

VFDR-BENCH: A Multi-Lingual, Multi-Domain Benchmark for Visually Faithful Document Reconstruction

Anonymous Submission

Abstract

PDF-to-HTML conversion improves accessibility but often compromises layout fidelity, as existing systems prioritize text accuracy over visual structure, typography, and spatial organization. We introduce VFDR-BENCH, a benchmark for *Visually Faithful Document Reconstruction* (VFDR) that assesses both structural and stylistic preservation in document-to-HTML generation. The benchmark consists of 5000 multilingual documents across 22 languages and diverse layouts. We also introduce REPLICa, a layout-aware conversion engine that produces FID-HTML - a semantically enriched HTML format unifying content, style, and geometry - further refined with human corrections to ensure high-quality ground truth. VFDR-BENCH evaluates systems along four dimensions: text extraction, logical structure, physical structure, and visual fidelity. To support this, we introduce four new metrics - Logical Structure Score (LSS), Global Position Score (GPS), Local Position Score (LPS), and Visual Fidelity Score (VFS), in addition to standard existing metrics. Benchmarking 17 diverse methods shows that even advanced VLMs such as GPT-5 and Gemini 2.5 Pro reach only $\sim 50\text{--}60\%$ fidelity, underscoring the difficulty of VFDR. VFDR-BENCH provides the first unified evaluation suite for layout-aware document rendering, establishing a foundation for visually grounded document understanding and web-native document representation.

Keywords: Document Layout, Multilingual OCR, Vision-Language Models, Document Parsing

Contents

| | | | |
|---|----------|--|-----------|
| 1 Introduction | 2 | 5.2 Localize: Layout-Aware HTML Generation (2) | 7 |
| 2 Related Work | 4 | 5.3 Assemble: Position- and Reading-Aware Merge (3) | 7 |
| 2.1 Logical Reconstruction | 4 | 5.4 Refine: Visual Fidelity Enhancements (4) | 9 |
| 2.2 Physical Reconstruction | 4 | 5.5 Implementation Details | 9 |
| 2.3 Benchmarks | 5 | 6 Benchmark Construction | 9 |
| 3 VFDR Requirements | 6 | 7 Evaluation Metrics | 9 |
| 4 FID-HTML: A High-Fidelity Document Representation. | 6 | 8 Results | 14 |
| 5 REPLICa: High-Fidelity Conversion Engine | 6 | 9 Discussion | 19 |
| 5.1 Segment: Robust Layout Detection (1) | 6 | 10 Limitations | 19 |
| | | 11 Conclusion | 19 |

| | |
|--|-----------|
| 12 REPLICA Implementation Details | 21 |
| 12.1 Segment | 21 |
| 12.1.1 Canonical Layout Classes . | 21 |
| 12.1.2 Hybrid Layout Prediction . | 21 |
| 12.2 Localize | 21 |
| 12.2.1 Stage 1: Raw sub-HTML generation | 21 |
| 12.2.2 Stage 2: Semantic sub-HTML generation | 24 |
| 12.3 Assemble | 30 |
| 12.3.1 Reading Order-Aware Merging | 30 |
| 12.4 Refine | 32 |
| 12.4.1 Font Size Fixing Algorithm | 32 |
| 12.4.2 Reflection | 33 |
| 12.4.3 Background Restoration . . | 34 |
| 13 Design choices behind REPLICA | 34 |
| 13.1 Module-specific Metrics | 34 |
| 13.2 Ablations | 34 |
| 13.3 REPLICA helps annotation . . . | 37 |
| 14 REPLICA as a baseline for VFDR | 39 |
| 15 Evaluations | 40 |
| 15.1 Human-centric Evaluations | 40 |
| 15.2 Prompts used for evaluations of methods | 40 |
| 15.3 Prompts used for VLM-as-a-judge metrics | 40 |
| 15.4 Overall Score Calculation | 40 |
| 16 Additional Results | 41 |
| 17 Qualitative Samples of REPLICA conversions | 41 |

1 Introduction

Portable Document Format (PDF) files are designed to preserve the visual integrity of documents through precise control over layout, typography, and styling. However, this fidelity often deteriorates when such documents are adapted for the web. Conventional optical character recognition (OCR) engines [4–6] recover textual content but discard spatial and stylistic information such as layout hierarchy, alignment, or font emphasis. Other workflows, including PDF-to-Markdown conversion [7, 8] or parser-based extraction [9–11], provide lightweight markup yet fail to capture the

document’s visual organization. As a result, these pipelines yield flattened or visually distorted representations that lack the semantic and aesthetic richness of the source.

PDF-to-HTML conversion remains a cornerstone in web publishing, digital archiving, and accessibility workflows. Tools such as pdf2htmlEX [12] produce minimal HTML that retains basic positioning but sacrifices typographic nuance. Features that convey meaning—font weight, size, color, and alignment—are often reduced to coarse approximations or omitted entirely. Consequently, the converted HTML maintains a logical skeleton of the document but not its intended visual semantics or reading experience.

Recent advances in vision–language models (VLMs) [13–16] and document-specific architectures have introduced the possibility of generating structured outputs directly from page images. Yet, despite promising results, these systems remain constrained by limited layout diversity, incomplete script coverage, and a tendency to prioritize textual extraction over visual fidelity. This gap motivates the need to systematically study *Visually Faithful Document Reconstruction* (VFDR).

Our key contributions are as follows:

- FID-HTML: a standardized, web-native document format in high-fidelity HTML that is both semantically enriched and visually faithful, preserving fine-grained spatial and stylistic cues.
- REPLICA: a layout-aware rendering engine that segments, localizes, and hierarchically assembles document elements to generate FID-HTML as high-fidelity ground-truth HTML, which are subsequently refined through human annotations.
- VFDR-BENCH: a multilingual, multi-domain benchmark comprising 5000 documents spanning over 22 languages and 17 domains, designed to evaluate document-to-web conversion fidelity, robustness, and cross-lingual generalization.
- Evaluation suite: a novel set of metrics—Logical Structure Score (LSS), Global Position Score (GPS), Local Position Score (LPS), and Visual Fidelity Score (VFS)—introduced to jointly quantify logical structure, physical layout alignment,

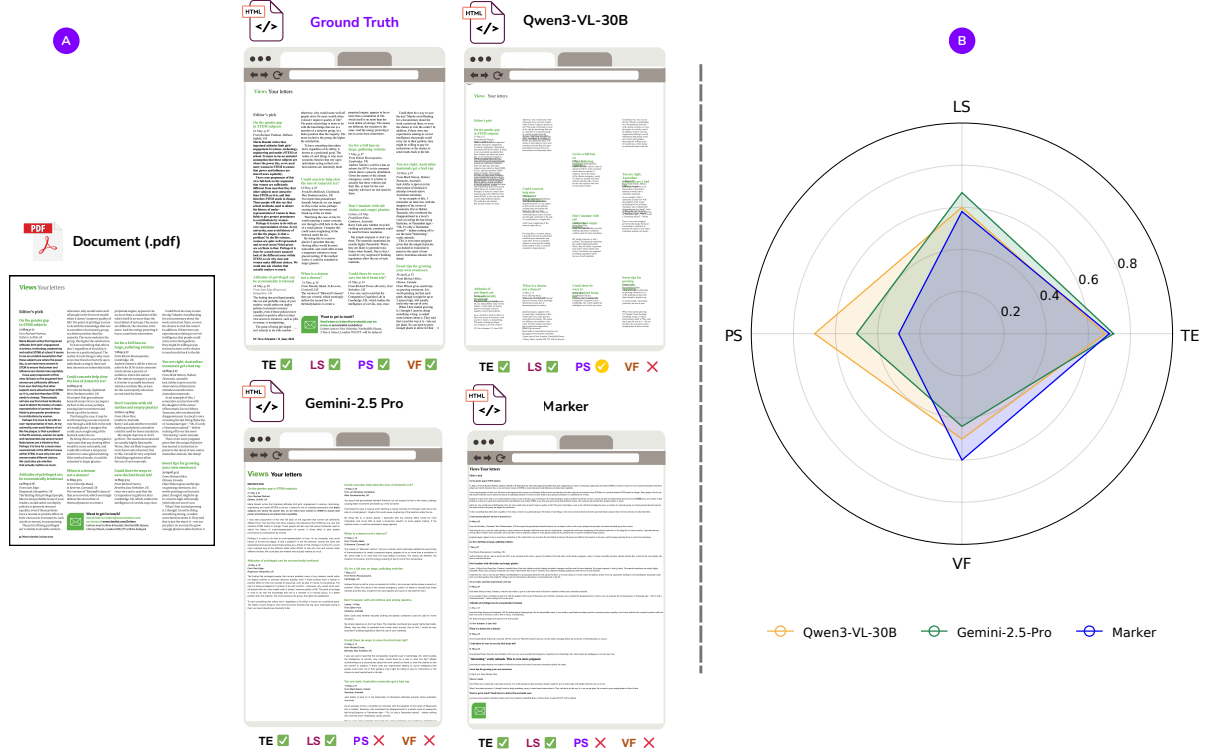


Fig. 1: A Comparing representative document reconstruction methods. The VFDR-BENCH benchmark includes open-source Vision-Language Models (e.g., Qwen3-VL-30B [1]), commercial systems (e.g., Gemini-2.5 Pro [2]), and pipeline-based approaches (e.g., Marker [3]). The benchmark helps measure fine-grained differences in visual fidelity, such as preservation of font attributes (bold, italics), font size, text color, paragraph indentation, and column structure. **B Quantitative radar plot** illustrating comparative performance across key evaluation dimensions—TE (Text Extraction), PS (Physical Structure), LS (Logical Structure), and VF (Visual Fidelity). These results demonstrate the challenge and comprehensiveness of VFDR-BENCH, motivating progress in visually faithful document reconstruction. Additional examples are provided in fig. 2, fig. 17, Supplementary

and visual fidelity in document-to-HTML reconstruction.

Using VFDR-BENCH, we evaluate 17 methods spanning multiple categories—including pipeline-based tools, expert VLMs, general-purpose VLMs (both open- and closed-source), and Screenshot2HTML models. Beyond overall performance, we conduct fine-grained analyses across languages, domains, column layouts, and document conditions, providing a comprehensive characterization of model behavior under diverse real-world settings.

By formalizing the VFDR task and introducing a reproducible benchmark with standardized metrics, VFDR-BENCH establishes a unified framework for visually grounded document understanding—bridging text-centric reconstruction and layout-aware rendering. This framework extends the utility of enriched HTML outputs beyond readability, enabling accessibility, digital archiving, style-preserving content reuse, and high-quality translation - see examples in fig. 2. The inclusion of semantic tags, preserved reading order, and enriched figure alt-text in FID-HTML creates possibilities for seamless screen-reader compatibility and web accessibility.



Fig. 2: Downstream applications with high fidelity preservation enabled by VFDR: **A** Document Preservation, **B** In-browser translation (English to Chinese).

2 Related Work

Document reconstruction can be broadly categorized into two complementary goals: *logical reconstruction*, which recovers semantic and hierarchical structure, and *physical reconstruction*, which restores layout geometry and visual styling. Most prior systems emphasize one aspect at the expense of the other. VFDR-BENCH unifies these perspectives by enabling quantitative evaluation of both through the lens of *Visually Faithful Document Reconstruction* (VFDR). Refer table 2 for more details.

2.1 Logical Reconstruction

Pipeline Parsing: Modular systems decompose the task into layout analysis, reading-order prediction, OCR, math/table parsing, etc., assembling outputs via rules or learned glue [30–37]. Open tools such as Marker and MinerU integrate state-of-the-art detectors and parsers to convert PDFs to Markdown/JSON/HTML with strong coverage of content and hierarchy [3, 26, 38]. While effective at logical structure, these pipelines typically linearize layouts (e.g., single-column Markdown) and under-specify style and exact positions, making high-fidelity rendering difficult.

End-to-End VLMs: Generalist VLMs (e.g., GPT-4V/Claude/Gemini, Qwen-VL, MiniCPM, InternVL, DeepSeek-VL2) demonstrate zero-shot document understanding and can emit free-form descriptions or lightweight HTML/JSON [2, 39–47]. However, they often miss fine layout/style details, yielding outputs that preserve logical structure but diverge from the original appearance. Specialist models—Donut/Nougat, LayoutLM-series, GOT, SPTS, UReader, mPLUG-DocOwl, MonkeyOCR/TextMonkey, etc.—are trained on document-centric corpora to emit structured markup with better hierarchy, tables, and formulas [27, 48–65]. Recent variants (e.g., Qwen2.5-VL, olmOCR) produce HTML-like outputs and are finetuned for structured conversion [14, 43, 66]. Despite clear gains in semantics and reading order, most still under-deliver on visual fidelity (fonts/spacing/multi-column placement).

2.2 Physical Reconstruction

Approaches for Document Images: Methods such as PDFMathTranslate and image-to-markup/LaTeX regeneration focus on scientific documents, preserving math, tables, and two-column layouts [67–70]. Beyond literal reassembly, recent systems like DREAM emphasize high-fidelity digital reconstruction and autoregressively encode both content and layout, recovering element positions [71].

Approaches for UI Screenshots: Screenshot-to-HTML approaches primarily target UI-centric, text-light layouts. Prior works have explored diverse approaches to screenshot-to-HTML generation [29, 72–84]. Many systems refine outputs

Table 1: Comparison of recent document parsing benchmarks. VFDR-BENCH offers the broadest and most diverse coverage with 22 languages and 17 domains, and uniquely provides high-fidelity FID-HTML annotations. It is also the only document benchmark that captures both physical structure (PS) and visual fidelity (VF)—the two most critical components for VFDR.

| Benchmark | Images | Languages | # Domains | Output Format | TE | LS | PS | VF |
|---------------------------|-------------|-----------|-----------|-----------------|----------|----------|----------|----------|
| OmniAI OCR Benchmark [17] | 1000 | 1 | NA | md | ✓ | ✓ | × | × |
| OmniDocBench [18] | 981 | 3 | 9 | txt | ✓ | × | × | × |
| CC-OCR [19] | 800 | 10 | 6 | txt | ✓ | × | × | × |
| olmOCR-Bench [14] | 1403 | 1 | NA | md | ✓ | ✓ | × | × |
| OCRBench-v2 [20] | 10000 | 2 | 26 | md | ✓ | ✓ | × | × |
| READoc [21] | 3576 | 27 | 6 | md | ✓ | ✓ | × | × |
| dp-bench [22] | 200 | 1 | 4 | md | ✓ | ✓ | × | × |
| KITAB-Bench [23] | 917 | 1 | 9 | md | ✓ | ✓ | × | × |
| Image2Struct [24] | 2100 | 1 | 3 | LaTeX, HTML | × | ✓ | × | × |
| VFDR-Bench (Ours) | 5000 | 22 | 17 | FID-HTML | ✓ | ✓ | ✓ | ✓ |

Table 2: Comparison of representative document-to-HTML approaches. ‘Semantic Tags’ (ST) indicates whether semantic HTML tags are preserved; ‘Absolute Position’ (AP) denotes support for absolute layout rendering; ‘Languages’ (Lang) represents the number of supported languages; ‘Font Size’ (FS) and ‘Font Attributes’ (FA) measure fidelity in maintaining size, style (bold, italics, underline), and color; ‘Visual Stitching’ (VS) evaluates retention of non-textual visual elements. ‘P’ indicates partial support. Methods are grouped into: Pipeline Tools, Specialized VLMs, General VLMs, and Screenshot2HTML categories.

| Method | HTML | CSS | ST | AP | Lang | FS | FA | VS |
|--------------------------|------|-----|----|----|------|----|----|----|
| MineruPDF [25] | – | – | – | – | 1 | – | – | – |
| Marker [3] | ✓ | – | – | – | 90 | – | P | ✓ |
| Docling [26] | ✓ | P | – | – | 1 | – | – | – |
| GTOCR [27] | ✓ | P | – | – | 2 | – | P | – |
| SmolDocling [13] | ✓ | P | ✓ | – | 1 | – | P | – |
| MonkeyOCR [15] | ✓ | – | – | – | 2 | – | – | – |
| Qwen-3-VL-30B [1] | ✓ | ✓ | – | P | 29 | – | ✓ | – |
| Gemma-3-27B-it [28] | ✓ | ✓ | – | – | 100+ | – | P | – |
| Gemini 2.5 Pro [2] | ✓ | ✓ | ✓ | – | 100+ | – | P | – |
| Waffle VLM Websight [29] | ✓ | ✓ | – | – | 1 | – | – | – |

iteratively via rendered feedback through reflection [81, 82, 85], and auxiliary work explores GUI detection [86, 87] and natural language-driven refinement [88]. However, these methods struggle to generalize to complex, text-heavy documents due to distribution shifts.

2.3 Benchmarks

Document Parsing Benchmarks. Recent document parsing benchmarks [14, 17–24] typically evaluate models through document-to-text or document-to-Markdown generation, emphasizing *text extraction* (TE) and, to a limited extent, *logical structure* (LS). However, these benchmarks

do not explicitly assess *physical structure* (PS) or *visual fidelity* (VF), both of which are essential for Visually Faithful Document Reconstruction (VFDR). Our benchmark addresses this gap by jointly evaluating TE, LS, PS, and VF, offering a more holistic assessment of document-to-HTML systems. Refer to table 1

Screenshot-to-HTML Benchmarks. Several benchmarks target screenshot-to-HTML generation [89–94]. Some of these efforts explicitly evaluate *visual fidelity*, but their inputs are predominantly web or mobile UI screenshots. Consequently, they do not address the challenges posed by *visually rich documents*—such

as dense layouts, complex reading orders, multilingual content, and heterogeneous typographic structures—which require fundamentally different modeling and evaluation strategies. Our benchmark fills this gap by focusing on visually faithful reconstruction of real-world document layouts rather than UI screenshots.

3 VFDR Requirements

VFDR-BENCH aims to evaluate the capability of models to achieve *Visually Faithful Document Reconstruction*, i.e., to reproduce both logical hierarchy and physical layout in web-native HTML. A good reconstruction must exhibit four key properties: (1) accurate text extraction (TE) across diverse layouts and languages; (2) faithful preservation of logical structure (LS), including section headers, captions, and nested lists; (3) precise maintenance of physical layout (PS), capturing spatial positioning, indentation, and placement of figures and tables; and (4) overall visual fidelity and stylistic consistency (VF), preserving fonts, colors, and emphasis. With these properties in mind, we introduce a standardized HTML-based format FID-HTML which can represent fine-grained semantic, spatial and stylistic properties of the original document. We describe this format next.

4 FID-HTML: A High-Fidelity Document Representation.

FID-HTML is a high-fidelity, semantically enriched HTML representation produced by the REPLICIA framework. It is designed to capture both the semantic structure and the visual presentation of a document, enabling downstream tasks that require faithful reconstruction as well as machine-readable semantics.

Core Representation. FID-HTML encodes document content along multiple dimensions: (1) *Textual content*, where full text is extracted and preserved at both word and block level; (2) *Hierarchical structure*, represented with nested `<div>` elements that capture document hierarchy such as sections and subsections; (3) *Semantic tags*, including `` for lists, `<table>`, `<tr>`, `<td>` for tables, and semantic class names for `<div>`s, providing rich semantic grounding; (4) *Positional*

information, with global placement encoded via absolute positioning and local alignment through relative positioning, where textual segments are wrapped in `<p>` tags inside higher-level `<div>` containers; (5) *Styling metadata*, retaining font size, color, and text attributes such as bold, italics, underline, and strikethrough as inline CSS for high-fidelity style preservation; and (6) *Figures and backgrounds*, where images, figures, and page backgrounds are layered into the HTML, and alt-text for figures is automatically generated by VLMs to improve accessibility. Refer fig. 3 for more details.

Advantages: Unlike traditional parsers that discard layout, style, and semantics, FID-HTML retains the full spectrum of document signals. Semantic tags and logical grouping benefit downstream LLM/VLM tasks such as understanding, summarization, and retrieval; absolute and relative positioning preserve layout fidelity; styling metadata ensures faithful rendering; and VLM-generated alt-text improves accessibility for screen readers—capabilities often absent in conventional systems.

5 REPLICIA: High-Fidelity Conversion Engine

Existing approaches for document-to-html conversion fail to produce FID-HTML or its equivalent representational aspects (table 2). Therefore, we build REPLICIA, a high-fidelity layout-aware document-to-html conversion engine to generate FID-HTML.

The pipeline adopts a modular, multi-stage architecture that integrates layout segmentation, OCR, and vision-language prompting across four key stages (fig. 17): **Segment** (1), **Localize** (2), **Assemble** (3), and **Refine** (4).

5.1 Segment: Robust Layout Detection (1)

Effective document-to-HTML conversion begins with identifying the most logical way to divide a document: its layout structure. Also, real-world documents are inherently nested (e.g., a title within a header, a footnote within a footer), unlike traditional methods that assume flat classes. To capture this, we construct a nested layout tree

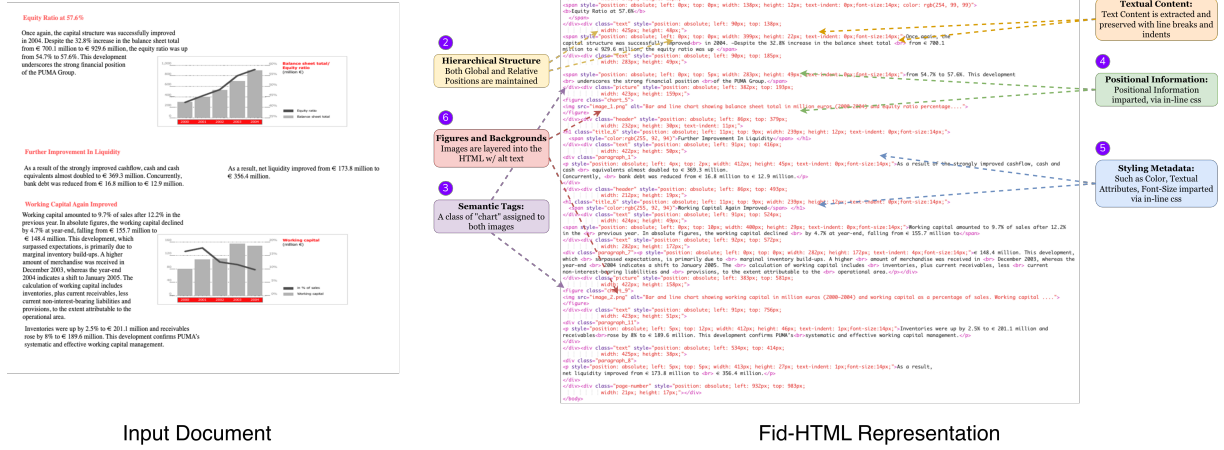


Fig. 3: FID-HTML: High-Fidelity Document Representation. FID-HTML is the enriched output format of REPLICa, designed to preserve both the semantic structure and visual presentation of a document. The representation encodes: (1) *Textual content*, where extracted text is preserved with line breaks and indents; (2) *Hierarchical structure*, maintaining global and relative positions with nested containers; (3) *Semantic tags*, which impart semantic information to each html tag (4) *Positional information*, captured with inline CSS for absolute and relative placement; (5) *Styling metadata*, including font size, color, and attributes (bold, italics, underline, strikethrough); and (6) *Figures and backgrounds*, where images and page backgrounds are integrated with alt-text for accessibility.

(1c in fig. 17), where parent-child relationships follow spatial containment. Because no single layout detection model generalizes across diverse documents and scripts, we employ an ensemble of complementary detectors: flat layout models (1a) and hierarchical text segmentation models (1b) to merge semantically related regions and provide nesting cues. This ensemble yields robust boundary detection, accurate element nesting, and well-structured layout trees - crucial for faithful reconstruction.

5.2 Localize: Layout-Aware HTML Generation (2)

Full-document HTML generation often fails on long, complex layouts. Hence, we use the nested layout tree from previous stage and generate local **sub-HTMLs** for each of the nodes via a two-stage, layout-aware prompting process:

Stage 1 – Raw sub-HTML generation (2a): We apply *Set-of-Mark (SoM) Prompting* [95], where the layout crop image is overlaid with indexed paragraph bounding boxes. This composite input, together with a text prompt, is provided

to a VLM, which then generates a flat sub-HTML. The output ensures OCR coverage while preserving the document’s physical structure through explicit bounding box retention.

Stage 2 - Refining sub-HTMLs (2b): We provide the layout crop, the raw sub-HTML, and a text prompt to a VLM, which produces a refined sub-HTML enriched with semantic tags, improved OCR, and styling such as color and font attributes (bold, italics, underline) through dynamic CSS. Auxiliary context is provided in the prompt through OCR ensembles and attribute recognizers to boost fidelity.

5.3 Assemble: Position- and Reading-Aware Merge (3)

We adopt a position- and reading-aware merge strategy, directly assembling sub-HTMLs from the localization stage into the final HTML. Each sub-HTML is placed using its absolute bounding box while aligned with the predicted reading order, and visual elements (e.g., images) are stitched back at their positions. This ensures the merged output preserves layout semantics, structural organization, and natural textual flow.

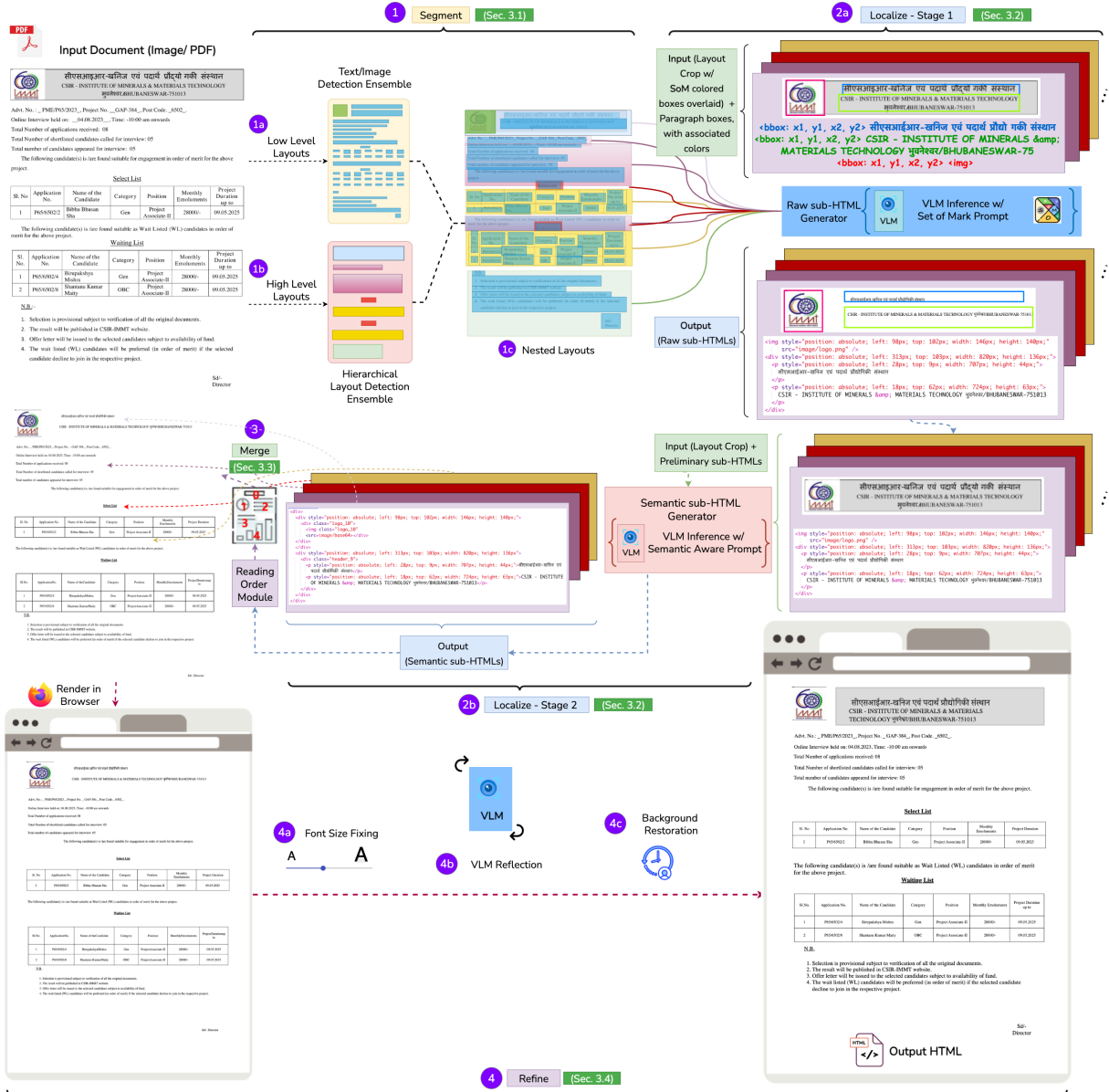


Fig. 4: Overview of REPLICA, our four-stage framework for high-fidelity document-to-HTML conversion. (1 **Segment**) Build a nested layout tree using an ensemble of flat, hierarchical, and text-segmentation models to capture both fine-grained and structural regions. (2 **Localize**) Generate region-level sub-HTMLs via two-stage layout-aware Set-of-Mark prompting, enriched with OCR and fine-grained attributes, enabling parallel and semantically rich generation. (3 **Assemble**) Merge sub-HTMLs by aligning absolute bounding boxes with predicted reading order, preserving layout fidelity and document flow. (4 **Refine**) Enhance visual quality via font-size fixing, background restoration, color correction, and an iterative VLM reflection loop for progressive alignment with the source document. Refer supplementary for more details.

5.4 Refine: Visual Fidelity Enhancements (4)

Font size encodes both visual and semantic cues—larger fonts often signal headings or emphasis. Preserving accurate sizing (4a) is thus essential for faithful rendering and interpretation. To ensure fidelity, font sizes are fixed at the leaf level within each sub-HTML: a binary search selects the maximum size that fits within the layout crop, preventing overflow and clutter while keeping text well-aligned. Visual quality is further improved through *background restoration* (4c), which reconstructs graphic elements and recovers global tints and shades. Finally, a vision-language model (VLM) drives a *reflection loop* (4b), comparing rendered HTML snapshots with the source image, identifying discrepancies, and iteratively refining the HTML until visual alignment is achieved.

Through this multi-stage design, REPLICA guarantees reproducible, layout-consistent, and visually faithful HTML reconstructions—forming the foundation for high-quality benchmark supervision and reliable evaluation. Refer to supplementary for more details on the design choices.

5.5 Implementation Details

Segment: We ensemble *Doclayout-YOLO-v10* [96] (flat layouts), *IndicDLP-YOLO-v10* [97] (hierarchical layouts), and *Hi-SAM* [98] (text segmentation). Doclayout-YOLO and Hi-SAM yield precise region boundaries, while IndicDLP-YOLO merges semantically related zones for coherent sub-HTMLs.

Localize: A two-stage Set-of-Mark prompting pipeline employs Hi-SAM paragraph- and line-level boxes, color-indexed for grounding. The VLM backbone is *Gemini 2.5 Pro* [2], augmented with *Google OCR* [99], *DocTr* [100] for word localization, and *TexTAR* [101] for font-style attributes. Robustness is ensured through automated retries and schema-level type validation.

Assemble: Sub-HTMLs are merged in reading order using absolute coordinates, with all visual elements re-stitched at original positions to preserve layout semantics.

Refine: Font sizes are optimized per paragraph via a binary search in `textFit.js` [102] using Hi-SAM boxes. Backgrounds are restored

by OpenCV inpainting over region masks, and a VLM-driven reflection loop (*Gemini 2.5 Pro*, `max.iterations=6`) iteratively aligns rendered and source layouts for final visual fidelity.

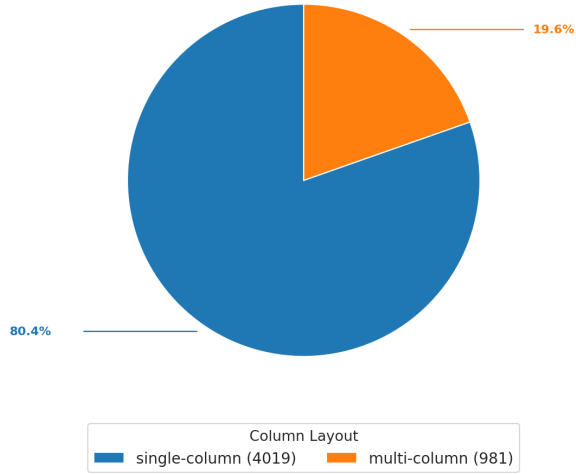
Additional implementation details and rationale behind design choices can be found in the supplementary.

