AMBIHTMLBENCH: A BENCHMARK FOR FAITHFUL DOCUMENT-TO-HTML RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Document parsing systems often produce Markdown, which captures linear text but offers limited support in many scenarios (e.g., for fine-grained spatial structure, table geometry, visual hierarchy, and rendering semantics). HTML provides a richer representation for heterogeneous, layout-sensitive documents while supporting editing, inspection, and reuse. However, document-to-HTML reconstruction lacks a dedicated benchmark for systematic evaluation. Thus, we introduce **AmbiHTMLBench** to evaluate models' ability to reconstruct real-world documents as renderable HTML that preserves their content, structure, and visual appearance. AmbiHTMLBench comprises 1,115 diverse pages with structured annotations and two complementary evaluation protocols: (1) The objective protocol aligns rendered blocks between reference and predicted outputs to measure table fidelity, page-level text error, and two-dimensional layout consistency; and (2) the reference-anchored subjective protocol evaluates perceived fidelity using LLM judges calibrated against human preferences. Using AmbiHTMLBench to evaluate ten multimodal systems, we find different systems lead under the two protocols (with the highest objective score as 0.8092). We further demonstrate that task-specific fine-tuning enables a compact expert model to perform effective document-to-HTML reconstruction with 4.41 weighted points on Qwen-9B base.

1 INTRODUCTION

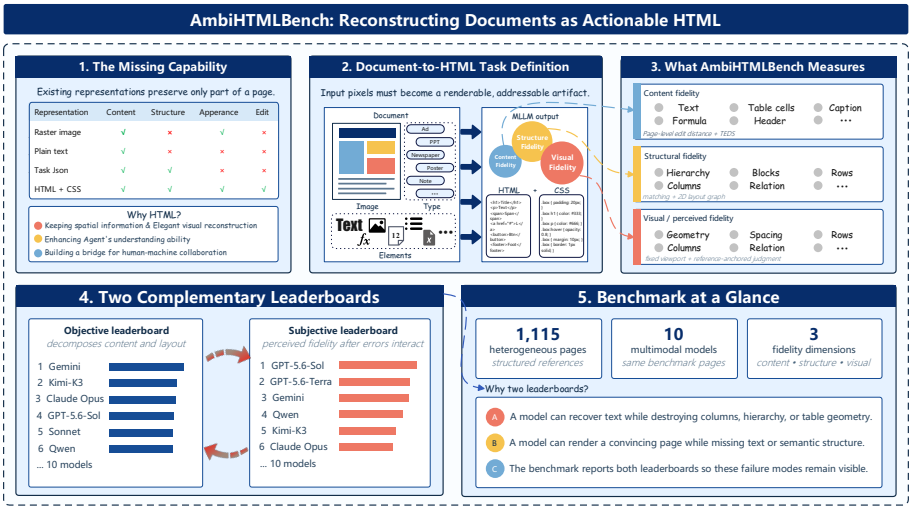

Figure 1: AmbiHTMLBench for evaluating document-to-HTML reconstruction.

Documents combine diverse elements (e.g., headings, tables, figures, typography) and spatial layouts that jointly shape how information is interpreted. Recent multimodal systems can extract and reason over document images (Xu et al., 2021; Wang et al., 2024; Hu et al., 2024), but document

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

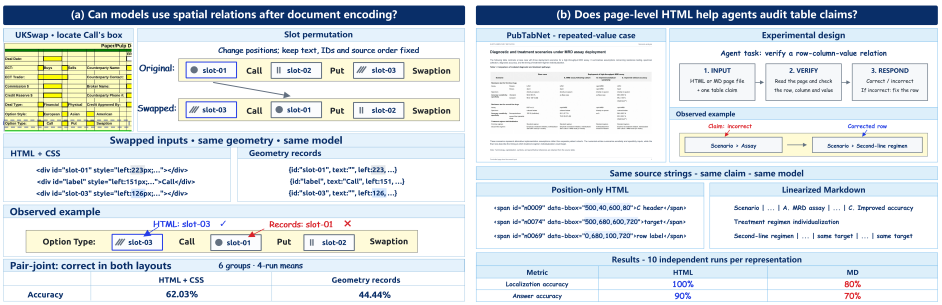

Figure 2: Pilot evidence that HTML representations support spatial binding and table verification.

agents require a stronger capability. They must be able to store evidence, locate the source of an answer, modify a document, and reuse the same representation across retrieval, reasoning, and interaction tasks (Han et al., 2025a). Understanding a page once is therefore insufficient; agents need a persistent and actionable representation of the page itself.

Existing representations expose only part of this information. (1) Plain text and Markdown provide compact and efficient representations for reading and generation, but their largely linear form does not explicitly preserve the two-dimensional spatial relations and visual organization that are central to document understanding (Xu et al., 2021; Wang et al., 2024). (2) Structured representations such as JSON can explicitly encode selected document semantics, but existing document parsing systems often organize these semantics into task-specific fields, labels, or schemas, rather than preserving the document as a unified visual and structural representation (Zhong et al., 2019; Pfitzmann et al., 2022). (3) Raster images retain the visual appearance of a document, but expose neither editable semantic units nor stable machine-addressable references to individual elements. These limitations force downstream agents to repeatedly infer structure from pixels, creating a gap between document perception and document interaction.

HTML provides a natural representation for closing this gap. Its Document Object Model (DOM) expresses hierarchical and addressable relationships among headings, paragraphs, tables, figures, and captions, allowing humans and agents to locate and manipulate individual elements (Li et al., 2022; Deng et al., 2022). At the same time, HTML and CSS provide an executable rendering mechanism for alignment, spacing, typography, positioning, and table geometry. This interface property is also reflected in contemporary web-agent systems. Frameworks such as BrowserGym expose HTML/DOM structures, accessibility trees, screenshots, and element-level properties as agent observations, while supporting direct interaction with individual page elements (de Chezelles et al., 2025), while Anthropic's browser use emphasizes the significant advantage of element-based targeting via DOM references compared to pixel coordinates.

We also obtain direct evidence for this interface property in two small pilots. On six field groups from a UKSwap form, HTML/CSS yields 62.03% pair-joint accuracy for field-to-slot binding, compared with 44.44% for geometry-matched non-HTML records. On PubTabNet tables, HTML-based evidence summaries make row, column, and value nodes explicitly addressable, enabling more reliable correction of an erroneous row assignment than linearized Markdown (see Figure 2). These pilots are strong evidence that HTML has outstanding ability to maintain information on complex pages

The missing capability. Despite progress in OCR and document parsing, existing benchmarks largely measure isolated components such as text, tables, formulas, and reading order (Ouyang et al., 2025; Zhang et al., 2026; Li et al., 2026). Screenshot-to-code benchmarks study visual reconstruction, but primarily target webpages with relatively regular DOM-oriented layouts. Neither setting directly tests whether a multimodal model can recover a heterogeneous real-world document as complete, executable HTML. This task is difficult because fidelity has several interacting dimensions: a prediction may contain the correct words but destroy columns or table geometry; it may preserve layout while hallucinating or omitting content; or it may encode a valid DOM whose CSS fails to reproduce the visual hierarchy that guides human reading.

These failure modes also motivate complementary evaluation. Content fidelity should measure whether text, numbers, formulas, and table cells are preserved. Structural fidelity should measure whether blocks and their relationships are organized as in the source. Visual or perceived fidelity should account for the fact that humans may tolerate harmless line-break differences but strongly penalize lost emphasis, displaced media, or broken table structure. A single metric cannot reli-

ably expose all of these cases, so a useful benchmark should report decomposed automatic signals together with a reference-anchored perceptual measure.

Our benchmark. We present **AmbiHTMLBench**, a benchmark of 1,115 diverse pages with expert annotations. We render each generated HTML page and align its semantic blocks with the ground truth to measure page-level text error, table fidelity, and two-dimensional layout consistency. We additionally use human-calibrated VLM judgments to estimate perceived fidelity under a separate subjective protocol. This design lets us ask not only whether a model obtains a high score, but also which reconstruction capability is responsible for success or failure.

