETHER operates through a two-stage neuro-symbolic mechanism. In the first stage, a Large Multimodal Model (LMM) guided by spatio-semantic constraints ex-

tracts a rigorous document skeleton from the raw ToC. In the second stage, an anchor-based alignment strategy tethers content blocks to this skeleton to resolve boundary ambiguities.

To evaluate TETHER's effectiveness, generalizability, and architectural necessity, we curate a multilingual benchmark (English, Simplified, and Traditional Chinese) comprising annual reports (proxying rigidly structured, compliance-driven documents), ESG reports (representing layout-flexible & unstructured publications), and product brochures (mirroring fragmented technical literature). This heterogeneity (Appendix A) benchmarks zero-shot structural recovery across unseen layouts.

Empirically, TETHER outperforms deep hierarchical recovery models (MTD (Hu et al., 2022) and CMM (Wang et al., 2023)), dedicated parsers (MinerU2-VLM (Niu et al., 2025), Marker (Datalab, 2025), Docling (Livathinos et al., 2025), and TextIn (Hehe Information and TextIn)), and general-purpose LMMs (Qwen3.5-397B-A17B, GPT-5.5, Gemini3.1-Pro, and DeepSeekOCR). Ablations confirm its architectural necessity, while efficiency analyses reveal 90% token and 80% runtime reductions versus end-to-end baselines, striking a balance of precision and scalability. Moreover, stress tests highlight TETHER's resilience to sparse omissions and textual noise, demonstrating that its precise semantic alignment is fundamentally anchored in topological and depth fidelity.

Our main contributions are as follows:

(1) Paradigm Shift. Reversing the reliance on bottom-up local aggregation, we propose TETHER to reformulate document parsing as ToC-anchored structural grounding. Driven by top-down constraint satisfaction, this paradigm strictly confines outputs within the document's native hierarchy, neutralizing the risks of unconstrained extraction.

(2) Methodological Novelty. Central to our approach is a training-free neuro-symbolic architecture with adaptive anchoring. Synergizing LMM-based ToC extraction and iterative content alignment, it ensures topological consistency, conquers spatial blindness, and achieves zero-shot universality across diverse domains.

(3) Empirical Superiority. Evaluated across three diverse categories of highly irregular multilingual documents, TETHER demonstrates SOTA superiority, resilience to intermediate skeleton perturbations, and architectural irreducibility, while achieving drastic token and runtime reductions against end-to-end baselines.

2 Related Work

Recent work increasingly treats long documents not as flat OCR streams, but as structured objects whose logical hierarchy determines whether downstream LLM and RAG systems can navigate them reliably (Ma et al., 2024; Livathinos et al., 2025; Jeong et al., 2025). Existing structure recovery methods broadly fall into two paradigms.

Content-Driven Hierarchy Prediction. Bypassing ToC priors, these methods extract structural hierarchies directly from page content. First, specialized heuristic parsers (e.g., Docling (Livathinos et al., 2025), MinerU (Wang et al., 2024), Marker (Datalab, 2025), TextIn (Hehe Information and TextIn)) couple layout analysis with OCR engines (JaidedAI and contributors; Cui et al., 2025; Smith, 2007), deriving logical hierarchy by applying predefined rules to visual typographic cues (e.g., font size, indentation). Second, deep hierarchical models (e.g., MTD (Hu et al., 2022), CMM (Wang et al., 2023)) train neural networks to infer hierarchical sequences from learned typographic and layout priors. Third, training-free generative approaches such as DocsRay (Jeong et al., 2025) bypass explicit layout analysis entirely, prompting large multimodal models (LMMs) to synthesize a pseudo-ToC solely from raw page content. Crucially, under high-entropy layout conditions, visual prominence frequently decouples from logical structural roles, leading to systematic failures in both heuristic and neural parsers (Pfitzmann et al., 2022; Da et al., 2023). Notably, CMM reports that over 10.8% of headers deviate from learned typographic patterns, indicating typographic overfitting and limited out-of-distribution generalization (Wang et al., 2023). Meanwhile, lacking spatial grounding, LMM-based generative methods inevitably hallucinate hierarchical boundaries when processing extended contexts (Jeong et al., 2025).

Metadata-Driven Extraction. This paradigm avoids noisy content-level prediction by recovering hierarchy from pre-existing digital signals, such as PDF tags, outlines, or XML-derived heading features (Iso, 2017; Adobe, 2026). For example, HiPS segments legal textbooks by matching PDF outline metadata to the full text (Wehnert et al., 2025). However, this assumes noise-free metadata and fails to handle discrepancies between the metadata and the main text. Notably, constrained by extreme layout heterogeneity, parsing the visual ToC