6 Benchmark Construction

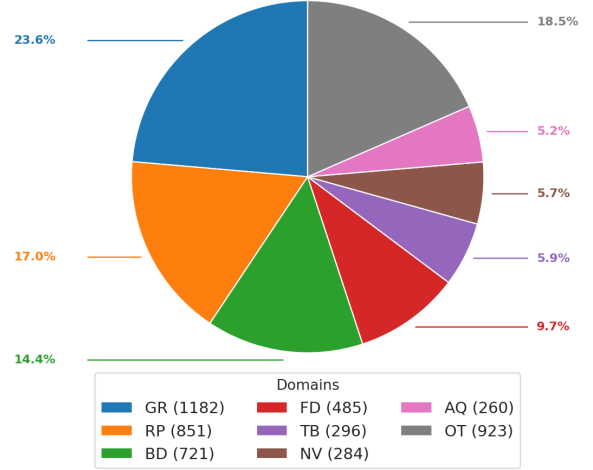
The benchmark comprises 5000 multilingual documents sourced from both web-scraped data and samples drawn from multiple public datasets—IndicDLP [97], RVL-CDIP [103], PubLayNet [104], DocLayNet [105], M⁶Doc [106], IDL [107], PRImA [108], D4LA [109], SROIE [110], FUNSD [111], and OJ4OCRMT [112] and CCpdf [105]. Collectively, these sources span 22 languages and over 20 distinct layout categories. This composition balances breadth and reproducibility, ensuring fair comparison across models. The distribution of documents across languages and domains is shown in fig. 5, and representative samples illustrating the diversity of VFDR-BENCH appear in fig. 6. Each document in the benchmark is first processed through our REPLICA pipeline to produce a high-fidelity reference HTML. The automatically generated outputs were largely accurate, requiring only light human verification. These were manually refined by trained annotators to guarantee accurate alignment between the rendered HTML and the source, with particular care for complex layouts and fine-grained details in standard documents. The strong synthetic supervision from REPLICA dramatically reduced manual editing compared to zero-shot *Gemini 2.5 Pro*, enabling scalable high-quality ground truth creation. Supplementary material includes the REPLICA baseline and qualitative examples of initial synthetic outputs and their refinements.

7 Evaluation Metrics

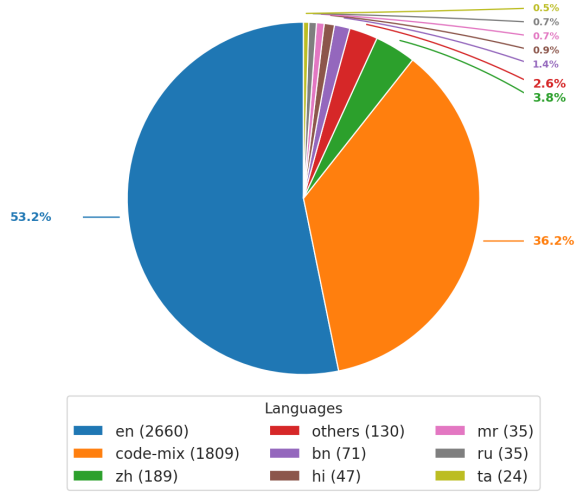
We evaluate Visually Faithful Document Reconstruction VFDR across four dimensions using a comprehensive suite of metrics. Refer to fig. 7



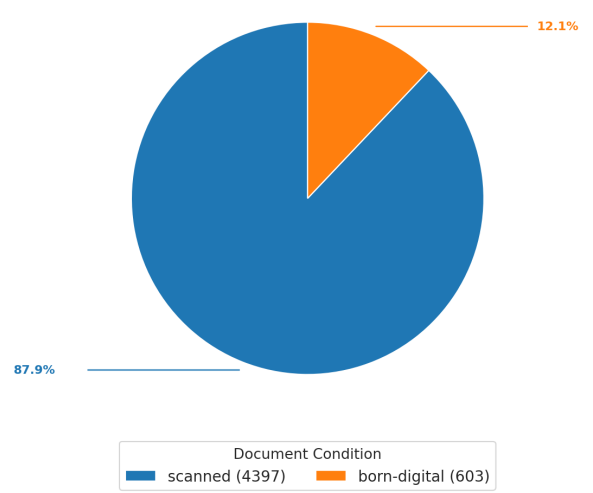
(a) Distribution Across Column Layout



(c) Distribution Across Domains



(b) Distribution Across Languages



(d) Distribution Across Document Condition

Fig. 5: Distribution of the VFDR-Bench dataset across four dimensions: (a) column layout (single- vs. multi-column), (b) languages (including code-mixed content), (c) document domains such as Government Regulatory Documents (GR), Financial Documents (FD), Research Papers (RP), Business Documents (BD), Textbooks (TB), Novels (NV), Assignments and Question Papers / Answer Keys (AQ), and Others (OT), and (d) document condition (scanned vs. born-digital). These charts summarize the structural, linguistic, and semantic diversity of VFDR-BENCH.

Text Extraction (TE): Predicted HTML text is compared using Word Recognition Rate (WRR) and Character Recognition Rate (CRR).

Logical Structure (LS): For *Normalized Tree Edit Distance* (NTED), both ground truth and predicted HTMLs are parsed into Document Object Model (DOM) trees and standardized

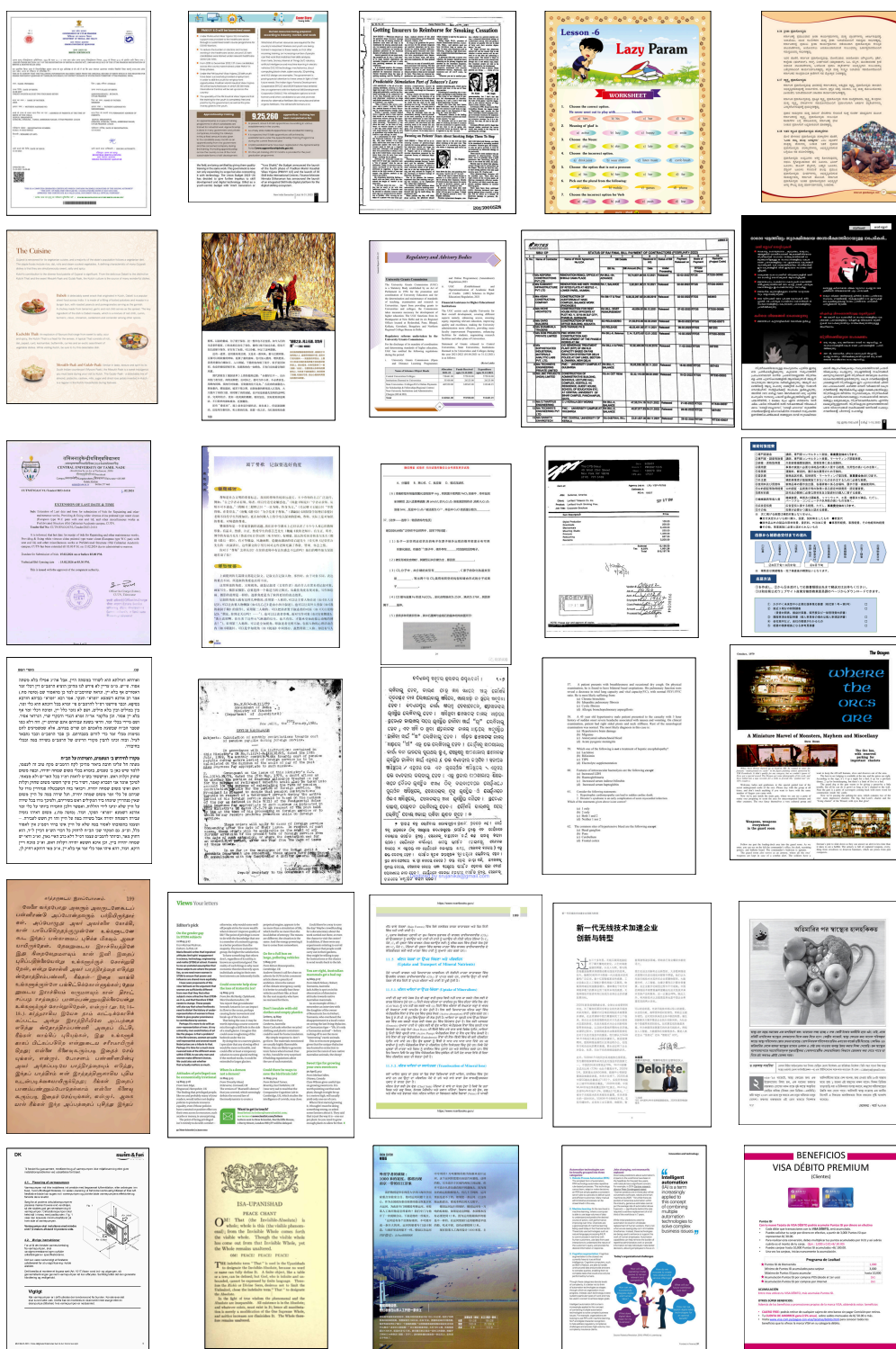


Fig. 6: Few samples from the VFDR-BENCH showcasing the diversity of the benchmark set.

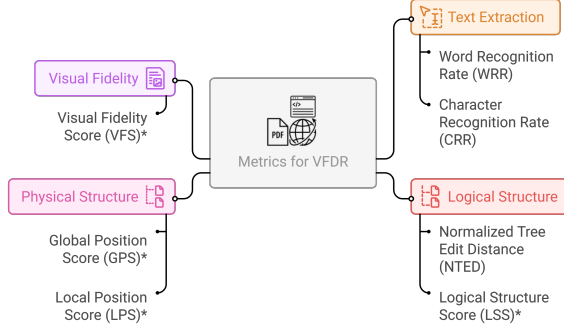


Fig. 7: Metrics for the VFDR Task. Overview of the evaluation suite used in VFDR-BENCH: Text Extraction (TE) measured via WRR and CRR; Logical Structure (LS) via NTED and the newly introduced LSS*; Physical Structure (PS) via the newly introduced GPS* and LPS*; and Visual Fidelity (VF) via the newly introduced VFS*. Metrics marked with * are proposed in this work.

for comparison. In the process, each HTML element is represented as a node with hierarchical parent-child relationships. This standardization ensures consistent computation of structural similarity. We measure structural preservation using Normalized Tree Edit Distance (NTED) [113] between the standardized representations.

Logical Structure Score (LSS): While the Normalized Tree Edit Distance (NTED) remains a widely used measure for structural similarity, its direct application to HTML-based document representations is limited. HTML is inherently flexible—multiple valid HTML representations can describe the same logical hierarchy—causing NTED to over-penalize semantically equivalent yet syntactically divergent structures. Moreover, since each method produces HTMLs in different formats and tag conventions, strict tree matching becomes unreliable. To address this, we introduce the Logical Structure Score (LSS), a VLM-as-a-judge metric that compares the DOM trees of predicted and reference HTMLs. LSS jointly evaluates (1) the hierarchical nesting of document sections, subsections, and elements, and (2) the correctness of semantic tags assigned to each node (e.g., `section`, `table`, `caption`). This allows robust, model-agnostic assessment of structural

fidelity without requiring exact syntactic alignment. Refer to fig. 8 for more details. The prompt can be found in the supplementary.

For Tables, we use the standard Tree Edit Distance Similarity (TEDS) [114]. For Formulae, we use Character Detection Matching (CDM) [115]. For Lists, we introduce *List Structure Penalty (LSP)*. This penalizes models that hard-code bullets as text instead of using proper `` tags:

$$Penalty = 1 - \frac{\#LI_{pred}}{\#LI_{gt}},$$

where $\#LI_{pred}$ and $\#LI_{gt}$ denote list items in predicted and ground-truth HTML. Lower penalty indicates better structural correctness.

Physical Structure (PS): For physical structure (PS), we introduce two new metrics: Global Position Score (GPS), which measures overlap between layout regions in the source document and the rendered HTML, and Local Position Score (LPS), which extends this to line-level alignment and indentation.

Global Position Score (GPS). The ground truth (GT) is obtained by running layout models on the original document image, while predictions are produced by running the same models on the rendered HTML output. The ground truth is subsequently refined through manual corrections. This yields layout-level bounding boxes (e.g., headers, tables, figures, paragraph blocks) for both the ground truth and predicted renderings. GPS measures structural alignment at the layout/block level. For each ground truth prediction bounding box pair (B_{gt}, B_{pred}) , we compute

$$GPS(B_{gt}, B_{pred}) = \frac{|B_{gt} \cap B_{pred}|}{|B_{gt}|},$$

where $|B|$ denotes the area of box B . The final GPS is the average across all detected layout blocks, with higher values indicating stronger preservation of global layout. Refer to fig. 9 for more details.

Local Position Score (LPS). The ground truth (GT) is obtained by applying Hi-SAM [98] to the original document image, while predictions are extracted by applying Hi-SAM to the rendered HTML, yielding line-level bounding boxes within each paragraph region. The ground truth is subsequently refined through manual corrections. LPS

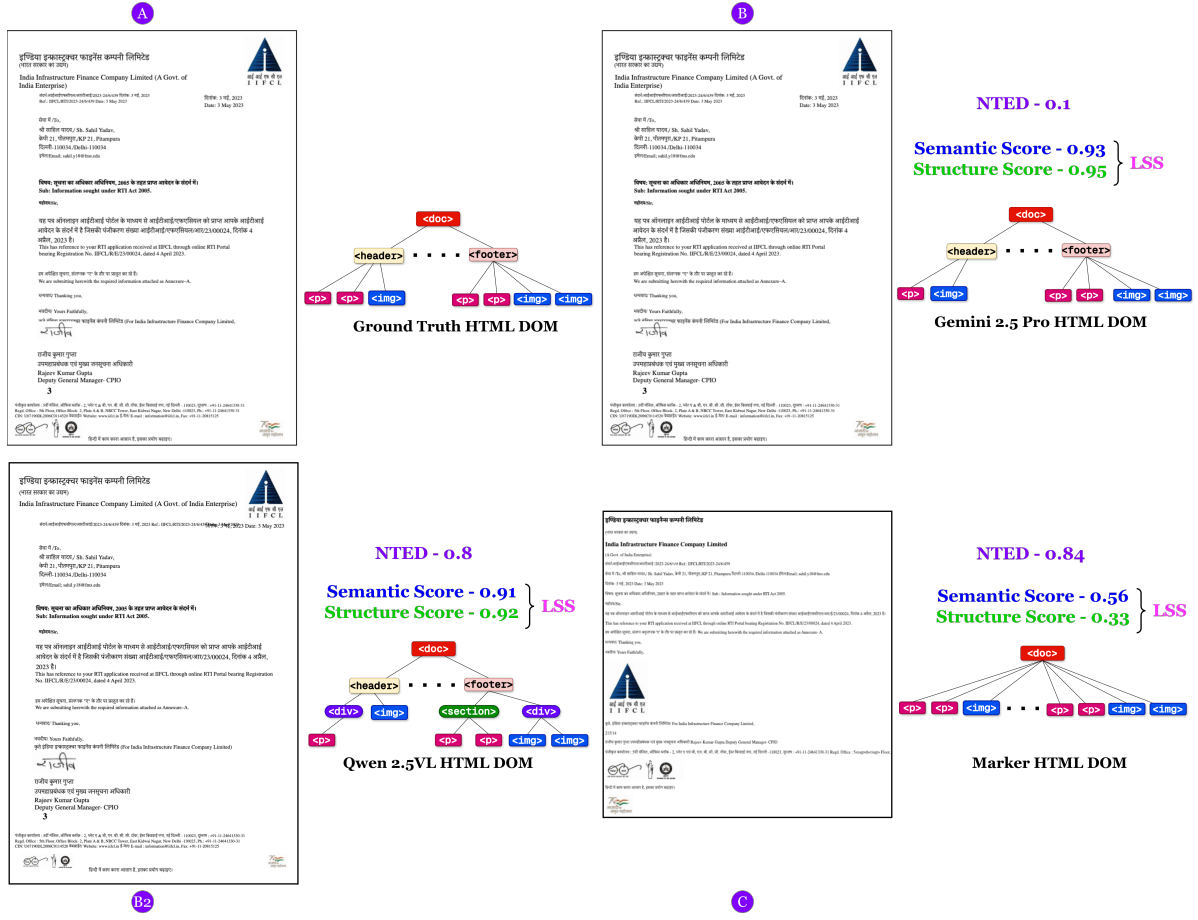


Fig. 8: Logical Structure Score (LSS). We illustrate the ground-truth DOM tree derived from FID-HTML in **A**. Examples **B** and **B2** represent alternative but semantically correct HTML structures; although both preserve the intended hierarchy, NTED penalizes **B2** for deviating too far from the reference FID-HTML. In contrast, LSS which uses VLM-as-a-judge does not penalize such valid structural variations, focusing instead on semantic and logical correctness. Example **C** demonstrates a clearly incorrect structure, and both NTED and LSS appropriately penalize it.

measures fine-grained structural alignment at the paragraph level. For each ground truth prediction line-level bounding box pair (B_{gt}, B_{pred}) , we compute

$$\text{LPS}(B_{gt}, B_{pred}) = \frac{|B_{gt} \cap B_{pred}|}{|B_{gt}|},$$

and report the average across all detected lines. Higher LPS reflects more faithful preservation of local alignment, indentation, and intra-paragraph structure. Refer to fig. 10 for more details.

Visual Fidelity (VF): Visual Fidelity Score (VFS) is computed using Gemini 2.5 Pro as a VLM-as-a-judge. The model is prompted (refer to supplementary for the prompt) with paired renderings of the prediction and the ground truth and instructed to score them across multiple dimensions of visual fidelity, including layout consistency, font attributes, styling, image placement, and overall appearance. This approach captures semantic and stylistic alignment beyond pixel-level similarity, ensuring that fidelity reflects perceptual quality rather than raw image overlap. Refer to fig. 11 for more details.

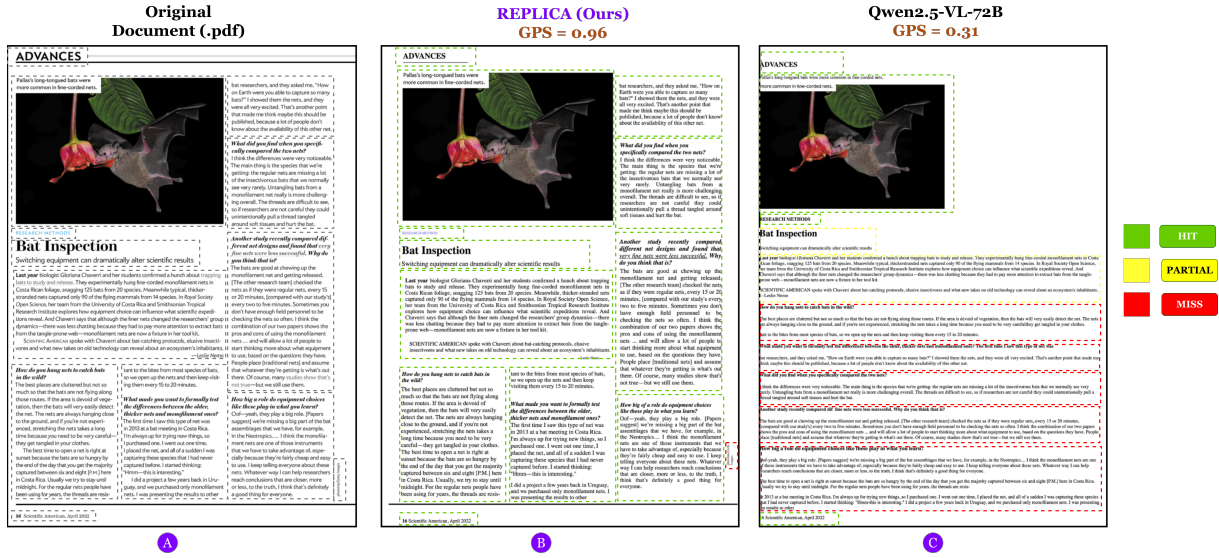


Fig. 9: Global Position Score (GPS). Ground-truth layout bounding boxes are overlaid on text regions in **A**. Example **B** is correctly aligned and therefore receives a high GPS, whereas in **C** most bounding boxes are misplaced due to text linearization, resulting in a low score.

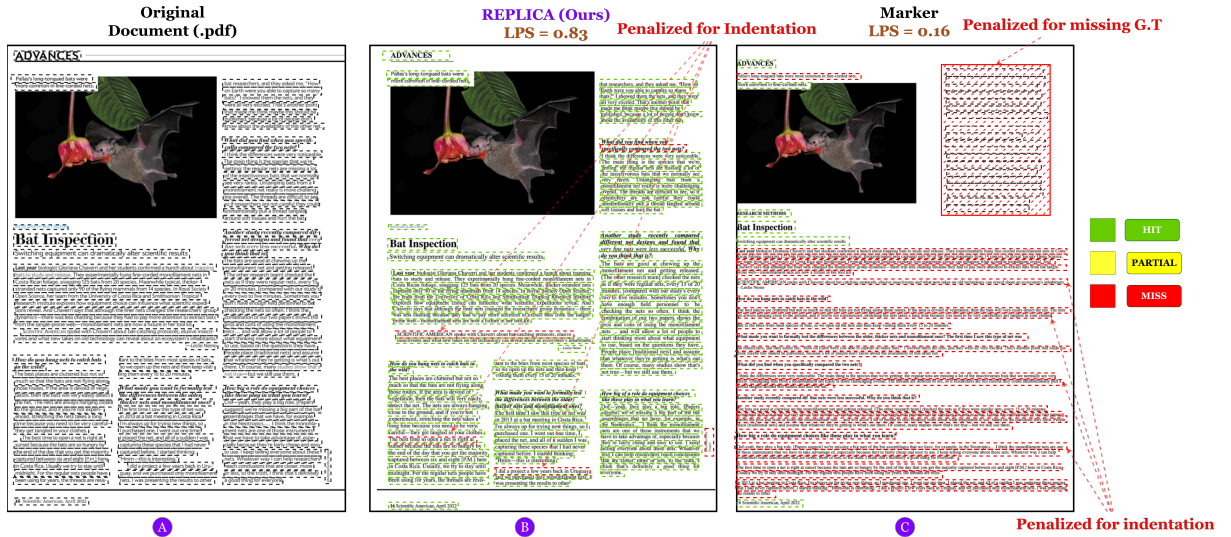


Fig. 10: Local Position Score (LPS). Ground-truth line-level bounding boxes are overlaid on text regions in **A**. Example **B** is correctly aligned and therefore receives a higher LPS, however, it is still penalized for missing indentation and alignment. Whereas in **C** most bounding boxes are misplaced due to text linearization, resulting in a low score.

8 Results

We evaluate a diverse set of document-to-HTML systems on VFDR-BENCH to establish reference performance and to characterize the challenges

of achieving visually faithful document rendering. All models are assessed using our unified metric suite spanning four fidelity dimensions: Text Extraction (TE), Logical Structure (LS), Physical Structure (PS), and Visual Fidelity (VF).

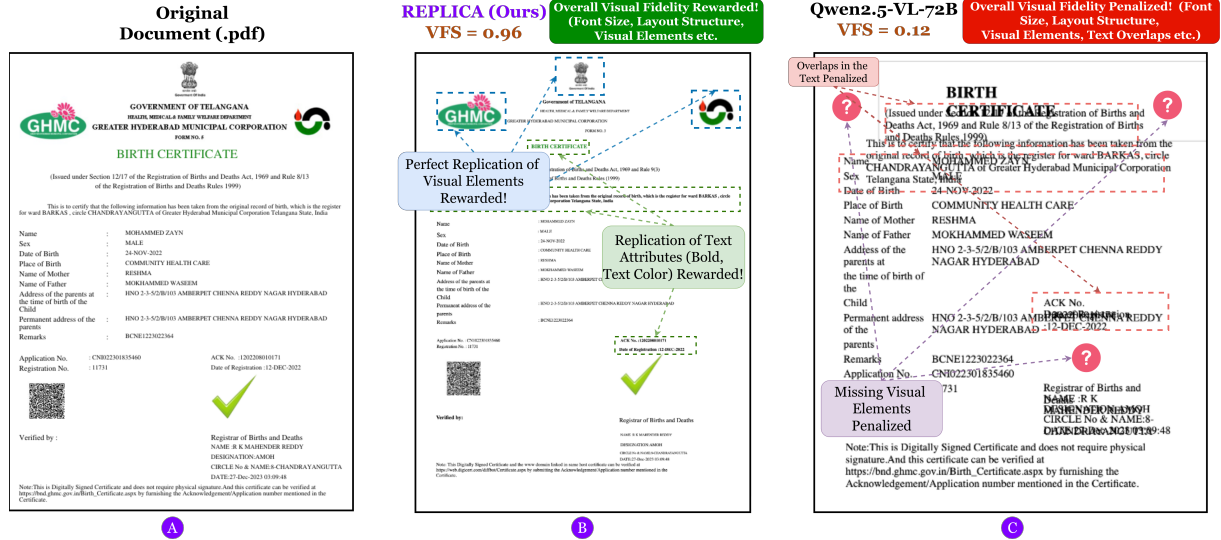


Fig. 11: Visual Fidelity Score (VFS). Visual fidelity is assessed using VLM-as-a-judge. Example **B** preserves text attributes, color, and visual elements, yielding a high score. In contrast, **C** contains overlaps and poorly reconstructed regions, leading to reduced fidelity.

Beyond overall scores, we introduce a fine-grained evaluation protocol that leverages benchmark metadata to surface deeper insights. For TE, we analyze performance across languages, reflecting the strong language dependence of text extraction quality. For LS, we examine results across document domains, since semantic structure varies significantly across categories. For PS, we highlight the impact of column layout, which distinguishes between naive linearization and true spatial preservation. Finally, for VF, we contrast born-digital and scanned documents to determine how rendering fidelity varies across these two modes.

Overall Performance. The overall score of the best performing method being mere 60% highlights the complexity of VFDR task and the gap in the existing methods (table 3). Traditional parsers and pipeline-based converters, while strong in text extraction, perform poorly in preserving physical layout. Open-source VLMs achieve better stylistic fidelity but still struggle with accurate positioning. Closed-source VLMs show more consistent strength across fidelity dimensions, yet they also remain far from reliably capturing full document structure. Screenshot2HTML-style models, despite promising results on web-oriented data, generalize poorly to complex document settings. The closed-source

VLMs perform better at generating fine-grained semantic structures for logical reconstruction. Expert VLMs—such as logics-parsing trained to generate Qwen-HTML—do retain spatial relationships to some extent, but their robustness is still limited. Refer fig. 12a for more details.

Performance across languages (TE). We evaluate text extraction quality across all languages represented in the benchmark (table 14). Marker and Gemini 2.5 Pro exhibit strong and stable performance across the multilingual spectrum. Specialist VLMs—such as logics-parsing or olmOCR—excel in English and often outperform general-purpose VLMs on English content, but their performance drops sharply on other languages. In Chinese, the Qwen family leads, with Qwen2.5-VL and logics-parsing achieving the highest scores. Screenshot2HTML-style systems offer limited multilingual support and perform poorly beyond English. Notably, both Marker and GPT-5 remain robust even under significant code-mixing within documents. Refer fig. 12b for more details.

Performance across Domains (LS). GPT-5, logics-parsing, and Marker demonstrate strong and stable performance on logical structure across most document domains (table 13). Marker benefits from its pipeline design—extracting layout elements first—which helps it anchor its structural

Table 3: Overall comparison of document-to-HTML methods across four evaluation dimensions. Text Extraction (TE: Word recognition rate (WRR), Character recognition rate (CRR)), Logical Structure (LS: Normalized Tree Edit Distance (NTED)), Physical Structure (PS: Global Position Score (GPS), Local Position Score (LPS)), and Visual Fidelity (VF: Visual Fidelity Score (VFS) via VLM-as-a-judge) are reported. The details about overall score computation can be found in the supplementary. Baseline methods are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML

| Method | TE | | LS | | PS | | VF | Overall |
|-------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | WRR (↑) | CRR (↑) | NTED (↓) | LSS (↑) | LPS (↑) | GPS (↑) | VFS (↑) | OS (↑) |
| Marker [3] | 0.65 | 0.80 | <u>0.80</u> | 0.68 | 0.25 | 0.35 | 0.76 | 0.62 |
| Docling [26] | 0.38 | 0.47 | <u>0.86</u> | 0.63 | 0.12 | 0.26 | 0.48 | 0.43 |
| RolmOCR [116] | 0.54 | 0.70 | 0.95 | 0.23 | 0.26 | 0.07 | 0.47 | 0.37 |
| smolDocling [13] | 0.34 | 0.40 | 0.90 | 0.42 | 0.13 | 0.24 | 0.45 | 0.36 |
| olmOCR-7B [66] | 0.54 | 0.67 | 0.93 | 0.26 | <u>0.28</u> | 0.10 | 0.28 | 0.34 |
| Nanonets-OCR-s [117] | 0.41 | 0.54 | 0.95 | 0.20 | 0.25 | 0.10 | 0.48 | 0.33 |
| GOT-OCR-2.0 [27] | 0.42 | 0.48 | 0.93 | 0.21 | 0.17 | 0.25 | 0.60 | 0.37 |
| OCRFlux-3B [118] | 0.45 | 0.57 | 0.88 | 0.51 | 0.25 | 0.25 | 0.37 | 0.41 |
| Logics-Parsing [119] | <u>0.64</u> | <u>0.77</u> | 0.84 | 0.57 | 0.32 | <u>0.42</u> | 0.65 | <u>0.57</u> |
| Qwen2.5-VL-7B-Instruct [43] | 0.50 | 0.61 | <u>0.80</u> | 0.60 | 0.14 | 0.24 | 0.37 | 0.43 |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.47 | 0.54 | 0.78 | 0.44 | 0.15 | 0.33 | 0.29 | 0.37 |
| gemma-3-27b-it [28] | 0.57 | 0.62 | 0.85 | 0.59 | 0.16 | 0.25 | 0.32 | 0.43 |
| InternVL3-14B [46] | 0.28 | 0.46 | 0.83 | <u>0.67</u> | 0.22 | 0.29 | 0.27 | 0.39 |
| Gemini-2.5-Pro [120] | 0.63 | 0.69 | 0.81 | <u>0.67</u> | 0.19 | 0.31 | 0.44 | 0.50 |
| GPT-5 [121] | <u>0.64</u> | <u>0.77</u> | <u>0.80</u> | 0.68 | 0.26 | 0.44 | <u>0.73</u> | 0.62 |
| WebCoder-1.3B [79] | 0.33 | 0.42 | 0.82 | 0.40 | 0.20 | 0.27 | 0.34 | 0.34 |
| WebSight-VLM-7B [29] | 0.14 | 0.42 | 0.85 | 0.57 | 0.19 | 0.29 | 0.16 | 0.31 |

Table 4: Text Extraction (TE): Comparison of language-wise scores (CRR, WRR) across methods. The table shows model performance across languages. Baseline methods are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML . Languages are abbreviated as: en = English, hi = Hindi, bn = Bengali, ta = Tamil, mr = Marathi, zh = Chinese, ru = Russian, mixed = Code Mixed, oth = Others. Certain values are excluded when error rates become excessively large, typically arising from languages unfamiliar to the model.

| Model | en | hi | bn | ta | mr | zh | mixed | oth |
|-------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Docling [26] | 0.58 | 0.03 | 0.01 | 0.02 | 0.01 | 0.01 | 0.30 | 0.09 |
| Marker [3] | <u>0.82</u> | <u>0.81</u> | 0.64 | <u>0.79</u> | <u>0.78</u> | 0.21 | 0.65 | 0.73 |
| Nanonets-OCR-s [117] | 0.66 | 0.31 | 0.30 | – | 0.18 | 0.37 | 0.27 | 0.14 |
| RolmOCR [116] | 0.77 | 0.66 | 0.49 | – | 0.47 | 0.34 | 0.48 | 0.34 |
| olmOCR-7B [66] | 0.77 | 0.71 | 0.29 | – | 0.63 | 0.33 | 0.45 | 0.20 |
| smolDocling [13] | 0.55 | 0.00 | 0.00 | 0.00 | – | – | 0.22 | 0.19 |
| OCRFlux-3B [118] | 0.69 | – | 0.33 | – | 0.64 | 0.31 | 0.35 | 0.11 |
| GOT-OCR-2.0 [27] | 0.69 | – | – | 0.03 | – | – | 0.31 | – |
| Logics-Parsing [119] | <u>0.82</u> | 0.60 | 0.37 | 0.15 | 0.66 | <u>0.50</u> | <u>0.61</u> | 0.35 |
| Gemini-2.5-Pro [120] | 0.74 | 0.88 | <u>0.53</u> | 0.84 | 0.82 | 0.31 | 0.58 | <u>0.68</u> |
| GPT-5 [121] | 0.84 | 0.47 | 0.19 | 0.12 | 0.57 | 0.28 | <u>0.61</u> | 0.61 |
| Gemma-3-27b-it [28] | 0.68 | 0.56 | 0.27 | 0.14 | 0.72 | 0.23 | 0.53 | 0.55 |
| Qwen2.5-VL-7B-Instruct [43] | 0.67 | 0.66 | 0.42 | 0.38 | 0.65 | 0.55 | 0.42 | 0.25 |
| InternVL3-14B [46] | 0.48 | – | – | – | – | 0.41 | 0.33 | – |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.63 | 0.57 | 0.26 | 0.34 | 0.30 | 0.49 | 0.36 | 0.26 |
| WebCoder-1.3B [79] | 0.44 | – | – | – | – | – | 0.18 | – |
| WebSight-VLM-7B [29] | 0.43 | – | – | – | – | – | 0.20 | – |

predictions to the underlying layout analysis. In contrast, GPT-5 and logics-parsing, being more flexible VLM-based generators, produce detailed

and semantically rich HTML structures directly. Across all methods, however, certain categories

Table 5: Logical Structure (LS): Comparison of domain-wise scores (LSS) across methods.