Our contributions are threefold: (1) A diverse document-to-HTML benchmark covering heterogeneous page types, content, and layouts; (2) A reproducible multi-dimensional evaluation protocol that separates content, structure, rendering, and perceived fidelity; and (3) An empirical study of ten recent outstanding MLLMs (gpt-5.6 series and Claude Opus, etc.)and task-specific fine-tuning. The objective and subjective leaderboards select different winners, and adaptation improves smaller models, showing that AmbiHTMLBench is useful both as a diagnostic evaluation and as a training target for agent-ready document representations.

2 Related Work

Document Parsing and Understanding. Early document understanding focused on OCR and layout analysis, which extract text and identify regions such as tables and figures but treat textual and spatial information separately. With large-scale multimodal benchmarks and pretrained models, the field moved toward layout-aware representation learning. LayoutLM integrated text with 2D coordinates (Xu et al., 2020), and follow-up models such as LayoutLMv2, LayoutLMv3, DocFormer, and LiLT added visual features, spatial relationships, and multilingual capabilities (Xu et al., 2022; Huang et al., 2022; Appalaraju et al., 2021; Wang et al., 2022). However, these models output tokens, regions, or task-specific labels rather than complete, reusable document structures.

Multimodal large language models (MLLMs) such as Donut, DocLLM, and mPLUG-DocOwl further enabled general-purpose document reasoning, including question answering and information extraction (Kim et al., 2022; Wang et al., 2023; Ye et al., 2023). More recently, agentic systems have extended document understanding to multi-step reasoning and tool interaction. MDocAgent coordinates specialized agents for cross-modal retrieval (Han et al., 2025b), DocAgent constructs document outlines for interactive reading (Yang et al., 2025), and SlideAgent builds hierarchical representations for multi-page visual documents (Jin et al., 2026).

These approaches reveal a key requirement: document representations should remain organized, addressable, and grounded in the original structure. Flat text loses spatial relationships, hierarchy, and cross-modal associations. Yet existing agentic systems focus on reasoning over extracted information, leaving open the problem of reconstructing the document itself into a faithful, editable, reusable form.

Structured Document Representation and HTML-based Generation. Structured representations explicitly encode hierarchy and relationships among elements. Markup-based formats such as XML and HTML have been used as intermediate representations: DOM-LM jointly models text and DOM structure (Deng et al., 2022), and MarkupLM incorporates markup into pretraining (Li et al., 2022).

HTML is particularly suitable for document reconstruction because it combines semantic organization with executable visual rendering. XML offers flexible schemas but does not specify rendering, whereas HTML provides standardized elements for headings, lists, tables, and paragraphs, with CSS controlling typography, positioning, and spacing. This duality lets a single representation expose semantics to machines while remaining directly renderable. Pix2Struct predicts simplified HTML from webpage screenshots (Lee et al., 2023), and AnyDoc adopts HTML/CSS for intention-conditioned document generation (Lin et al., 2026). However, these works target webpage understanding or controllable generation rather than faithful document-to-HTML reconstruction.

HTML also serves as a natural agent interface. WebAgent compresses HTML into task-relevant representations and grounds actions in webpage structure (Gur et al., 2024). But such systems assume a DOM already exists—an assumption that fails for PDFs and scanned pages, whose content is available only as pixels and text positions.

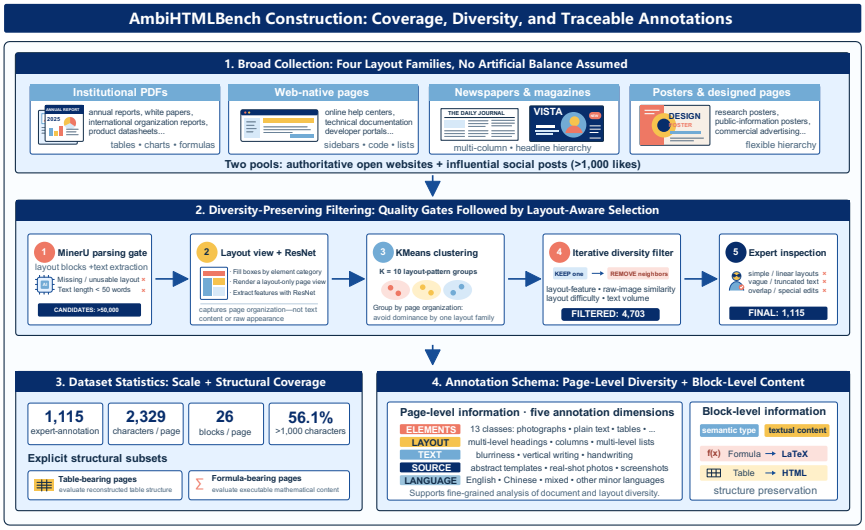

Figure 3: The overall AmbiHTMLBench Construction process.

AmbiHTMLBench targets this missing step: evaluating whether multimodal models can reconstruct agent-compatible HTML from visually rendered documents while preserving content, semantic organization, and visual appearance.

Evaluation of Document Reconstruction and Understanding. Traditional document-parsing evaluation uses local metrics such as edit distance for text, tree-based metrics for tables and formulas, page IOU for layout, and sequence matching for reading order. OmniDocBench consolidates these signals with fine-grained layout categories (Ouyang et al., 2025), and PureDocBench derives traceable annotations from HTML/CSS (Li et al., 2026). Both evaluate component-level parsing rather than the document as an organic whole.

Visual reconstruction benchmarks assess whether models recover executable representations from visual inputs. WebSight provides HTML/screenshot pairs for webpage reconstruction (Laurençon et al., 2024), and Design2Code combines automatic and human evaluation for screenshot-to-code generation (Si et al., 2025). These target webpages, which align naturally with HTML layout, whereas real-world documents pose greater layout and content challenges. ParseBench evaluates document parsing for downstream agent usability (Zhang et al., 2026), but focuses on parsed outputs rather than complete HTML reconstruction.

Existing evaluations thus cover complementary but incomplete aspects: parsing benchmarks measure component-level extraction, while screenshot-to-code benchmarks measure visual reproduction. AmbiHTMLBench bridges them by evaluating document-to-HTML reconstruction as a unified task requiring faithfulness in content, structure, and visual realization.

3 Benchmark Construction

Data Collection. From different document pools on the internet and institutions where the total volume can reach one billion, we sampled over 50,000 pages from a four source categories: institutional PDFs (e.g., reports and datasheets), web-native documentation, newspapers and magazines, and posters or other highly designed pages. These sources cover conventional and nonstandard layouts containing dense text, tables, figures, and mixed media.

Filtering and Quality Control. After data collection, there are above 50,000 pages remaining. Before producing the final benchmark subset, we applied several quality control steps, as is shown in figure 3. First, we use MinerU(Niu et al., 2025) to conduct basic document parsing, and (1) Remove samples whose layout parsing result is missing or unusable. (2) Remove samples with less than 50 words. This step ensures that every retained sample has both a clear layout structure and abundant text content.

Next, based on MinerU layout parsing result, we provide a layout view of the documents by filling the bounding boxes with different colors according to their categories, and then use a layout-aware clustering to reduce redundancy. All candidate samples are clustered into ten groups using KMeans over the features obtained after feature extraction of this layout view using ResNet-50. This process groups documents primarily by page organization rather than by raw appearance or textual content.

Within each cluster, we applied an iterative diversity-preserving filtering procedure. At each iteration, we retained one candidate and removed nearby pages according to layout-feature similarity, visual similarity, estimated layout difficulty, and text volume. This procedure was repeated until the retained set reached the desired diversity level.

This procedure balances diversity and quality by limiting dominant layout families, removing near-duplicates, and retaining informative pages across difficulty levels.

After filtering 50,000+ candidates to 4,703 pages, expert review removed overly simple, truncated, overlapping, or specially edited samples, yielding 1,115 benchmark pages.