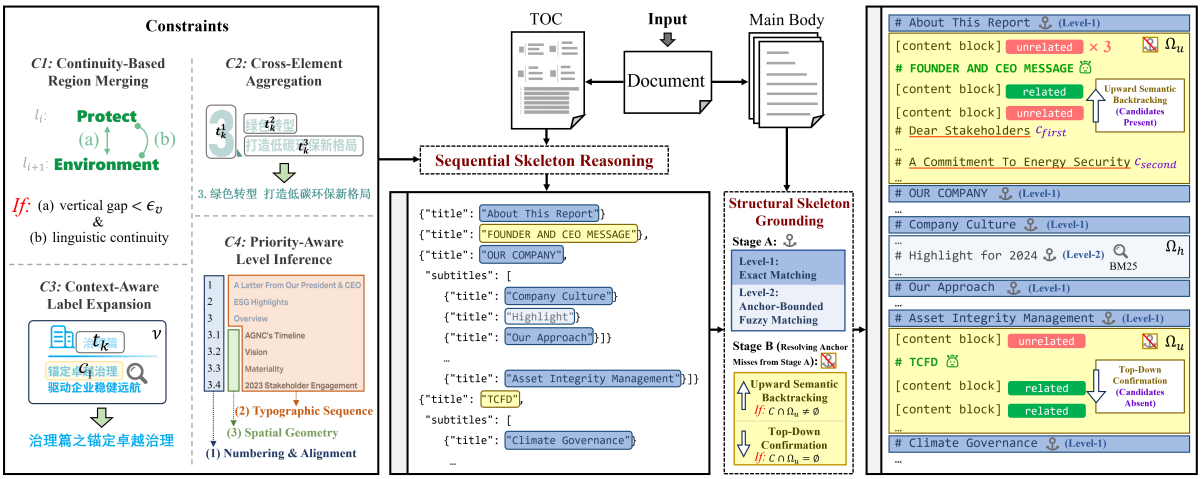

Figure 1: Overview of the TETHER. (1) *Sequential Skeleton Reasoning* distills ToC structural priors under spatio-semantic constraints (C1–C4). (2) *Structural Skeleton Grounding* aligns this skeleton to the body via a coarse-to-fine neuro-symbolic pipeline of symbolic matching (Stage A) and agentic inference (Stage B).

as a structural prior to guide body alignment has been a widely avoided paradigm (Wang et al., 2023; Jeong et al., 2025; Wehnert et al., 2025). TETHER pioneers this approach, addressing two core challenges: (1) Hierarchy extraction against severe ToC layout entropy, and (2) Hierarchy recovery to resolve complex ToC–body textual discrepancies.

3 Methodology

TETHER's pipeline (Figure 1) comprises *Sequential Skeleton Reasoning* to repurpose the ToC as a deterministic structural prior, and *Structural Skeleton Grounding* for precise content alignment.

3.1 Sequential Skeleton Reasoning

We formulate the extraction of the target skeleton $\mathcal{H}$ from raw ToCs as an LMM-driven sequential reasoning task. This process is governed by universal structural priors, formally defined as an orthogonal framework of four constraints that systematically decouple layout anomalies across spatial (C1, C2), semantic (C3), and hierarchical (C4) dimensions:

***C1**: Continuity-Based Region Merging.* To resolve header fragmentation in narrow-column layouts, adjacent lines l_i and l_{i+1} are merged if they satisfy spatial proximity (vertical gap $< \epsilon_v$, scaled by dominant line height) and linguistic continuity constraints, yielding unified semantic units.

***C2**: Cross-Element Aggregation.* To mitigate visual scattering in complex layouts, distributed title segments (e.g., numbers, core texts, and tags) are unified into a composite title t_k based on horizontal alignment and layout grouping signals (e.g., shared backgrounds).

***C3**: Context-Aware Label Expansion.* To resolve semantic sparsity in labels t_k (e.g., "Part A"), descriptive candidates are harvested from a proximal visual window $\mathcal{V}$. Candidates are validated via composite Jaccard and normalized Levenshtein metrics to synthesize a semantically complete title.

***C4**: Priority-Driven Level Inference.* Subordinating unstable visual cues (e.g., the "larger-is-higher" rule) to structural logic, we infer hierarchical level ℓ_k via a priority-ordered cascade: (1) *Numbering & Alignment* to validate nesting and sibling relationships; (2) *Typographic Sequence* incorporating weight, size, and color to resolve ambiguous ties; and (3) *Spatial Geometry* utilizing indentation and gaps to delineate parent boundaries.

Skeleton Formalization. Ultimately, this reasoning chain yields a purified hierarchical sequence:

$$\mathcal{H} = (h_k = (t_k, \ell_k))_{k=1}^{K} \quad (1)$$

where each entry h_k couples the reconstructed text t_k (via C1–C3) with its inferred depth $\ell_k \in \mathbb{N}$ (via C4). Crucially, the index k strictly follows the document's natural reading order thanks to the LMM's sequential generation. This ordered skeleton $\mathcal{H}$ then serves as the deterministic anchor for the downstream alignment phase.

3.2 Structural Skeleton Grounding

To ground the structured skeleton $\mathcal{H}$ onto the linear document stream D, we propose a two-stage neuro-symbolic protocol: (1) *Symbolic Deterministic Anchoring* establishes a definitive structural

backbone, followed by (2) *Neural Agentic Fine Insertion* for context-aware boundary calibration.

3.2.1 Task Formulation: Topology-Constrained Insertion

Formally, given the document stream $D = (b_n)_{n=1}^{N}$ and visual candidates $\mathcal{C} \subset D$, we formulate alignment as a latent structure insertion task. Conventional extraction methods only work with observed candidates, while we cast alignment into a topology-constrained insertion process. Each node $h_k \in \mathcal{H}$ is anchored within a restricted search space, where parent-child hierarchy and sequential order serve as hard constraints. Benefiting from this formulation, our method infers latent boundaries within valid topological windows to locate ghost nodes, leveraging structural reasoning to compensate for insufficient visual signals.