Baseline methods are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML . Document categories are abbreviated as: GR = Government Regulatory, FD = Financial, BD = Business, HC = Healthcare, NV = Novels, LG = Legal, AQ = Assignments/Question Papers, TB = Textbooks, RP = Research Papers, MN = Manuals, BR = Brochures, FM = Forms, RS = Resumes, RI = Receipts/Invoices, ID = ID Documents/Certificates, LN = Lecture Notes, OT = Others

| Model | GR | FD | BD | HC | NV | LG | AQ | TB | RP | MN | BR | FM | OT | RS | RI | ID | LN |
|-------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Marker [3] | 0.72 | 0.66 | 0.70 | 0.72 | 0.72 | 0.69 | 0.61 | 0.60 | 0.67 | 0.62 | 0.77 | 0.57 | 0.71 | 0.60 | <u>0.65</u> | 0.50 | 0.80 |
| Docling [26] | 0.61 | 0.65 | <u>0.69</u> | 0.68 | 0.61 | 0.73 | 0.52 | 0.58 | 0.63 | 0.61 | 0.53 | 0.56 | 0.65 | <u>0.68</u> | <u>0.65</u> | 0.50 | 0.68 |
| RolmOCR [116] | 0.28 | 0.28 | 0.28 | 0.08 | 0.29 | 0.19 | 0.24 | 0.11 | 0.13 | 0.18 | 0.33 | 0.31 | 0.26 | 0.28 | <u>0.65</u> | 0.37 | 0.12 |
| smolDocling [13] | 0.37 | 0.31 | 0.46 | 0.37 | 0.58 | 0.51 | 0.31 | 0.40 | 0.46 | 0.50 | 0.42 | 0.33 | 0.53 | 0.33 | <u>0.50</u> | 0.43 | 0.52 |
| olmOCR-7B [66] | 0.26 | 0.36 | 0.28 | 0.13 | 0.22 | 0.23 | 0.34 | 0.28 | 0.17 | 0.22 | 0.48 | 0.28 | 0.45 | 0.28 | 0.10 | 0.18 | 0.50 |
| Nanonets-OCR-s [117] | 0.27 | 0.26 | 0.20 | 0.13 | 0.25 | 0.12 | 0.17 | 0.15 | 0.14 | 0.11 | 0.26 | 0.32 | 0.21 | 0.10 | 0.10 | 0.05 | 0.12 |
| OCRFlux-3B [118] | 0.51 | 0.51 | 0.55 | 0.52 | 0.46 | 0.39 | 0.54 | 0.45 | 0.48 | 0.59 | 0.52 | 0.52 | 0.52 | 0.68 | 0.25 | 0.25 | 0.72 |
| Logics-Parsing [119] | 0.57 | 0.62 | 0.57 | 0.46 | 0.69 | 0.52 | 0.50 | 0.46 | 0.55 | 0.54 | 0.66 | 0.62 | 0.62 | 0.40 | 0.50 | 0.50 | 0.62 |
| GOT-OCR-2.0 [27] | 0.20 | 0.18 | 0.19 | 0.20 | 0.30 | 0.21 | 0.19 | 0.21 | 0.21 | 0.18 | 0.22 | 0.20 | 0.33 | 0.22 | 0.25 | 0.20 | 0.20 |
| Qwen2.5-VL-7B-Instruct [43] | 0.57 | 0.53 | 0.57 | 0.61 | 0.67 | 0.64 | 0.63 | 0.64 | 0.65 | 0.55 | <u>0.68</u> | 0.55 | 0.65 | 0.77 | 0.85 | 0.65 | <u>0.75</u> |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.41 | 0.34 | 0.43 | 0.36 | 0.56 | 0.52 | 0.42 | 0.49 | 0.51 | 0.39 | 0.46 | 0.35 | 0.56 | 0.27 | 0.30 | 0.25 | 0.70 |
| gemma-3-27b-it [28] | 0.57 | 0.60 | 0.60 | 0.67 | 0.61 | 0.53 | 0.61 | 0.62 | 0.59 | 0.58 | 0.62 | 0.42 | 0.59 | 0.58 | 0.65 | <u>0.65</u> | 0.45 |
| InternVL3-14B [46] | 0.65 | <u>0.71</u> | 0.67 | 0.55 | <u>0.71</u> | 0.61 | 0.69 | 0.68 | 0.67 | <u>0.66</u> | <u>0.68</u> | 0.71 | <u>0.68</u> | <u>0.65</u> | <u>0.65</u> | <u>0.45</u> | <u>0.75</u> |
| Gemini-2.5-Pro [120] | <u>0.69</u> | 0.68 | 0.66 | 0.69 | 0.68 | 0.62 | 0.63 | <u>0.69</u> | 0.64 | 0.65 | 0.63 | 0.76 | <u>0.67</u> | 0.55 | <u>0.65</u> | 0.42 | 0.60 |
| GPT-5 [121] | 0.68 | 0.73 | <u>0.69</u> | <u>0.70</u> | 0.72 | <u>0.68</u> | <u>0.67</u> | 0.72 | <u>0.66</u> | 0.67 | 0.64 | 0.49 | <u>0.67</u> | 0.58 | 0.50 | 0.68 | 0.58 |
| WebCoder-1.3B [79] | 0.16 | 0.15 | 0.15 | 0.04 | 0.21 | 0.11 | 0.15 | 0.09 | 0.09 | 0.07 | 0.18 | 0.18 | 0.20 | 0.09 | 0.07 | 0.03 | 0.12 |
| WebSight-VLM-7B [29] | 0.17 | 0.14 | 0.17 | 0.09 | 0.19 | 0.12 | 0.19 | 0.14 | 0.11 | 0.04 | 0.15 | 0.16 | 0.17 | 0.13 | 0.09 | 0.00 | 0.07 |

Table 6: Physical Structure (PS): Comparison of column-layout wise scores (GPS, LPS) across methods. Baseline methods are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML . Document conditions are abbreviated as: SC = Single-column, MC = Multi-column

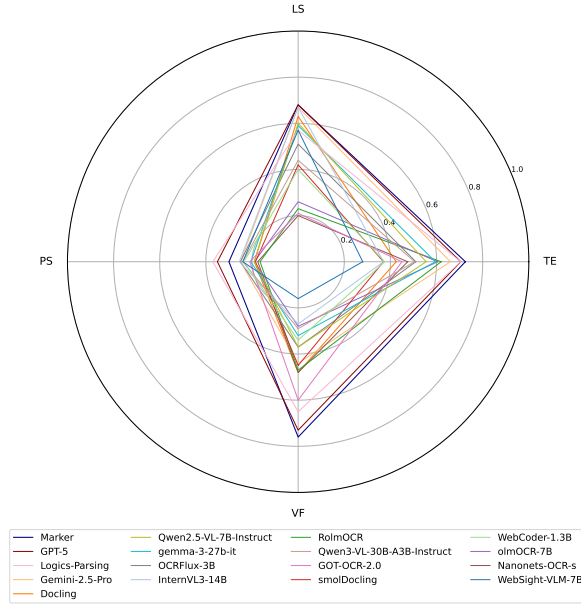
| Model | SC | MC |
|-------------------------------|-------------|-------------|
| Marker [3] | 0.29 | 0.25 |
| Docling [26] | 0.20 | 0.17 |
| RolmOCR [116] | 0.15 | 0.12 |
| smolDocling [13] | 0.19 | 0.19 |
| olmOCR-7B [66] | 0.17 | 0.14 |
| Nanonets-OCR-s [117] | 0.16 | 0.13 |
| OCRFlux-3B [118] | 0.17 | 0.11 |
| Logics-Parsing [119] | 0.37 | <u>0.34</u> |
| GOT-OCR-2.0 [27] | 0.21 | 0.21 |
| Qwen2.5-VL-7B-Instruct [43] | 0.32 | 0.28 |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.34 | 0.31 |
| gemma-3-27b-it [28] | 0.21 | 0.18 |
| InternVL3-14B [46] | 0.26 | 0.24 |
| Gemini-2.5-Pro [120] | 0.24 | 0.27 |
| GPT-5 [121] | <u>0.36</u> | 0.39 |
| WebCoder-1.3B [79] | 0.07 | 0.11 |
| WebSight-VLM-7B [29] | 0.10 | 0.13 |

Table 7: Visual Fidelity (VF): Comparison of document condition wise scores (VFS) across methods. Baseline methods are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML . Document conditions are abbreviated as: SC = Scanned, BD = Born Digital

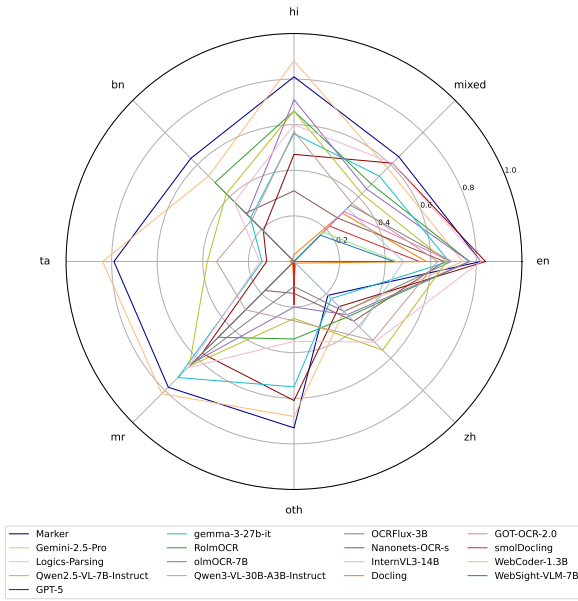
| Model | SC | BD |
|-------------------------------|-------------|-------------|
| Marker [3] | 0.76 | 0.75 |
| Docling [26] | 0.46 | <u>0.68</u> |
| RolmOCR [116] | 0.47 | 0.46 |
| smolDocling [13] | 0.44 | 0.54 |
| olmOCR-7B [66] | 0.27 | 0.39 |
| Nanonets-OCR-s [117] | 0.48 | 0.47 |
| OCRFlux-3B [118] | 0.36 | 0.48 |
| Logics-Parsing [119] | 0.65 | 0.61 |
| GOT-OCR-2.0 [27] | 0.58 | 0.75 |
| Qwen2.5-VL-7B-Instruct [43] | 0.35 | 0.50 |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.27 | 0.43 |
| gemma-3-27b-it [28] | 0.31 | 0.39 |
| InternVL3-14B [46] | 0.25 | 0.48 |
| Gemini-2.5-Pro [120] | 0.42 | 0.55 |
| GPT-5 [121] | <u>0.72</u> | 0.75 |
| WebCoder-1.3B [79] | 0.22 | 0.38 |
| WebSight-VLM-7B [29] | 0.24 | 0.40 |

consistently remain challenging: Forms, ID Documents, and Receipts/Invoices yield notably lower

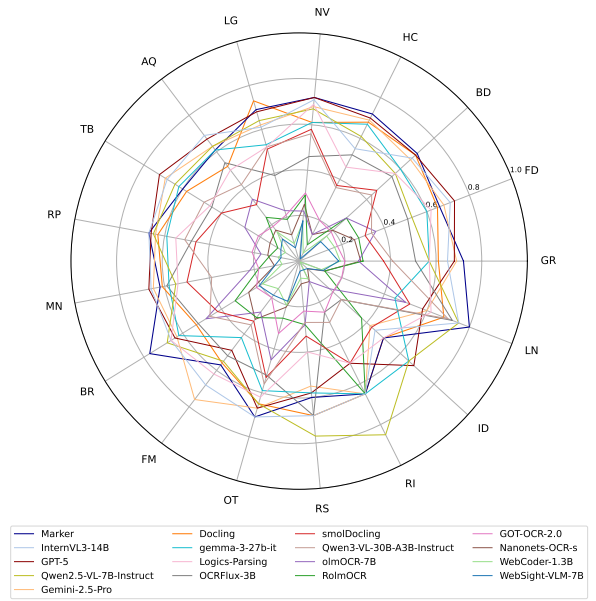
scores due to their dense, irregular, and highly structured layouts. Refer fig. 12c for more details.



(a) Performance across 4 dimensions: Text Extraction (TE), Logical Structure (LS), Physical Structure (PS) and Visual Fidelity (VF).



(b) Performance across languages.



(c) Performance across document domains.

Fig. 12: Comprehensive comparison of model performance. Radar plots provide complementary evaluation perspectives: (a) overall performance across 4 crucial dimensions of VFDR, (b) text extraction robustness across multiple languages, (c) logical structure generalization across diverse document domains.

Performance across Column Layouts (PS). All methods show noticeable difficulty with absolute positioning (table 15). The position score is

generally higher on single-column documents than

on multi-column layouts. The Qwen family stands out slightly in multi-column settings, aided by its position-aware Qwen-HTML output format, with logics-parsing achieving the strongest results among them. GPT-5 being closed-source, demonstrates strong visual understanding and is able to reconstruct multi-column layouts with reasonable reliability. It is also important to note that the metrics in this setting are particularly strict, as they expect near-perfect reconstruction and the Hi-SAM bounding boxes used as reference are typically very tight around the text.

Performance across document conditions (VF). Across all methods, fidelity is consistently higher on born-digital documents than on scanned ones (table 16). Since the evaluation compares the original document to the rendered output, the lack of background restoration in current systems naturally reduces visual fidelity for scanned inputs. GOT2.0 and Docling perform particularly well on born-digital documents but show a substantial drop of roughly 20% when applied to scanned material. For visual fidelity, Marker achieves the highest score (76%) due to its visual-stitching pipeline. GPT-5 follows closely—within 4%—despite not relying on any stitching mechanism, underscoring its strong visual understanding and style-consistent generation.

9 Discussion

VLM sensitivity to prompts and adaptations. VLMs are sensitive to prompts and our standardized text prompts may impact model evaluations. Beyond zero-shot prompting, there exist a variety of adaptation methods—specific procedures for invoking a model—such as chain-of-thoughts [122] or auto-prompt [123] that can enhance the performance of the models. We leave measuring the performance of VLMs under other adaptations as future work.

Why HTML for VFDR? Achieving strong performance on VFDR-BENCH requires VLMs to reason not only over visual content but also over formal markup languages (e.g., HTML or LaTeX). While LaTeX could, in principle, serve as an intermediate rendering format for VFDR, we adopt HTML due to its web-native nature and its substantially stronger support within modern LLM ecosystems. Consequently, HTML provides better

compatibility, richer tooling, and greater downstream utility for high-fidelity document reconstruction.

10 Limitations

Although VFDR-BENCH offers a unified framework for visually faithful document rendering, it has inherent limitations. The current release focuses on single-page documents, and thus does not evaluate cross-page consistency or long-range layout dependencies. Future versions will extend to multi-page corpora to capture inter-page continuity. Reference HTMLs, generated by REPLICA, emphasize layout and visual accuracy over precise font classification, leaving fine-grained typographic recognition as an open research challenge. While the benchmark spans 22 languages, it remains skewed toward Latin and high-resource scripts, reflecting biases in publicly available corpora. Finally, the reliance on vision-language models for LSS and VFS introduces potential perceptual bias; subsequent iterations will integrate human preference studies to mitigate this.

11 Conclusion

VFDR-BENCH establishes a foundation for the study of *Visually Faithful Document Reconstruction* (VFDR) by introducing a large-scale benchmark, a standardized metric suite, and a reproducible evaluation framework. By unifying textual, structural, spatial, and perceptual dimensions, it exposes limitations in existing OCR and document-generation pipelines, providing actionable insights for the development of more holistic document understanding models.

While modern vision-language models demonstrate strong semantic reconstruction, maintaining coherent layout and style across multilingual and visually rich documents remains a significant challenge. VFDR-BENCH is intended to catalyze progress toward this goal by encouraging layout-aware, multilingual, and visually grounded document modeling. Beyond its research scope, visually faithful rendering is central to accessibility, digital preservation, and open educational resources. By fostering accurate, inclusive, and style-preserving digital representations, the

benchmark advances equitable access to information and contributes to a more representative and inclusive digital ecosystem.

12 REPLICA Implementation Details

REPLICA adopts a modular, layout-aware approach that emphasizes both precision and flexibility. Instead of processing the document in a single-shot, it employs a divide-and-conquer strategy centered on four key questions:

1. *How to divide the document into meaningful components?* (**Segment**) ①
2. *How to convert each component into equivalent HTML?* (**Localize**) ②
3. *How to assemble components into a coherent structure?* (**Assemble**) ③
4. *How to refine structure for visual and structural fidelity?* (**Refine**) ④

12.1 Segment

12.1.1 Canonical Layout Classes

Layout Classes after Segment

text, list, title, header, sub-header, picture, table, formula, footer, page-number

Layout Classes after Semantic Tagging from Localize Stage 2

header, sub-header, table-of-contents, references, contact-info, placeholder-text, footer, sidebar, title, heading, paragraph, caption, page-number, dateline, table, form, ordered-list, unordered-list, figure, picture, logo, chart, formula, code-block, handwriting, signature, form, question, options

12.1.2 Hybrid Layout Prediction

Document layout models: These can be divided into (1) *atomic layout predictors*, which classify coarse regions (text, table, figure, title), and (2) *hierarchical layout predictors*, which identify finer structures (headers, footers, lists, placeholder text). Each has complementary strengths for VLM-based HTML generation. Atomic models such as DocLayout-YOLO [96] operate with fewer classes, yielding higher confidence and stronger text coverage, but they lack semantic grouping, leading to flat HTML. Hierarchical models such as IndicDLP-DocLayoutYOLO [97] enable richer clustering of related blocks, producing more semantically meaningful HTML, but their larger class space increases the risk of misclassification or missing text.

Ensemble strategy: To leverage both, we ensemble DocLayout-YOLO and IndicDLP-DocLayoutYOLO—where the former provides high-confidence atomic predictions for coverage and the latter contributes higher-level grouping for semantics. We further augment this ensemble with Hi-SAM [98] text detections, which improve overall text coverage.

12.2 Localize

12.2.1 Stage 1: Raw sub-HTML generation

Goal. The goal of this stage is to ensure accurate positioning and text recovery. Achieving reliable positional grounding is challenging: heuristic OCR matching or passing bounding boxes directly as prompts to VLMs are fragile. Instead, we adapt the Set-of-Mark (SoM) prompting technique [124], which assigns unique marks to image regions to guide grounding.

Color-coded SoM Prompting for Document Reconstruction. Standard SoM with alphanumeric overlays fails on text-rich documents, as indices leak into generated HTML. We address this by assigning

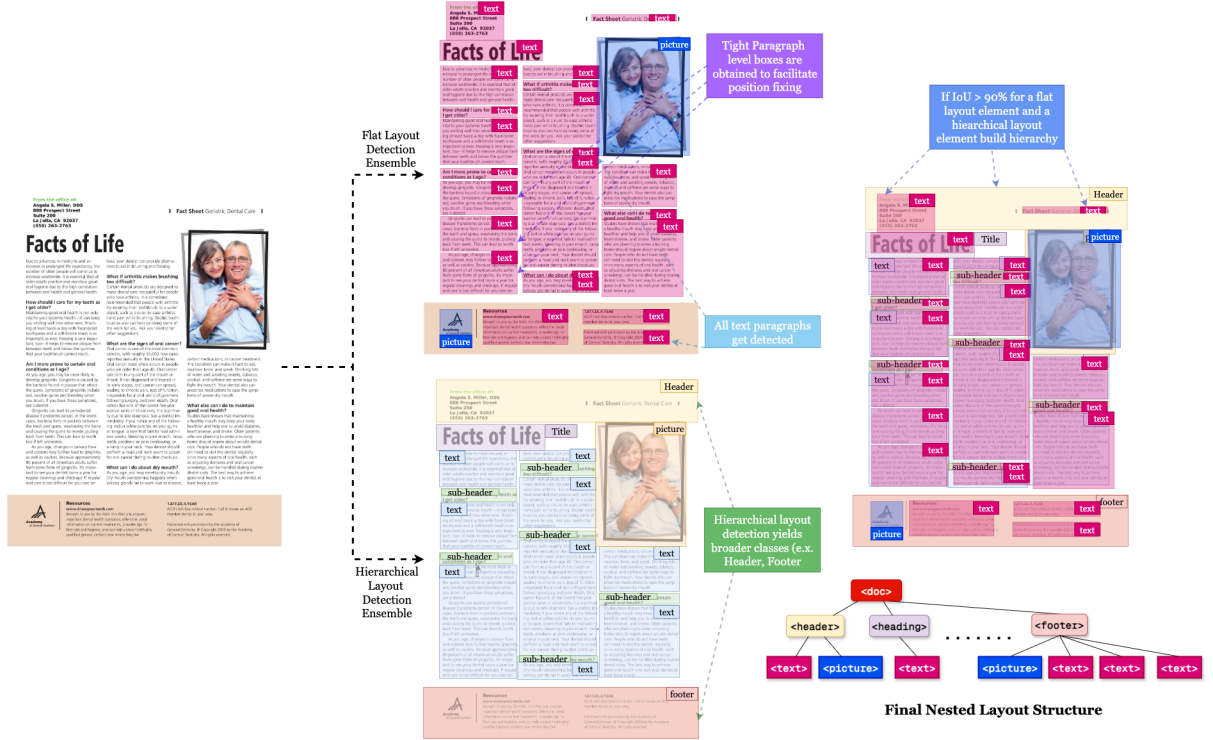


Fig. 13: Hybrid Layout Prediction via Model Ensembling. Flat layout detectors (e.g., DocLayout-YOLO) provide dense paragraph-level boxes with strong text coverage but limited semantics. Hierarchical layout detectors (e.g., IndicDLP-DocLayoutYOLO) produce broader semantic classes such as *Header* and *Footer*, enabling structured grouping but with higher risk of missed detections. By ensembling both predictions and augmenting with Hi-SAM [98] paragraph detections, we obtain a complementary representation: flat predictions ensure coverage, hierarchical predictions supply semantic hierarchy, and IoU-based alignment integrates them into a final nested layout tree.

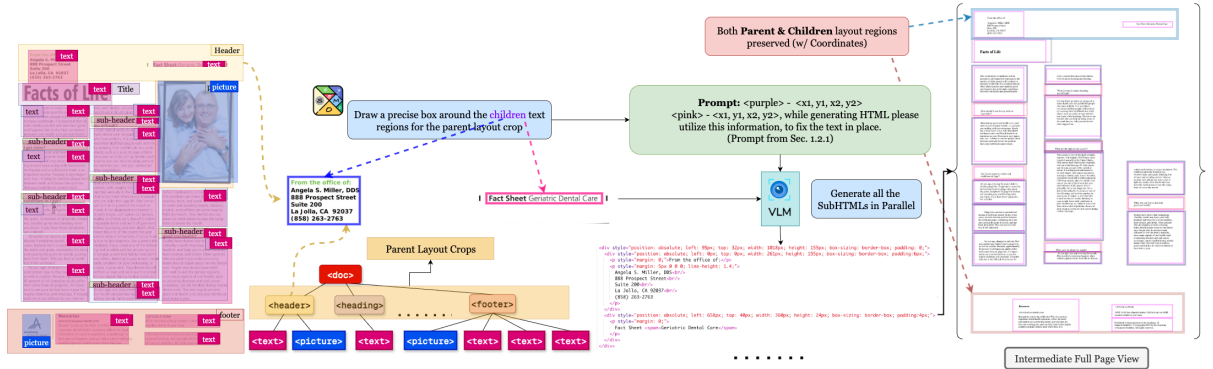


Fig. 14: Localize Stage: 1 Parent and child layout regions are localized and annotated with distinct color-coded bounding boxes on the cropped image. These color-coordinate mappings are passed to the VLM, which generates sub-HTMls in parallel while preserving structural hierarchy and positional alignment. (An intermediate full-page view is provided to illustrate the output from this stage)

each bounding box a distinct color (absent from the source document) and providing a color-coordinate mapping in the prompt. This forces the VLM to ground text regions by matching colors to coordinates, leading to accurate bounding-box assignments and faithful OCR integration into HTML. Our color-coded SoM adaptation provides precise grounding for document-to-HTML conversion while avoiding textual interference from index overlays, making it a key enabler for accurate localization.

Stage 1 Prompt

Task:

Convert the given image into HTML.

The generated HTML must:

1. Preserve structure of the original document.
2. Precisely position each content block according to its bounding box coordinates.
3. Group related text segments that belong to the same logical point into a single `<p>` tag, and insert `
` tags wherever the OCR text contains `\n` to preserve line breaks.
4. Strictly use the OCR text as given, without corrections or modifications.

Bounding Box Format

All bounding boxes are provided in the format:

`x1, y1, x2, y2`

where:

- `(x1, y1)` = top-left corner of the block
- `(x2, y2)` = bottom-right corner of the block

HTML Tagging Rules

1. Text Paragraphs

`<p data-bbox='x1 y1 x2 y2'>...</p>`

- Use `data-bbox` with the exact bounding box values.
- Keep all OCR text exactly as extracted (no corrections).
- Preserve line breaks with `
`.
- Maintain alignment (left, right, center).

2. Images (including signatures, logos, figures, etc.)

``

- Do not put text inside ``.
- Do not wrap `` in other tags.

Mappings Provided

- Image size: `{{image_size}}`
- Page Layout: `{{parent_layout}}`
- Child Picture Layout Bounding Boxes (format : (R,G,B):[x1,y1,x2,y2]) : `{{child_layout_picture}}`
- Text Bounding Boxes (format : (R,G,B):[x1,y1,x2,y2]):

`{{JSON}}`

Extracted OCR Text

`{{OCR}}`

Your output must be correct HTML with correct data-bbox values.
Do not generate text that was not present in OCR.

12.2.2 Stage 2: Semantic sub-HTML generation

Goal. The goal of this stage is to recover richer HTML semantics by assigning appropriate tags and capturing hierarchical logical structure. Beyond plain text, this includes inserting semantic tags (e.g., headings, lists, tables), recovering text attributes (bold, italics, underline, strikethrough), preserving font colors and background colors, and refining overall HTML semantics for faithful document representation.

Color Detection. Preserving textual attributes such as color, bold, italic, and underline is critical for both semantics and visual fidelity. We implement a heuristic pipeline to detect word-level colors from Hi-SAM [98] polygons. Each word is processed in three steps: (1) polygon masking and cropping, (2) foreground stroke isolation via grayscale conversion and Otsu thresholding, and (3) extraction of text pixels with median RGB estimation. To ensure consistency, near-black artifacts are removed and remaining colors are clustered with K-Means (cluster count via elbow method). The resulting representative colors are appended as auxiliary hints (Prompt 12.2.2) to the Stage-2 prompt.

Extra Instructions - Color

Colour Information

The document has words with unique RGB colors (format: [[R,G,B]...]):
`{color_information}`

Instructions:

- Match each word to its RGB value.
- Apply `inline style='color:rgb(r,g,b)'` for words.
- If a line has multiple colors, wrap words in `` with their color.

Font Attributes. To capture bold, italic, underline, strikeout, and their combinations, we use TexTAR—a multi-task, context-aware Transformer trained on MMTAD with DocTr for OCR. Predictions are post-processed to retain only useful word-level attributes and OCR text. This information is then provided as structured hints (Prompt 12.2.2) for Stage-2, enabling faithful reconstruction of semantic and stylistic cues.

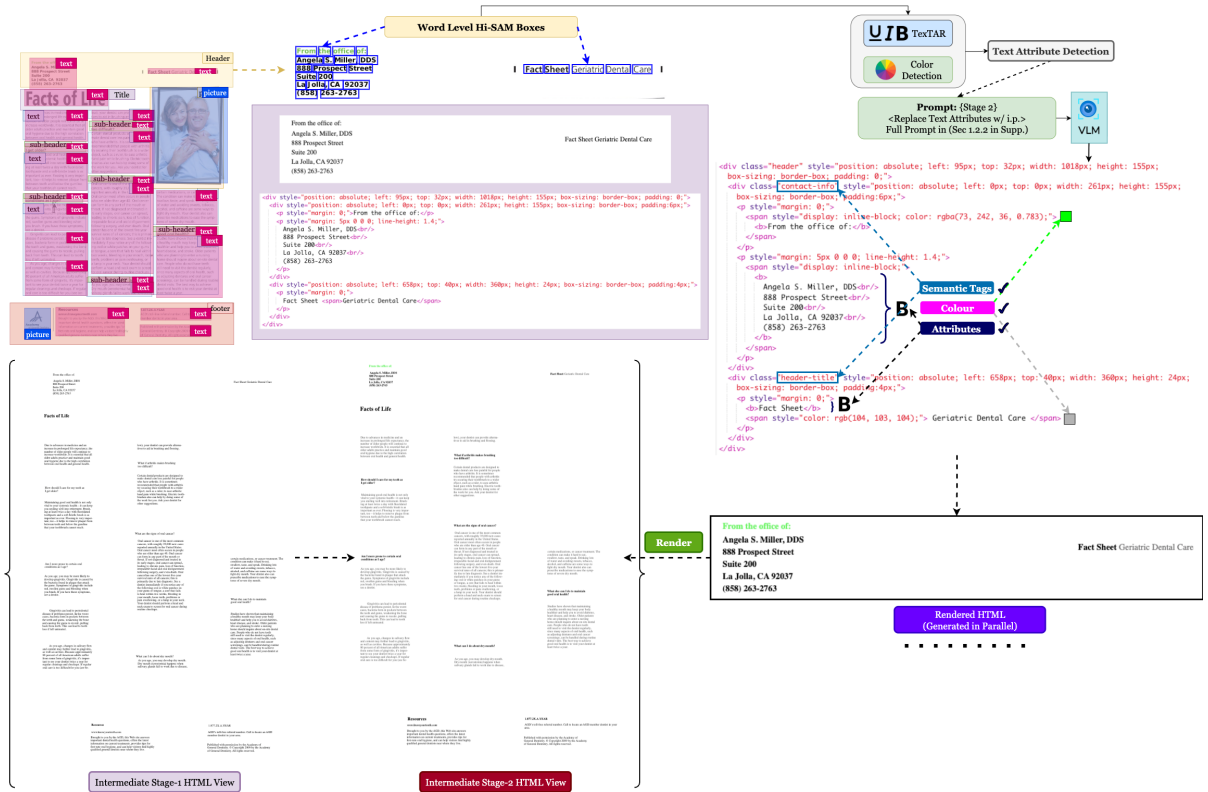


Fig. 15: Localize Stage: 2 This stage enriches raw sub-HTMLs by adding semantic tags, textual attributes, and stylistic details. Word-level Hi-SAM [98] polygons provide fine-grained text crops, which are analyzed for font attributes (via TextTAR) and color cues (via heuristic color detection with K-Means clustering). These attributes are incorporated into layout-aware prompts that guide the VLM to insert correct inline styles (bold, italics, underline, strikethrough, and color) and semantic tags (e.g., headings, lists, tables, equations, figures). An intermediate full-page view is also shown to illustrate the results of this stage and highlight the improvements compared to Stage-1.