Dataset Statistics. An annotation team of 15 experts that is familiar with OCR-related business annotated all 1,115 pages at page and block levels. Page-level labels cover 13 element types, layout (headings, columns, lists), text appearance (blur, vertical writing, handwriting), source type, and language. Block annotations record type and text; formulas use LaTeX and tables use HTML. Pages average 2,329 characters and 26 blocks, and 56.1% exceed 1,000 characters. Together with 5 different senior researchers' quality inspection, the process took 44 days.

4 Evaluation Protocol

Objective Evaluation. We render each prediction in headless Chromium at a fixed viewport and extract visible text, lists, tables, figures, formulas, and normalized bounding boxes. Ground-truth annotations are converted to the same block representation, enabling source-independent comparison of content and geometry.

Each object is normalized as $o = (k, t, b)$, where k is its category, t its textual or structural content, and $b = (x, y, w, h)$ its page-normalized bounding box.

• **Block Matching Algorithm.** Tables are matched first as table objects. Regular text, formulas, and cells from unmatched tables then share a text pool, giving text credit to linearized tables while reserving structural credit for preserved table objects.

Table candidates combine content, structure, text, and geometry scores and are assigned one-to-one by thresholded Hungarian matching.

Because HTML may split or merge annotated text blocks, partial text similarity defines a bipartite candidate graph; connected components support one-to-many alignments, with residuals retained as insertions or deletions. The resulting shared alignment feeds both page-level text scoring and layout analysis. Details are in Appendix A.

• **Content consistency.** For matched tables, we report normalized row/cell edit scores: TEDS includes cell text, while TEDS-S measures structure only (an approximation, not full DOM tree edit), with $S_{\text{table}} = 1 - \frac{d_{\text{row/cell}}}{\max(|T_g|,|T_p|,1)}$. Scores are reported on the 263 pages containing tables.

Page-level text fidelity concatenates aligned groups, including one-sided residuals, and computes normalized character edit distance $D_{\text{text}} = \frac{\text{Lev}(T_g,T_p)}{\max(|T_g|,|T_p|,1)}$.

• **Layout consistency.** Layout uses matched tables and text groups as nodes, compares normalized center and size, and evaluates corresponding edges in a ground-truth 3-nearest-neighbor graph. Empty-sided or unsupported text matches do not receive layout credit.

For an edge, similarity combines direction, relative displacement vector, and length as $s_{\text{edge}} = 0.4s_{\text{dir}} + 0.3s_{\text{len}} + 0.3s_{\text{vec}}$. The page-level layout score is $S_{\text{layout}} = 0.55\,\overline{s}_{\text{edge}} + 0.30\,\overline{s}_{\text{node}} + 0.15F_1^{\text{coverage}}$. The score retains two-dimensional direction and displacement; coverage penalizes omitted reference content.

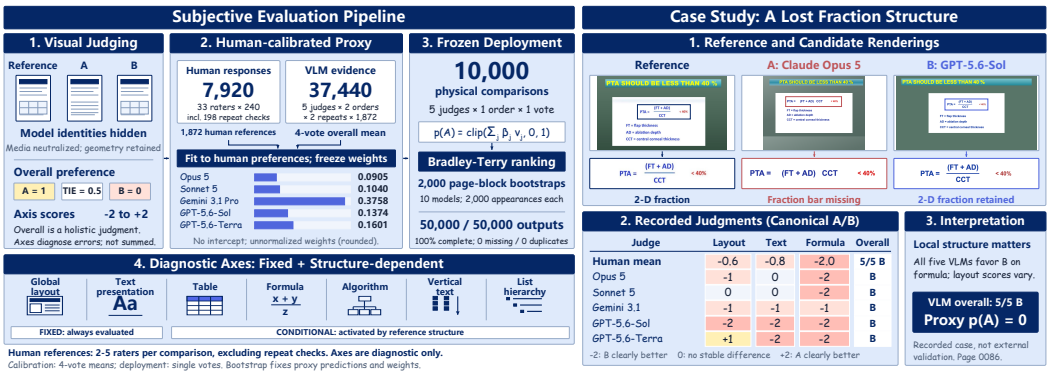

Figure 4: Human calibration and deployment of the frozen preference proxy.

Subjective Evaluation. Objective metrics cannot fully capture how content and layout errors interact in perception, so we add reference-anchored pairwise judgments.

Each trial presents a reference page and two reconstructions and asks which is more faithful. Figure 4 summarizes the protocol.

• **Judgment protocol and human reference.** Human annotators provide pairwise preference labels. The VLM judges return an overall A/B/TIE decision together with confidence, a brief rationale, and diagnostic axis scores. For images and charts, judges are instructed to ignore internal semantic content and assess only presence, geometry, size, placement, alignment, and relationships to captions and neighboring regions.

Diagnostic axes cover global layout, text presentation, and element-specific properties (tables, formulas, algorithms, vertical text, and list hierarchy); scores range from -2 to 2 and are diagnostic rather than inputs to the frozen preference proxy or final leaderboard.

Thirty-three participants completed 7,920 pairwise judgments (7,722 main annotations and 198 hidden repeats; repeats are excluded from calibration; see Appendix B.1). After displayed candidates are mapped to their canonical identities, `CANNOT_JUDGE` responses are excluded and valid A, TIE, and B votes are encoded as 1, 0.5, and 0, respectively. Their mean defines the human target $h_i = \frac{n_{i,A}+0.5n_{i,T}}{n_{i,A}+n_{i,T}+n_{i,B}}$.

• **Frozen proxy and deployment.** For calibration, each of five VLM judges evaluates every human comparison in two display orders and two repetitions. The four canonicalized overall votes are averaged for each judge, and a no-intercept ordinary least-squares regression is fit to the human target; the full-data coefficients are then frozen. At deployment, each four-vote training mean is replaced by one canonicalized overall vote $x_{i,j}$ per judge, giving $\hat{p}_i(A) = \mathrm{clip}_{[0,1]}\left(\sum_{j=1}^{5} \beta_j x_{i,j}\right)$. All 10,000 scheduled comparisons are scored by this frozen proxy. The schedule retains the 1,872 human-study comparison identities and adds 8,128 further comparisons; human annotations are used for calibration and internal validation rather than substituted directly into leaderboard outcomes. A centered Bradley–Terry model treats $\hat{p}_i(A)$ as a fractional preference outcome, and page-block bootstrap estimates ranking uncertainty. Prompts, frozen coefficients, audits, candidate-version details, and implementation choices appear in Appendix B.

5 Experiments

We organize the experiments around two complementary leaderboards. The objective leaderboard decomposes reconstruction quality into content and layout fidelity, whereas the subjective leaderboard captures perceived fidelity after these errors interact.

5.1 Experimental Setup

We evaluate ten multimodal models on the same 1,115-page benchmark set. Each model receives the page image under its standard image-to-HTML prompting interface. Rendering, matching thresh-

Table 1: Objective leaderboard on 1,115 pages for the ten primary systems. Text ED is lower-is-better; all other columns are higher-is-better. The weighted score uses table weight 0.3 and text and layout weights 1.0. TEDS and TEDS-S are averaged over the 263 table-bearing pages.

Rank	Model	Text ED ↓	TEDS ↑	TEDS-S ↑	Layout ↑	Obj. score ↑
1	Gemini 3.1 Pro Preview	0.1754	0.6518	0.6650	0.8410	80.92
2	Kimi-K3	0.2399	0.6253	0.6389	0.8307	77.32
3	Claude Opus 5	0.2537	0.6290	0.6465	0.8387	77.12
4	Qwen3.7-Plus	0.2659	0.6113	0.6250	0.8088	75.05
5	GPT-5.6-Sol	0.3055	0.5967	0.6103	0.8425	74.61
6	GPT-5.6-Terra	0.2998	0.5953	0.6131	0.8335	74.45
7	Claude Sonnet 5	0.2935	0.6042	0.6218	0.8218	74.33
8	ERNIE-5.0	0.2871	0.5982	0.6137	0.8120	74.10
9	GLM-5V-Turbo	0.3081	0.6121	0.6316	0.8103	73.30
10	Grok-4.5	0.6754	0.2874	0.3024	0.3950	35.04

olds, and aggregation are fixed before comparing models; implementation details are provided in Appendix A.