To avoid large-scale annotation, we propose a training-free framework that models this insertion as a constraint satisfaction problem (CSP) anchored by $\mathcal{H}$. This formulation fundamentally decouples topological integrity from semantic reasoning: symbolic logic provides mathematical guardrails, ensuring the neural agent operates exclusively within a provably safe sub-space. Consequently, this decoupled architecture yields three system-level guarantees:

- *Zero-Shot Generalization:* Symbolic logic and few-shot reasoning enable direct domain adaptation without fine-tuning.
- *Bounded Inference:* Anchor-defined local windows reduce search complexity from linear scanning to localized queries.
- *Topology-Preserving Integrity:* Global hierarchy constraints strictly prevent logical violations during boundary inference.

3.2.2 Stage A: Symbolic Anchoring

This stage establishes a definitive structural backbone via a precision-first, two-level symbolic protocol to anchor high-confidence matches in $\mathcal{C}$.

Level-1: Global Exact Matching. We initialize the anchor set $\mathcal{M}$ with normalized exact matches ($\text{Norm}(c_j) = \text{Norm}(h_k)$) from $\mathcal{C}$ for each $h_k \in \mathcal{H}$.

Level-2: Anchor-Bounded Fuzzy Matching. For each unmatched h_k, we define a window Ω_h bounded by its nearest anchors in $\mathcal{M}$. Within Ω_h, we apply BM25 retrieval over $\mathcal{C}$ to resolve noisy variations (e.g., "Env." to "Environmental"). These matches augment $\mathcal{M}$, leaving the residual *ghost nodes* $\mathcal{U}$ for the neural phase.

Crucially, this stage reduces intractable global search to Ω-bounded inference. This topological firewall mathematically restricts the subsequent neural reasoning space, preventing local errors from cascading into structural hallucinations.

3.2.3 Stage B: Neural Agentic Fine Insertion

For each ghost node $u \in \mathcal{U}$, the neural agent pinpoints its optimal semantic onset p^* within an $\mathcal{M}$-anchored constraint window Ω_u in two steps:

Step 1: Topology-Constrained Initialization. We establish a tentative anchor p_{init} for u, adapting to the presence of visual candidates ($\mathcal{C} \cap \Omega_u$):
Scenario I ($\mathcal{C} \cap \Omega_u \neq \emptyset$): The agent scans the candidate subset for semantic matches, anchoring p_{init} to the upper boundary of the first match c_{first}.
Scenario II ($\mathcal{C} \cap \Omega_u = \emptyset$): Lacking candidates, p_{init} deterministically defaults to immediately below the window's upper bound Ω_{start}.

Step 2: Semantic Boundary Calibration. From p_{init}, the agent eschews unconstrained generation for Ω-gated, state-aware traversals to pinpoint the semantic onset p^* via scenario-specific strategies:

Calibration for Scenario I (Upward Semantic Backtracking): To capture introductory content strictly preceding the candidate heading, the agent iteratively scans text blocks upward from p_{init}. Upon detecting semantic relevance to u, p^* is re-anchored to the current block's upper boundary. Backtracking terminates upon encountering three consecutive unrelated blocks from the previous section, finalizing p^* at their lower boundary.

Calibration for Scenario II (Top-Down Confirmation): The agent executes a semantic scan downstream from p_{init}, bypassing transitional noise or irrelevant segments belonging to the previous section. The process terminates upon validating two consecutive blocks relevant to u, positioning p^* at the upper boundary of the first valid block.

Finally, recovered nodes anchored at p^* augment $\mathcal{M}$ to complete the mapping. Crucially, bounded by Ω, the agent transcends exact string matching to achieve robust semantic alignment. By analyzing contextual semantics, it successfully resolves complex paraphrasing (e.g., aligning the body's "Leader's Speech" with the ToC's "Chairman's Address") without risking the downstream structural hallucinations typically triggered by unbounded LLM extraction.

Table 1: Hierarchy extraction performance (%). Main values reflect our proposed ***constrained extraction*** (an orthogonal 4-constraint framework); bracketed values (+Δ) indicate gains over prior-free ***naive extraction***. ***Ours*** applies TETHER on the compact Qwen2.5-VL, enabling secure, local deployment on limited resources.

Model	Annual Reports			Product Manuals			ESG Reports		
	CC	OC	HC	CC	OC	HC	CC	OC	HC
Qwen3.5-397B-A17B	**100** (+5.0)	**100** (+8.8)	97.0 (+7.9)	**100** (+5.9)	**98.0** (+2.0)	97.0 (+4.9)	97.2 (+4.0)	98.5 (+4.2)	**99.0** (+9.4)
GPT-5.5	97.2 (+10.8)	98.3 (+7.0)	**98.5** (+19.1)	93.0 (+12.0)	94.1 (+7.0)	**98.0** (+20.8)	**99.0** (+13.8)	**100** (+11.8)	96.7 (+25.4)
Gemini3.1-Pro	**100** (+3.3)	**100** (+4.5)	98.2 (+2.6)	97.0 (+3.9)	96.0 (+2.9)	94.1 (+3.0)	98.5 (+3.6)	98.1 (+6.8)	98.1 (+4.7)
DeepSeekOCR	96.2 (+10.7)	95.3 (+11.0)	96.1 (+13.0)	94.2 (+9.6)	95.1 (+18.7)	92.1 (+12.8)	96.4 (+12.1)	94.0 (+11.8)	96.1 (+17.8)
Ours (Qwen2.5-VL)	95.6 (+8.7)	96.9 (+15.1)	96.0 (+21.2)	93.2 (+11.8)	95.5 (+17.8)	93.8 (+25.6)	95.9 (+10.0)	93.5 (+12.1)	95.9 (+19.1)