Extra Instructions - Other Font Attributes

Text Attribute Information

Here are the font attributes for each word, provided as a list of dictionaries mapping OCR text to its attributes: {attr_information}

Instructions:

- Map the correct text to its correct text attribute.
- Use inline CSS styles to represent these attributes.
- If words within a line have different attributes, wrap each word in a `` with its corresponding style.

Layout-aware prompting. Rather than relying on a single generic prompt for all regions, REPLICA employs layout-aware prompts during Stage 2 of Localize. Each sub-HTML fragment is generated with crop-specific instructions tailored to the layout elements identified in the Segment stage. This approach optimizes token usage by focusing the VLM’s attention on the structural requirements of the detected

Algorithm 1 Unique Text Color Detection

Require: Cropped region size D , OCR text file T , maximum clusters K_{max}

Ensure: Set of representative unique colors C

```
1: Load OCR polygons and image from  $T$ 
2: Filter polygons within crop dimension  $D$ 
3: for each word polygon  $p$  do
4:    $c_p \leftarrow \text{GETWORDCOLOR}(p, \text{image})$ 
5:   Store  $(p, c_p)$ 
6: end for
7: Remove background/near-black colors
8: Collect all remaining word colors into set  $W$ 
9: if  $W$  is empty then
10:   return  $\emptyset$ 
11: end if
12: Determine optimal cluster count  $k \leq K_{max}$  using elbow method
13: Apply K-Means on  $W$  with  $k$  clusters
14: Let  $C$  be the cluster centers (representative unique colors)
15: return  $C$ 
16: function GETWORDCOLOR( $p, I$ )
17:   Create mask for polygon  $p$  in  $I$ 
18:   Crop region inside bounding box of  $p$ 
19:   Apply mask to keep only pixels within  $p$ 
20:   Convert to grayscale and threshold to isolate text
21:   Extract text pixels
22:   if no text pixels found then
23:     return  $(0, 0, 0)$ 
24:   else
25:     Compute median RGB of text pixels
26:     return median color
27:   end if
28: end function
```

element. For example, list regions receive explicit instructions to use ordered or unordered list tags (``, ``, ``); table crops are prompted to follow correct row–cell hierarchies (`<tr>`, `<td>`); mathematical regions are instructed to output MathML for equations and formulas; and figure or image crops are directed to generate `` tags with alt-text metadata. Similarly, headings are emphasized with the appropriate `<h>` hierarchy to reflect document semantics. By aligning the prompt design with detected layout types, layout-aware prompting ensures that the generated HTML not only preserves visual fidelity but also encodes semantic correctness, reducing hallucinations and improving structural consistency compared to generic prompting.

“ ”

Layout Aware Prompt - Formulas (must use MathML)

Example:

```
<div data-bbox='x1 y1 x2 y2' class='formula'>
  <math xmlns='http://www.w3.org/1998/Math/MathML'>
    <msup>
```

```

      <mi>E</mi>
      <mn>2</mn>
    </msup>
    <mo>=</mo>
    <mi>mc</mi>
    <msup>
      <mi>c</mi>
      <mn>2</mn>
    </msup>
  </math>
</div>

```

Layout Aware Prompt - Headings

Example:

```

<h1 data-bbox='x1 y1 x2 y2' class='heading'>...</h1>
...
<h6 data-bbox='x1 y1 x2 y2' class='heading'>...</h6>

```

Notes:

- Style for font size, weight, and color.
- Use appropriate heading level (<h1> to <h6>) based on hierarchy.

Layout Aware Prompt - Images

Example:

```

<img data-bbox='x1 y1 x2 y2' class='(semantic rich name)'/>

```

Notes:

- Do not wrap in unnecessary containers.
- Ensure no text is generated within images.

Layout Aware Prompt - Lists (Ordered and Unordered)

Example:

```

<ol class='ordered-list'>
  <li data-bbox='x1 y1 x2 y2'>...</li>
</ol>

<ul class='unordered-list'>
  <li data-bbox='x1 y1 x2 y2'>...</li>

```


Notes:

- Style for indentation and marker formatting.
- Do not hardcode numbering or bullet symbols.
- Combine multiline list items into a single .

Layout Aware Prompt - Tables

Example:

```
<table class='table'>
  <tr>
    <th data-bbox='x1 y1 x2 y2'>..</th>
  </tr>
  <tr>
    <td data-bbox='x1 y1 x2 y2'>..</td>
  </tr>
</table>
```

Notes:

- Style for borders, padding, and alignment.
- Use colspan and rowspan where applicable.
- Combine multiline cells into a single <td> with merged bbox.

Stage 2 Prompt

Objective: Refine the provided Qwen-HTML to visually and semantically match the source image.

Your Task:

Given an image and its corresponding Qwen-HTML (Input HTML), you must improve the HTML by applying CSS styling and semantic structure.

Strict Instructions:

1. Use only the bounding boxes provided in the HTML exactly as they are. Do not modify, resize, or adjust them in any way.
2. Visual Styling: Add CSS to replicate the source image's visual appearance. This includes, but is not limited to, fonts weight, borders, and shading. Do not give font size and font family, background color.
3. Semantic Grouping: Identify semantically related elements in the Qwen-HTML. Merge them into a single parent <div> tag.
4. Class Assignment: Assign a single, appropriate semantic class to each new parent <div>.

5. Strictly preserve the given text. Do NOT generate new text or alter the existing alignment. Ensure all `
` tags are retained.
6. Output format: The final output must contain only the refined HTML and CSS code. Do not include explanations, markdown, or extra commentary.
7. Crucially, retain original text colors, background colors, and text attributes (bold, italic, underline, etc.) for all text elements, and preserve center, left, or right alignment where present.

HTML Tag Usage and Styling Guidelines

1. Text Paragraphs

`<p data-bbox='x1 y1 x2 y2' class='(semantic rich name)'\>...</p>`

Must ensure every paragraph includes a data-bbox attribute.

`{{heading_prompt}}`

`{{images_prompt}}`

`{{List_prompt}}`

`{{tables_prompt}}`

`{{formula_prompt}}`

Input HTML:

`{{HTML}}`

Approved Class List:

Use the most fitting class from this list for each semantic group.

header
sub-header
table-of-contents
references
contact-info
placeholder-text
footer
sidebar
title
heading
paragraph
caption

```

page-number
dateline
table
form
ordered-list
unordered-list
figure
picture
logo
chart
formula
code-block
handwriting
signature
form
question
options

```

12.3 Assemble

12.3.1 Reading Order-Aware Merging

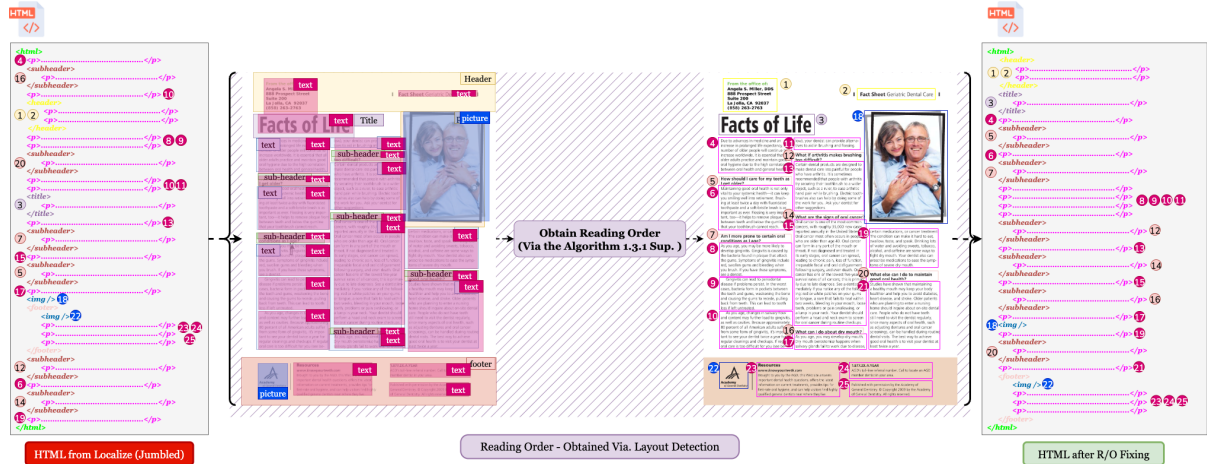


Fig. 16: Stage 3: Reading Order-Aware Merging. After the two-stage *Localize* step, each document region has been converted into a styled sub-HTML fragment. In this stage, the fragments are merged into a page-level HTML by aligning their absolute coordinates for visual fidelity and arranging them into the correct logical sequence. REPLICa’s algorithm integrates spatial alignment with logical sequencing, ensuring that the final HTML is both visually accurate and semantically coherent. The figure illustrates how initially jumbled sub-HTMLs are reordered through layout-driven reading order detection to produce structured, accessible HTML.

After the 2-stage *Localize* step, each document region is refined into a styled sub-HTML fragment. These fragments are merged into a page-level HTML by aligning them with absolute coordinates for

visual fidelity and arranging them in the correct reading sequence for semantic coherence. Reading order here goes beyond simple left-to-right or top-to-bottom traversal, capturing the intended logical flow of text, images, figures, and tables. This ordering is critical for accessibility (e.g., screen readers), downstream tasks (indexing, search, translation), and faithful conversions to formats like LaTeX or Markdown. REPLICIA’s reading order algorithm integrates spatial alignment with logical sequencing to produce HTML that is both visually accurate and structurally coherent (Algorithm 2).

Algorithm 2 Reading Order-Aware Merging of sub-HTMLs

Require: Set of sub-HTMLs $\{H_1, H_2, \dots, H_n\}$ with positional styles

Ensure: Full HTML document H_{full} with correct reading order

```

1:
2: Initialize empty list  $B$  ▷ stores body elements with coordinates
3: for each sub-HTML  $H_i$  do
4:   Parse  $H_i$  with HTML parser
5:   Extract absolutely positioned divs from  $H_i.body$ 
6:   for each div  $d$  in  $H_i.body$  do
7:     Convert inline CSS style style( $d$ ) to coordinates  $[x_1, y_1, x_2, y_2]$ 
8:     Store tuple  $(d, [x_1, y_1, x_2, y_2])$  in list  $B$ 
9:   end for
10: end for
11:
12: Let  $C = \{[x_1, y_1, x_2, y_2] \mid (d, [x_1, y_1, x_2, y_2]) \in B\}$ 
13:  $C_{sorted} \leftarrow \text{GetReadingOrder}(C)$ 
14:
15: Initialize new HTML document  $H_{full}$ 
16: for each bbox  $b$  in  $C_{sorted}$  do
17:   Find div  $d$  in  $B$  such that  $d$  has coordinates  $b$ 
18:   Append  $d$  to  $H_{full}.body$ 
19: end for
20:
21: return  $H_{full}$ 
22:
23: function GETREADINGORDER( $C$ )
24:   Sort  $C$  by  $x_1$  (left coordinate) to get  $C_x$ 
25:   Compute mean width  $\mu = \frac{1}{|C_x|} \sum_{b \in C_x} (b.x_2 - b.x_1)$ 
26:   Initialize empty list of vertical lines  $V$ 
27:   Initialize temporary line  $L \leftarrow []$ 
28:   Set current baseline  $x_{cur} \leftarrow C_x[0].x_1$ 
29:   for each bbox  $b$  in  $C_x$  do
30:     if  $b.x_1 \geq x_{cur} + \mu$  then
31:       Append  $L$  to  $V$ 
32:       Reset  $L \leftarrow [b]$ 
33:       Update  $x_{cur} \leftarrow b.x_1$ 
34:     else
35:       Append  $b$  to  $L$ 
36:     end if
37:   end for
38:   Append last  $L$  to  $V$ 
39:   for each vertical line  $L \in V$  do
40:     Sort  $L$  by  $y_1$  (top coordinate) ▷ top-down order
41:   end for
42:   Flatten all lines  $V$  into a single ordered list  $C_{sorted}$ 
43:   return  $C_{sorted}$ 
44: end function

```

12.4 Refine

12.4.1 Font Size Fixing Algorithm

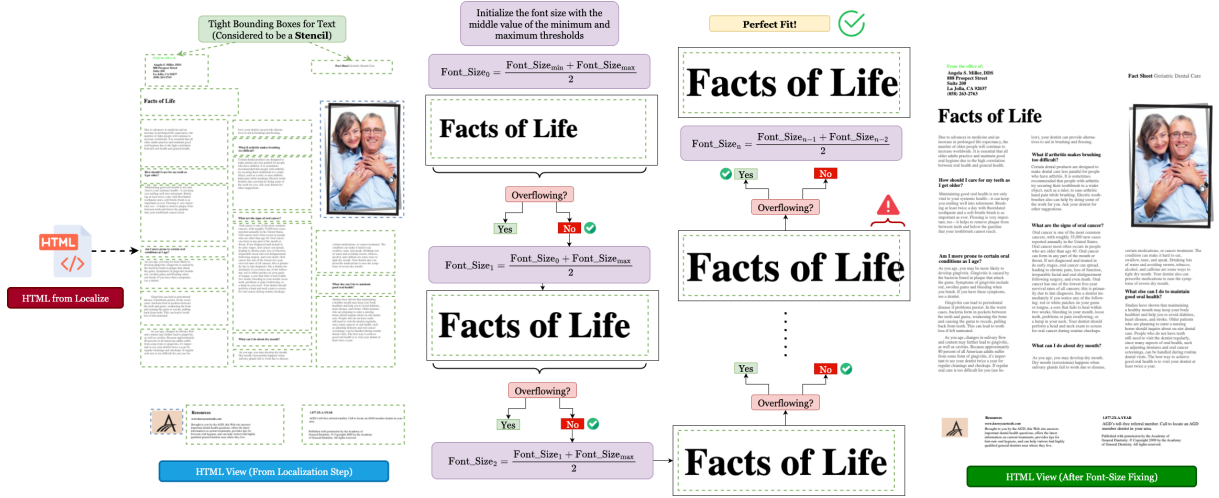


Fig. 17: Stage 4.1: Font Size Fixing. Following the segmentation, localization, and assembly stages, the HTML font sizes are adjusted to faithfully reproduce the visual appearance of the source document. Tight bounding boxes from the flat layouts obtained in the segmentation stage act as stencils, defining the container dimensions for each text element. Font sizes are then optimized via a binary search between minimum and maximum thresholds: starting from the midpoint, candidate sizes are iteratively tested for overflow until the largest fitting value is identified. The figure illustrates this stencil-based fitting process, the iterative adjustment loop, and the final full-page HTML view after font size correction.

The *Font Size* algorithm (Algorithm 3) dynamically adjusts text size to fit within fixed container dimensions. Each element's text is wrapped in a span, container width/height are measured, and alignment or multiline options applied. A binary search between minimum and maximum font sizes iteratively tests fit, converging on the largest size that avoids overflow. The final value is assigned to the span, maximizing readability while respecting spatial constraints.

Algorithm 3 Fit Text to Container Element

```
1: procedure PROCESSITEM(el, Settings) ▷ Initial setup and validation
2:   if not IsElement(el) or (el is already processed and not Settings.reProcess) then
3:     return
4:   end if
5:   containerWidth  $\leftarrow$  inner width of el
6:   containerHeight  $\leftarrow$  inner height of el
7:   if containerWidth = 0 or (containerHeight = 0 and not Settings.widthOnly) then
8:     throw Error("Container requires set dimensions.")
9:   end if
10:  innerSpan  $\leftarrow$  wrap content of el in a new <span>
11:  APPLYPRELIMINARYSTYLES(el, innerSpan, Settings) ▷ Handles text-align, multi-line, etc.
12:  ▷ Binary search for the optimal font size
13:  low  $\leftarrow$  Settings.minFontSize
14:  high  $\leftarrow$  Settings.maxFontSize
15:  optimalSize  $\leftarrow$  low
16:  while low < high do
17:    mid  $\leftarrow$   $\lfloor (low + high) / 2 \rfloor$ 
18:    Set font size of innerSpan to mid pixels
19:    currentWidth  $\leftarrow$  measured width of innerSpan
20:    currentHeight  $\leftarrow$  measured height of innerSpan
21:    if currentWidth  $\leq$  containerWidth and (Settings.widthOnly or currentHeight  $\leq$  container-
    Height) then
22:      optimalSize  $\leftarrow$  mid ▷ This size fits, save it
23:      low  $\leftarrow$  mid + 1 ▷ Try for an even larger size
24:    else
25:      high  $\leftarrow$  mid - 1 ▷ This size is too large, try smaller
26:    end if
27:  end while
28:  Set final font size of innerSpan to optimalSize pixels
29:  if Settings.alignVert then
30:    APPLYVERTICALALIGNMENTSTYLES(el, innerSpan)
31:  end if
32: end procedure
```

12.4.2 Reflection

Reflection-based Iterative Refinement. The reflection stage introduces a review loop where a Vision-Language Model (VLM) inspects the rendered HTML alongside the original document image and its HTML source. The VLM identifies discrepancies—such as misaligned elements, missing or incorrect attributes, and structural inconsistencies—and proposes corrections. This process repeats iteratively until no further errors are detected or a maximum iteration limit is reached.

Reflection Loop. We implement an iterative refinement cycle using state-of-the-art reasoning VLMs (e.g., Gemini 2.5 Pro [2], Qwen QVQ-72B [125]) with `max_iterations=6`. At each step, the VLM generates a structured list of errors in the current HTML and revises the markup accordingly. The loop progressively improves fidelity by correcting issues such as text indentation, line breaks, color mismatches, and subtle positional misalignments.

Reasoning vs. Non-Reasoning VLMs. Reasoning-oriented VLMs demonstrate superior performance in reflection. They converge to error-free HTML in fewer iterations compared to non-reasoning counterparts, owing to their stronger multi-step reasoning and correction capabilities. This makes them more effective at refining fine-grained details and achieving consistent alignment between the rendered HTML and the original document.

12.4.3 Background Restoration

Importance of Background preservation. Backgrounds in documents—such as decorative elements, highlights, or subtle aging artifacts—carry stylistic and contextual information critical for visual fidelity. Preserving them ensures authenticity in HTML reconstructions.

Pipeline. We implement a two-stage adaptive inpainting pipeline using OpenCV. Unlike traditional use in photo restoration [126–129], inpainting is repurposed here for controlled erasure: text and layout elements (tables, figures) are removed to reconstruct the underlying background.

Mask generation. Line-level polygons from Hi-SAM [98] and layout boxes from the Segment stage are used to create binary masks. These are expanded with adaptive morphological dilation (`cv2.dilate`), where kernel size and iterations scale with bounding box dimensions. Dilation ensures anti-aliased strokes and scan artifacts are included, preventing residuals while avoiding over-expansion into background areas.

Inpainting. We apply `cv2.inpaint` with Telea’s Fast Marching Method [126] (`INPAINT_TELEA`), which is efficient and produces smooth fills for text-sized regions. The Navier–Stokes method [127], suited for thin cracks and scratches, was avoided as it introduced streaking artifacts on large text regions. Restored backgrounds are saved and integrated seamlessly into the HTML rendering pipeline. Figure 18 demonstrates the importance of background in maintaining visual fidelity.

13 Design choices behind REPLICA

13.1 Module-specific Metrics

For the *Segment* stage, we propose two new metrics - Text Coverage (TC) and Visual Coverage (VC) - which emphasize completeness by measuring whether all textual and visual elements in the source are detected as some class. Unlike raw mAP, which may be high despite missing elements, TC and VC prioritize coverage since misclassified regions can still be retagged in Localize Stage 2 via VLMs, whereas undetected elements are irrecoverable. For the *Localize* stage, we introduce the List Structure Penalty (LSP) to assess proper list formatting, while established metrics such as Tree Edit Distance Similarity (TEDS) [130] for tables and Character Detection Matching (CDM) [131] for formulas are also employed. Full metric definitions and fine-grained evaluations are provided in the supplementary material.

13.2 Ablations

Effect of layout ensembles: Ensembling flat and hierarchical layout detectors yields the strongest performance (table 8a), reducing the variance of single models and improving robustness across document types. Adding a hierarchical segmentation model such as Hi-SAM provides the largest boost, substantially increasing element coverage.

Effect of 2-stage layout-aware Set-of-Mark (SoM) VLM prompts: We observe consistent gains across all design choices (table 9a). Moving from single-stage to two-stage prompting improves both structural accuracy and position retention. Using layout-aware prompting instead of generic prompts further boosts performance, with particularly large benefits on complex multi-column layouts such as newspapers. Finally, SoM prompts outperform traditional OCR-enriched bounding-box prompts, delivering higher structural and positional fidelity while also reducing hallucinations. Overall, the two-stage, layout-aware SoM design consistently yields the best results. Refer supplementary for additional ablations.

Effect of font-size fixing: Retaining font size improves both readability and semantics, as failing to do so often leads to overlapping text and reduced visual fidelity (table 8c). Our method significantly enhances fidelity by explicitly fixing font sizes, whereas closed-source VLMs attempt to infer them but remain noticeably less accurate.

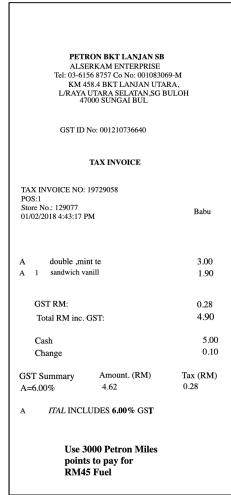
Effect of SoM Prompting for Position Injection into HTML. While models such as Qwen-2.5-VL support native document grounding, they are inconsistent in reliably giving positional information through bounding boxes. Hi-SAM [98] provides accurate text bounding boxes, but the way this information is conveyed to the VLM is critical. Directly providing bounding box coordinates often leads to



Document (.pdf)



Without Background



With Background

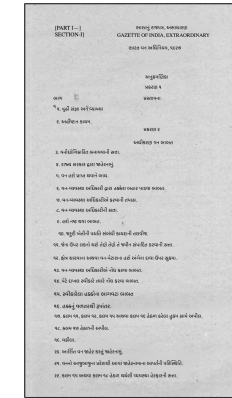
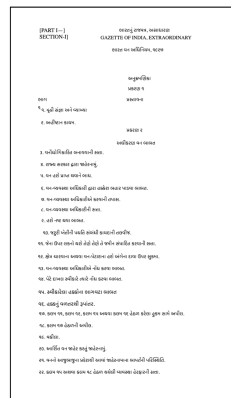
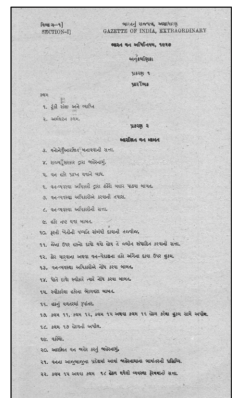
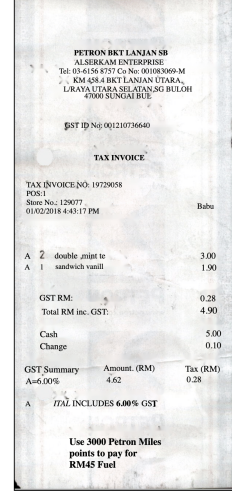


Fig. 18: Demonstration of the role of background in preserving visual fidelity.

Table 8: Framework Ablations. Default settings are marked in gray. See the Supplementary for additional ablations.

(a) **Segment Module Ablations.** Text Coverage (TC), Visual Coverage (VC), and mean Average Precision (mAP) for layout models and ensembles. **DLN**: DocLayout-YOLO [96], **IDLP-DLN**: IndicDLP-trained DocLayout [97], **HS**: Hi-SAM [98].

| Layout Model | TC (↑) | VC (↑) | mAP (↑) |
|----------------------------|-------------|-------------|-------------|
| DLN | 0.87 | 0.88 | 0.27 |
| IDLP-DLN | 0.96 | 0.96 | 0.89 |
| DLN + IDLP-DLN | 0.97 | 0.99 | 0.99 |
| DLN + IDLP-DLN + HS | 0.99 | 0.99 | 0.99 |

(b) **Localize Module Ablations.** Effect of two-stage sub-HTML generation, layout-aware prompting (LA), and Set of Mark (SoM) prompting on Overall Score (OS) [95].

| Method | OS (↑) |
|-------------------------------|-------------|
| Stage 1, w/o SoM | 0.55 |
| Stage 2, w/o LA, w/o SoM | 0.72 |
| Stage 2, w/ LA, w/o SoM | 0.75 |
| Stage 2, w/o LA, w/ SoM | 0.78 |
| Stage 2, w/ LA, w/ SoM | 0.86 |

(c) **Refine Module Ablations.** Effect of Stage 2 with Set of Mark (SoM) and layout-aware prompting (LA) (Base), font size fixing (fs), and background stitching (bg) on Visual Fidelity Score (VFS).

| Method | VFS (↑) |
|--|-------------|
| Base | 0.84 |
| Base, w/ bg | 0.85 |
| Base, w/ fs | 0.87 |
| Base, w/ fs, w/ bg | 0.90 |
| Base, w/ fs, w/ bg, w/ reflection | 0.93 |

hallucinations, and conventional Set-of-Mark (SoM) prompting with numeric indices introduces another issue—numbers are mistakenly carried over into the final HTML. To address this, we adapt SoM by using colors as indices, supplying the VLM with a color-to-bounding box mapping. This color-coded strategy avoids numerical leakage and enables robust injection of positional information into the generated HTML (Table 9a).

Effect of choice of VLM in localize: Among the models evaluated, *GPT-5* consistently outperforms all alternatives, achieving the strongest localization accuracy and position retention. Within the open-source space, *Qwen-2.5-VL-72B* emerges as the most competitive option (Table 9b), surpassing other large open-source VLMs such as Gemma 3 and InternVL. This highlights that while GPT-5 remains the most reliable choice overall, Qwen-2.5-VL-72B offers the best trade-off for open-source deployments.

Effect of Using Auxiliary Information. As shown in Table 9c, each auxiliary signal incrementally improves reconstruction quality. OCR establishes a strong base, with color and text attributes adding semantic and stylistic fidelity. Font size and background restoration further enhance visual consistency. Iterative reflection proves most impactful, correcting residual errors and consolidating earlier gains into the highest-quality HTML.

Effect of Choice of VLM for Reflection. Open-source VLMs provide limited benefits in reflection, often producing inconsistent results. Improvements, when present, emerge only after many iterations and plateau quickly. In contrast, reasoning-oriented VLMs such as Gemini-2.5-Pro and GPT-5 yield substantial gains in visual fidelity within a few iterations (Table 9d), while open-source reasoning models like Qwen QVQ-72B also contribute positively but with more gradual improvements. These trends highlight the importance of strong reasoning capabilities for effective reflection-based refinement.

Effect of Number of Reflection Iterations. For general open-source models, iterative reflection shows diminishing returns—improvements are inconsistent, often marginal, and can even degrade with more iterations. In contrast, reasoning-capable models benefit clearly from additional iterations, with GPT-5 delivering the most stable and effective gains (Table 9e). Qwen QVQ-72B provides moderate improvements among open-source reasoning models, though the gains are more gradual compared to closed-source counterparts.

Table 9: Additional Framework Ablations. Best settings are marked in gray. OS denotes Overall Score, VFS denotes Visual Fidelity Score, and GPS denotes Global Position Score. For, **Auxiliary features** such as text attributes (e.g., bold, underline), color, font size (fs), background (bg), and reflection are incorporated. These attributes enhance visual fidelity by improving the clarity and distinctiveness of the visual elements, with reflection providing additional visual refinement. Also, n is number of iterations of the reflection loop.

(a) **Effect of SoM Prompting for Position Injection into HTML**

| Method | GPS (\uparrow) |
|------------------------------|--------------------|
| Qwen-HTML, w/o SoM | 0.55 |
| Hi-SAM bboxes in prompt [98] | 0.72 |
| SoM (number indexing) | 0.68 |
| SoM (colour indexing) | 0.77 |

(b) **Effect of choice of VLM in localize**

| Model | OS (\uparrow) |
|-----------------------------|-------------------|
| Intern-S1 [132] | 0.24 |
| InternVL3.5 [46] | 0.21 |
| gemma-3-27b-it [28] | 0.25 |
| Qwen2.5-VL-32B-Instruct[43] | 0.51 |
| Qwen2.5-VL-72B-Instruct[43] | 0.86 |
| Gemini-2.5-Pro [120] | 0.91 |
| GPT-5[121] | 0.97 |

(c) **Effect of Using Auxiliary Information.**

| Method | VFS (\uparrow) |
|--|--------------------|
| Base w/o OCR | 0.54 |
| Base w/ OCR | 0.59 |
| Base w/ OCR, w/ Color | 0.60 |
| Base w/ OCR, w/ Text Attr, w/ Color | 0.75 |
| Base w/ OCR, w/ Text Attr, w/ Color, w/ fs | 0.87 |
| Base w/ OCR, w/ Text Attr, w/ Color, w/ fs, w/ bg | 0.90 |
| Base w/ OCR, w/ Text Attr, w/ Color, w/ fs, w/ bg w/ reflection | 0.93 |

(d) **Effect of Choice of VLM for Reflection**

| Model | VFS (\uparrow) |
|-----------------------------|--------------------|
| Intern-S1 [132] | 0.79 |
| InternVL3.5 [46] | 0.78 |
| gemma-3-27b-it [28] | 0.73 |
| Qwen2.5-VL-32B-Instruct[43] | 0.80 |
| Qwen2.5-VL-72B-Instruct[43] | 0.83 |
| Qwen QVQ-72B [125] | 0.86 |
| Gemini-2.5-Pro [120] | 0.91 |
| GPT-5[121] | 0.97 |

(e) **Effect of Number of Reflection Iterations**

| Method | $n = 1$ | $n = 6$ | $n = 10$ | $n = 15$ |
|------------------------------|-------------|-------------|-------------|-------------|
| Intern-S1 [132] | 0.79 | 0.75 | 0.76 | 0.75 |
| InternVL3.5 [46] | 0.78 | 0.77 | 0.75 | 0.74 |
| gemma-3-27b-it [28] | 0.73 | 0.71 | 0.69 | 0.66 |
| Qwen2.5-VL-32B-Instruct [43] | 0.80 | 0.77 | 0.75 | 0.75 |
| Qwen2.5-VL-72B-Instruct [43] | 0.83 | 0.81 | 0.80 | 0.80 |
| Qwen QVQ-72B [125] | 0.86 | 0.88 | 0.89 | 0.89 |
| Gemini-2.5-Pro [120] | 0.89 | 0.91 | 0.92 | 0.93 |
| GPT-5 [121] | 0.91 | 0.93 | 0.94 | 0.97 |

13.3 REPLICA helps annotation

Due to the carefully designed choices, the REPLICA framework substantially streamlines the annotation process for the VFDR task in the FID-HTML format. As illustrated in fig. 19, the number of manual edits required after generation with REPLICA (using Gemini 2.5 Pro) is markedly lower compared to directly using Gemini 2.5 Pro. A similar trend is observed for Qwen3-VL, highlighting the generality of our approach. A key advantage of the modular pipeline is that the *absolute positioning* of document

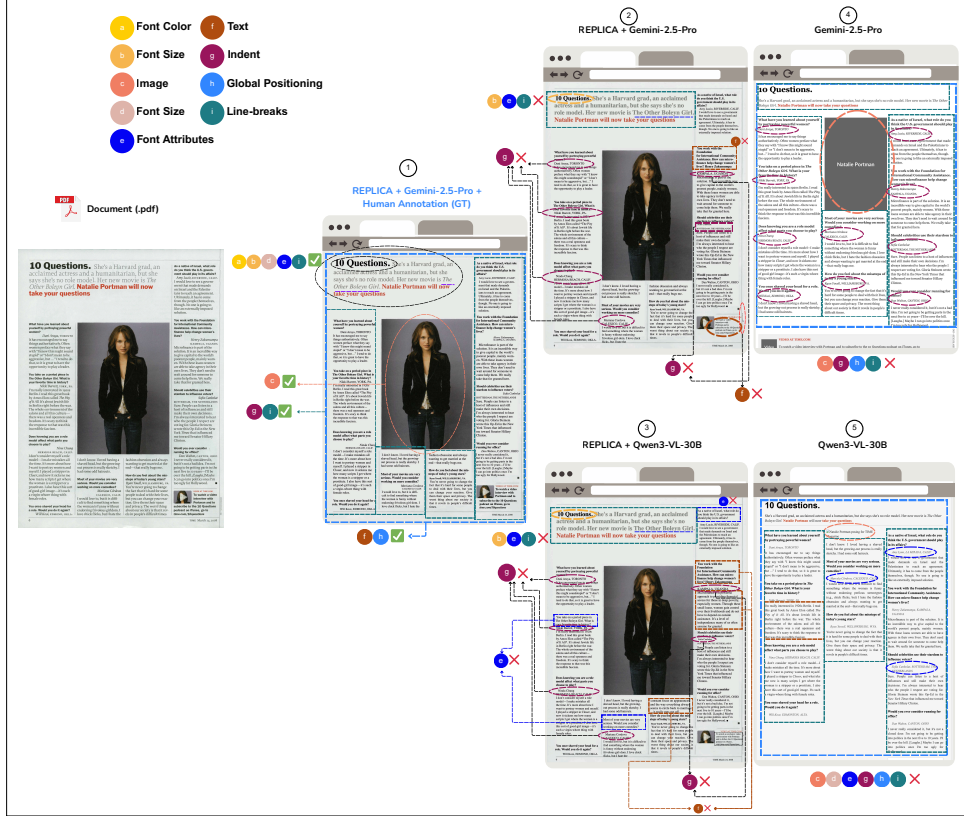


Fig. 19: Comparison of annotation difficulty across systems. Replica-based pipelines require substantially less human correction than VLM-only baselines. The relative ease-of-annotation follows: REPLICA + Gemini-2.5-Pro + human annotation > REPLICA + Gemini-2.5-Pro > REPLICA + Qwen3-VL-30B > Gemini-2.5-Pro > Qwen3-VL-30B, highlighting REPLICA’s advantage for producing stable and verifiable annotations.

substructures is handled automatically. This is particularly important for preserving visual coherence: without such control, models like Qwen3-VL and Gemini 2.5 Pro often produce overlaid or intersecting elements, resulting in visually inconsistent renderings. Furthermore, REPLICA reliably manages *font size*, a critical factor in perceived visual fidelity. Other challenging aspects—such as color, indentation, and font attributes—are also largely preserved, with only minor corrections needed during annotation. Overall, REPLICA provides strong synthetic supervision that significantly accelerates high-fidelity ground-truth construction, reducing annotation burden while improving consistency and visual alignment.