5.2 Objective Leaderboard

Gemini 3.1 Pro Preview leads the objective ranking, primarily because it has the lowest page-level text error and the strongest table score while maintaining a competitive layout score. Kimi-K3 and Claude Opus 5 form the next tier: Kimi-K3 is slightly stronger on text, whereas Claude Opus 5 is slightly stronger on layout and table structure. Qwen3.7-Plus ranks fourth by balancing moderate text error with stable table and layout scores. GPT-5.6-Sol obtains the highest layout score, but its larger text error and weaker table score place it below the content-leading systems. Grok-4.5 is separated from the remaining models on every axis, especially text, table fidelity, and layout. These results motivate reporting the decomposition rather than interpreting a single score as a complete description of reconstruction quality.

Figure 5 demonstrated the ability for objective evaluation. In comparison, Claude has a lower rate of text hallucinations, and the relative structure of each block is more similar to the original layout. They give it an advantage in the objective evaluation protocol.

5.3 Task-Specific Fine-tuning

To test whether AmbiHTMLBench can also serve as a training signal, we fine-tune Qwen3.5-9B and Qwen3.5-4B with transcription-style distillation. We use 1,230 accepted examples from 1,475 images in the PureDocBench `real_degraded` track, with no overlap with the benchmark pages. Gemini 3.1 Pro Preview produces HTML using the same transcription instruction as inference; outputs are retained using a block-recall filter (83.4% acceptance), with best-of-four resampling for initially low-scoring pages. This construction favors faithful transcription of the observed page rather than authored or normalized HTML.

Both model sizes use the same one-epoch LoRA recipe: rank and alpha 16, all linear language-model layers trainable, frozen vision and alignment modules, learning rate 10^{-5} with cosine decay, and a maximum sequence length of 21,504. Checkpoints are first screened with inexpensive coverage and output-discipline diagnostics and then selected by the full objective evaluator. Table 2 reports the best checkpoints on all 1,115 pages.

Table 2: Task-specific fine-tuning results on AmbiHTMLBench. All reported metrics are higher-is-better; "Text" is the normalized text score and the weighted score is the primary objective metric.

Model	Method	Weighted ↑	Text ↑	TEDS ↑	Layout ↑
Qwen3.5-9B	Base	66.23	0.6321	0.5440	0.7281
Qwen3.5-9B	Distill SFT	**70.64**	**0.6761**	**0.5827**	**0.7737**
Qwen3.5-4B	Base	67.86	0.6323	0.5382	0.7670
Qwen3.5-4B	Distill SFT	**69.34**	**0.6556**	**0.5722**	**0.7675**

Distillation improves the 9B and 4B bases by 4.41 and 1.48 weighted points, respectively, with gains on every 9B axis and on text and table fidelity for 4B. The smaller adapted model also exceeds the unadapted 9B base (0.6934 versus 0.6623). A matched-budget 27B ablation isolates the target style: transcription-style targets reach 0.7373 from a 0.7320 base, whereas cleaned, normalized,

and original authored-HTML targets reach 0.7102, 0.7141, and 0.6850. These results indicate that supervision aligned with the reconstruction objective is substantially more effective than cleaner or more normalized alternatives. A subsequent DAPO-style RL stage adds only 0.16 and 0.27 points over the best 9B and 4B SFT checkpoints, respectively, which is within the observed evaluation variability; we therefore use SFT as the primary adapted-model result.

5.4 Reliability of Subjective Evaluation

Human reliability. Valid A/B/TIE labels account for 98.85% of the 7,722 main annotations and yield a human preference target for all 1,872 comparisons. Pairwise exact inter-rater agreement is 62.30% (95% CI: 60.79–63.82), while exact agreement on the 198 within-rater hidden repeats is 73.23%. These statistics quantify different sources of variation and are not directly interchangeable. At the aggregate level, removing any one of the 33 participants leaves the ten-model human ranking unchanged (Spearman $= 1.0$; maximum rank movement $= 0$). Thus, although individual comparisons are not unanimous, the aggregate human ordering is insensitive to removal of any single participant.

Human–machine alignment. We evaluate the preference proxy with leave-one-annotation-block-out validation: coefficients are fit on seven blocks and predictions are generated for the held-out eighth, rotating through all eight blocks. The resulting 1,872 out-of-fold predictions achieve Brier/MSE 0.0778 (95% CI: 0.0722–0.0836) and MAE 0.2115 (0.2031–0.2200). The induced ten-model ranking has Spearman correlation 0.915 (0.867–0.952) with the human ranking. These results distinguish comparison-level fit from system-level ordering: individual preference estimates remain imperfect, while the aggregate model ranking is strongly aligned with the human ordering. The validation uses four-vote means per judge, whereas deployment uses one vote per judge, so these internal estimates do not independently validate the expanded 10,000-comparison leaderboard. Bootstrap intervals keep the out-of-fold predictions fixed. Appendix B provides the full agreement measures, fitting protocol, committee audit, and deployment limitations.

5.5 Subjective Leaderboard

The subjective leaderboard measures perceived fidelity, which depends on the interaction among content, structure, and rendering errors. We therefore report it separately from the objective leaderboard rather than treating human preference as another automatic metric.

Table 3: Final subjective leaderboard scored by the frozen preference proxy. BT scores are centered; soft preference is the mean proxy preference point over each model's 2,000 scheduled appearances. Intervals use page-block bootstrap. All 10,000 comparisons are proxy-scored; human annotations are used for calibration and internal validation rather than directly replacing proxy outcomes.

Rank	Model	BT score (95% CI)	Soft pref.	Rank CI
1	GPT-5.6-Sol	154.70 [140.59, 168.82]	71.64%	[1,1]
2	GPT-5.6-Terra	102.87 [90.18, 116.25]	64.76%	[2,3]
3	Gemini 3.1 Pro Preview	100.86 [89.88, 112.74]	64.49%	[2,3]
4	Qwen3.7-Plus	56.04 [45.63, 66.45]	58.12%	[4,4]
5	Kimi-K3	9.62 [−2.16, 21.71]	51.32%	[5,6]
6	Claude Opus 5	−2.47 [−14.01, 8.96]	49.51%	[5,6]
7	Grok-4.5	−55.49 [−68.68, −43.23]	41.77%	[7,8]
8	Claude Sonnet 5	−58.17 [−68.74, −47.74]	41.38%	[7,8]
9	GLM-5V-Turbo	−117.28 [−129.38, −106.41]	33.12%	[9,9]
10	ERNIE-5.0	−190.68 [−203.33, −178.59]	23.88%	[10,10]

Across the 2,000 page-block bootstrap resamples, GPT-5.6-Sol ranks first in every draw, indicating stability under this resampling scheme. Its BT score exceeds the next system by 51.83 points, and its soft preference is 6.88 percentage points above GPT-5.6-Terra. Several adjacent systems, however, are not cleanly separated by the reported bootstrap rank intervals: Terra and Gemini both occupy $[2, 3]$, Kimi-K3 and Claude Opus 5 occupy $[5, 6]$, and Grok-4.5 and Claude Sonnet 5 occupy $[7, 8]$. We therefore avoid interpreting the point estimate as a definitive ordering within these overlapping pairs. Qwen3.7-Plus, GLM-5V-Turbo, and ERNIE-5.0 retain singleton rank intervals under the same bootstrap. Overall, the leaderboard provides a proxy-based estimate of perceived fidelity rather than 10,000 independently human-validated judgments.