4 Experimental Setup

4.1 Dataset Construction

We introduce TETHER-Gold, a benchmark comprising 150 recent documents that represent three structural archetypes: (1) Strictly Regulated (50 annual reports, 6,392 pages): Proxying formal, compliance-driven documents like government regulations and industry standards; (2) Stylized & Unconstrained (50 ESG reports, 3,651 pages): Characterized by high layout entropy, creative typography, and non-linear reading orders, typical of design-centric publications like corporate white papers and visual magazines; and (3) Highly Fragmented (50 product manuals, 11,193 pages): Defined by disjointed text flows frequently interrupted by dense tables, warning blocks, and multimodal diagrams, mirroring complex technical documentation such as medical equipment guides and hardware specifications. Furthermore, to ensure cross-lingual adaptability, the corpus is evenly balanced across English, Simplified, and Traditional Chinese.

Five specialists and a meta-reviewer executed structure-to-content mapping, parsing ToCs into full-depth skeletons to localize each node's semantic onset p_k. Ghost node boundaries were inferred via latent semantic transitions and topological constraints. Independent double-annotation yielded an overall block-level Cohen's Kappa (κ) of 0.91 (0.98 for explicit anchors, 0.83 for ghost nodes), with discrepancies adjudicated by the meta-reviewer.

4.2 Baselines

TETHER is benchmarked against three system categories: (1) *Deep Hierarchical Recovery Models* (MTD (Hu et al., 2022), CMM (Wang et al., 2023)), predicting hierarchies via supervised visual monotonicity biases; (2) *Dedicated Document Parsers* (MinerU2-VLM (Niu et al., 2025), Marker (Datalab, 2025), Docling (Livathinos et al., 2025), TextIn (Hehe Information and TextIn)), engineered for end-to-end extraction; and (3) *General-Purpose Multimodal Models* (Qwen3.5-397B-A17B, GPT-5.5, Gemini3.1-Pro, DeepSeekOCR), serving as strong zero-shot visual reasoning baselines.

4.3 Metrics

4.3.1 Hierarchy Extraction

As TETHER's ToC prior suppresses hallucinations and guarantees high precision, we prioritize recall-oriented metrics for structural completeness:

(1) *Content Completeness (CC)*: The ratio of correctly extracted titles to the ground truth total: $\text{CC} = |T_{\text{correct}}|/|T_{\text{total}}|$.

(2) *Order Consistency (OC)*: $\text{OC} = |\{(t_i, t_j) \in T_{\text{pairs}} \mid \text{idx}(t_i) < \text{idx}(t_j)\}|/|T_{\text{pairs}}|$, where T_{pairs} are ground-truth ordered title pairs and $\text{idx}(\cdot)$ is the extracted index. Pairs involving unextracted nodes strictly count as errors.

(3) *Hierarchical Consistency (HC)*: Structural topology fidelity: $\text{HC} = |R_{\text{correct}}|/|R_{\text{total}}|$, representing the validly recovered fraction of ground-truth parent-child edges.

4.3.2 Hierarchy Recovery

We employ a coarse-to-fine protocol:

(1) *Order Accuracy (OA)*: Alignment sequential fidelity: $\text{OA} = |\{(i, j) \in P_{\mathcal{G}} \mid \text{pos}(i) < \text{pos}(j)\}|/|P_{\mathcal{G}}|$, where $P_{\mathcal{G}}$ are ground-truth ordered pairs and $\text{pos}(\cdot)$ is the predicted vertical index.

(2) *Tree Edit Distance Similarity (TEDS)*: It quantifies topological transformation costs between predicted ($\mathcal{P}$) and ground-truth ($\mathcal{G}$) trees: $\text{TEDS} = 1 - \text{TED}(\mathcal{G}, \mathcal{P})/\max(|\mathcal{G}|, |\mathcal{P}|)$.

(3) *Insertion Accuracy (IA)*: Boundary alignment precision via a successor signature $\sigma(h)$ (the immediate non-empty downstream unit). IA measures the exact-match proportion for gold headings h^G:

Table 2: Hierarchy recovery performance (%).

Category	Model	ESG Reports			Annual Reports			Product Manuals		
		OA	TEDS	IA	OA	TEDS	IA	OA	TEDS	IA
Deep Hierarchical Recovery	MDT	11.0	13.1	36.3	1.4	17.4	16.5	3.6	21.4	20.8
	CMM	7.8	22.8	35.2	4.2	26.1	29.0	7.3	29.9	32.4
Dedicated Parsers	Marker	82.6	-	37.6	52.3	-	3.8	65.3	-	15.5
	Docling	83.7	-	36.4	83.7	-	16.4	78.2	-	14.0
	MinerU2-VLM	81.7	-	32.9	86.9	-	6.9	80.5	-	11.8
	TextIn	87.3	-	39.7	89.6	-	35.3	87.9	-	15.3
General-Purpose LMMs	Qwen3.5-397B-A17B	80.7	56.0	73.0	87.7	53.1	66.4	76.9	58.4	72.0
	Gemini3.1-Pro	75.4	48.3	58.0	75.3	45.2	53.6	67.3	43.0	51.3
	GPT-5.5	87.6	55.1	73.7	85.4	50.7	64.3	82.4	63.1	75.3
	DeepSeekOCR	79.5	39.7	54.4	56.2	52.3	61.0	66.9	42.2	55.8
Ours	**TETHER**	**96.5**	**89.1**	**94.5**	**98.4**	**88.0**	**97.7**	**98.0**	**84.7**	**92.4**

$\text{IA} = |\mathcal{H}^G|^{-1} \sum_{h^G \in \mathcal{H}^G} \mathbb{1}[\sigma(h^G) \equiv \sigma(m(h^G))]$.