To quantify the reduction in manual post-editing enabled by our pipeline, we sampled 100 documents and asked a group of 7 trained annotators to correct the generated HTML. For each document, we recorded the number of atomic edits—OCR fixes, positional adjustments, style corrections, and HTML hierarchy repairs—and report the average edits per document aggregated across annotators. REPLICA + Gemini 2.5 Pro requires the fewest corrections, followed by REPLICA + Qwen3VL, whereas using the base VLMs alone leads to substantially higher editing effort. This monotonic trend underscores the effectiveness of REPLICA’s layout-aware generation in reducing annotation cost, while also highlighting the benefits of coupling VLMs with a constrained HTML rendering pipeline. Refer to table 10 for more details.

Table 10: Average number of manual edits per document across 100 sampled documents, aggregated over 7 annotators. Lower is better. Each edit corresponds to an OCR, positional, styling, or HTML hierarchy correction.

| Method | Avg. Edits / Document |
|--------------------------|-----------------------|
| REPLICA + Gemini 2.5 Pro | 2.24 |
| REPLICA + Qwen3VL | 2.81 |
| Gemini 2.5 Pro | 4.32 |
| Qwen3VL | 7.11 |

Table 11: Overall comparison of document-to-HTML methods across four evaluation dimensions. Text Extraction (TE: Word recognition rate (WRR), Character recognition rate (CRR)), Logical Structure (LS: Normalized Tree Edit Distance (NTED)), Physical Structure (PS: Global Position Score (GPS), Local Position Score (LPS)), and Visual Fidelity (VF: Visual Fidelity Score (VFS) via VLM-as-a-judge) are reported. The details about overall score computation can be found in the supplementary. Baseline methods are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML with REPLICA representing our method.

| Method | TE | | LS | | PS | | VF | Overall |
|----------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| | WRR (↑) | CRR (↑) | NTED (↓) | LSS (↑) | LPS (↑) | GPS (↑) | VFS (↑) | |
| Marker [3] | 0.65 | 0.80 | <u>0.80</u> | 0.68 | 0.25 | 0.35 | 0.76 | 0.62 |
| Docling [26] | 0.38 | 0.47 | 0.86 | 0.63 | 0.12 | 0.26 | 0.48 | 0.43 |
| RolmOCR [116] | 0.54 | 0.70 | 0.95 | 0.23 | 0.26 | 0.07 | 0.47 | 0.37 |
| smolDocling [13] | 0.34 | 0.40 | 0.90 | 0.42 | 0.13 | 0.24 | 0.45 | 0.36 |
| olmOCR-7B [66] | 0.54 | 0.67 | 0.93 | 0.26 | <u>0.28</u> | 0.10 | 0.28 | 0.34 |
| Nanonets-OCR-s [117] | 0.41 | 0.54 | 0.95 | 0.20 | 0.25 | 0.10 | 0.48 | 0.33 |
| GOT-OCR-2.0 [27] | 0.42 | 0.48 | 0.93 | 0.21 | 0.17 | 0.25 | 0.60 | 0.37 |
| OCRFlux-3B [118] | 0.45 | 0.57 | 0.88 | 0.51 | 0.25 | 0.25 | 0.37 | 0.41 |
| Logics-Parsing [119] | 0.64 | 0.77 | 0.84 | 0.57 | 0.32 | <u>0.42</u> | 0.65 | 0.57 |
| Qwen2.5-VL-7B-Instruct [43] | 0.50 | 0.61 | <u>0.80</u> | 0.60 | 0.14 | 0.24 | 0.37 | 0.43 |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.47 | 0.54 | 0.78 | 0.44 | 0.15 | 0.33 | 0.29 | 0.37 |
| gemma-3-27b-it [28] | 0.57 | 0.62 | 0.85 | 0.59 | 0.16 | 0.25 | 0.32 | 0.43 |
| InternVL3-14B [46] | 0.28 | 0.46 | 0.83 | <u>0.67</u> | 0.22 | 0.29 | 0.27 | 0.39 |
| Gemini-2.5-Pro [120] | 0.63 | 0.69 | 0.81 | <u>0.67</u> | 0.19 | 0.31 | 0.44 | 0.50 |
| GPT-5 [121] | 0.64 | 0.77 | <u>0.80</u> | 0.68 | 0.26 | 0.44 | <u>0.73</u> | 0.62 |
| WebCoder-1.3B [79] | 0.33 | 0.42 | 0.82 | 0.40 | 0.20 | 0.27 | 0.34 | 0.34 |
| WebSight-VLM-7B [29] | 0.14 | 0.42 | 0.85 | 0.57 | 0.19 | 0.29 | 0.16 | 0.31 |
| REPLICA (+Qwen3-VL) (Ours) | 0.86 | 0.90 | 0.24 | 0.71 | 0.73 | 0.86 | 0.90 | 0.83 |
| REPLICA (+Gemini 2.5 Pro) (Ours) | 0.88 | 0.93 | 0.20 | 0.74 | 0.73 | 0.86 | 0.93 | 0.86 |

14 REPLICA as a baseline for VFDR

REPLICA achieves substantial gains over all zero-shot baselines across method categories. Its modular pipeline—combining SoM-style prompting, Hi-SAM-based region refinement, and dedicated modules for font-size normalization, color and font-attribute correction, and image stitching—leads to markedly improved absolute positioning and consistently high visual fidelity. Comprehensive results are provided in table 11.

Table 12: Pearson correlation between human rankings (over five model outputs) and our VFDR metrics. All metrics show strong correlation (> 0.8).

| Human Annotators (7) | LSS | VFS | GPS | LPS | Overall |
|----------------------|-------------|-------------|-------------|-------------|-------------|
| Annotator-1 | 0.84 | 0.87 | 0.82 | 0.85 | 0.88 |
| Annotator-2 | 0.81 | 0.86 | 0.80 | 0.84 | 0.86 |
| Annotator-3 | 0.83 | 0.88 | 0.83 | 0.86 | 0.89 |
| Annotator-4 | 0.85 | 0.90 | 0.84 | 0.87 | 0.91 |
| Annotator-5 | 0.82 | 0.85 | 0.81 | 0.83 | 0.86 |
| Annotator-6 | 0.86 | 0.89 | 0.85 | 0.88 | 0.92 |
| Annotator-7 | 0.84 | 0.87 | 0.82 | 0.85 | 0.88 |
| Average | 0.83 | 0.87 | 0.82 | 0.85 | 0.89 |

15 Evaluations

15.1 Human-centric Evaluations

To evaluate the reliability of our evaluation metrics, we perform a human–metric correlation study involving seven experts with backgrounds in document design and layout analysis. We randomly select 100 documents from VFDR-BENCH and ask each expert to rank the HTML reconstructions produced by five representative models (anonymized as M1–M5). For each annotator, this yields a 5-way ranking vector.

Our metrics - Logical Structure Score (LSS), Visual Fidelity Score (VFS), Global Position Score (GPS), and Local Position Score (LPS)—produce their own corresponding 5-way ranking vectors for the same set of models. We compute the Pearson correlation between each annotator’s ranking and each metric’s ranking. Since correlation is computed over model rankings, the effect of multiple models is already incorporated into each coefficient; thus, the table reports per-annotator correlations rather than per-model values. As shown in table 12, all four metrics exhibit strong correlation with human judgment (correlation > 0.8). Furthermore, the aggregated *Overall Score* (mean of LSS, VFS, GPS, and LPS) correlates even more strongly with human preference, indicating that jointly evaluating logical, visual, and spatial fidelity provides a more faithful proxy for human assessment.

15.2 Prompts used for evaluations of methods

Refer to fig. 20 for details about the prompts used for evaluations. For QwenVL family of models - Qwen2.5-VL [43], Qwen3-VL [1] and logics-parsing [119], we get the output in Qwen HTML. For Smoldocling [13], we use the model to convert to docling format, and postprocess it into a HTML. For all other models, we use a standard prompt.

15.3 Prompts used for VLM-as-a-judge metrics

The prompts used for Visual Fidelity Score (VFS) can be found in fig. 22 and Logical Structure Score (LSS) in fig. 21. All VLM-as-a-judge evaluations are conducted using Gemini 2.5 Pro.

15.4 Overall Score Calculation

We report an *Overall Score* that aggregates all four evaluation dimensions—Text Extraction (TE), Logical Structure (LS), Physical Structure (PS), and Visual Fidelity (VF)—using a simple and interpretable

Qwen HTML Prompt

Models:

1. Qwen3-VL-30B-A3B-Instruct
2. Qwen2.5-VL-7B-Instruct
3. Logics-Parsing

Prompt: For the given document, give its Qwen HTML.

Plain HTML Prompt

Prompt: For the given document, give its HTML. Make sure to include the accurate text, nested hierarchical structure, styling and positions of the boxes. (Return clean, well-indented, semantic HTML, retaining style, color, and position.)

SmolDocling Prompt

Prompt: Convert this page to docling.

Fig. 20: Prompt used for different models for HTML generation.

formulation:

$$\text{Overall} = \frac{\text{TE} + \text{LS} + \text{PS} + \text{VF}}{4}, \quad \text{TE} = \frac{\text{CRR} + \text{WRR}}{2}, \quad \text{PS} = \frac{\text{GPS} + \text{LPS}}{2}.$$

Because VFDR-BENCH is multilingual and includes scripts with highly variable OCR difficulty, extremely poor model outputs can yield $\text{CER}/\text{WER} > 1$, resulting in negative CRR/WRR values. To prevent such extreme failures from disproportionately influencing the final score, we apply a lower bound of -1 to CRR , WRR for each sample, and consequently to TE for the aggregate. This clipping ensures that the Overall Score remains stable and comparable across languages and models, while still penalizing severe errors appropriately. By limiting the impact of pathological cases, the aggregated score provides a fair, robust summary of model performance across all VFDR dimensions.

16 Additional Results

These results mirror those reported in the main paper. The overall score exhibits trends similar to the individual metrics because the same methods that perform well on each evaluation axis also achieve strong aggregate performance across all four axes. Additional details can be found in table 14, table 13, table 15 and table 16.

17 Qualitative Samples of REPLICA conversions

Sample conversions across diverse datasets, covering multiple languages and document types, are provided in figs. 23 to 40. These examples have source image in the left, and where both the rendered output in corresponding FID-HTML representations are available.

Logical Structure Score: VLM-as-a-Judge Prompt

You are an HTML evaluation assistant.

You will be given two HTML documents:

- Ground-Truth HTML (GT) | the reference HTML
- Predicted HTML (PRED) | the model-generated HTML

Your task is to output two numeric scores in valid JSON:

```
{
  "structure_score": <float>,
  "semantic_score": <float>
}
```

1. structure_score (0.0 to 1.0)

Quantifies similarity of the document's overall hierarchy and layout organization.

Focus on the arrangement, nesting, and hierarchy of major elements, while remaining tolerant to minor structural variations such as extra wrappers, flattened hierarchies, or slight reordering. Penalize only when the logical organization is clearly disrupted.

Score Guidelines:

```
1.0   → nearly identical structures
0.8 to 0.9 → same high-level structure with few differences
0.5 to 0.7 → generally similar, partial hierarchy loss
0.2 to 0.4 → weak structural resemblance
0.0 to 0.1 → unrelated or unrecognizable structure
```

2. semantic_score (0.0 to 1.0)

Evaluates preservation of semantic meaning and tag usage.

Check for the correct use of key tags such as <p>, <h1>, , and <table>, while remaining lenient toward minor tag substitutions like <div> vs. <section> when the intent is preserved. Ignore whitespace, attributes, and other trivial stylistic variations.

Score Guidelines:

```
1.0   → all key semantics correct
0.8 to 0.9 → most tags/roles correct, few errors
0.5 to 0.7 → about half of key meanings preserved
0.2 to 0.4 → weak semantic resemblance
0.0 to 0.1 → major semantic loss or mismatch
```

Important: Always output strictly valid JSON in the format:

```
{
  "structure_score": <float>,
  "semantic_score": <float>
}
```

Fig. 21: Prompt used for computing the Logical Structure Score (LSS) via a VLM-as-a-Judge evaluation of structural and semantic fidelity.

Visual Fidelity Score: VLM-as-a-Judge Prompt

You are an expert in document rendering evaluation.
You will be given two images:

- Image A = Predicted rendering (model output) (First Image)
- Image B = Ground truth rendering (reference) (Second Image)

Your task is to evaluate how visually faithful Image A is compared to Image B.

Instructions

1. Compare Image A and Image B across all aspects:
 - Positioning accuracy (High priority)
 - Color schemes and visual styling (High priority)
 - Graphics and image elements
 - Text content and typography fidelity
 - Table structures and list formatting
 - General visual appearance and presentation
2. Based on the overall comparison, provide:
 - A single final fidelity score (0.0 to 1.0)
 - 1.0 = perfect match
 - 0.0 = complete mismatch
 - Intermediate values proportionally reflect quality.
 - A list of reasons describing mismatches and differences, grouped by aspect.

Output Format (Strict)

Final Fidelity Score: <score between 0.0 and 1.0>

Reasons:

- <reason 1>
- <reason 2>
- <reason 3>
- ...

Fig. 22: Prompt used for computing the Visual Fidelity Score (VFS) via a VLM-as-a-Judge evaluation of visual fidelity.

Table 13: Domain-wise Overall Scores (OS) across different document categories. Baseline methods are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML . Document categories are abbreviated as: GR = Government Regulatory, FD = Financial, BD = Business, HC = Healthcare, NV = Novels, LG = Legal, AQ = Assignments/Question Papers, TB = Textbooks, RP = Research Papers, MN = Manuals, BR = Brochures, FM = Forms, RS = Resumes, RI = Receipts/Invoices, ID = ID Documents/Certificates, LN = Lecture Notes, OT = Others.

| Model | GR | FD | BD | HC | NV | LG | AQ | TB | RP | MN | BR | FM | RS | RI | ID | LN | OT |
|-------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Marker [3] | 0.62 | <u>0.61</u> | <u>0.64</u> | 0.65 | 0.61 | 0.68 | <u>0.62</u> | <u>0.60</u> | <u>0.57</u> | <u>0.66</u> | 0.65 | <u>0.58</u> | <u>0.68</u> | <u>0.50</u> | <u>0.54</u> | 0.67 | 0.62 |
| Docling [26] | 0.36 | <u>0.50</u> | <u>0.50</u> | 0.51 | 0.29 | 0.49 | 0.35 | 0.40 | 0.49 | <u>0.50</u> | 0.40 | <u>0.44</u> | <u>0.58</u> | 0.28 | <u>0.19</u> | 0.54 | 0.41 |
| RolmOCR [116] | 0.33 | 0.34 | 0.42 | 0.38 | 0.41 | 0.41 | 0.39 | 0.34 | 0.39 | 0.36 | 0.36 | 0.29 | 0.49 | 0.44 | 0.29 | 0.52 | 0.40 |
| smolDocling [13] | 0.31 | 0.30 | 0.40 | 0.36 | 0.32 | 0.45 | 0.21 | 0.33 | 0.44 | 0.55 | 0.40 | 0.20 | 0.48 | 0.32 | 0.24 | 0.40 | 0.37 |
| olmOCR-7B [66] | 0.29 | 0.34 | 0.38 | 0.25 | 0.29 | 0.40 | 0.41 | 0.35 | 0.31 | 0.40 | 0.48 | 0.24 | 0.26 | 0.03 | 0.28 | 0.52 | 0.43 |
| Nanonets-OCR-s [117] | 0.27 | 0.35 | 0.39 | 0.26 | 0.35 | 0.45 | 0.38 | 0.31 | 0.36 | 0.36 | 0.22 | 0.25 | 0.34 | 0.03 | 0.01 | 0.52 | 0.37 |
| OCRFlux-3B [118] | 0.34 | 0.43 | 0.46 | 0.39 | 0.34 | 0.48 | 0.49 | 0.40 | 0.43 | 0.45 | 0.38 | 0.38 | 0.65 | 0.07 | 0.09 | 0.63 | 0.44 |
| Logics-Parsing [119] | 0.54 | 0.59 | 0.62 | 0.54 | 0.61 | 0.64 | 0.58 | 0.56 | 0.52 | 0.64 | <u>0.60</u> | <u>0.58</u> | 0.65 | 0.51 | 0.33 | 0.60 | <u>0.58</u> |
| GOT-OCR-2.0 [27] | 0.29 | 0.31 | 0.45 | 0.40 | 0.32 | 0.50 | 0.39 | 0.34 | 0.42 | 0.45 | 0.40 | 0.09 | 0.48 | 0.12 | 0.35 | 0.53 | 0.36 |
| Qwen2.5-VL-7B-Instruct [43] | 0.38 | 0.37 | 0.43 | 0.44 | 0.49 | 0.45 | 0.46 | 0.44 | 0.45 | 0.42 | 0.57 | 0.47 | 0.60 | 0.47 | 0.41 | 0.49 | 0.47 |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.29 | 0.27 | 0.40 | 0.29 | 0.40 | 0.54 | 0.35 | 0.46 | 0.43 | 0.36 | 0.42 | 0.26 | 0.40 | 0.08 | 0.12 | 0.61 | 0.45 |
| gemma-3-27b-it [28] | 0.40 | 0.40 | 0.44 | 0.50 | 0.44 | 0.53 | 0.47 | 0.43 | 0.40 | 0.40 | 0.48 | 0.33 | 0.60 | 0.37 | 0.36 | 0.38 | 0.53 |
| InternVL3-14B [46] | 0.36 | 0.41 | 0.43 | 0.34 | 0.32 | 0.43 | 0.49 | 0.46 | 0.34 | 0.47 | 0.45 | 0.37 | 0.49 | 0.38 | 0.27 | 0.23 | 0.41 |
| Gemini-2.5-Pro [120] | 0.50 | 0.51 | 0.54 | 0.47 | 0.48 | 0.56 | 0.57 | 0.54 | 0.43 | 0.53 | 0.46 | 0.61 | 0.63 | 0.46 | 0.35 | 0.51 | <u>0.56</u> |
| GPT-5 [121] | <u>0.60</u> | 0.63 | 0.65 | <u>0.64</u> | <u>0.58</u> | <u>0.67</u> | 0.65 | 0.63 | 0.55 | 0.69 | 0.65 | 0.57 | 0.72 | <u>0.48</u> | 0.67 | <u>0.65</u> | 0.62 |
| WebCoder-1.3B [79] | 0.25 | 0.25 | 0.36 | 0.20 | 0.24 | 0.38 | 0.17 | 0.30 | 0.26 | 0.35 | 0.18 | 0.05 | 0.23 | 0.00 | — | 0.22 | 0.35 |
| WebSight-VLM-7B [29] | 0.26 | 0.29 | 0.37 | 0.24 | 0.29 | 0.39 | 0.16 | 0.25 | 0.25 | 0.36 | 0.15 | 0.07 | 0.24 | — | — | 0.21 | 0.30 |

Table 14: Language-wise Overall Scores across models. The table shows model performance across languages. Baseline methods are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML . Languages are abbreviated as: en = English, hi = Hindi, bn = Bengali, ta = Tamil, mr = Marathi, zh = Chinese, ru = Russian, mixed = Code Mixed, oth = Others.

| Model | en | hi | bn | ta | mr | zh | mixed | oth |
|-------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Docling [26] | 0.52 | 0.15 | 0.23 | 0.16 | 0.22 | 0.16 | 0.38 | 0.24 |
| Marker [3] | <u>0.64</u> | 0.70 | 0.60 | <u>0.61</u> | 0.55 | 0.46 | 0.60 | 0.63 |
| Nanonets-OCR-s [117] | 0.39 | 0.43 | 0.35 | 0.03 | 0.24 | 0.30 | 0.26 | 0.24 |
| RolmOCR [116] | 0.42 | 0.43 | 0.36 | 0.11 | 0.29 | 0.29 | 0.32 | 0.28 |
| olmOCR-7B [66] | 0.39 | 0.28 | 0.25 | - | 0.42 | 0.30 | 0.29 | 0.17 |
| smolDocling [13] | 0.44 | 0.27 | 0.27 | 0.18 | 0.13 | 0.11 | 0.27 | 0.35 |
| OCRFlux-3B [118] | 0.48 | 0.13 | 0.29 | 0.06 | 0.41 | 0.30 | 0.35 | 0.30 |
| GOT-OCR-2.0 [27] | 0.46 | 0.04 | 0.09 | 0.05 | - | 0.18 | 0.30 | 0.13 |
| Logics-Parsing [119] | 0.61 | <u>0.62</u> | <u>0.47</u> | 0.41 | <u>0.55</u> | 0.49 | <u>0.54</u> | 0.48 |
| Gemini-2.5-Pro [120] | 0.52 | 0.61 | 0.46 | 0.69 | 0.57 | 0.44 | 0.47 | 0.51 |
| GPT-5 [121] | 0.66 | 0.53 | 0.38 | 0.46 | 0.58 | 0.49 | 0.58 | <u>0.58</u> |
| Gemma-3-27b-it [28] | 0.46 | 0.30 | 0.23 | 0.16 | 0.47 | 0.28 | 0.42 | 0.42 |
| Qwen2.5-VL-7B-Instruct [43] | 0.46 | 0.48 | 0.39 | 0.55 | 0.46 | <u>0.46</u> | 0.38 | 0.38 |
| InternVL3-14B [46] | 0.43 | 0.22 | 0.06 | 0.06 | 0.05 | 0.43 | 0.37 | 0.26 |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.43 | 0.40 | 0.21 | 0.28 | 0.21 | 0.30 | 0.31 | 0.26 |
| WebCoder-1.3B [79] | 0.36 | 0.01 | 0.03 | - | - | 0.06 | 0.21 | 0.09 |
| WebSight-VLM-7B [29] | 0.40 | - | 0.06 | - | - | 0.11 | 0.26 | 0.13 |

Table 15: Comparison of **Overall Score (OS)** on single-column and multi-column document sets. Models are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML .

| Model | SC | MC |
|-------------------------------|-------------|-------------|
| Marker [3] | 0.62 | 0.61 |
| Docling [26] | 0.45 | 0.36 |
| RolmOCR [116] | 0.39 | 0.31 |
| smolDocling [13] | 0.37 | 0.32 |
| olmOCR-7B [66] | 0.35 | 0.29 |
| Nanonets-OCR-s [117] | 0.35 | 0.26 |
| OCRFlux-3B [118] | 0.42 | 0.36 |
| Logics-Parsing [119] | 0.59 | 0.49 |
| GOT-OCR-2.0 [27] | 0.39 | 0.27 |
| Qwen2.5-VL-7B-Instruct [43] | 0.44 | 0.37 |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.39 | 0.28 |
| gemma-3-27b-it [28] | 0.45 | 0.34 |
| InternVL3-14B [46] | 0.40 | 0.37 |
| Gemini-2.5-Pro [120] | 0.51 | 0.47 |
| GPT-5 [121] | 0.63 | 0.55 |
| WebCoder-1.3B [79] | 0.32 | 0.22 |
| WebSight-VLM-7B [29] | 0.31 | 0.23 |

Table 16: Comparison of **Overall Score (OS)** on Scanned and Born-Digital document sets. Models are grouped into four categories: Pipeline Tools , Expert VLMs , General VLMs , and Screenshot2HTML .

| Model | SC | BD |
|-------------------------------|-------------|-------------|
| Marker [3] | 0.61 | 0.64 |
| Docling [26] | 0.41 | 0.61 |
| RolmOCR [116] | 0.37 | 0.42 |
| smolDocling [13] | 0.35 | 0.43 |
| olmOCR-7B [66] | 0.32 | 0.45 |
| Nanonets-OCR-s [117] | 0.33 | 0.37 |
| OCRFlux-3B [118] | 0.40 | 0.51 |
| Logics-Parsing [119] | 0.57 | 0.59 |
| GOT-OCR-2.0 [27] | 0.35 | 0.48 |
| Qwen2.5-VL-7B-Instruct [43] | 0.41 | 0.53 |
| Qwen3-VL-30B-A3B-Instruct [1] | 0.35 | 0.51 |
| gemma-3-27b-it [28] | 0.42 | 0.46 |
| InternVL3-14B [46] | 0.37 | 0.51 |
| Gemini-2.5-Pro [120] | 0.50 | 0.54 |
| GPT-5 [121] | 0.61 | 0.66 |
| WebCoder-1.3B [79] | 0.29 | 0.36 |
| WebSight-VLM-7B [29] | 0.27 | 0.41 |

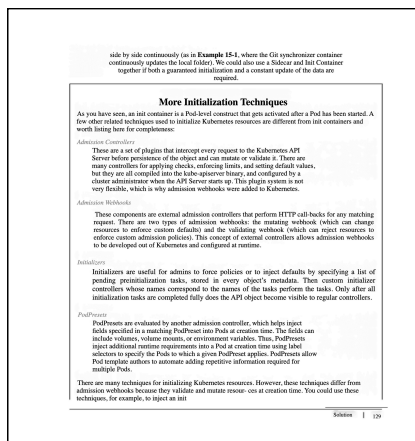
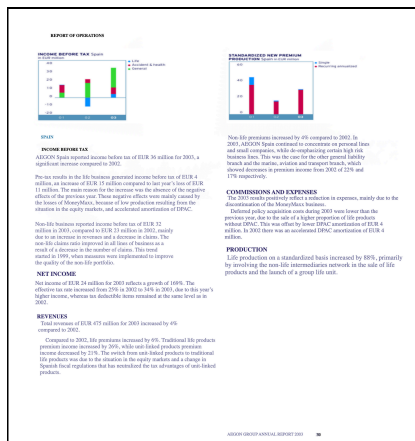


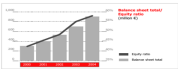
Fig. 23: Samples from DocLayNet dataset

PDF

Document (.pdf)

Equity Ratio at 57.6%

Once again, the capital structure was successfully improved in 2004. Despite the 22.8% increase in the balance sheet total from € 201.1 million to € 250.6 million, the equity ratio rose up from 54.7% to 57.6%. This development underlines the strong financial position of the PUMA Group.



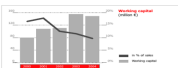
Further Improvement in Liquidity

As a result of the strongly improved cashflow, cash and cash equivalents almost doubled to € 368.9 million. Conversely, bank debt was reduced from € 16.8 million to € 12.0 million.

As a result, net liquidity improved from € 173.8 million to € 356.4 million.

Working Capital Again Improved

Working capital amounted to 9.7% of sales after 12.2% in the previous year. In absolute figures, the working capital declined by 4.2% at year-end, falling from € 155.7 million to € 148.6 million. This development, which is related to increased receivables, is primarily due to marginal inventory build-ups. A higher amount of receivables was recorded in December 2003, whereas the year-end 2004 indicates a shift to January 2005. The calculation of working capital includes inventory, plus current receivables, less current liabilities. It is not a measure of long-term solvency and provisions to the extent attributable to the operating area.



Inventory were up by 2.5% to € 201.1 million and receivables rose by 19.1 to € 188.1 million. This development confirms PUMA's systematic and effective working capital management.

Revenue | Mission | Management Report | Sales | Marketing | Consolidated Financial Statements | IFRS | Report of Supervisory Board | Board of Management

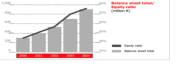
25

HTML

Rendered Fid-HTML

Equity Ratio at 57.6%

Once again, the capital structure was successfully improved in 2004. Despite the 22.8% increase in the balance sheet total from € 201.1 million to € 250.6 million, the equity ratio rose up from 54.7% to 57.6%. This development underlines the strong financial position of the PUMA Group.



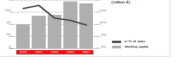
Further Improvement in Liquidity

As a result of the strongly improved cashflow, cash and cash equivalents almost doubled to € 368.9 million. Conversely, bank debt was reduced from € 16.8 million to € 12.0 million.

As a result, net liquidity improved from € 173.8 million to € 356.4 million.

Working Capital Again Improved

Working capital amounted to 9.7% of sales after 12.2% in the previous year. In absolute figures, the working capital declined by 4.2% at year-end, falling from € 155.7 million to € 148.6 million. This development, which is related to increased receivables, is primarily due to marginal inventory build-ups. A higher amount of receivables was recorded in December 2003, whereas the year-end 2004 indicates a shift to January 2005. The calculation of working capital includes inventory, plus current receivables, less current liabilities. It is not a measure of long-term solvency and provisions to the extent attributable to the operating area.



Inventory were up by 2.5% to € 201.1 million and receivables rose by 19.1 to € 188.1 million. This development confirms PUMA's systematic and effective working capital management.

Revenue | Mission | Management Report | Sales | Marketing | Consolidated Financial Statements | IFRS | Report of Supervisory Board | Board of Management

25

Step 4: For a Precured Patch

Apply film adhesive or paste adhesive to the damaged area and place the precured patch on top. Vacuum bag the repair and cure at the correct temperature for the film adhesive or paste adhesive. Most film adhesives cure at either 250 °F or 350 °F. Some paste adhesives cure at room temperature, although an elevated temperature could be used to speed the curing process.

| Bonded versus bonded repair | Bonded | Bonded |
|--|--------|--------|
| Lightly loaded structures – laminate thickness less than 0.1" | X | X |
| Lightly loaded structures – laminate thickness between 0.125" – 0.5" | X | X |
| Lightly loaded structures – laminate thickness larger than 0.5" | X | |
| High loading situations | | X |
| Honeycomb structure | | X |
| Dry surfaces | X | X |
| Wet and/or contaminated surfaces | X | |
| Discontinuity required | X | |
| Restores correct strength | | X |

Bonded versus Bonded Repair

Bonded repair concepts have found applicability in both types of manufacturing assembly methods. They have the advantage of not introducing stress concentrations by drilling fastener holes for patch installation and can be stronger than original part material. The disadvantage of bonded repairs is that most repair materials require special storage, handling, and curing procedures.

Bonded repairs are quicker and easier to fabricate than bonded repairs. They are normally used on composite skin thicker than 0.125 inch to ensure sufficient fastener bearing area is available for load transfer. They are prohibited in honeycomb sandwich assemblies due to the potential for moisture intrusion from the fastener holes and the resulting core degradation. Bonded repairs are heavier than comparable bonded repairs, limiting their use on weight sensitive flight control surfaces.

Honeycomb sandwich parts often have thin face sheets and are most effectively repaired by using a bonded scarf type repair. A bonded external step patch can be used as an alternative. Bonded repairs are not effective for thin laminates, because of the low bearing stress of the composite laminate. Thicker solid laminates tend on larger internal core to be up to an inch thick in highly loaded areas and these types of laminates cannot be effectively repaired using a bonded scarf type repair (Figure 7-74).