5.6 Comparing the Two Leaderboards

The two leaderboards are positively but only moderately correlated (Spearman $\rho = 0.588$; Kendall $\tau = 0.378$). GPT-5.6-Sol ranks first subjectively but fifth objectively, whereas Gemini ranks first objectively and third subjectively. Across all ten models, the mean absolute rank shift is 2.2 positions, and seven models move by at least two positions. GPT-5.6-Terra rises from sixth objectively to second subjectively, while Kimi-K3 and Claude Opus 5 each move down three positions; Qwen3.7-Plus and GLM-5V-Turbo retain the same ranks. Thus, the two protocols produce meaningfully different system orderings despite their positive association, consistent with their different targets: the objective score aggregates decomposed content and geometric fidelity, whereas the subjective proxy measures holistic visual preference relative to the reference.

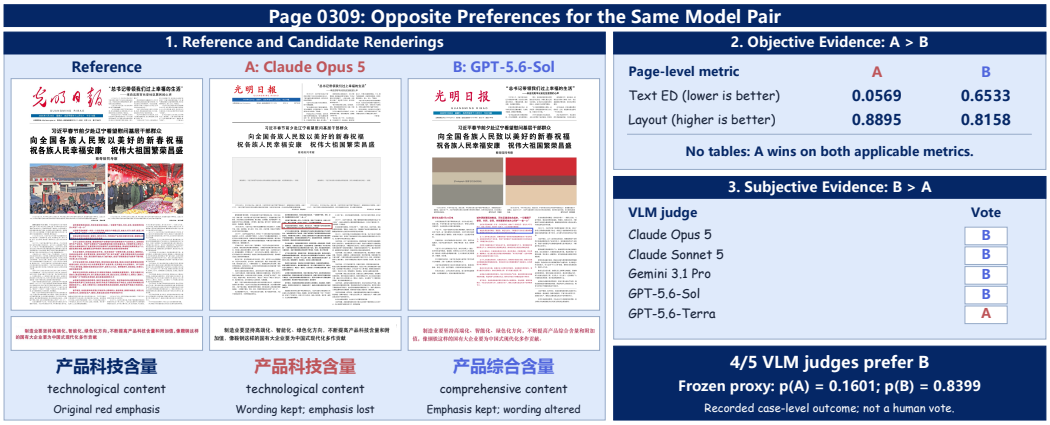

Figure 5: Page-level reversal on page 0309: objective metrics favor Opus (A), whereas the frozen subjective proxy favors Sol (B). Insets contrast faithful wording with preserved visual emphasis.

On page 0309 (Figure 5), Opus performs better on both applicable objective metrics, whereas the frozen proxy assigns Sol 0.8399 preference points, with four of the five VLM judges selecting Sol. Sol preserves the red emphasis while changing *technological content* to *comprehensive content*; Opus preserves the wording but loses the emphasis. This example should not be interpreted as evidence that visual emphasis generally outweighs textual fidelity. Rather, it illustrates how holistic preference can diverge from component metrics when several error types coexist on the same page.

6 Conclusion

We introduced **AmbiHTMLBench**, a 1,115-page benchmark for reconstructing heterogeneous documents as executable HTML. Its annotations and evaluation protocol treat content, structure, and rendered appearance as related but distinct requirements. The objective leaderboard reveals the model's ability to restore the layout structure, while the subjective study demonstrates that perceived fidelity cannot be recovered from any single automatic component. They have clearly and comprehensively pointed out the dual aspects of the development direction of document parsing in the agent era.

AI USE STATEMENT

We used AI tool in three ways. (1) AI are used to help generate main code, and ten MLLM are tested via the code. (2) AI are used for paper language polishing only, not to generate our core innovation. (3) AI is used for literature collection and organization. All work carried out with the help of AI are carefully checked, and we accept full responsibility for the paper's content.

ETHICS STATEMENT

Data from AmbiHTMLBench is collected from the Internet. We follow the processing practices recognized by the academic community and use data solely for non-commercial academic research. The dataset contains only publicly available documents, checks are made to prevent the leakage of personal information.

REPRODUCIBILITY STATEMENT

We provide the code of the benchmark as well as all model inferences in the Supplementary Material to make AmbiHTMLBench reproducible.

REFERENCES

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. Docformer: End-to-end transformer for document understanding, 2021. URL https://arxiv.org/abs/2106.11539.

Thibault Le Sellier de Chezelles, Maxime Gasse, Alexandre Lacoste, Massimo Caccia, Alexandre Drouin, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Graham Neubig, Quentin Cappart, Russ Salakhutdinov, and Nicolas Chapados. The browsergym ecosystem for web agent research. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=5298fKGmv3. Expert Certification.

Xiang Deng, Prashant Shiralkar, Colin Lockard, Binxuan Huang, and Huan Sun. DOM-LM: Learning generalizable representations for HTML documents. *arXiv preprint arXiv:2201.10608*, 2022. URL https://arxiv.org/abs/2201.10608.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. A real-world WebAgent with planning, long context understanding, and program synthesis. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e91bf7dfba0477554994c6d64833e9d8-Abstract-Conference.html.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025a.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding, 2025b. URL https://arxiv.org/abs/2503.13964.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3096–3120. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.175. URL https://aclanthology.org/2024.findings-emnlp.175/.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022. URL https://arxiv.org/abs/2204.08387.

Yiqiao Jin, Rachneet Kaur, Zhen Zeng, Sumitra Ganesh, and Srijan Kumar. Slideagent: Hierarchical agentic framework for multi-page visual document understanding, 2026. URL https://arxiv.org/abs/2510.26615.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer, 2022. URL https://arxiv.org/abs/2111.15664.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset, 2024. URL https://arxiv.org/abs/2403.09029.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18893–18912. PMLR, 2023. URL https://proceedings.mlr.press/v202/lee23g.html.

Junlong Li, Yiheng Xu, Lei Cui, and Furu Wei. MarkupLM: Pre-training of text and markup language for visually rich document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6078–6087. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.420. URL https://aclanthology.org/2022.acl-long.420/.

Zhiheng Li, Zongyang Ma, Jiaxian Chen, Jianing Zhang, Zhaolong Su, Yutong Zhang, Zhiyin Yu, Ruiqi Liu, Xiaolei Lv, Bo Li, Jun Gao, Ziqi Zhang, Chunfeng Yuan, Bing Li, and Weiming Hu. How far is document parsing from solved? PureDocBench: A source-traceable benchmark across clean, degraded, and real-world settings. *arXiv preprint arXiv:2605.07492*, 2026. URL https://arxiv.org/abs/2605.07492.

Jiawei Lin, Wanrong Zhu, Vlad I. Morariu, and Christopher Tensmeyer. AnyDoc: Enhancing document generation via large-scale HTML/CSS data synthesis and height-aware reinforcement optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Lin_AnyDoc_Enhancing_Document_Generation_via_Large-Scale_HTMLCSS_Data_Synthesis_and_CVPR_2026_paper.pdf.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, Zirui Tang, Boyu Niu, Ziyang Miao, Hejun Dong, Siyi Qian, Junyuan Zhang, Jingzhou Chen, Fangdong Wang, Xiaomeng Zhao, Liqun Wei, Wei Li, Shasha Wang, Ruiliang Xu, Yuanyuan Cao, Lu Chen, Qianqian Wu, Huaiyu Gu, Lindong Lu, Keming Wang, Dechen Lin, Guanlin Shen, Xuanhe Zhou, Linfeng Zhang, Yuhang Zang, Xiaoyi Dong, Jiaqi Wang, Bo Zhang, Lei Bai, Pei Chu, Weijia Li, Jiang Wu, Lijun Wu, Zhenxiang Li, Guangyu Wang, Zhongying Tu, Chao Xu, Kai Chen, Yu Qiao, Bowen Zhou, Dahua Lin, Wentao Zhang, and Conghui He. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing, 2025. URL https://arxiv.org/abs/2509.22186.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. OmniDocBench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24838–24848, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ouyang_OmniDocBench_Benchmarking_Diverse_PDF_Document_Parsing_with_Comprehensive_Annotations_CVPR_2025_paper.html.