4.4 Implementation Details

Eliminating multi-model overhead, TETHER executes both skeleton reasoning and agentic insertion via a single Qwen3.5-27B model to ensure secure local deployment on consumer-grade hardware.

5 Experimental Results & Analysis

We demonstrate TETHER's multi-domain robustness, utilizing ablations and stress tests to validate its architectural irreducibility and operational boundaries, respectively.

5.1 Hierarchy Extraction

As shown in Table 1, our constrained extraction demonstrates strong generalization across three domains, achieving a 10.5% average gain over unconstrained baselines. The framework is particularly effective on the most complex structural metric, HC, delivering a 13.8% average gain (up to +25.6%). Functioning as a performance equalizer, it bridges the gap between smaller models and proprietary giants, elevating the overall mean accuracy to >96% and restricting the standard deviation to 1.37%. Ultimately, despite marginally trailing proprietary models, our 27B backbone enables secure, privacy-preserving local extraction without prohibitive costs.

5.2 Hierarchy Recovery

We demonstrate TETHER's SOTA performance, analyze its stage-wise efficacy, and elucidate the mechanisms underlying its noise robustness.

Table 3: Stage-wise breakdown of ToC-body alignment. $|\mathcal{H}|$: Reference ToC nodes; $|\mathcal{C}|$: Candidate body headings. ESG: ESG reports, AR: annual reports, PM: product manuals.

Domains	$\|\mathcal{H}\|$ H1/H2/H3	$\|\mathcal{C}\|$	$\frac{\|\mathcal{C}\|}{\|\mathcal{H}\|}$	Alignment Count		
				Exact	Fuzzy	Agentic
PM	1,465 355/977/133	9,335	6.37	813 (55.5%)	467 (31.9%)	185 (12.6%)
ESG	2,213 535/1,452/226	11,028	4.98	1,438 (65.0%)	569 (25.7%)	206 (9.3%)
AR	1,383 365/887/131	3,647	2.64	1,034 (74.8%)	246 (17.8%)	103 (7.4%)

Hierarchy Recovery Performance. Table 2 confirms TETHER achieves SOTA performance (OA $> 96\%$, IA $> 92\%$) across all domains. This highlights our success in overcoming three systemic bottlenecks in prior art: (1) Deep hierarchical methods rely on rigid heuristics that fail in complex layouts; (2) Dedicated parsers excel at sequential scanning (high OA) but lack hierarchical reasoning (low IA); (3) Standard LMMs suffer from structural drift in long contexts (e.g., annual reports IA drops to 54%–66%). TETHER overcomes these limitations by replacing rigid heuristics with four structural constraints, and employing Ω-bounded inference to prevent local errors from cascading into global hallucinations.

Stage-wise Contribution & Synergy. Table 3 deconstructs the recall contribution of each processing stage. Symbolic anchoring (exact and fuzzy) establishes a robust baseline, resolving 87.4% (product manuals) to 92.6% (annual reports) of unambiguous mappings. Subsequently, the agentic inference layer targets the remaining 7.4%–12.6% of

Table 4: Ablation Study of TETHER. Performance and per-document efficiency (total tokens and latency (s)).

#	Method	ESG Reports					Annual Reports					Product Manuals				
		OA	TEDS	IA	Tok.	Lat.	OA	TEDS	IA	Tok.	Lat.	OA	TEDS	IA	Tok.	Lat.
1	w/o Exact Anchoring	95.5	78.9	81.3	433	67.8	93.1	74.1	80.9	583	44.3	95.9	64.2	87.3	612	90.1
2	w/o Fuzzy Anchoring	83.5	63.8	53.7	266	44.1	86.2	63.7	71.3	577	56.0	80.9	62.4	42.9	536	91.7
3	Fuzzy via String Sim.	92.5	67.8	83.8	205	28.7	88.3	63.0	87.6	538	33.4	91.7	61.8	85.1	420	39.7
4	Fuzzy via Dense Retr.	95.6	68.1	84.4	167	24.6	89.3	71.6	88.4	502	38.2	94.7	64.2	86.8	415	41.3
5	w/o Window (Ω_h)	86.5	70.8	88.5	192	35.8	89.3	70.6	86.9	485	46.1	89.7	63.8	89.3	454	51.8
6	w/o Agentic Insertion	78.3	54.5	77.9	-	0.0	79.3	60.6	77.3	-	0.0	64.6	63.8	62.1	-	0.0
7	w/o Upward Backtrack	83.7	71.0	83.9	172	25.4	86.7	61.6	85.0	353	35.1	84.6	64.3	80.4	307	36.5
8	w/o Top-Down Confirm	95.0	74.0	74.9	138	24.3	91.0	71.1	81.3	198	19.9	93.7	72.0	83.9	157	19.7
9	Direct Context Feed	66.1	43.3	40.4	1140	82.8	47.4	47.9	40.0	2236	97.2	57.5	35.2	33.5	2901	89.1
10	**TETHER**	**96.5**	**89.1**	**94.5**	**94.0**	**16.3**	**98.4**	**88.0**	**97.7**	**442**	**32.0**	**98.0**	**84.7**	**92.4**	**366**	**30.0**

intractable cases, such as those involving semantic abstraction or ghost nodes. This synergy ensures that costly LLM reasoning is deployed only when symbolic logic falls short.