Bonded Repairs

Alcock designed in the 1970s used composite sandwich honeycomb structure for lightly loaded secondary structure, but now large aircraft use thick solid laminates for primary structure instead of sandwich honeycomb. These thick solid laminate structures are quite different from the traditional sandwich honeycomb structures used for flight controls, landing gear doors, flap, and spoilers of today's aircraft. They present a challenge to repair and are difficult to repair with a bonded repair method. Bonded repair methods have been developed to repair thicker solid laminates.

Bonded repairs are not desirable for honeycomb sandwich structures due to the limited bearing strength of the thin face sheets and weakened honeycomb structure from drilling

Step 1: Inspection of the Damage

The tap test is not effective to detect delamination in thick laminates unless the damage is close to the surface. An

Step 4: For a Precured Patch

Apply film adhesive or paste adhesive to the damaged area and place the precured patch on top. Vacuum bag the repair and cure at the correct temperature for the film adhesive or paste adhesive. Most film adhesives cure at either 250 °F or 350 °F. Some paste adhesives cure at room temperature, although an elevated temperature could be used to speed the curing process.

| Bonded versus bonded repair | Bonded | Bonded |
|--|--------|--------|
| Lightly loaded structures – laminate thickness less than 0.1" | X | X |
| Lightly loaded structures – laminate thickness between 0.125" – 0.5" | X | X |
| Lightly loaded structures – laminate thickness larger than 0.5" | X | |
| High loading situations | | X |
| Honeycomb structure | | X |
| Dry surfaces | X | X |
| Wet and/or contaminated surfaces | X | |
| Discontinuity required | X | |
| Restores correct strength | | X |

Bonded versus Bonded Repair

Bonded repair concepts have found applicability in both types of manufacturing assembly methods. They have the advantage of not introducing stress concentrations by drilling fastener holes for patch installation and can be stronger than original part material. The disadvantage of bonded repairs is that most repair materials require special storage, handling, and curing procedures.

Bonded repairs are quicker and easier to fabricate than bonded repairs. They are normally used on composite skin thicker than 0.125 inch to ensure sufficient fastener bearing area is available for load transfer. They are prohibited in honeycomb sandwich assemblies due to the potential for moisture intrusion from the fastener holes and the resulting core degradation. Bonded repairs are heavier than comparable bonded repairs, limiting their use on weight sensitive flight control surfaces.

Honeycomb sandwich parts often have thin face sheets and are most effectively repaired by using a bonded scarf type repair. A bonded external step patch can be used as an alternative. Bonded repairs are not effective for thin laminates because of the low bearing stress of the composite laminate. Thicker solid laminates tend on larger internal core to be up to an inch thick in highly loaded areas and these types of laminates cannot be effectively repaired using a bonded scarf type repair (Figure 7-74).

Bonded Repairs

Alcock designed in the 1970s used composite sandwich honeycomb structure for lightly loaded secondary structure, but now large aircraft use thick solid laminates for primary structure instead of sandwich honeycomb. These thick solid laminate structures are quite different from the traditional sandwich honeycomb structures used for flight controls, landing gear doors, flap, and spoilers of today's aircraft. They present a challenge to repair and are difficult to repair with a bonded repair method. Bonded repair methods have been developed to repair thicker solid laminates.

Bonded repairs are not desirable for honeycomb sandwich structures due to the limited bearing strength of the thin face sheets and weakened honeycomb structure from drilling

Step 1: Inspection of the Damage

The tap test is not effective to detect delamination in thick laminates unless the damage is close to the surface. An

Fig. 24: Samples from DocLayNet dataset



Document (.pdf)



Rendered Fid-HTML

CASE FORM

CASE NAME: Wanda G. Robinson and Carroll Robinson v. Raytheon-Machefran, et al.

COURT: San Francisco Superior Court - No. 990378

LORELLARD ENTITIES: Lorillard Tobacco Company

DATE FILED: July 23, 1998

DATE SERVED: August 3, 1998

CASE TYPE: Adversus

PLAINTIFFS COUNSEL: Wanda G. Robinson, Harwitz, Smith & Tigerman
Madelyn J. Chabot
101 California Street, Suite 2200
San Francisco, California 94111
415/965-5566

LORELLARD COUNSEL:

JUDGE:

TRIAL DATE:

82491256

CASE FORM

CASE NAME: Wanda G. Robinson and Carroll Robinson v. Raytheon-Machefran, et al.

COURT: San Francisco Superior Court - No. 990378

LORELLARD ENTITIES: Lorillard Tobacco Company

DATE FILED: July 23, 1998

DATE SERVED: August 3, 1998

CASE TYPE: Adversus

PLAINTIFFS COUNSEL: Wanda G. Robinson, Harwitz, Smith & Tigerman
Madelyn J. Chabot
101 California Street, Suite 2200
San Francisco, California 94111
415/965-5566

LORELLARD COUNSEL:

JUDGE:

TRIAL DATE:

XC: R.B.S.
12-9-99
R.B. SPELL
DEC - 9 1999

COVINGTON & BURLING

1201 Pennsylvania Avenue, N.W.
P.O. Box 7566
Washington, D.C. 20044-7566
Fax Numbers (202) 662-4291 or (202) 737-0528
Fax Operator (202) 662-4280

THIS FACSIMILE TRANSMISSION IS INTENDED ONLY FOR THE ADDRESSEE SHOWN BELOW. IF YOU RECEIVE INFORMATION THAT IS UNLAWFUL, CONFIDENTIAL OR OTHERWISE PROTECTED FROM DISCLOSURE, ANY REVIEW, REPRODUCTION OR USE OF THIS TRANSMISSION OR ITS CONTENTS BY PERSONS OTHER THAN THE ADDRESSEE IS STRICTLY PROHIBITED. IF YOU HAVE RECEIVED THIS TRANSMISSION IN ERROR, PLEASE NOTIFY US IMMEDIATELY AND MAIL THE ORIGINAL TO US AT THE ABOVE ADDRESS.

Date: December 9, 1999

To: Haney H. Bell, Esq.

From: David H. Remes
(202) 778-5212 - direct fax

Room: 803E

// Pages (including cover)

MESSAGE:

82573101

XC: R.B.S.
12-9-99
R.B. SPELL
DEC - 9 1999

COVINGTON & BURLING

1201 Pennsylvania Avenue, N.W.
P.O. Box 7566
Washington, D.C. 20044-7566
Fax Numbers (202) 662-4291 or (202) 737-0528
Fax Operator (202) 662-4280

THIS FACSIMILE TRANSMISSION IS INTENDED ONLY FOR THE ADDRESSEE SHOWN BELOW. IF YOU RECEIVE INFORMATION THAT IS UNLAWFUL, CONFIDENTIAL OR OTHERWISE PROTECTED FROM DISCLOSURE, ANY REVIEW, REPRODUCTION OR USE OF THIS TRANSMISSION OR ITS CONTENTS BY PERSONS OTHER THAN THE ADDRESSEE IS STRICTLY PROHIBITED. IF YOU HAVE RECEIVED THIS TRANSMISSION IN ERROR, PLEASE NOTIFY US IMMEDIATELY AND MAIL THE ORIGINAL TO US AT THE ABOVE ADDRESS.

Date: December 9, 1999

To: Haney H. Bell, Esq.

From: David H. Remes
(202) 778-5212 - direct fax

Room: 803E

// Pages (including cover)

MESSAGE:

82573101

Fig. 25: Samples from FUNSD dataset



Document (.pdf)



Rendered Fid-HTML

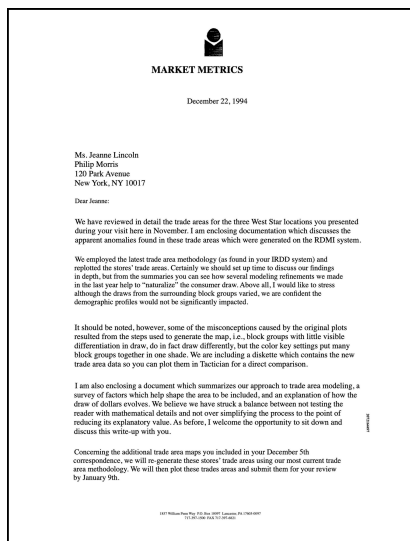
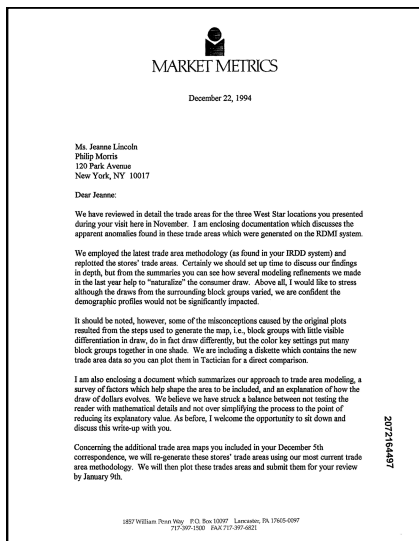
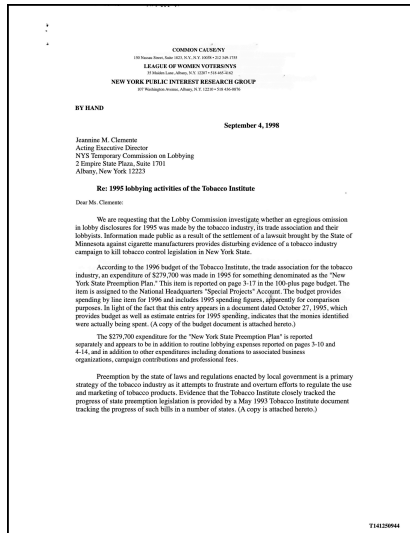
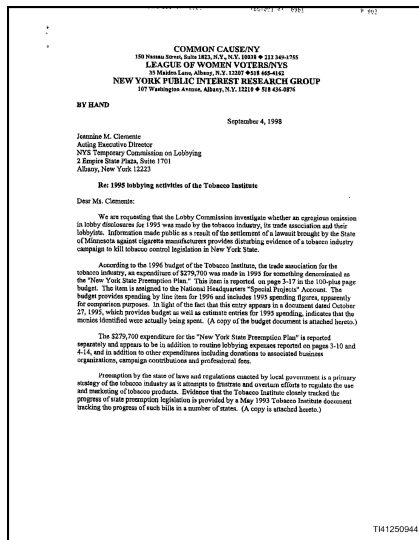


Fig. 26: Samples from Industry Document Library dataset



Document (.pdf)



Rendered Fid-HTML

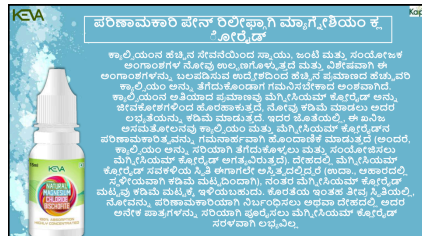
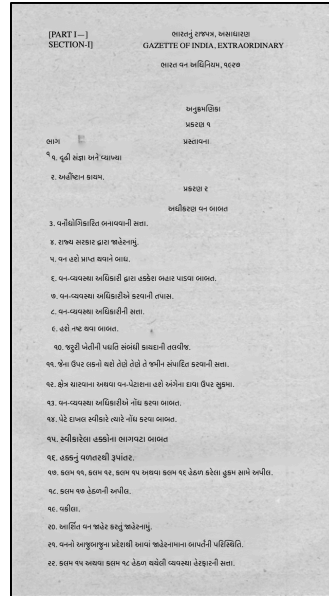
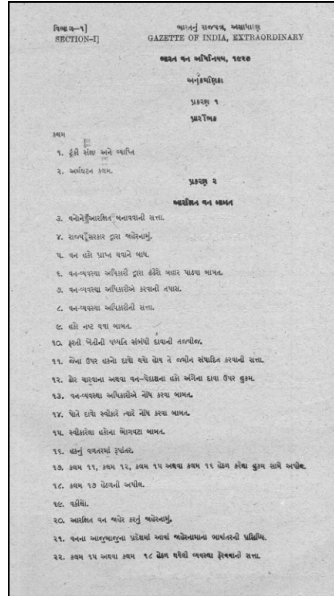
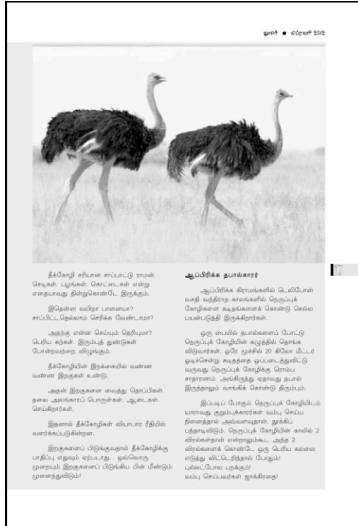


Fig. 27: Samples from IndicDLP dataset



Document (.pdf)



Rendered Fid-HTML

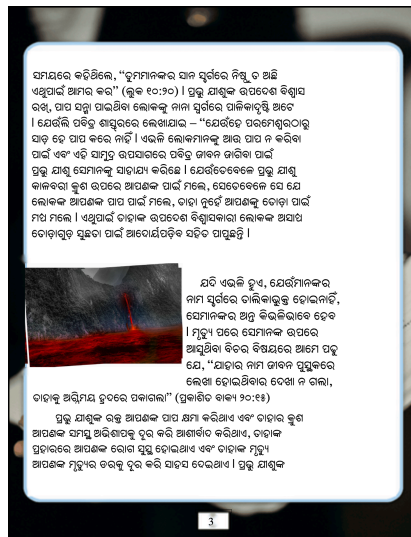
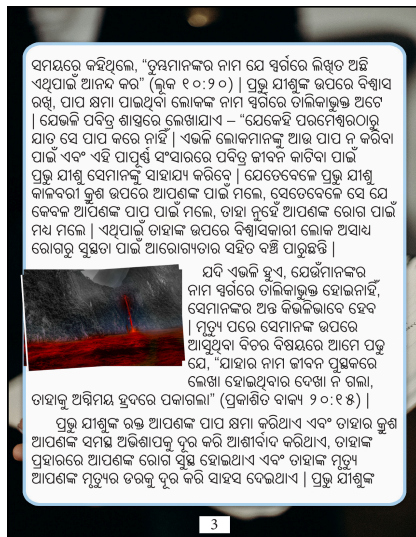
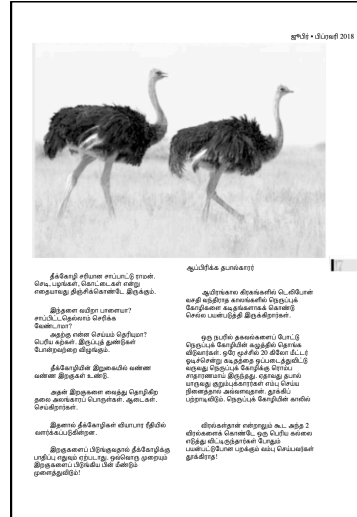


Fig. 28: Samples from IndicDLP dataset



- [illegible]

- पाकिस्तान के जमीन-परीक्षा को अर्धतंत्र रूप से, पूरा करना ऐसे किसी भी कारण पर विचार किए बिना, पाकिस्तान में विवेक नहीं लेना चाहिए। विदेशी का सम्बन्ध सम्बन्धित क्षेत्रों के नेताओं, अधिकारियों और अन्य लोगों के सम्बन्ध में अत्यन्त सावधान रहना चाहिए और उनके संविधिक अधिकारों के साथ सम्बन्धित क्षेत्रों की प्रजातन्त्रिक विधि का सम्मान करना चाहिए।
- खानों की सम्पत्ति यास्तविक अधिकारों में दी जानी चाहिए, जबकि नगरपालिकाओं की राशियाँ जाल साधक सम्पत्ति में देवानी जानी चाहिए। जस्टिस निवेदित करने वाले में पृथक्गीर्ण की जाना चाहिए। उपरान्त 4,56,100 रुपये की राशि 5 देवानी जमीन पर प्रति हे. 4.6 या 5,00,00,000 रुपये। उसी प्रस्ताव 61,49,500 रुपये की राशि को 61 देवानी जाल, 61.5 या 61,00,000 नती।
- उक्त विधायक सम्पत्ति (कम्पनी अधिनियम, 1956 की धारा 2 में यथा परिभाषित) द्वारा हस्ताक्षरित होने चाहिए और यदि ऐसा कोई प्रस्ताव नहीं है, तो यथा निर्देश अथवा संश्लिष्ट कम्पनी के किसी ऐसे अधिकारी को हस्ताक्षर करना चाहिए जिसे निदेशक समूह ने अधिक प्राधिकृत किया है और निम्न कम्पनी हस्ताक्षर इस प्रमाणों के लिए पर्याप्त दिवस तक को जना गया है। यदि मनुष्य हस्ताक्षर विधिक अधिकारों में प्रस्तुत नहीं करता है, तो उस व्यक्ति को प्राधिकृत अधिकारी द्वारा हस्ताक्षरित किया जाना चाहिए और उसका मनुष्य हस्ताक्षर अलग से प्रस्तुत करना चाहिए।
- विदेशी के किसी भाग में, "कृषि क्षेत्रों हेतु विधि कृषि नहीं है" से संबंधित जाल पर "दर" खाता की संख्या के लिए रखे स्तंभ में "रिपोर्ट नहीं" लिखा जाना और राशि के लिए रखे जाल पर "00" दर्शाया जाना चाहिए।
- इन विधायकों में उल्लिखित "हस्ताक्षर कम्पनी" और "कृषि ही मनुष्य की कम्पनी" का आशय कम्पनी अधिनियम, 1956 की धाराएं 400 और 372(1)(1) में दिए गए अर्थ में है जो कि 31 अक्टूबर 1998 को कम्पनी अधिनियम में दिए गए संशोधन के पहले दिवस तक का है।
- यदि यह विदेशी इन्वेंटरी भाष्य (इंटेलिजेन्स से निर्दिष्ट विषय संरक्षित) का यास्त्री हिस्क साजुन 3.5 द्वारा भी जा रही है तो उसकी विधिपर हस्ताक्षर मुद्रित प्रति संबंधित क्षेत्रीय कानूनन्य को प्रस्तुत की जानी चाहिए।



Fig. 29: Samples from IndicDLP dataset

Document (.pdf)

Rendered Fid-HTML

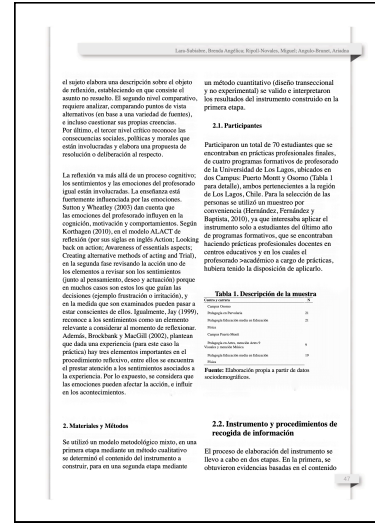
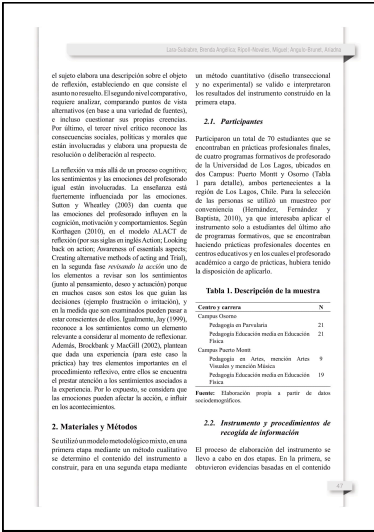
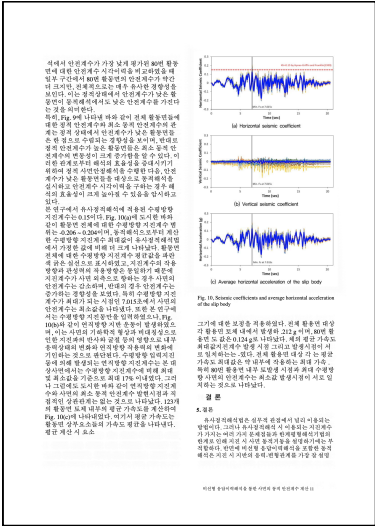
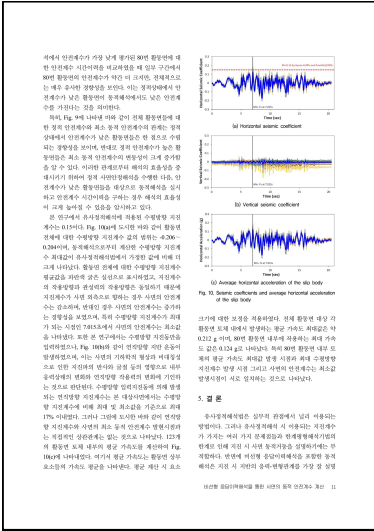


Fig. 30: Samples from LoRaLay dataset [133]



54



Document (.pdf)



Rendered Fid-HTML



Fig. 32: Samples from M⁶Doc dataset

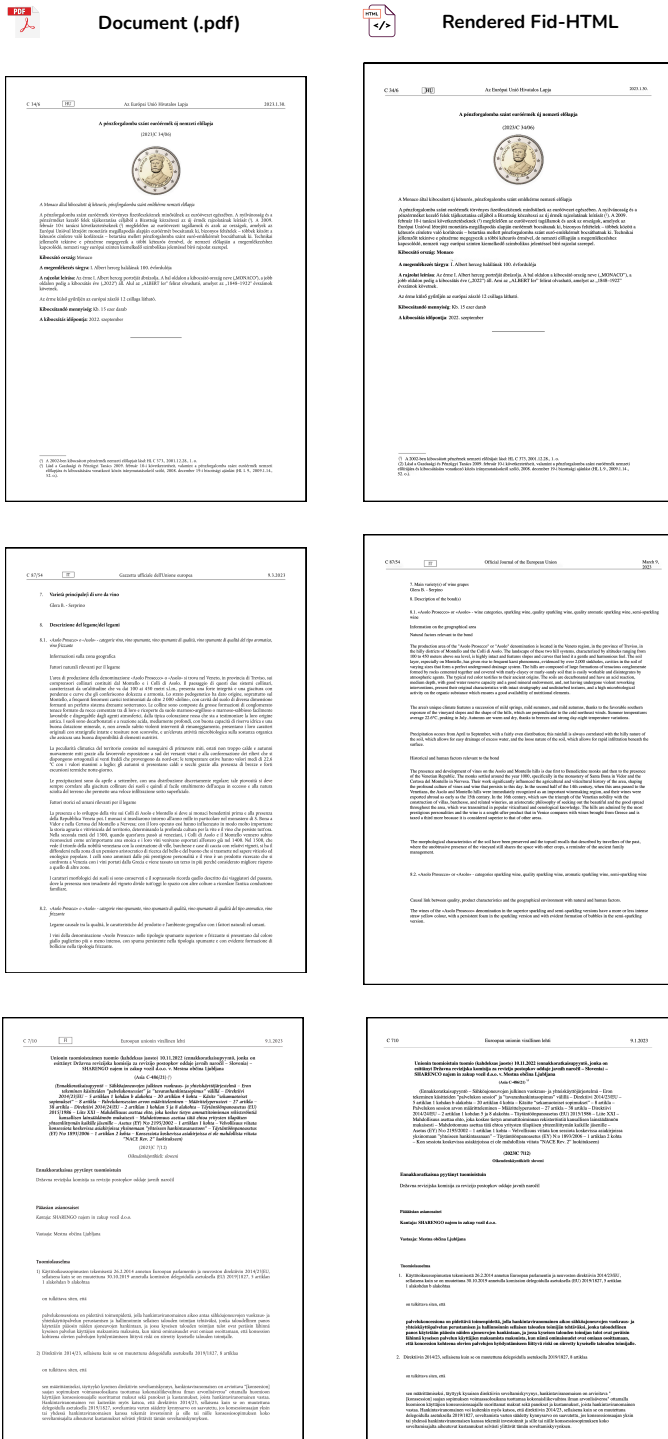


Fig. 33: Samples from OJ4OCRMT dataset



Rendered Fid-HTML

[illegible]

Fig. 34: Samples from OJ4OCRMT dataset



Document (.pdf)



Rendered Fid-HTML

10 Questions. She's a Harvard grad, an acclaimed actress and a humanitarian, but she says she's no role model. Her new movie is *The Other Boleyn Girl*. **Natalie Portman will now take your questions**

What have you learned about yourself by portraying powerful women? Dani Arpa, TORONTO It has encouraged me to say things authoritatively. Often women prefer what they say with "I know this might sound stupid" or "I don't mean to be aggressive, but..." I tend to do that, so it's great to have the opportunity to play a leader.

You take on a pivotal place in *The Other Boleyn Girl*. What is your favorite time in history? Nikki Barnes, YONK, NY I'm really interested in Anne Boleyn. I read this great book by Antonia Fraser called *The Six Wives of Henry VIII*. It's about how life in Britain right before the war. The whole environment of the slow and all this culture—there was a real openness and freedom. It was my wish that the response to that was this incredible fashion.

Does knowing you are a role model affect what parts you choose to play? Nina Cheng, HERNDON, VA I don't consider myself a role model—I make mistakes all the time. It's more about how I want to portray women and myself. I played a stripper in *Clover* and now I know me how many scripts I get where the woman is a stripper or a prostitute. I also have had a lot of good girl things—it's such a virgin where things with female roles.

You once shared your head for a role. Would you do it again? Will Kent, CINCINNATI, OH

As a native of Israel, what role do you think the U.S. government should play in the conflict? Amy Sachs, EVANSTON, ILL. I would love to see a government that made demands on Israel and the Palestinians to reach an agreement. I think it's hard to come from the people themselves, though. No one is going to take an externally imposed solution.

You work with the Foundation for International Community Assistance. How can politicians help change women's lives? Henry Zakem, KANSAS CITY, MO Microfinance is part of the solution. It is an incredible way to give capital to the world's poorest people, mainly women. With these loans, women are able to take agency in their own lives. They don't need to wait around for someone to come help them. We really take that for granted here.

Should celebrities use their clout to influence voters? Spike Condit, BOSTON, MA I think it's important that people can listen to a host of influences and still make their own decisions. I'm always interested to hear what the people respect and voting for. Gloria Steinem wrote this Op Ed in the New York Times that influenced me toward Senator Hillary Clinton.

Would you ever consider running for office? Dan Wallin, CANTON, OH I don't know. I love having a shared head, but the growing our process is really busy. I had some odd haircuts.

How do you feel about the role of today's young stars? Ryan Hunt, WILLOW, IL You're never going to change the fact that it's hard for some people to deal with their lives, but you can change your music. Give them their space and privacy. The worst thing about our society is that it's not in people's difficult times.

To watch a close interview with Natalie Portman and to see the 10 Questions podcast on iTunes, go to time.com/20080321/portman

TIME March 21, 2008

10 Questions. She's a Harvard grad, an acclaimed actress and a humanitarian, but she says she's no role model. Her new movie is *The Other Boleyn Girl*. **Natalie Portman will now take your questions**

What have you learned about yourself by portraying powerful women? Dani Arpa, TORONTO It has encouraged me to say things authoritatively. Often women prefer what they say with "I know this might sound stupid" or "I don't mean to be aggressive, but..." I tend to do that, so it's great to have the opportunity to play a leader.

You take on a pivotal place in *The Other Boleyn Girl*. What is your favorite time in history? Nikki Barnes, YONK, NY I'm really interested in Anne Boleyn. I read this great book by Antonia Fraser called *The Six Wives of Henry VIII*. It's about how life in Britain right before the war. The whole environment of the slow and all this culture—there was a real openness and freedom. It was my wish that the response to that was this incredible fashion.

Does knowing you are a role model affect what parts you choose to play? Nina Cheng, HERNDON, VA I don't consider myself a role model—I make mistakes all the time. It's more about how I want to portray women and myself. I played a stripper in *Clover* and now I know me how many scripts I get where the woman is a stripper or a prostitute. I also have had a lot of good girl things—it's such a virgin where things with female roles.

You once shared your head for a role. Would you do it again? Will Kent, CINCINNATI, OH

As a native of Israel, what role do you think the U.S. government should play in the conflict? Amy Sachs, EVANSTON, ILL. I would love to see a government that made demands on Israel and the Palestinians to reach an agreement. I think it's hard to come from the people themselves, though. No one is going to take an externally imposed solution.

You work with the Foundation for International Community Assistance. How can politicians help change women's lives? Henry Zakem, KANSAS CITY, MO Microfinance is part of the solution. It is an incredible way to give capital to the world's poorest people, mainly women. With these loans, women are able to take agency in their own lives. They don't need to wait around for someone to come help them. We really take that for granted here.

Should celebrities use their clout to influence voters? Spike Condit, BOSTON, MA I think it's important that people can listen to a host of influences and still make their own decisions. I'm always interested to hear what the people respect and voting for. Gloria Steinem wrote this Op Ed in the New York Times that influenced me toward Senator Hillary Clinton.

Would you ever consider running for office? Dan Wallin, CANTON, OH I don't know. I love having a shared head, but the growing our process is really busy. I had some odd haircuts.

How do you feel about the role of today's young stars? Ryan Hunt, WILLOW, IL You're never going to change the fact that it's hard for some people to deal with their lives, but you can change your music. Give them their space and privacy. The worst thing about our society is that it's not in people's difficult times.

To watch a close interview with Natalie Portman and to see the 10 Questions podcast on iTunes, go to time.com/20080321/portman

TIME March 21, 2008

David Schuff and Robert St. Louis

CENTRALIZATION VS. DECENTRALIZATION OF APPLICATION SOFTWARE

Whichever way the IT department chooses, the result should never lose sight of the user.

Historically, information technology departments have cycled between centralized and decentralized application software distribution, although modular program design and enterprise management software may break that cycle. Meanwhile, IT departments that want to manage the distribution and configuration of software across their networks are searching for an acceptable balance of control, reliability, and speed. Distributing application files on individual PCs maximizes network performance, but makes it much more difficult to enforce configuration standards and maintain control. Placing application files in a few central locations gives an IT department significant control over software configuration but may degrade network performance and lead to user dissatisfaction.

As the components and software in corporate networks become increasingly complex, simplification of their management and administration becomes essential. The network-attached PC has a high cost of ownership. CIO Magazine has estimated the cost to be \$10,000 per desktop per year (3, 4). The costs of a network-attached PC mostly cover maintenance, not installation or actually buying the equipment. These costs can be cut by reducing the number of hours spent on network administration tasks like implementing new software, distributing new software versions, applying patches, and troubleshooting problems with the individual software applications installed on each workstation.

Centralization of application software is one

David Schuff and Robert St. Louis

CENTRALIZATION VS. DECENTRALIZATION OF APPLICATION SOFTWARE

Whichever way the IT department chooses, the result should never lose sight of the user.

Historically, information technology departments have cycled between centralized and decentralized application distribution, although modular program design and enterprise management software may break that cycle. Meanwhile, IT departments that want to manage the distribution and configuration of software across their networks are searching for an acceptable balance of control, reliability, and speed. Distributing application files on individual PCs maximizes network performance, but makes it much more difficult to enforce configuration standards and maintain control. Placing application files in a few central locations gives an IT department significant control over software configuration but may degrade network performance and lead to user dissatisfaction.

As the components and software in corporate networks become increasingly complex, simplification of their management and administration becomes essential. The network-attached PC has a high cost of ownership. CIO Magazine has estimated the cost to be \$10,000 per desktop per year (3, 4). The costs of a network-attached PC mostly cover maintenance, not installation or actually buying the equipment. These costs can be cut by reducing the number of hours spent on network administration tasks like implementing new software, distributing new software versions, applying patches, and troubleshooting problems with the individual software applications installed on each workstation.

Centralization of application software is one

Fig. 35: Samples from Prima dataset



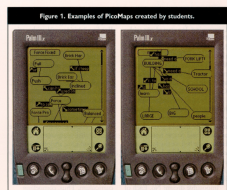
Document (.pdf)



Rendered Fid-HTML



Log On Education



that enable learning and teaching. And there is the really important bit: link those applications to existing curricular materials so educators essentially already know how to use the handheld devices. Having enough applications that have educational utility is a strong advantage in arguing for a switch from graphing calculators, a definite one-trick pony, to handheld devices. Here, then, are our candidates for some effective handheld applications.