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*, KDD '22, pp. 3743–3751. ACM, August 2022. doi: 10.1145/3534678.3539043. URL http://dx.doi.org/10.1145/3534678.3539043.

Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2Code: Benchmarking multimodal code generation for automated front-end engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3956–3974, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.199. URL https://aclanthology.org/2025.naacl-long.199/.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding, 2023. URL https://arxiv.org/abs/2401.00908.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pp. 8529–8548. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.463. URL https://aclanthology.org/2024.acl-long.463/.

Jiapeng Wang, Lianwen Jin, and Kai Ding. Lilt: A simple yet effective language-independent layout transformer for structured document understanding, 2022. URL https://arxiv.org/abs/2202.13669.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multimodal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 2579–2591. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.201. URL https://aclanthology.org/2021.acl-long.201/.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding, 2022. URL https://arxiv.org/abs/2012.14740.

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*, KDD '20, pp. 1192–1200. ACM, August 2020. doi: 10.1145/3394486.3403172. URL http://dx.doi.org/10.1145/3394486.3403172.

Dayu Yang, Antoine Simoulin, Xin Qian, Xiaoyi Liu, Yuwei Cao, Zhaopu Teng, and Grey Yang. Docagent: A multi-agent system for automated code documentation generation, 2025. URL https://arxiv.org/abs/2504.08725.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-docowl: Modularized multimodal large language model for document understanding, 2023. URL https://arxiv.org/abs/2307.02499.

Boyang Zhang, Sebasti'an G. Acosta, Preston Carlson, Sacha Bron, Pierre-Lo"ic Doulcet, and Simon Suo. ParseBench: A document parsing benchmark for AI agents. *arXiv preprint arXiv:2604.08538*, 2026. URL https://arxiv.org/abs/2604.08538.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis, 2019. URL https://arxiv.org/abs/1908.07836.

A Objective Evaluation Details

This appendix records implementation choices that are fixed for reproducibility but are not central to the paper's conceptual contribution.

A.1 Rendered block extraction and normalization

Each predicted HTML file is rendered in headless Chromium at the benchmark viewport. The evaluator traverses visible DOM content and records the semantic type, visible text, raw table HTML when available, LaTeX-like markup when exposed by the rendered DOM, and the rendered box. A predicted table is kept as one table object; its cells are not independently promoted to page-level blocks unless the table remains unmatched and is transferred to the text pool. Elements whose class attribute contains `placeholder` are excluded from matching. Invisible text, transparent text, and extremely small text are also ignored. This prevents accessibility or scene-description placeholders from contributing text that is not visibly rendered as document content.

Ground-truth boxes and predicted boxes are mapped to the common page coordinate system. For each object, the outermost coverage box is used. The overlap used by matching is $s_{\mathrm{ov}}(g, p) =$

$\frac{|b_g \cap b_p|}{\min(|b_g|,|b_p|)}$, with zero returned for disjoint boxes. Thus, the denominator is not the union area: a small block fully covered by a larger block can obtain an overlap of one. This is useful for detecting splits and merges, but the final text and layout scores still penalize missing or excessive content.

A.2 Table assignment and TEDS approximation

For every ground-truth/predicted table pair, the evaluator computes TEDS, TEDS-S, partial text similarity, and the smaller-area overlap. The candidate score is $0.45\,\mathrm{TEDS} + 0.25\,\mathrm{TEDS\text{-}S} + 0.20\,s_{\mathrm{partial}} + 0.10\,s_{\mathrm{ov}}$. Pairs below the table candidate threshold 0.40 are unavailable to the assignment. The remaining score matrix is solved with a Hungarian one-to-one assignment, followed by threshold validation. This is a global assignment over table candidates, not a greedy sort.

The table parser converts HTML `tr`/`td`/`th` elements into ordered rows and cells. The implemented score is a normalized row/cell sequence edit approximation rather than a general-purpose DOM tree edit distance. TEDS retains cell strings; TEDS-S replaces them with structural cell tokens. Both are reported in $[0, 1]$, with one denoting exact agreement. Cell substitutions are weighted by the normalized character-level text distance, so a one-character difference has a smaller cost than a substantially different cell. Row, insertion, deletion, and cell-structure differences are also represented. Pages without a ground-truth table are omitted from the table means.

A.3 Text normalization and grouping

Text matching uses a conservative normalization. Whitespace and dollar delimiters are removed, while punctuation, mathematical symbols, CJK characters, Japanese, Korean, and other Unicode letters are retained. A ground-truth ξ is interpreted as a wildcard character during edit-distance computation. Placeholder elements are removed before this stage. Formula objects from both sides are converted to text objects by removing dollar delimiters, and no separate formula CDM score is computed in the current objective leaderboard.

For every ground-truth and predicted text pair, the evaluator computes `partial_ratio` after the above normalization. A bipartite edge is created when $s_{\mathrm{partial}} \geq 0.66$. Unlike earlier overlap-driven variants, this candidate graph is text-led: geometry is used later for layout scoring rather than for admitting text edges. Connected components of the bipartite graph form local candidate regions. Singleton blocks without edges are kept as residual insertions or deletions, so they still affect page text distance but do not create layout nodes.

For each connected component, the evaluator constructs several deterministic order hypotheses: visual reading order, horizontal order, reversed reading order, and a graph-guided order that places a block near its strongest partners on the other side. It chooses the pair of orders with the lowest concatenated component edit distance. For components whose total number of blocks is small enough, the evaluator additionally tries a bounded local reorder pass over the recovered groups.

Given fixed ground-truth and predicted orders, the evaluator computes the final groups with a monotone segmentation dynamic program. Each transition may consume a short ground-truth segment, a short predicted segment, or both. Two-sided transitions are supported by the bipartite edges in the component, while one-sided transitions represent insertions or deletions and receive full edit cost. The recovered groups therefore support one-to-one, one-to-many, many-to-one, and many-to-many matches while preventing unsupported distant text from being grouped solely to reduce local cost.

After all components and residual blocks are processed, the evaluator performs one page-level ordering pass over the recovered groups and computes a single normalized edit distance $D_{\mathrm{text}} = \frac{\mathrm{Lev}(T_g, T_p)}{\max(|T_g|,|T_p|,1)}$. This max denominator makes the score symmetric with respect to which side contains more characters. A lower score is better. The same recovered groups, excluding groups with an empty side, are passed to the layout scorer.

A.4 Layout graph similarity

The layout graph is built only from matched objects. A matched table contributes one node. A text group contributes one node whose box is the axis-aligned envelope of all GT or predicted boxes in that group. Text groups with an empty GT side or an empty prediction side are not inserted into the

graph. When non-text anchors such as matched tables are available, very weak text groups whose normalized text distance exceeds 0.75 are also excluded from graph construction. Node similarity combines normalized center displacement with width and height agreement. For the ground-truth nodes, the evaluator computes the $k = 3$ nearest neighbors by Euclidean center distance. The same node identities are used to read the corresponding predicted relations, so the graph comparison does not require a second text matching step.

For an edge $i \to j$, direction similarity compares the normalized two-dimensional vectors, length similarity compares the relative Euclidean distances, and vector similarity compares the normalized displacement after prediction. The edge score is $s_{\text{edge}} = 0.4s_{\text{dir}} + 0.3s_{\text{len}} + 0.3s_{\text{vec}}$, and the final score is $S_{\text{layout}} = 0.55\,\overline{s}_{\text{edge}} + 0.30\,\overline{s}_{\text{node}} + 0.15F_1^{\text{coverage}}$. This construction preserves two-dimensional information: a relation that has the correct distance but points in a different direction receives a lower score. A group with many source blocks still contributes one graph node because its envelope represents the layout region that was matched. The coverage term is computed from the matched graph nodes relative to the objects eligible for layout scoring, so pages that recover only a small subset of meaningful matched objects receive lower layout scores even when the recovered subset is geometrically accurate.