Mechanism Insight: Robustness to Extreme Noise. A key advantage of TETHER is its resilience in highly fragmented domains (e.g., product manuals), which are characterized by a high $\frac{|\mathcal{C}|}{|\mathcal{H}|}$ ratio (Table 3) where spurious text artifacts mimic structural headings. While such dense noise induces severe structural drift in baselines (Table 2), TETHER exhibits remarkable robustness. Counterintuitively, it repurposes this structural noise by utilizing extraneous candidate headings as frequent, tightly spaced validation anchors. This strategy converts layout density into structural precision, maintaining high alignment accuracy (IA $> 92\%$) even within highly irregular product manuals.

5.3 Ablation Studies

Table 4 validates the architectural irreducibility and computational efficiency of TETHER.

5.3.1 Performance Analysis

This section analyzes individual module contributions to alignment accuracy.

Symbolic Anchoring: The Alignment Foundation (rows 1–2). Symbolic matching forms the robust alignment skeleton. ***(1)*** *Exact Anchoring (Row 1):* Its ablation universally degrades structure, confirming its role as an error-cascade stabilizer. ***(2)*** *Fuzzy Anchoring (Row 2):* Its removal causes catastrophic failure, proving it essential for bridging gaps in texts with inconsistent expressions.

Fuzzy Matcher Selection: Why BM25? (Rows 3–4). ***(1)*** *String Similarity (Row 3):* Fails due to high-frequency word confusion. ***(2)*** *Dense Retrieval (Row 4):* Yields topological errors via semantic collisions among generic headers. In contrast, BM25 delivers noise-tolerant lexical granularity to resist semantic collapse, ensuring precise topological gap-filling.

Bounded Inference: Constraints Against Drift (Row 5). Removing the candidate window significantly drops OA, validating that bounded mutual exclusion, enforced by the local candidate window, prevents cross-level mismatches of generic titles, proving essential for topological stability.

Agentic Fine Insertion: Surgical Refinement (Rows 6–8). ***(1)*** *Agentic Insertion (Row 6):* Disabling the agent leads to the lowest TEDS scores among pipeline variants, confirming that ghost nodes require semantic inference. ***(2)*** *Upward Backtracking (Row 7):* Its absence results in significant TEDS degradation, as introductory text is misclassified into preceding chapters without this corrective mechanism to rectify boundary shifts. ***(3)*** *Top-Down Confirmation (Row 8):* Critical for precision in data-sparse conditions. Removing this verification step leads to a sharp decline in IA, demonstrating its function in preventing false insertions when local anchor evidence is insufficient.

The Failure of Direct Feeding (Row 9). Directly feeding raw content between anchors to the LLM results in a catastrophic collapse across all metrics. This demonstrates that unconstrained generation within long contexts is severely exacerbated by layout noise and semantic ambiguity, making topological consistency unattainable. In contrast, by refining boundaries via agentic inference with upward backtracking and top-down confirmation, our training-free framework achieves controllable,

traceable, and generalizable structure recovery.

5.3.2 Efficiency Analysis

Our framework demonstrates superior token and runtime efficiency, as detailed in Table 4.

Hierarchical Acceleration. Removing *Symbolic Anchoring* (Rows 1–2) significantly increases runtime and token consumption, highlighting our resource-efficient design: the lightweight symbolic layer quickly resolves high-confidence anchors, thereby bypassing the unnecessary overhead of large models.

Impact of Window Constraints on Speed. Applying the constraint window Ω_h reduces runtime by approximately 50% compared to Row 5 (w/o Window). This demonstrates that exact matching constraints effectively prune the retrieval search space for fuzzy matching, avoiding unnecessary computational overhead.

Targeted Processing vs. Brute-Force Baseline. The Direct Feeding baseline (Row 9) incurs excessive overhead, consuming $> 10\times$ tokens (ESG Reports: 1140 vs. 94) and $5\times$ runtime (82.8s vs. 16.3s) versus TETHER. This indicates that unconstrained generation forces the model to process all content between anchors. In contrast, by optimizing boundaries through intelligent reasoning with upward backtracking and top-down confirmation, our framework inputs only essential content: Scenario I evaluates only candidate headings, while Scenario II adds a minimal text stream for backtracking correction.

5.4 Error Propagation Analysis

Since our pipeline relies on the document skeleton $\mathcal{H}$ generated during hierarchy extraction, we conducted a stress test to evaluate how these errors propagate to the final hierarchy recovery and prove our metrics (OA, IA, and TEDS) independently isolate distinct structural errors. We injected four types of synthetic noise into $\mathcal{H}$ prior to hierarchy recovery (Table 5):

Textual Corruption: Injecting OCR-style errors (e.g., character substitutions, random insertions, or accidental deletions) noticeably impacts TEDS (17.72 pp) but barely affects structural accuracy (ΔOA = 2.15 pp, ΔIA = 1.95 pp). This confirms that superficial mismatches penalize TEDS solely via its text-substitution mechanics, rather than actual structural degradation.

Table 5: Performance degradation (in percentage points, pp) under various types of synthetic noise injected into the intermediate document skeleton $\mathcal{H}$. Δ denotes the absolute drop in the respective metric.