PicMap. PicMap is a concept mapping tool for handheld devices. Figure 1 presents several PicMaps created by 10- to 12-year-olds in Denver during a unit on the physics of heavy machinery. PicMap goes beyond paper-and-pencil concept maps in the

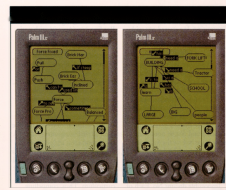
the critically important process of creating and revising their documents.

Children can print out their PicMaps, allowing the teacher to track what the children are doing and give feedback, while parents use the printers for refrigerator decorations. Classroom management issues are not to be taken lightly; making printing straightforward makes teachers comfortable with having each of their 30 students equipped with a handheld device. (If only every 30 handheld device to one desktop computer was as simple.)

Palm sheets. For better or worse, worksheets are a fixture in K-12 classrooms. But a handheld device's worksheet has numerous advantages over its paper cousin: the handheld device can immediately check a student's input and provide feedback. After the data from the worksheet are transferred to a desktop computer, the data can be automatically aggregated and pie charts generated that depict all the students' answers. Figure 2 shows an 11-year-old filling out an Air Quality Inventory worksheet.

Coates. How do germs spread? Drawing on the work at the MIT Media Lab with SmartBadges, we developed a socio-kinesthetic simulation on handheld devices to help children understand this process. Children "meet" each other by walking around a classroom with a handheld device and beaming each other a digital

Log On Education



that enable learning and teaching. And there is the really important bit: link those applications to existing curricular materials so educators essentially already know how to use the handheld devices. Having enough applications that have educational utility is a strong advantage in arguing for a switch from graphing calculators, a definite one-trick pony, to handheld devices. Here, then, are our candidates for some effective handheld applications.

PicMap. PicMap is a concept mapping tool for handheld devices. Figure 1 presents several PicMaps created by 10- to 12-year-olds in Denver during a unit on the physics of heavy machinery. PicMap goes beyond paper-and-pencil concept maps in the

the critically important process of creating and revising their documents.

Children can print out their PicMaps, allowing the teacher to track what the children are doing and give feedback, while parents use the printers for refrigerator decorations. Classroom management issues are not to be taken lightly; making printing straightforward makes teachers comfortable with having each of their 30 students equipped with a handheld device. (If only every 30 handheld device to one desktop computer was as simple.)

Palm sheets. For better or worse, worksheets are a fixture in K-12 classrooms. But a handheld device's worksheet has numerous advantages over its paper cousin: the handheld device can immediately check a student's input and provide feedback. After the data from the worksheet are transferred to a desktop computer, the data can be automatically aggregated and pie charts generated that depict all the students' answers. Figure 2 shows an 11-year-old filling out an Air Quality Inventory worksheet.

Coates. How do germs spread? Drawing on the work at the MIT Media Lab with SmartBadges, we developed a socio-kinesthetic simulation on handheld devices to help children understand this process. Children "meet" each other by walking around a classroom with a handheld device and beaming each other a digital

Fig. 36: Samples from Prima dataset



Document (.pdf)



Rendered Fid-HTML

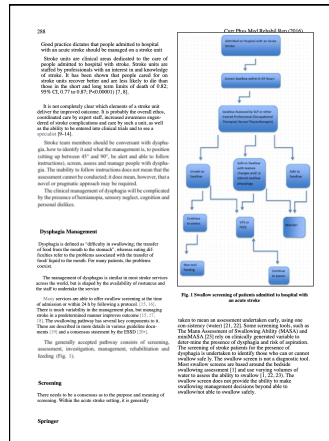
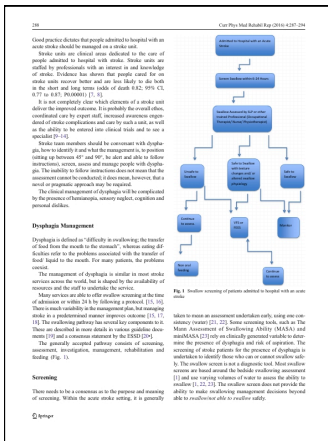
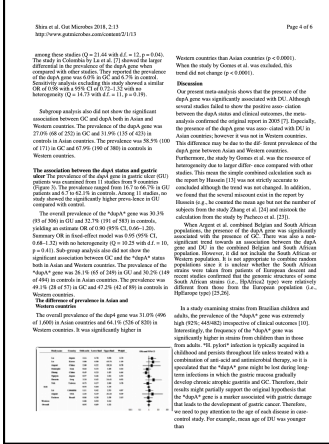
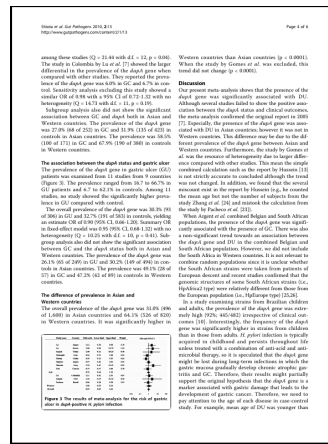
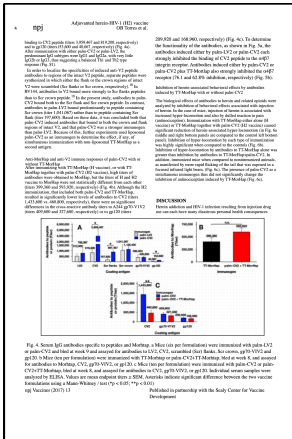
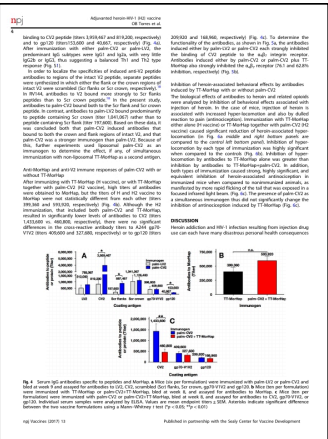


Fig. 37: Samples from PubLayNet dataset



Document (.pdf)



Rendered Fid-HTML

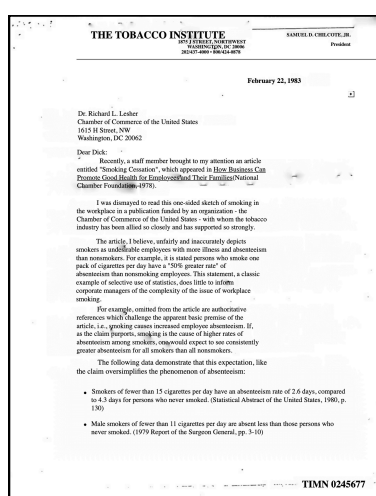
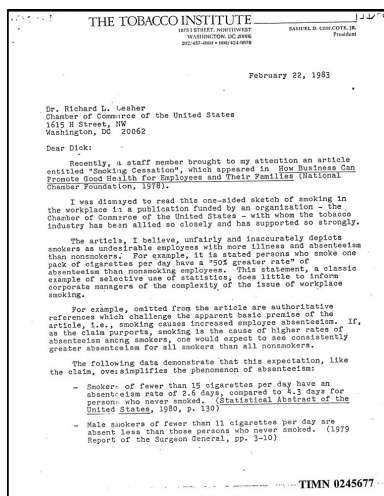
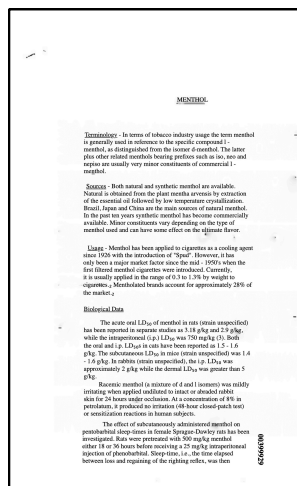
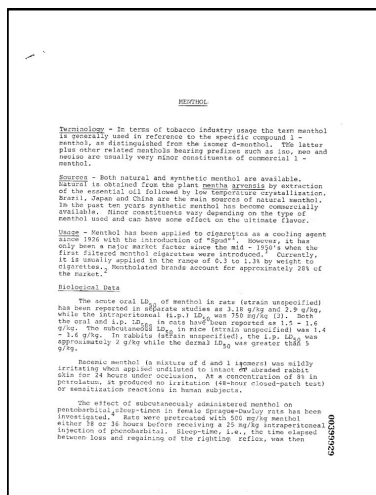


Fig. 38: Samples from RVLCDIP dataset

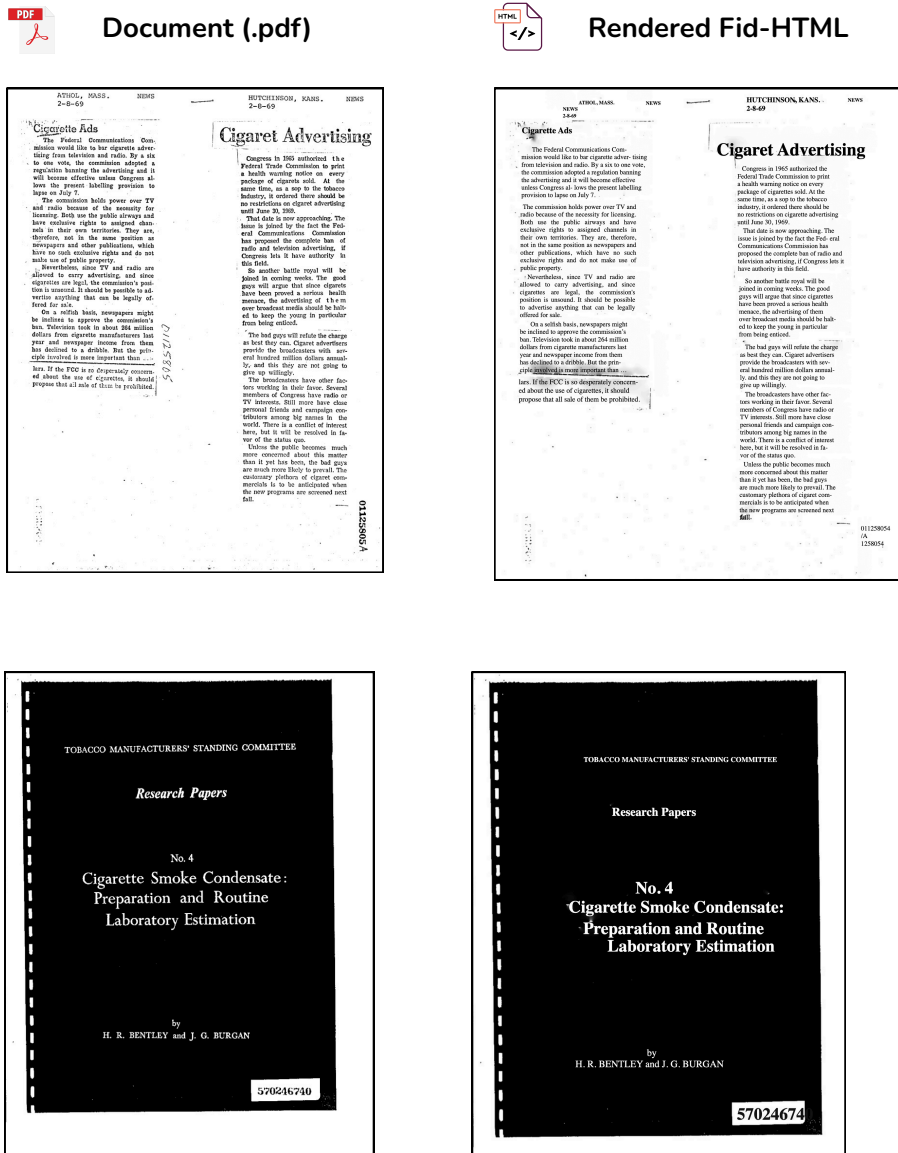


Fig. 39: Samples from RVLCDIP dataset



Document (.pdf)

PREMIO STATIONERY SDN BHD
(Co. No. 123789-W)
GST Reg. No. 00102384846
No 57, Jalan SS 3/29,
47300 Petaling Jaya, Selangor
03-7674 8605

TAX INVOICE

Invoice No : 953-154459
Date : 20/03/2018 3:20:16 PM
Cashier : IVY

| Description | Qty | Price | Amount |
|-----------------------------------|-----|-------|--------|
| 1 SR ENERGY BATTERY AA 4'S | 1 | 13.40 | 13.40 |
| Total: | | | 13.40 |
| Discount: | | | 0.00 |
| Total Sales Inclusive GST @6.00%: | | | 13.40 |
| Cash Received: | | | 20.00 |
| Change: | | | 6.60 |

Starting 1st April, we will no longer be issuing handwritten invoices. Your service will be in electronic form and will be issued by the system during each step of the receipt.

All goods and services are subjected to the 6% GST during checkout.

All goods sold are not refundable/exchangeable. Exchanges are only in merchandise within 3 days with a receipt of receipt.

No cash refund for credit card purchases.

No refund on exchanges will be entertained without proof of receipt.

Terms and conditions apply.

Thank you for Shopping at PREMIO
Have a Nice Day!

For the latest news and updates
like us on Facebook.com/premiostationery



Rendered Fid-HTML

PREMIO STATIONERY SDN BHD
(Co. No. 123789-W)
GST Reg. No. 00102384846
No 57, Jalan SS 3/29,
47300 Petaling Jaya, Selangor
03-7674 8605

TAX INVOICE

Invoice No : 953-154459
Date : 20/03/2018 3:20:16 PM
Cashier : IVY

| Description | Qty | Price | Amount |
|-----------------------------------|-----|-------|--------|
| 1 SR ENERGY BATTERY AA 4'S | 1 | 13.40 | 13.40 |
| Total: | | | 13.40 |
| Discount: | | | 0.00 |
| Total Sales Inclusive GST @6.00%: | | | 13.40 |
| Cash Received: | | | 20.00 |
| Change: | | | 6.60 |

Starting 1st April, we will no longer be issuing handwritten invoices. Your service will be in electronic form and will be issued by the system during each step of the receipt.

All goods and services are subjected to the 6% GST during checkout.

All goods sold are not refundable/exchangeable. Exchanges are only in merchandise within 3 days with a receipt of receipt.

All cashbacks & items sold are final and are non-refundable.

No cash refund for credit card purchases.

No refund on exchanges will be entertained without proof of receipt.

Terms and conditions apply.

Thank you for Shopping at PREMIO
Have a Nice Day!

For the latest news and updates
like us on Facebook.com/premiostationery

VIVOPAC MARKETING SDN BHD (1070687-M)
14, JALAN MANIS 4 TAMAN SEGAR 56100 KL
TEL: 03-9134364 FAX: 03-9131022
www.vivopac.com
GST REG: 00050608056

TAX INVOICE

Doc # : C12112049
Date : 17/04/2017 10:30:27 PM
Printer : T22 Cashier : A2

| DESCRIPTION | PRICE | QTY | AMOUNT | TAX | TAX CODE |
|--------------------|-------|------|--------|------|----------|
| PP - 14x20 0.04mm | 14.00 | 1.00 | 14.00 | 0.84 | SR |
| PP - 16x26 0.04mm | 14.00 | 1.00 | 14.00 | 0.84 | SR |
| 24 x 28 (250T) Red | 6.00 | 1.00 | 6.00 | 0.36 | SR |
| 26 x 32 (850T) Red | 8.00 | 1.00 | 8.00 | 0.48 | SR |
| 1 OPP Tapes | 0.80 | 1.00 | 0.80 | 0.05 | SR |

Item Count: 5 Item Qty: 5.00

Sub Total (Exclusive GST): 44.00
GST 6%: 2.64
Rounded Total: 46.65

Cash: 51.65
Cash Change: 5.00

GST Summary: Amount (RM) 46.65 Tax (RM) 2.80

Terms & Conditions
Goods sold are not refundable with cash.
For exchange of goods used the following apply:
1. The original receipt must be presented.
2. Exchange is done within 7 days from date of receipt.
3. In good condition and in its original packing.

VIVOPAC MARKETING SDN BHD (1070687-M)
14, JALAN MANIS 4 TAMAN SEGAR 56100 KL
TEL: 03-9134364 FAX: 03-9131022
www.vivopac.com
GST REG: 00050608056

TAX INVOICE

Doc # : C12112049
Date : 17/04/2017 10:30:27 PM
Printer : T22 Cashier : A2

| DESCRIPTION | PRICE | QTY | AMOUNT | TAX | TAX CODE |
|--------------------|-------|------|--------|------|----------|
| PP - 14x20 0.04mm | 14.00 | 1.00 | 14.00 | 0.84 | SR |
| PP - 16x26 0.04mm | 14.00 | 1.00 | 14.00 | 0.84 | SR |
| 24 x 28 (250T) Red | 6.00 | 1.00 | 6.00 | 0.36 | SR |
| 26 x 32 (850T) Red | 8.00 | 1.00 | 8.00 | 0.48 | SR |
| 1 OPP Tapes | 0.80 | 1.00 | 0.80 | 0.05 | SR |

Item Count: 5 Item Qty: 5.00

Sub Total (Exclusive GST): 44.00
GST 6%: 2.64
Rounded Total: 46.65

Cash: 51.65
Cash Change: 5.00

GST Summary: Amount (RM) 46.65 Tax (RM) 2.80

Terms & Conditions
Goods sold are not refundable with cash.
For exchange of goods used the following apply:
1. The original receipt must be presented.
2. Exchange is done within 7 days from date of receipt.
3. In good condition and in its original packing.

PETRON BKT LANJAN SB
ALSERKAM ENTERPRISE
Tel: 03-6156 8757 Co No: 001083069-M
KM 458.4 BKT LANJAN UTARA,
L/RAYA UTARA SELATAN, SG BULOH
47000 SUNGAI BULOH

GST ID No: 001210736640

TAX INVOICE

TAX INVOICE NO: 19729058
POS: 129077
Store No.: 129077
01/02/2018 4:43:17 PM Babu

| | |
|----------------------|------|
| A 2 double mint te | 3.00 |
| A 1 sandwich vanilla | 1.90 |
| GST RM: | 0.28 |
| Total RM inc GST: | 4.90 |
| Cash | 5.00 |
| Change | 0.10 |

GST Summary: Amount (RM) 4.62 Tax (RM) 0.28
A=6.00%

Use 3000 Petron Miles
points to pay for
RM45 Fuel

PETRON BKT LANJAN SB
ALSERKAM ENTERPRISE
Tel: 03-6156 8757 Co No: 001083069-M
KM 458.4 BKT LANJAN UTARA,
L/RAYA UTARA SELATAN, SG BULOH
47000 SUNGAI BULOH

GST ID No: 001210736640

TAX INVOICE

TAX INVOICE NO: 19729058
POS: 129077
Store No.: 129077
01/02/2018 4:43:17 PM Babu

| | |
|----------------------|------|
| A 2 double mint te | 3.00 |
| A 1 sandwich vanilla | 1.90 |
| GST RM: | 0.28 |
| Total RM inc GST: | 4.90 |
| Cash | 5.00 |
| Change | 0.10 |

GST Summary: Amount (RM) 4.62 Tax (RM) 0.28
A=6.00%

Use 3000 Petron Miles
points to pay for
RM45 Fuel

Fig. 40: Samples from SROIE dataset

References

- [1] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 Technical Report (2025). <https://arxiv.org/abs/2505.09388>
- [2] Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
- [3] Paruchuri, V.: Marker (2024)
- [4] Smith, R.: An overview of the tesseract ocr engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, pp. 629–633 (2007). IEEE
- [5] J. AI: EasyOCR: Ready-to-use OCR with 80+ Supported Languages. <https://github.com/JaidedAI/EasyOCR>. Version: 1.7.0, Accessed: 2024-06-24 (2024)
- [6] Authors, P.: PaddleOCR, Awesome multilingual OCR toolkits based on PaddlePaddle. <https://github.com/PaddlePaddle/PaddleOCR> (2020)
- [7] Developers, P.: Pandoc User’s Guide. Pandoc, (2025). Pandoc. Comprehensive documentation of Pandoc’s input/output format capabilities. <https://pandoc.org/MANUAL.html>
- [8] contributors, W.: Pandoc. Wikipedia, The Free Encyclopedia (2025). Article on Pandoc, highlighting its role as a universal document converter
- [9] pypdf Maintainers: pypdf: A Pure-Python PDF Library. <https://github.com/py-pdf/pypdf>. Accessed: 2024-06-24 (2024)
- [10] P. Team: PyPDFium2: Python bindings for PDFium. <https://github.com/pypdfium2-team/pypdfium2>. Accessed: 2024-06-24 (2024)
- [11] A. S. Inc.: PyMuPDF. <https://github.com/pymupdf/PyMuPDF>. Accessed: 2024-06-24 (2024)
- [12] Documentation, U.: pdf2htmlEX Manual (Ubuntu Xenial Manpage). (2016). Describes the pdf2htmlEX utility for preserving PDF visual fidelity in HTML. <https://manpages.ubuntu.com/manpages/xenial/man1/pdf2htmlEX.1.html>
- [13] Nassar, A., Marafioti, A., Omenetti, M., Lysak, M., Livathinos, N., Auer, C., Morin, L., Lima, R.T., Kim, Y., Gurbuz, A.S., et al.: Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. arXiv preprint arXiv:2503.11576 (2025)
- [14] Poznanski, A., al.: olmocr: Open layout modeling for robust ocr-free document understanding. arXiv preprint arXiv:2502.01983 (2025)
- [15] Li, Z., Liu, Y., Liu, Q., Ma, Z., Zhang, Z., Zhang, S., Guo, Z., Zhang, J., Wang, X., Bai, X.: Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. arXiv preprint arXiv:2506.05218 (2025)

- [16] Feng, H., Wei, S., Fei, X., Shi, W., Han, Y., Liao, L., Lu, J., Wu, B., Liu, Q., Lin, C., et al.: Dolphin: Document image parsing via heterogeneous anchor prompting. arXiv preprint arXiv:2505.14059 (2025)
- [17] getomni-ai: getomni-ai/ocr-benchmark Dataset. Hugging Face. Accessed from the Hugging Face Datasets repository (n.d.). <https://huggingface.co/datasets/getomni-ai/ocr-benchmark#omniai-ocr-benchmark>
- [18] Ouyang, L., Qu, Y., Zhou, H., Zhu, J., Zhang, R., Lin, Q., Wang, B., Zhao, Z., Jiang, M., Zhao, X., Shi, J., Wu, F., Chu, P., Liu, M., Li, Z., Xu, C., Zhang, B., Shi, B., Tu, Z., He, C.: Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24838–24848 (2025)
- [19] Yang, Z., Tang, J., Li, Z., Wang, P., Wan, J., Zhong, H., Liu, X., Yang, M., Wang, P., Bai, S., Jin, L., Lin, J.: Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21744–21754 (2025)
- [20] Fu, L., Kuang, Z., Song, J., Huang, M., Yang, B., Li, Y., Zhu, L., Luo, Q., Wang, X., Lu, H., Li, Z., Tang, G., Shan, B., Lin, C., Liu, Q., Wu, B., Feng, H., Liu, H., Huang, C., Tang, J., Chen, W., Jin, L., Liu, Y., Bai, X.: OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning (2025). <https://arxiv.org/abs/2501.00321>
- [21] Li, Z., Abulaiti, A., Lu, Y., Chen, X., Zheng, J., Lin, H., Han, X., Jiang, S., Dong, B., Sun, L.: READoc: A unified benchmark for realistic document structured extraction. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025, pp. 21889–21905. Association for Computational Linguistics, Vienna, Austria (2025). <https://doi.org/10.18653/v1/2025.findings-acl.1128> . <https://aclanthology.org/2025.findings-acl.1128/>
- [22] Upstage: DP-Bench: Document Parsing Benchmark. <https://huggingface.co/datasets/upstage/dp-bench>. Accessed: 2025-11-13
- [23] Heakl, A., Sohail, M.A., Ranjan, M., Elbadry, R., Ahmad, G.S., El-Geish, M., Maher, O., Shen, Z., Khan, F.S., Khan, S.: KITAB-bench: A comprehensive multi-domain benchmark for Arabic OCR and document understanding. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025, pp. 22006–22024. Association for Computational Linguistics, Vienna, Austria (2025). <https://doi.org/10.18653/v1/2025.findings-acl.1135> . <https://aclanthology.org/2025.findings-acl.1135/>
- [24] Roberts, J.S., Lee, T., Wong, C.H., Yasunaga, M., Mai, Y., Liang, P.: Image2struct: Benchmarking structure extraction for vision-language models. In: Advances in Neural Information Processing Systems, vol. 37, pp. 115058–115097. Curran Associates, Inc., ??? (2024). <https://doi.org/10.52202/079017-3653>
- [25] Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et al.: Mineru: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839 (2024)
- [26] Livathinos, N., Auer, C., Lysak, M., Nassar, A., Dolfi, M., Vagenas, P., Ramis, C.B., Omenetti, M., Dinkla, K., Kim, Y., et al.: Docling: An efficient open-source toolkit for ai-driven document

- [27] Wei, H., Liu, C., Chen, J., Wang, J., Kong, L., Xu, Y., Ge, Z., Zhao, L., Sun, J., Peng, Y., et al.: General ocr theory: Towards ocr-2.0 via a unified end-to-end model (2024)
- [28] Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J.-b., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., Kenealy, K., Beyer, L., Zhai, X., Tsitsulin, A., Busa-Fekete, R., Feng, A., Sachdeva, N., Coleman, B., Gao, Y., Mustafa, B., Barr, I., Parisotto, E., Tian, D., Eyal, M., Cherry, C., Peter, J.-T., Sinopalnikov, D., Bhupatiraju, S., Agarwal, R., Kazemi, M., Malkin, D., Kumar, R., Vilar, D., Brusilovsky, I., Luo, J., Steiner, A., Friesen, A., Sharma, A., Sharma, A., Gilady, A.M., Goedeckemeyer, A., Saade, A., Feng, A., Kolesnikov, A., Bendebury, A., Abdagic, A., Vadi, A., György, A., Pinto, A.S., Das, A., Bapna, A., Miech, A., Yang, A., Paterson, A., Shenoy, A., Chakrabarti, A., Piot, B., Wu, B., Shahriari, B., Petrini, B., Chen, C., Lan, C.L., Choquette-Choo, C.A., Carey, C., Brick, C., Deutsch, D., Eisenbud, D., Cattle, D., Cheng, D., Paparas, D., Sreepathihalli, D.S., Reid, D., Tran, D., Zelle, D., Noland, E., Huizenga, E., Kharitonov, E., Liu, F., Amirkhanyan, G., Cameron, G., Hashemi, H., Klimczak-Plucińska, H., Singh, H., Mehta, H., Lehri, H.T., Hazimeh, H., Ballantyne, I., Szpektor, I., Nardini, I., Pouget-Abadie, J., Chan, J., Stanton, J., Wieting, J., Lai, J., Orbay, J., Fernandez, J., Newlan, J., Ji, J.-y., Singh, J., Black, K., Yu, K., Hui, K., Vodrahalli, K., Greff, K., Qiu, L., Valentine, M., Coelho, M., Ritter, M., Hoffman, M., Watson, M., Chaturvedi, M., Moynihan, M., Ma, M., Babar, N., Noy, N., Byrd, N., Roy, N., Momchev, N., Chauhan, N., Sachdeva, N., Bunyan, O., Botarda, P., Caron, P., Rubenstein, P.K., Culliton, P., Schmid, P., Sessa, P.G., Xu, P., Stanczyk, P., Tafti, P., Shivanna, R., Wu, R., Pan, R., Rokni, R., Willoughby, R., Vallu, R., Mullins, R., Jerome, S., Smoot, S., Girgin, S., Iqbal, S., Reddy, S., Sheth, S., Pöder, S., Bhatnagar, S., Panyam, S.R., Eiger, S., Zhang, S., Liu, T., Yacovone, T., Liechty, T., Kalra, U., Evci, U., Misra, V., Roseberry, V., Feinberg, V., Kolesnikov, V., Han, W., Kwon, W., Chen, X., Chow, Y., Zhu, Y., Wei, Z., Egyed, Z., Cotruta, V., Giang, M., Kirk, P., Rao, A., Black, K., Babar, N., Lo, J., Moreira, E., Martins, L.G., Sanseviero, O., Gonzalez, L., Gleicher, Z., Warkentin, T., Mirrokni, V., Senter, E., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Matias, Y., Sculley, D., Petrov, S., Fiedel, N., Shazeer, N., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Alayrac, J.-B., Anil, R., Dmitry, Lepikhin, Borgeaud, S., Bachem, O., Joulin, A., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L.: Gemma 3 Technical Report (2025). <https://arxiv.org/abs/2503.19786>
- [29] Liang, S., Jiang, N., Qian, S., Tan, L.: Waffle: Multi-modal model for automated front-end development. arXiv preprint arXiv:2410.18362 (2024)
- [30] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 16965–16974 (2024)
- [31] Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083–4091 (2022)
- [32] Wang, Z., Xu, Y., Cui, L., Shang, J., Wei, F.: Layoutreader: Pre-training of text and layout for reading order detection (2021)
- [33] Jaided AI: Easyocr: Ready-to-use ocr with 80+ supported languages. <https://github.com/JaidedAI/EasyOCR> (2024)
- [34] Li, C., Liu, W., Guo, R., Yin, X., Jiang, K., Du, Y., Du, Y., Zhu, L., Lai, B., Hu, X., et al.: Pp-ocrv3:

More attempts for the improvement of ultra lightweight ocr system. arXiv preprint arXiv:2206.03001 (2022)

- [35] Wang, B., Gu, Z., Liang, G., Xu, C., Zhang, B., Shi, B., He, C.: Unimernet: A universal network for real-world mathematical expression recognition. arXiv preprint arXiv:2404.15254 (2024)
- [36] Padhi, I., Schiff, Y., Melnyk, I., Rigotti, M., Mroueh, Y., Dognin, P., Ross, J., Nair, R., Altman, E.: Tabular transformers for modeling multivariate time series. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3565–3569 (2021). IEEE
- [37] He, Y., Qi, X., Ye, J., Gao, P., Chen, Y., Li, B., Tang, X., Xiao, R.: Pingan-vcgroup’s solution for icdar 2021 competition on scientific table image recognition to latex. arXiv preprint arXiv:2105.01846 (2021)
- [38] Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., et al.: Mineru: An open-source solution for precise document content extraction. arXiv preprint arXiv:2409.18839 (2024)
- [39] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., Wang, L.: The dawn of LMMs: Preliminary explorations with GPT-4V (ision). arXiv preprint arXiv:2309.17421 (2023)
- [40] Anthropic: Claude 3.5 Sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>. Accessed: 2025-09-01 (2024)
- [41] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
- [42] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
- [43] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
- [44] Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: MiniCPM-V: A GPT-4V level MLLM on your phone. arXiv preprint arXiv:2408.01800 (2024)
- [45] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 24185–24198 (2024)
- [46] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24185–24198 (2024)
- [47] Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)

- [48] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision, pp. 498–517 (2022). Springer
- [49] Kim, G., Park, T., Lee, J.: Ocr-free document understanding transformer. In: European Conference on Computer Vision, pp. 630–646 (2022). Springer
- [50] Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.: Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418 (2023)
- [51] Blecher, L., Krithivasan, K., Pfeiffer, J., Koehn, P., Ruder, S.: Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418 (2023)
- [52] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., al.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1192–1200 (2020)
- [53] Xu, Y., Xu, T., Cui, L., al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740 (2020)
- [54] Huang, Y., Xu, Y., Gong, Y., Cui, L., Zhang, D., Liu, S., Wang, B., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. arXiv preprint arXiv:2204.08387 (2022)
- [55] Wei, J., Xiao, K., Chen, Z., Wang, X., Lin, X., He, R., et al.: General object understanding in structured documents. arXiv preprint arXiv:2401.13312 (2024)
- [56] Peng, D., Wang, X., Liu, Y., Zhang, J., Huang, M., Lai, S., Li, J., Zhu, S., Lin, D., Shen, C., et al.: Spts: single-point text spotting, 4272–4281 (2022)
- [57] Liu, Y., Zhang, J., Peng, D., Huang, M., Wang, X., Tang, J., Huang, C., Lin, D., Shen, C., Bai, X., et al.: Spts v2: single-point scene text spotting. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(12), 15665–15679 (2023)
- [58] Ye, D., Huang, Y., Xu, Y., Cui, L., Wang, B., Wei, F.: Ureader: Read like humans with unified semantic-visual pretraining. arXiv preprint arXiv:2305.14167 (2023)
- [59] Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., Zhou, J.: mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. arXiv preprint arXiv:2409.03420 (2024)
- [60] Ye, D., Xu, Y., Cui, L., Wei, F., Wang, B.: mplug-docowl: Modular pre-training for multimodal understanding and generation. arXiv preprint arXiv:2306.12607 (2023)
- [61] Hu, W., Ye, D., Cui, L., Wang, B., Wei, F.: mplug-docowl v2: Generalist multimodal document understanding with modular visual expert. arXiv preprint arXiv:2403.01277 (2024)
- [62] Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26763–26773 (2024)
- [63] Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., Bai, X.: Textmonkey: An ocr-free large multimodal model for understanding document. arXiv preprint arXiv:2403.04473 (2024)