A.5 Leaderboard aggregation

The reported table means use only the 263 pages with ground-truth tables. Text error and layout similarity use all 1,115 pages. We convert text error to quality by $1 - D_{\text{text}}$ and use $S_{\text{obj}} = (0.3\,\overline{\text{TEDS}} + \overline{(1 - D_{\text{text}})} + \overline{S_{\text{layout}}})/2.3$. TEDS-S is reported as a diagnostic and is not an additional term in the weighted aggregate, avoiding double-counting table structure. The final leaderboard incorporates the rerun outputs for the previously identified low-scoring pages before aggregation.

B Subjective Evaluation Supplement

B.1 Human study and preference targets

The human study comprises eight disjoint blocks of 234 physical comparisons. Blocks 1–5 have five participants each, block 6 has four, and blocks 7–8 have two, for 33 participants in total. Each participant answers 234 main trials and six hidden repeats. Thus, the 7,920 responses comprise 7,722 main annotations and 198 repeated annotations; repeats are used to assess short-term consistency and are excluded from the calibration targets. Of the main annotations, 7,633 contain a valid A, B, or TIE preference and 89 are `CANNOT_JUDGE`, giving 98.85% annotation coverage. All 1,872 physical comparisons retain a human target.

For comparison i, $n_{i,A}$, $n_{i,T}$, and $n_{i,B}$ count valid votes after mapping the displayed candidates to their canonical identities. Their encoded mean h_i measures preference for canonical A. A tie contributes half a preference point to each candidate. For example, three A votes, one tie, and one B vote give $h_i = 0.7$. A value of 0.5 can arise from unanimous ties or opposing A/B votes, so it does not by itself measure agreement. We concatenate the eight 234-entry vectors to obtain the 1,872-entry human reference; we do not average unrelated comparisons across blocks.

The reported pairwise inter-rater agreement uses 13,265 valid within-comparison rating pairs. Exact agreement is 62.30% (95% CI: 60.79–63.82). Fleiss-type κ is 0.357 (0.334–0.379), and ordinal Krippendorff-type α is 0.459 (0.433–0.484). These quantify agreement, not error against an objective ground truth. Among the 1,872 comparisons, 39.85% are unanimous among valid raters and 82.00% have a unique plurality; a plurality need not exceed half the valid votes. Hidden-repeat agreement is 145/198 (73.23%). Removing each of the 33 participants in turn leaves all ten human ranks unchanged. These checks address different sources of variation; they do not establish that individual human judgments are noise-free. Block content and participant composition are partly confounded, particularly for the two-rater blocks.

B.2 JUDGE INPUTS, AXES, AND PROMPT

Each VLM receives the reference image followed by the two candidate images, with model identities omitted. The reference is the visual standard. For images and charts, judges are instructed not to evaluate internal semantic content; they assess presence, geometry, size, placement, alignment, and relationships to captions and neighboring regions. Text is judged by visible presentation and substantial omissions or garbling rather than character-by-character OCR correctness. Table 4 gives the active comparison counts in the human set; multiple axes may occur in one comparison.

Table 4: Diagnostic axes in the 1,872-comparison human set. Counts denote comparisons, not distinct pages. The rarest axes support case analysis but not precise population comparisons.

Activation	Axis	n	Visual criteria
Fixed	Global layout	1,872	Regions, columns, spacing, grouping, reading order, proportions.
Fixed	Text presentation	1,872	Wrapping, typography, density, alignment, hierarchy, legibility.
Conditional	Table	497	Visible row/column organization and cell alignment.
Conditional	Formula	61	Fractions, scripts, radicals, matrices, and mathematical alignment.
Conditional	Algorithm	12	Block organization, indentation, and step hierarchy.
Conditional	Vertical text	90	Vertical flow, orientation, and reading organization.
Conditional	List hierarchy	9	Nesting, indentation, markers, and grouping.

The following is the shared English prompt, with line wrapping adjusted for readability:

> You are an independent visual reconstruction judge for AmbiHTMLBench.
>
> You receive exactly three page images in this order: REFERENCE, CANDIDATE A, CANDIDATE B. The Reference is the sole visual standard. Determine which candidate more faithfully and completely reconstructs it as a human viewer would.
>
> Images and charts: do not judge their internal semantic content. Judge only presence or absence, geometry, size, aspect ratio, position, alignment, and relationships to captions and neighboring regions.
>
> Text: judge visible presentation, not character-by-character OCR correctness. Consider missing or extra text regions, large garbling, wrapping, line count, density, typography, size, weight, color, emphasis, spacing, alignment, hierarchy and legibility. Tiny character or punctuation differences that are difficult to notice should have little or no effect.
>
> `GLOBAL_LAYOUT_FIDELITY` covers page-level spatial organization, regions and columns, alignment, adjacency, spacing, grouping, nesting, reading order, proportions, aspect ratios, whitespace and visual balance.
>
> `TEXT_VISUAL_PRESENTATION` covers the visible presentation of text across the page.
>
> Adaptive axes are present only when the corresponding structure exists in the Reference. Use each supplied English description and `bbox_1000` target only to locate the Reference structure. The candidate structure may move; judge its corresponding visible organization rather than absolute coordinates.
>
> Score every active axis: 2 = A clearly better; 1 = A slightly better; 0 = no stable perceptible difference; -1 = B slightly better; -2 = B clearly better.
>
> Overall winner must be A, B, or TIE. It is a holistic visual judgment, not a mechanical sum of axis scores. Confidence 1–5 means expected reproducibility, not magnitude of difference.
>
> Return strict JSON only:
>
> ```
> {"winner":"A|B|TIE", "confidence":1,
>  "reason":"brief English reason, <=80 words",
>  "axes":[{"id":"AXIS_ID", "score":0}]}
> ```
>
> The axes array must exactly match the supplied active axes in membership and order. Do not mention model identities or infer which model produced a candidate.

An `ACTIVE_AXES` JSON array is appended to this prompt for each comparison. Every entry contains its axis ID, global/adaptive scope, English description, and any reference targets with their role and `bbox_1000` coordinates. This supplies location guidance without requiring the candidate structure to occupy the same coordinates. The response validator checks the overall label, confidence range, axis membership/order, and integer scores. Swapped displays are mapped back by exchanging A/B and negating axis scores; TIE is unchanged. Confidence and free-text rationales do not enter the preference proxy.

B.3 Preference-proxy fitting and internal validation

The five judges are Claude Opus 5, Claude Sonnet 5, Gemini 3.1 Pro Preview, GPT-5.6-Sol, and GPT-5.6-Terra. Each evaluates every human comparison in two orders and two repetitions, yielding 37,440 completed outputs. Let $v_{ijor} \in \{0, 0.5, 1\}$ be judge j's canonicalized overall vote under order o and repeat r. The training feature is $\bar{x}_{ij} = \frac{1}{4}\sum_{o=1}^{2}\sum_{r=1}^{2} v_{ijor}$, and the coefficients are estimated as $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^5} \sum_{i=1}^{1872}(h_i - \bar{\boldsymbol{x}}_i^{\mathsf{T}}\boldsymbol{\beta})^2$. We use ordinary least squares, giving every physical comparison equal weight. The fit has no intercept, ridge penalty, nonnegativity constraint, or sum-to-one constraint. The resulting coefficients happen to be positive; they are not normalized. Table 5 records the frozen full-data coefficients. Special-axis scores do not enter this fit. In the ten-parameter representation, the other five coefficients are zero.

Table 5: Frozen preference-proxy coefficients. The deployment applies clipping to the weighted sum; the coefficients are not normalized.