Noise Type	Ratio	ΔOA	ΔTEDS	ΔIA
Textual Corruption	10%	2.15	17.72	1.95
Node Deletion	20%	2.21	11.92	2.83
Depth Randomization	20%	3.44	17.72	16.72
Topological Swap	20%	6.30	30.97	13.45

Node Deletion: Simulating extraction omissions yields minor spatial degradation (ΔOA = 2.21 pp, ΔIA = 2.83 pp). Since OA evaluates relative ordering and IA anchors on the physical successor $\sigma(h)$, sparse missing nodes fail to disrupt the overarching layout logic, demonstrating considerable tolerance to data loss.

Depth Randomization: Perturbing hierarchical levels (e.g., Section to Subsection) severely degrades insertion accuracy (ΔIA = 16.72 pp). Although the relative sequence (OA) remains intact, erroneous depths inherently skew parent-child scoping. This causes the exact-match boundary indicator $\mathbb{1}[\sigma(h^G) \equiv \sigma(m(h^G))]$ to fail, highlighting a heavy reliance on accurate depth assignments.

Topological Swap: Exchanging structural relationships (e.g., swapping a parent node with its child) inflicts the maximum global degradation (ΔOA = 6.30 pp, ΔTEDS = 30.97 pp). This distortion intrinsically inverts the spatial ordered pairs $\text{pos}(i) < \text{pos}(j)$ essential for OA and forces expensive global tree-edit operations in TEDS, revealing vulnerability to topological errors, albeit largely mitigated by native priors in practice.

6 Conclusion

This study introduces TETHER, which fundamentally shifts the hierarchy restoration paradigm in design-driven long documents away from bottom-up local element aggregation and prediction. By repurposing the ToC into a deterministic structural prior via adaptive anchoring, the framework natively mitigates unconstrained parsing risks and guarantees topological alignment across heterogeneous layouts. Beyond achieving SOTA accuracy and zero-shot universality across three representative domains, extensive evaluations validate TETHER's architectural irreducibility and operational boundaries. Ultimately, with high precision and limited computational overhead, TETHER paves the way for scalable, real-world deployment.

Limitations

Fundamentally, this work strictly focuses on design-driven long documents featuring an explicit Table of Contents (ToC). We acknowledge that a small portion of real-world long documents naturally lacks a ToC, rendering TETHER inherently inapplicable to such formats. Furthermore, while our experiments encompass three representative document families—annual reports, ESG reports, and product manuals—across English, Simplified Chinese, and Traditional Chinese, these categories do not exhaust all long-document domains. Finally, although TETHER is a training-free framework, it intrinsically depends on large multimodal and language models for ToC skeleton reasoning and boundary calibration. As a result, its practical application remains bottlenecked by model availability and inference costs and latency constraints.

Potential Risks

Although TETHER is designed to recover document hierarchies rather than generate new factual content, its outputs may still pose potential risks. The primary risk lies in error propagation. To quantify this risk, as detailed in Section 5.4, we injected four types of synthetic noise into $\mathcal{H}$ prior to hierarchy recovery. These defects can be inherited by the final recovery stage. Consequently, such errors may propagate to downstream applications built upon the structured document, including retrieval, summarization, and decision-support systems. Therefore, when processing topologically degenerate or severely corrupted ToCs, particularly in high-stakes domains such as legal, regulatory, medical, or safety-critical settings, TETHER's structural outputs must not serve as the autonomous basis for decision-making.

Artifact Use, Licenses, and Data Considerations

Artifacts and Licenses. This work uses several existing scientific artifacts, including document parsers, OCR systems, hierarchical recovery models, and general-purpose multimodal models, as baselines or backbone components. We cite the corresponding creators or official references where applicable and use these artifacts only for their intended purposes, namely document conversion, OCR, hierarchy recovery, or multimodal/text inference. Open-source tools are used under their respective licenses, while proprietary or hosted models are accessed according to their service terms. Our created artifact, TETHER-Gold, is a manually annotated benchmark for evaluating document hierarchy recovery. It is intended for research and evaluation purposes only. For source documents whose redistribution is not permitted by their original access conditions, we do not redistribute the raw PDFs and instead release only permissible metadata, annotations, or links where allowed.

Intended Use. TETHER and TETHER-Gold are designed to support research on structure-aware long-document processing. They should be used for evaluating and improving document hierarchy extraction and recovery, especially for ToC-bearing long documents such as corporate annual reports, ESG disclosures, and technical/product manuals. When recovered hierarchies are consumed by downstream LLM or RAG applications, users should perform appropriate human verification or domain-specific validation.

PII and Offensive Content. The documents used in our benchmark are public-facing corporate reports, ESG disclosures, and technical/product manuals, rather than private user-generated content. During annotation and quality checking, annotators inspected the documents for obvious privacy-sensitive or offensive content. The annotation target is document hierarchy—section titles, levels, and boundaries—rather than personal attributes or individual-level records. Some public corporate reports may mention named executives, board members, or signatories in official roles; such names are part of the publicly released documents and are not used as protected-attribute labels or prediction targets. We do not intentionally collect private personal information, and any released annotations focus only on structural information required for hierarchy recovery.

References

Adobe. 2026. Page thumbnails and bookmarks in PDFs. https://helpx.adobe.com/acrobat/using/page-thumbnails-bookmarks-pdfs.html. Accessed: 2026-05-24.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, and 1 others. 2025. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*.

Cheng Da, Chuwei Luo, Qi Zheng, and Cong Yao. 2023.