Judge	Coefficient β_j
Claude Opus 5	0.09045204815695966
Claude Sonnet 5	0.10401091278902799
Gemini 3.1 Pro Preview	0.37576678499354577
GPT-5.6-Sol	0.13742077090225210
GPT-5.6-Terra	0.16012498207363500

For internal evaluation, each of the eight annotation blocks is held out in turn: 1,638 comparisons fit the coefficients and 234 receive predictions. The identifiers `PAPER-01` through `PAPER-08` denote these annotation blocks, not source research articles. The 1,872 out-of-fold predictions yield equal-case Brier/MSE $N^{-1}\sum_i(\hat{p}_i - h_i)^2 = 0.0778$ (95% CI: 0.0722–0.0836) and MAE 0.2115 (0.2031–0.2200), with 100% prediction coverage. Spearman 0.915 (0.867–0.952) compares the ten-model proxy and human rankings; it is not a case-level correlation. The full-data coefficients are used for deployment, not to compute these out-of-fold alignment estimates.

Deployment substitutes one canonicalized vote for each four-vote mean. The internal estimates therefore do not directly validate single-pass deployment on new comparisons. Also, $\sum_j \beta_j \simeq 0.8678$: re-encoding all inputs as $1 - x_j$ does not generally give $1 - \hat{p}(A)$. Canonical identities are consequently fixed throughout scoring. This asymmetry is a limitation of the frozen estimator, not corrected post hoc.

B.4 Discrete committee audit and cost ablations

The historical committee audit uses the same repeated evidence but a different aggregation rule. A judge is resolved when at least three of its four canonicalized overall labels agree. Three judges then use two-of-three voting. Unresolved comparisons receive $p(A) = 0.5$ when computing Brier and remain in its denominator. Resolved coverage is reported separately. The audit weights each comparison by its number of valid human votes divided by five, whereas the preference proxy gives each comparison equal weight; their reported losses should not be interpreted as a controlled comparison differing only in aggregation.

Eligible committees contain one judge from each vendor family and achieve at least 75% coverage. The pilot uses 320 comparisons; the remaining 1,552 form the holdout. Selection minimizes abstention-aware Brier, breaking ties by resolved-case Brier and then coverage. Table 6 reports all four eligible cross-family committees.

Table 6: Historical cross-family committee audit. Pilot $n = 320$; holdout $n = 1,552$. Coverage is the fraction resolved. O, S, G, L, and T denote Opus, Sonnet, Gemini, Sol, and Terra. These discrete committees are separate from the deployed preference proxy.

	Pilot			Holdout		
Committee	Brier	Coverage	Rank	Brier	Coverage	Rank
S+G+T (v1)	0.1258	77.81%	1	0.1293	78.99%	4
O+G+T	0.1302	78.44%	2	0.1263	79.90%	2
S+G+L	0.1369	79.06%	3	0.1267	79.45%	3
O+G+L (v2)	0.1404	79.06%	4	0.1240	80.35%	1

The pilot winner v1 falls to fourth on holdout. The paired holdout Brier difference (v1 minus v2) is $+0.0053$ (95% CI: -0.0001 to $+0.0106$). Among 10,000 bootstrap resamples, v2 ranks first 76.81% of the time. The interval includes zero, and v2 was selected after inspecting holdout; this is selection evidence rather than independent confirmation of v2. Furthermore, formula, algorithm, list, and vertical-text comparisons were concentrated in the pilot and are absent from holdout. On the full human set, descriptive macro Brier over the three special axes with at least 30 comparisons favors v1 (0.1268 versus 0.1324). Thus, the overall ordering does not establish universal superiority across structures. Gemini is present in every eligible committee, and family-specific preferences remain a potential confound.

A separate nested ablation tests all 31 nonempty judge subsets using a learned blend of overall and axis evidence. It fits nonnegative, sum-to-one weights with ridge shrinkage toward equal weights; subset selection uses only inner folds, and loss is weighted by valid human votes. With 100% coverage on 1,872 comparisons, three judges yield Brier 0.0653 (0.0609–0.0699); two judges yield 0.0663 (0.0616–0.0713). The paired two-minus-three difference is $+0.0010$ (-0.0002 to $+0.0021$), and the one-standard-error rule selects Gemini+Sol in all eight outer folds. Five judges are worse than three by $+0.0025$ ($+0.0009$ to $+0.0041$). Intervals use 10,000 page-block resamples. These results concern an axis-aware model with different weighting and fitting, so they neither establish five judges as the minimum nor replace the frozen overall-only regression estimator. With no prespecified noninferiority margin, overlapping intervals do not prove equivalence.

B.5 Deployment schedule and candidate versions

The fixed schedule contains 10,000 comparisons across 1,113 eligible pages (seed 20260831). Each of ten models appears 2,000 times: 35 model pairs occur 222 times and ten occur 223 times. It retains the 1,872 human-study page/model pairs and adds 8,128 pairs. Pages and document structures are not sampled equally.

Each comparison receives one output per judge in a single display order, totaling 50,000 outputs. Historical directions depend on completed-result availability, never on predicted winners; reuse may therefore preserve latency-related selection effects. New requests use the primary order. The initial deployment reuses 11,380 outputs and obtains 38,620 new ones. The final table has five outputs per comparison, with no missing or duplicated entries.

The reported leaderboard incorporates the subsequent candidate synchronization. Of 58 rerun page/model outputs touched by the schedule, 54 pass the existing visible-text and render-geometry checks; four retain their earlier candidates. The replacements affect 98 comparisons: 490 judge outputs are refreshed and 49,510 retained. Pair identities, quotas, orientations, and coefficients remain fixed. All ten point ranks are unchanged, although scores change; Sol moves from 154.34 to 154.70. This checks version sensitivity, not independent performance.

Twenty-one affected comparisons come from the human-study schedule; their earlier labels do not validate the replacement candidates. The final table therefore contains 1,851 unchanged human-study comparisons, 21 updated comparisons without new human labels, and 8,128 unlabeled additions. The original calibration set is unchanged. Neither its overlap with deployment nor the updated table supplies external human validation.

The initial deployment's final 22 failures comprise 14 request-size failures and eight timeouts. Eighteen repaired calls use high-quality JPEG reference copies; one further reduces the long edge from

6,320 to 6,000 pixels (5.1%), preserving aspect ratio, to meet the visual-patch limit. Original references remain unchanged. These repairs are separate from candidate replacements.

B.6 RANKING AND UNCERTAINTY

For comparison i between models a_i and b_i, the proxy outcome contributes $\hat{p}_i(A)$ preference points to a_i and $1 - \hat{p}_i(A)$ to b_i. The soft-preference column is the mean of these points over a model's 2,000 scheduled appearances. It is not accuracy against human labels.

Bradley–Terry fitting estimates centered log-abilities θ_m with $q_i = \sigma(\theta_{a_i} - \theta_{b_i})$ by maximizing $\sum_i [\hat{p}_i(A) \log q_i + (1 - \hat{p}_i(A)) \log(1 - q_i)] - \frac{10^{-3}}{2} \sum_m \theta_m^2$. The small ℓ_2 term stabilizes the ranking fit; it is not a penalty on the OLS coefficients. Centering sets the mean ability to zero. Reported scores are $400\theta_m / \log(10)$, an Elo-like relative scale. A negative score indicates below-average estimated ability within this comparison pool, not failed reconstruction.

For the deployment leaderboard, 2,000 resamples draw 1,113 page blocks with replacement (seed 20260831). Each selected page carries all of its scheduled comparisons. The Bradley–Terry model is refitted for every resample; 2.5th and 97.5th percentiles form the score and rank intervals. These are marginal intervals and do not constitute simultaneous tests of every model pair. "First in every bootstrap sample" describes the observed resampling frequency, not certainty under all sources of error.

Proxy alignment uses 5,000 page-block resamples stratified by annotation block (seed 20260830), with fixed out-of-fold predictions; the committee audit uses 10,000 similarly stratified resamples. Within each stratum, comparisons from the same page remain together. Bootstrap fixes proxy coefficients and predictions, omitting renewed judge sampling, regression fitting/selection, candidate generation/cleaning, and systematic bias. Alignment is internal to the human study; the expanded leaderboard provides no full-scale human Accuracy, Recall, or F1.