Vision grid transformer for document layout analysis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19462–19472.

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

Hehe Information and TextIn. Textin xparse: Intelligent document parsing (product documentation). Online documentation. Accessed: 2026-02-02.

Pengfei Hu, Zhenrong Zhang, Jianshu Zhang, Jun Du, and Jiajia Wu. 2022. Multimodal tree decoder for table of contents extraction in document images. In *2022 26th international conference on pattern recognition (ICPR)*, pages 1756–1762. IEEE.

I Iso. 2017. Document management—portable document format—part 2: Pdf 2.0. *International Organization for Standardization, Geneva, Switzerland*.

JaidedAI and contributors. Easyocr: Ready-to-use ocr with 80+ supported languages. GitHub repository. Accessed: 2026-02-02.

Hyeon Seong Jeong, Sangwoo Jo, Byeong Hyun Yoon, Yoonseok Heo, Haedong Jeong, and Taehoon Kim. 2025. Zero-shot document understanding using pseudo table of contents-guided retrieval-augmented generation. *arXiv preprint arXiv:2507.23217*.

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. 2025. Towards long context hallucination detection. *arXiv preprint arXiv:2504.19457*.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, and 1 others. 2025. Docling: An efficient open-source toolkit for ai-driven document conversion. *arXiv preprint arXiv:2501.17887*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, and 1 others. 2025. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, and 1 others. 2024. Lmdx: Language model-based document information extraction and localization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15140–15168.

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

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

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Xinyu Wang, Lin Gui, and Yulan He. 2023. A scalable framework for table of contents extraction from complex ESG annual reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13215–13229, Singapore. Association for Computational Linguistics.

Sabine Wehnert, Harikrishnan Changaramkulath, and Ernesto William De Luca. 2025. Hips: Hierarchical pdf segmentation of textbooks. *arXiv preprint arXiv:2509.00909*.

A Challenges in Hierarchical Parsing

Extracting hierarchies from design-driven long documents is hindered by three cascading bottlenecks:

Figure A.1: Intrinsic ToC Ambiguity. (1) Implicit Hierarchy Markers (■). (2) Low Inter-Level Variance (■). (3) Stylistic Heterogeneity (■). (4) Numerical and Semantic Noise (■).

Intrinsic ToC Ambiguity. Extreme formatting heterogeneity complicates the reliable extraction of ToC structures (Figure A.1): ***(1)*** *Implicit Hierarchy Markers*: Section numbering is frequently omitted. ***(2)*** *Low Inter-Level Variance*: Visual distinctions between levels are minimal. ***(3)*** *Stylistic Heterogeneity*: Same-level entries exhibit disparate styles. ***(4)*** *Numerical and Semantic Noise*: Numerical polysemy and decorative text induce interference.

Figure A.2: Intra-Body Layout Entropy. (1) Violation of Visual Monotonicity (■), where solid borders denote Level-1 headings and dashed borders mark non-heading content. (2) Visual Decoys (■). (3) Text Rasterization (■). (4) Structural Discontinuity (■).

Intra-Body Layout Entropy. Layout entropy decouples visual cues from logical structure (Figure A.2): ***(1)*** *Violation of Visual Monotonicity*: Typographical prominence contradicting logical depth eliminates feature invariance. ***(2)*** *Visual Decoys*: Salient decorative elements mimic logical headers. ***(3)*** *Text Rasterization*: Graphical headers evade OCR, causing node omissions. ***(4)*** *Structural Discontinuity*: Extreme parent-child spans sever lineage dependencies.

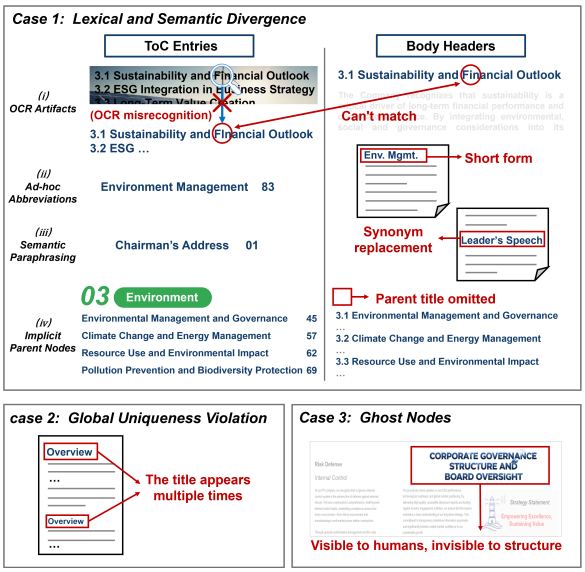

Figure A.3: ToC-to-Body Alignment Discrepancies.

ToC-to-Body Alignment Discrepancies. Both semantic and structural discrepancies confound alignment (Figure A.3): ***(1)*** *Lexical and Semantic Divergence*: *(i) OCR Artifacts*: Character misclassifications defy exact matching. *(ii) Ad-hoc Abbreviations*: Non-standard contractions expand lexical distance. *(iii) Semantic Paraphrasing*: Synonymous rewriting yields lexically disjoint headers. *(iv) Implicit Parent Nodes*: Omitted ToC parents force direct transitions to descendants. ***(2)*** *Global Uniqueness Violation*: Generic headers recur across sections, inducing incorrect anchoring. ***(3)*** *Ghost Nodes*: Rasterized headers lack text, becoming unanchored.