- [64] Li, X., Zhao, S., Zhang, C., al.: Monkey: Self-training multimodal language models for document ai. arXiv preprint arXiv:2402.12837 (2024)
- [65] Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., Bai, X.: TextMonkey: An OCR-free large multimodal model for understanding document. arXiv preprint arXiv:2403.04473 (2024)
- [66] Poznanski, J., Borchardt, J., Dunkelberger, J., Huff, R., Lin, D., Rangapur, A., Wilhelm, C., Lo, K., Soldaini, L.: olmocr: Unlocking trillions of tokens in pdfs with vision language models. arXiv preprint arXiv:2502.18443 (2025)
- [67] Ouyang, R., Chu, C., Xin, Z., Ma, X.: PDFMathTranslate: Scientific Document Translation Preserving Layouts (2025). <https://arxiv.org/abs/2507.03009>
- [68] Safnuk, B., Hu, G.: Reconstructing latex source files from generated pdfs - a neural network approach. In: 2018 IEEE 16th International Conference on Industrial Informatics (INDIN), pp. 890–895 (2018). <https://doi.org/10.1109/INDIN.2018.8472050>
- [69] Kayal, P., Anand, M., Desai, H., Singh, M.: Tables to latex: structure and content extraction from scientific tables. International Journal on Document Analysis and Recognition (IJDAR) **26**(2), 121–130 (2022) <https://doi.org/10.1007/s10032-022-00420-9>
- [70] Wang, Z., Liu, J.C.: Translating math formula images to latex sequences using deep neural networks with sequence-level training. International Journal on Document Analysis and Recognition (IJDAR) **24**, 63–75 (2021) <https://doi.org/10.1007/s10032-020-00360-2>
- [71] Li, X., Gong, M., Wu, Y., Dai, J., Guo, A., Jiang, X., Cao, H., Liu, Y., Jiang, D., Sun, X.: DREAM: Document Reconstruction via End-to-end Autoregressive Model (2025). <https://arxiv.org/abs/2507.05805>
- [72] Zhou, T., Zhao, Y., Hou, X., Sun, X., Chen, K., Wang, H.: Bridging design and development with automated declarative ui code generation. arXiv preprint arXiv:2409.11667 (2024)
- [73] Wan, Y., Wang, C., Dong, Y., Wang, W., Li, S., Huo, Y., Lyu, M.: Divide-and-conquer: Generating ui code from screenshots. Proceedings of the ACM on Software Engineering **2**(FSE), 2099–2122 (2025)
- [74] Wu, F., Gao, C., Li, S., Wen, X.-C., Liao, Q.: Mllm-based ui2code automation guided by ui layout information. Proceedings of the ACM on Software Engineering **2**(ISSTA), 1123–1145 (2025)
- [75] Nguyen, T.A., Csallner, C.: Reverse engineering mobile application user interfaces with remaui (t). In: 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 248–259 (2015). IEEE
- [76] Gui, Y., Wan, Y., Li, Z., Zhang, Z., Chen, D., Zhang, H., Su, Y., Chen, B., Zhou, X., Jiang, W., *et al.*: Uicopilot: Automating ui synthesis via hierarchical code generation from webpage designs. In: Proceedings of the ACM on Web Conference 2025, pp. 1846–1855 (2025)
- [77] Chen, Y., Ding, S., Zhang, Y., Chen, W., Du, J., Sun, L., Chen, L.: Designcoder: Hierarchy-aware and self-correcting ui code generation with large language models. arXiv preprint arXiv:2506.13663 (2025)
- [78] Ge, T., Liu, Y., Ye, J., Li, T., Wang, C.: Advancing vision-language models in front-end development via data synthesis. arXiv preprint arXiv:2503.01619 (2025)

- [79] Gui, Y., Li, Z., Wan, Y., Shi, Y., Zhang, H., Chen, B., Su, Y., Chen, D., Wu, S., Zhou, X., *et al.*: Webcode2m: A real-world dataset for code generation from webpage designs. In: Proceedings of the ACM on Web Conference 2025, pp. 1834–1845 (2025)
- [80] Yun, S., Lin, H., Thushara, R., Bhat, M.Q., Wang, Y., Jiang, Z., Deng, M., Wang, J., Tao, T., Li, J., *et al.*: Web2code: A large-scale webpage-to-code dataset and evaluation framework for multimodal llms. CoRR (2024)
- [81] Kolthoff, K., Kretzer, F., Fiebig, L., Bartelt, C., Maedche, A., Ponzetto, S.P.: Zero-shot prompting approaches for llm-based graphical user interface generation. arXiv preprint arXiv:2412.11328 (2024)
- [82] Guo, H., Zhang, W., Chen, J., Gu, Y., Yang, J., Du, J., Hui, B., Liu, T., Ma, J., Zhou, C., *et al.*: Iw-bench: Evaluating large multimodal models for converting image-to-web. arXiv e-prints, 2409 (2024)
- [83] Xiao, S., Chen, Y., Li, J., Chen, L., Sun, L., Zhou, T.: Prototype2code: End-to-end front-end code generation from ui design prototypes. In: International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, vol. 88353, pp. 02–02038 (2024). American Society of Mechanical Engineers
- [84] Vu, T.D., Hoang, C., Hy, T.-S.: Multimodal graph representation learning for website generation based on visual sketch. arXiv preprint arXiv:2504.18729 (2025)
- [85] Si, C., Zhang, Y., Li, R., Yang, Z., Liu, R., Yang, D.: Design2code: Benchmarking multimodal code generation for automated front-end engineering. arXiv preprint arXiv:2403.03163 (2024)
- [86] Xie, M., Feng, S., Xing, Z., Chen, J., Chen, C.: Uied: a hybrid tool for gui element detection. In: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 1655–1659 (2020)
- [87] Xiao, S., Chen, Y., Song, Y., Chen, L., Sun, L., Zhen, Y., Chang, Y., Zhou, T.: Ui semantic component group detection: Grouping ui elements with similar semantics in mobile graphical user interface. Displays **83**, 102679 (2024)
- [88] Zhang, T., Peiguo, F., Liu, J., Zhang, Y., Chen, X.: Nlidesign: A ui design tool for natural language interfaces. In: Proceedings of the ACM Turing Award Celebration Conference-China 2024, pp. 153–158 (2024)
- [89] Lin, Z., Zhou, Z., Zhao, Z., Wan, T., Ma, Y., Gao, J., Li, X.: WebUIBench: A Comprehensive Benchmark for Evaluating Multimodal Large Language Models in WebUI-to-Code (2025). <https://arxiv.org/abs/2506.07818>
- [90] Si, C., Zhang, Y., Li, R., Yang, Z., Liu, R., Yang, D.: Design2Code: Benchmarking multimodal code generation for automated front-end engineering. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) 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. Association for Computational Linguistics, Albuquerque, New Mexico (2025). <https://doi.org/10.18653/v1/2025.naacl-long.199> . <https://aclanthology.org/2025.naacl-long.199/>
- [91] Wan, Y., Dong, Y., Xiao, J., Huo, Y., Wang, W., Lyu, M.R.: MRWeb: An Exploration of Generating Multi-Page Resource-Aware Web Code from UI Designs (2024). <https://arxiv.org/abs/2412.15310>

- [92] Xiao, J., Wang, M., Lam, M.H., Wan, Y., Liu, J., Huo, Y., Lyu, M.R.: DesignBench: A Comprehensive Benchmark for MLLM-based Front-end Code Generation (2025). <https://arxiv.org/abs/2506.06251>
- [93] Guo, H., Zhang, W., Chen, J., Gu, Y., Yang, J., Du, J., Cao, S., Hui, B., Liu, T., Ma, J., Zhou, C., Li, Z.: IW-bench: Evaluating large multimodal models for converting image-to-web. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025, pp. 6449–6466. Association for Computational Linguistics, Vienna, Austria (2025). <https://doi.org/10.18653/v1/2025.findings-acl.334> . <https://aclanthology.org/2025.findings-acl.334/>
- [94] Wan, Y., Wang, C., Dong, Y., Wang, W., Li, S., Huo, Y., Lyu, M.: Divide-and-conquer: Generating ui code from screenshots. *Proc. ACM Softw. Eng.* **2**(FSE) (2025) <https://doi.org/10.1145/3729364>
- [95] Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441* (2023)
- [96] Zhao, Z., Kang, H., Wang, B., He, C.: Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628* (2024)
- [97] Nath, O., Kukkala, S., Khapra, M., Sarvadevabhatla, R.K.: Indicdlp: A foundational dataset for multi-lingual and multi-domain document layout parsing. In: Yin, X.-C., Karatzas, D., Lopresti, D. (eds.) Document Analysis and Recognition – ICDAR 2025, pp. 23–39. Springer, Cham (2026). Dataset spans 11 Indic languages plus English, covering 12 document domains. <https://indicdlp.github.io/>
- [98] Ye, M., Zhang, J., Liu, J., Liu, C., Yin, B., Liu, C., Du, B., Tao, D.: Hi-sam: Marrying segment anything model for hierarchical text segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)
- [99] Google Cloud: OCR with Google AI. <https://cloud.google.com/use-cases/ocr?hl=en>. Accessed: 2025-11-14 (n.d.)
- [100] Mindee: doctr: Document Text Recognition. <https://github.com/mindee/doctr>. GitHub repository (2021)
- [101] Kumar, R., Jinka, J.S., Sarvadevabhatla, R.K.: Textar: Textual attribute recognition in multi-domain and multi-lingual document images (2025)
- [102] Reed, S.: textFit. <https://github.com/STRML/textFit>. Accessed: 2025-09-01 (2014)
- [103] Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: International Conference on Document Analysis and Recognition (ICDAR)
- [104] Zhong, X., Tang, J., Jimeno Yepes, A.: Publaynet: Largest dataset ever for document layout analysis. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1015–1022 (2019). <https://doi.org/10.1109/ICDAR.2019.00166>
- [105] Pfizmann, B., Auer, C., Dolfi, M., Nassar, A.S., Staar, P.: 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. Association for

Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3534678.3539043> .
<https://doi.org/10.1145/3534678.3539043>

- [106] Cheng, H., Zhang, P., Wu, S., Zhang, J., Zhu, Q., Xie, Z., Li, J., Ding, K., Jin, L.: M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. *CoRR* **abs/2305.08719** (2023)
- [107] PixParse: IDL-WDS: A WebDataset of Interleaved Document Layouts and Webpage Screenshots. Accessed: 2025-09-17 (2025). <https://huggingface.co/datasets/pixparse/idl-wds>
- [108] Antonacopoulos, A., Clausner, C., Papadopoulos, C., Pletschacher, S.: Icdar 2013 competition on document image segmentation and layout analysis. In: *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1404–1408 (2013). <https://doi.org/10.1109/ICDAR.2013.280>
- [109] Da, C., Luo, C., Zheng, Q., Yao, C.: Vision grid transformer for document layout analysis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19462–19472 (2023)
- [110] Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., Jawahar, C.V.: Icdar2019 competition on scanned receipt ocr and information extraction. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1516–1520 (2019). <https://doi.org/10.1109/ICDAR.2019.00244>
- [111] Jaume, G., Kemal Ekenel, H., Thiran, J.-P.: Funsd: A dataset for form understanding in noisy scanned documents. In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, pp. 1–6 (2019). <https://doi.org/10.1109/ICDARW.2019.10029>
- [112] McNamee, P., Duh, K., Carpenter, C., Colaianni, R., King, N., Murray, K.: OJ4OCRMT: A large multilingual dataset for OCR-MT evaluation. In: Bouillon, P., Gerlach, J., Girletti, S., Volkart, L., Rubino, R., Sennrich, R., Farinha, A.C., Gaido, M., Daems, J., Kenny, D., Moniz, H., Szoc, S. (eds.) *Proceedings of Machine Translation Summit XX: Volume 1*, pp. 113–125. European Association for Machine Translation, Geneva, Switzerland (2025). <https://aclanthology.org/2025.mtsummit-1.9/>
- [113] Li, Y., Chenguang, Z.: A metric normalization of tree edit distance. *Frontiers of Computer Science in China* **5**(1), 119–125 (2011)
- [114] Pawlik, M., Augsten, N.: Tree edit distance. *Inf. Syst.* **56**(C), 157–173 (2016) <https://doi.org/10.1016/j.is.2015.08.004>
- [115] Wang, B., Wu, F., Ouyang, L., Gu, Z., Zhang, R., Xia, R., Zhang, B., He, C.: Image Over Text: Transforming Formula Recognition Evaluation with Character Detection Matching (2025). <https://arxiv.org/abs/2409.03643>
- [116] AI, R.: RolmOCR: A Faster, Lighter Open Source OCR Model (2025)
- [117] Mandal, S., Talewar, A., Ahuja, P., Juvatkar, P.: Nanonets-OCR-S: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging (2025)
- [118] chatdoc-com: OCRFlux: A multimodal toolkit for converting PDFs and images into Markdown. Accessed: 2025-08-29 (2025)

- [119] Chen, X., Li, S., Zhu, X., Chen, Y., Yang, F., Fang, C., Qu, L., Xu, X., Wei, H., Wu, M.: Logics-Parsing Technical Report (2025). <https://arxiv.org/abs/2509.19760>
- [120] Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T.C., Jacobsson, H., Szpektor, I., Jiang, N.-J., Haridasan, K., Omran, A., Saunshi, N., Bahri, D., Mishra, G., Chu, E., Boyd, T., Hekman, B., Parisi, A., Zhang, C., Kawintiranon, K., Bedrax-Weiss, T., Wang, O., Xu, Y., Purkiss, O., Mendlovic, U., Deutel, I., Nguyen, N., Langley, A., Korn, F., Rossazza, L., Ramé, A., Waghmare, S., Miller, H., Byrd, N., Sheshan, A., Bhardwaj, R.H.S., Janus, P., Rissa, T., Horgan, D., Silver, S., Wahid, A., Brin, S., Raimond, Y., Kloboves, K., Wang, C., Gundavarapu, N.B., Shumailov, I., Wang, B., Pajarskas, M., Heyward, J., Nikoltchev, M., Kula, M., Zhou, H., Garrett, Z., Kafle, S., Arik, S., Goel, A., Yang, M., Park, J., Kojima, K., Mahmoudieh, P., Kavukcuoglu, K., Chen, G., Fritz, D., Bulyenov, A., Roy, S., Paparas, D., Shemtov, H., Chen, B.-J., Strudel, R., Reitter, D., Roy, A., Vlasov, A., Ryu, C., Leichner, C., Yang, H., Mariet, Z., Vnukov, D., Sohn, T., Stuart, A., Liang, W., Chen, M., Rawlani, P., Koh, C., Co-Reyes, J., Lai, G., Banzal, P., Vytiniotis, D., Mei, J., Cai, M., Badawi, M., Fry, C., Hartman, A., Zheng, D., Jia, E., Keeling, J., Louis, A., Chen, Y., Robles, E., Hung, W.-C., Zhou, H., Saxena, N., Goenka, S., Ma, O., Fisher, Z., Taege, M.H., Graves, E., Steiner, D., Li, Y., Nguyen, S., Sukthankar, R., Stanton, J., Eslami, A., Shen, G., Akin, B., Guseynov, A., Zhou, Y., Alayrac, J.-B., Joulin, A., Farkash, E., Thapliyal, A., Roller, S., Shazeer, N., Davchev, T., Koo, T., Forbes-Pollard, H., Audhkhasi, K., Farquhar, G., Gilady, A.M., Song, M., Aslanides, J., Mendolicchio, P., Parrish, A., Blitzer, J., Gupta, P., Ju, X., Yang, X., Datta, P., Tacchetti, A., Mehta, S.V., Dibb, G., Gupta, S., Piccinini, F., Hadsell, R., Rajayogam, S., Jiang, J., Griffin, P., Sundberg, P., Hayes, J., Frolov, A., Xie, T., Zhang, A., Dasgupta, K., Kalra, U., Shani, L., Macherey, K., Huang, T.-K., MacDermed, L., Duddu, K., Zacchello, P., Yang, Z., Lo, J., Hui, K., Kastelic, M., Gasaway, D., Tan, Q., Yue, S., Barrio, P., Wieting, J., Yang, W., Nystrom, A., Demmessie, S., Levskaya, A., Viola, F., Tekur, C., Billock, G., Necula, G., Joshi, M., Schaeffer, R., Lokhande, S., Sorokin, C., Shenoy, P., Chen, M., Collier, M., Li, H., Bos, T., Wichers, N., Lee, S.J., Pouget, A., Thangaraj, S., Axiotis, K., Crone, P., Sterneck, R., Chinaev, N., Krakovna, V., Ferludin, O., Gemp, I., Winkler, S., Goldberg, D., Korotkov, I., Xiao, K., Mehrotra, M., Mariserla, S., Piratla, V., Thurk, T., Pham, K., Ma, H., Senges, A., Kumar, R., Meyer, C., Talius, E., Pierse, N.W., Sandhu, B., Toma, H., Lin, K., Nath, S., Stone, T., Sadigh, D., Gupta, N., Guez, A., Singh, A., Thomas, M., Duerig, T., Gong, Y., Tanburn, R., Zhang, L.L., Dao, P., Hammad, M., Xie, S., Rijhwani, S., Murdoch, B., Kim, D., Thompson, W., Cheng, H.-T., Sohn, D., Sprechmann, P., Xu, Q., Tadepalli, S., Young, P., Zhang, Y., Srinivasan, H., Aperghis, M., Ayyar, A., Fitoussi, H., Burnell, R., Madras, D., Dusenberry, M., Xiong, X., Oguntebi, T., Albrecht, B., Bornschein, J., Mitrović, J., Dimarco, M., Shamanna, B.K., Shah, P., Sezener, E., Upadhyay, S., Lacey, D., Schiff, C., Baur, S., Ganapathy, S., Schnider, E., Wirth, M., Schenck, C., Simanovsky, A., Tan, Y.-X., Fränken, P., Duan, D., Mankalale, B., Dhawan, N., Sequeira, K., Wei, Z., Goel, S., Unlu, C., Zhu, Y., Sun, H., Balashankar, A., Shuster, K., Umekar, M., Alnahlawi, M., Oord, A., Chen, K., Zhai, Y., Dai, Z., Lee, K.-H., Doi, E., Zilka, L., Vallu, R., Shrivastava, D., Lee, J., Husain, H., Zhuang, H., Cohen-Addad, V., Barber, J., Atwood, J., Sadovsky, A., Wellens, Q., Hand, S., Rajendran, A., Turker, A., Carey, C., Xu, Y., Soltau, H., Li, Z., Song, X., Li, C., Kemaev, I., Brown, S., Burns, A., Patraucean, V., Stanczyk, P., Aravamudhan, R., Blondel, M., Noga, H., Blanco, L., Song, W., Isard, M., Sharma, M., Hayes, R., Badawy, D.E., Lamp, A., Laish, I., Kozlova, O., Chan, K., Singla, S., Sunkara, S., Upadhyay, M., Liu, C., Bai, A., Wilkiewicz, J., Zlocha, M., Liu, J., Li, Z., Li, H., Barak, O., Raboshchuk, G., Choi, J., Liu, F., Jue, E., Sharma, M., Marzoca, A., Busa-Fekete, R., Korsun, A., Elisseeff, A., Shen, Z., Carthy, S.M., Lamerigts, K., Hosseini, A., Lin, H., Chen, C., Yang, F., Chauhan, K., Omernick, M., Jia, D., Zainullina, K., Hassabis, D., Vainstein, D., Amid, E., Zhou, X., Votel, R., Vértes, E., Li, X., Zhou, Z., Lazaridou, A., McMahan, B., Narayanan, A., Soyer, H., Basu, S., Lee, K., Perozzi, B., Cao, Q., Berrada, L., Arya, R., Chen, K., Katrina, Xu, Lochbrunner, M., Hofer, A., Sharifzadeh, S., Wu, R., Goldman, S., Awasthi, P.,

Wang, X., Wu, Y., Sha, C., Zhang, B., Mikula, M., Graziano, F., Mcloughlin, S., Giannoumis, I., Namiki, Y., Malik, C., Radebaugh, C., Hall, J., Leal-Cavazos, R., Chen, J., Sindhvani, V., Kao, D., Greene, D., Griffith, J., Welty, C., Montgomery, C., Yoshino, T., Yuan, L., Goodman, N., Michaely, A.H., Lee, K., Sawhney, K., Chen, W., Zheng, Z., Shum, M., Savinov, N., Pot, E., Pak, A., Zadimoghaddam, M., Bhatnagar, S., Lewenberg, Y., Kutzman, B., Liu, J., Katzen, L., Selier, J., Djolonga, J., Lepikhin, D., Xu, K., Liang, J., Tan, J., Schillings, B., Ersoy, M., Blois, P., Bandemer, B., Singh, A., Lebedev, S., Joshi, P., Brown, A.R., Palmer, E., Pathak, S., Jalan, K., Zubach, F., Lall, S., Parker, R., Gunjan, A., Rogulenko, S., Sanghai, S., Leng, Z., Egyed, Z., Li, S., Ivanova, M., Andriopoulos, K., Xie, J., Rosenfeld, E., Wright, A., Sharma, A., Geng, X., Wang, Y., Kwei, S., Pan, R., Zhang, Y., Wang, G., Liu, X., Yeung, C., Cole, E., Rosenberg, A., Yang, Z., Chen, P., Polovets, G., Nair, P., Saxena, R., Smith, J., Chang, S.-y., Mahendru, A., Grant, S., Iyer, A., Cai, I., McGiffin, J., Shen, J., Walton, A., Girgis, A., Woodman, O., Ke, R., Kwong, M., Rouillard, L., Rao, J., Li, Z., Xu, Y., Prost, F., Zou, C., Ji, Z., Magni, A., Liechty, T., Calian, D.A., Ramachandran, D., Krivokon, I., Huang, H., Chen, T., Hauth, A., Ilić, A., Xi, W., Lim, H., Ion, V.-D., Moradi, P., Toksoz-Exley, M., Bullard, K., Allamanis, M., Yang, X., Wang, S., Hong, Z., Gergely, A., Li, C., Mittal, B., Kovalev, V., Ungureanu, V., Labanowski, J., Wassenberg, J., Lacasse, N., Cideron, G., Dević, P., Marsden, A., Nguyen, L., Fink, M., Zhong, Y., Kiyono, T., Ivanov, D., Ma, S., Bain, M., Yalasangi, K., She, J., Petrushkina, A., Lunayach, M., Bromberg, C., Hodgkinson, S., Meshram, V., Vlastic, D., Kyker, A., Xu, S., Stanway, J., Yang, Z., Zhao, K., Tung, M., Odoom, S., Fujii, Y., Gilmer, J., Kim, E., Halim, F., Le, Q., Bohnet, B., El-Sayed, S., Neyshabur, B., Reynolds, M., Reich, D., Xu, Y., Moreira, E., Sharma, A., Liu, Z., Hosseini, M.J., Raisinghani, N., Su, Y., Lao, N., Formoso, D., Gelmi, M., Gueta, A., Dey, T., Gribovskaya, E., Čevič, D., Mudgal, S., Bingham, G., Wang, J., Kumar, A., Cullum, A., Han, F., Bousmalis, K., Cedillo, D., Chu, G., Magay, V., Michel, P., Hlavnova, E., Calandriello, D., Ariafar, S., Yao, K., Sehwag, V., Vezer, A., Lago, A.D., Zhu, Z., Rubenstein, P.K., Porter, A., Baddepudi, A., Riva, O., Istin, M.D., Yeh, C.-K., Li, Z., Howard, A., Jha, N., Chen, J., Liedekerke, R., Ahmed, Z., Rodriguez, M., Bhatia, T., Wang, B., Elqursh, A., Klinghoffer, D., Chen, P., Kohli, P., I, T., Zhang, W., Nado, Z., Chen, J., Chen, M., Zhang, G., Singh, A., Hillier, A., Lebron, F., Tao, Y., Liu, T., Dulac-Arnold, G., Zhang, J., Narayan, S., Liu, B., Firat, O., Bhowmick, A., Liu, B., Zhang, H., Zhang, Z., Rotival, G., Howard, N., Sinha, A., Grushetsky, A., Beyret, B., Gopalakrishnan, K., Zhao, J., He, K., Payrits, S., Nabulsi, Z., Zhang, Z., Chen, W., Lee, E., Fallen, N., Gollapudi, S., Zhou, A., Pavetić, F., Köppe, T., Huang, S., Pasumarthi, R., Fernando, N., Fischer, F., Čurko, D., Gao, Y., Svensson, J., Stone, A., Qureshi, H., Sinha, A., Kulshreshtha, A., Matysiak, M., Mao, J., Saroufim, C., Faust, A., Duan, Q., Fidel, G., Katircioglu, K., Kaufman, R.L., Shah, D., Kong, W., Bapna, A., Weisz, G., Dunleavy, E., Dutta, P., Liu, T., Chaabouni, R., Parada, C., Wu, M., Belias, A., Bissacco, A., Fort, S., Xiao, L., Huot, F., Knutsen, C., Blau, Y., Li, G., Prendki, J., Love, J., Chow, Y., Charoenpanit, P., Shimokawa, H., Coriou, V., Gregor, K., Izo, T., Akula, A., Pinto, M., Hahn, C., Paulus, D., Guo, J., Sharma, N., Hsieh, C.-J., Chukwuka, A., Hashimoto, K., Rauschmayr, N., Wu, L., Angermueller, C., Wang, Y., Gerlach, S., Pliskin, M., Mirylenka, D., Ma, M., Baugher, L., Gale, B., Bijwadia, S., Rakićević, N., Wood, D., Park, J., Chang, C.-C., Seal, B., Tar, C., Krasowiak, K., Song, Y., Stephanov, G., Wang, G., Maggioni, M., Lin, S.X., Wu, F., Paul, S., Jiang, Z., Agrawal, S., Piot, B., Feng, A., Kim, C., Doshi, T., Lai, J., Chuqiao, Xu, Vikram, S., Chelba, C., Krause, S., Zhuang, V., Rae, J., Denk, T., Collister, A., Weerts, L., Luo, X., Lu, Y., Garnes, H., Gupta, N., Spitz, T., Hassidim, A., Liang, L., Shafran, I., Humphreys, P., Vassigh, K., Wallis, P., Shejwalkar, V., Perez-Nieves, N., Hornung, R., Tan, M., Westberg, B., Ly, A., Zhang, R., Farris, B., Park, J., Kosik, A., Cankara, Z., Maksai, A., Xu, Y., Cassirer, A., Caelles, S., Abdolmaleki, A., Chiang, M., Fabrikant, A., Shetty, S., He, L., Giménez, M., Hashemi, H., Panthaplackel, S., Kulizhskaya, Y., Deshmukh, S., Pighin, D., Alazard, R., Jindal, D., Noury, S., S, P.K., Qin, S., Dotiwala, X., Spencer, S., Babaeizadeh, M., Chen, B.J., Mehta, V., Lees, J., Leach, A., Koanantakool, P., Akolzin, I., Comanescu, R., Ahn, J., Svyatkovskiy, A., Mustafa, B., D'Ambrosio, D., Garlapati, S.M.R., Lamblin, P., Agarwal, A., Song, S., Sessa, P.G., Coquinot, P., Maggs, J., Masoom, H., Pitta, D., Wang, Y., Morris-Suzuki, P., Porter,

B., Jia, J., Dudek, J., R, R., Paduraru, C., Ansell, A., Bolukbasi, T., Lu, T., Ganeshan, R., Wang, Z., Griffiths, H., Benenson, R., He, Y., Swirhun, J., Papamakarios, G., Chawla, A., Sengupta, K., Wang, Y., Milutinovic, V., Mordatch, I., Jia, Z., Smith, J., Ng, W., Nigam, S., Young, M., Vušak, E., Hechtman, B., Goenka, S., Zipori, A., Ayoub, K., Popat, A., Acharya, T., Yu, L., Bloxwich, D., Song, H., Roit, P., Li, H., Boag, A., Nayakanti, N., Chandra, B., Ding, T., Mehta, A., Hope, C., Zhang, J., Shtacher, I.H., Badola, K., Nakashima, R., Sozanschi, A., Comşa, I., Žužul, A., Caveness, E., Odell, J., Watson, M., Cesare, D., Lippe, P., Lockhart, D., Verma, S., Chen, H., Sun, S., Zhuo, L., Shah, A., Gupta, P., Muzio, A., Niu, N., Zait, A., Singh, A., Gaba, M., Ye, F., Ramachandran, P., Saleh, M., Popa, R.A., Dubey, A., Liu, F., Javanmardi, S., Epstein, M., Hemsley, R., Green, R., Ranka, N., Cohen, E., Fu, C.K., Ghemawat, S., Borovik, J., Martens, J., Chen, A., Shyam, P., Pinto, A.S., Yang, M.-H., Tifrea, A., Du, D., Gong, B., Agarwal, A., Kim, S., Frank, C., Shah, S., Song, X., Deng, Z., Mikhlap, A., Chatziprimou, K., Chung, T., Creswell, T., Zhang, S., Jun, Y., Lebsack, C., Truong, W., Andačić, S., Yona, I., Fornoni, M., Rong, R., Toropov, S., Soudagar, A.S., Audibert, A., Zaiem, S., Abbas, Z., Rusu, A., Potluri, S., Weng, S., Kementsietsidis, A., Tsitsulin, A., Peng, D., Ha, N., Jain, S., Latkar, T., Ivanov, S., McLean, C., GP, A., Venkataraman, R., Liu, C., Krishnan, D., D'sa, J., Yogev, R., Collins, P., Lee, B., Ho, L., Doersch, C., Yona, G., Gao, S., Ferreira, F.T., Ozturk, A., Muckenhirn, H., Zheng, C., Balasubramaniam, G., Bansal, M., Driessche, G., Eiger, S., Haykal, S., Misra, V., Goyal, A., Martins, D., Leung, G., Valfridsson, J., Flynn, F., Bishop, W., Pang, C., Halpern, Y., Yu, H., Moore, L., Yuvein, Zhu, Thiagarajan, S., Drori, Y., Xiao, Z., Dery, L., Jagerman, R., Lu, J., Ge, E., Aggarwal, V., Khare, A., Tran, V., Elyada, O., Alet, F., Rubin, J., Chou, I., Tian, D., Bai, L., Chan, L., Lew, L., Misiunas, K., Bilal, T., Ray, A., Raghuram, S., Castro-Ros, A., Carpenter, V., Zheng, C., Kilgore, M., Broder, J., Xue, E., Kallakuri, P., Dua, D., Yuen, N., Chien, S., Schultz, J., Agrawal, S., Tsarfaty, R., Hu, J., Kannan, A., Marcus, D., Kothari, N., Sun, B., Horn, B., Bošnjak, M., Naeem, F., Hirsch, D., Chiang, L., Fang, B., Han, J., Wang, Q., Hora, B., He, A., Lučić, M., Changpinyo, B., Tripathi, A., Youssef, J., Kwak, C., Schlattner, P., Graves, C., Leblond, R., Zeng, W., Andreassen, A., Rasskin, G., Song, Y., Cao, E., Oh, J., Hoffman, M., Skut, W., Zhang, Y., Stritar, J., Cai, X., Khanna, S., Wang, K., Sharma, S., Reisswig, C., Jun, Y., Prasad, A., Sholokhova, T., Singh, P., Rosenthal, A.G., Ruoss, A., Beaufays, F., Kirmani, S., Chen, D., Schalkwyk, J., Herzig, J., Kim, B., Jacob, J., Vincent, D., Reyes, A.N., Balazevic, I., Hussenot, L., Schneider, J., Barnes, P., Castro, L., Babbula, S.R., Green, S., Cabi, S., Duduta, N., Driess, D., Galt, R., Velan, N., Wang, J., Jiao, H., Mauger, M., Phan, D., Patel, M., Galić, V., Chang, J., Marcus, E., Harvey, M., Salazar, J., Dabir, E., Sheth, S.S., Mandhane, A., Sedghi, H., Willcock, J., Zandieh, A., Prabhakara, S., Amini, A., Miech, A., Stone, V., Nicosia, M., Niemczyk, P., Xiao, Y., Kim, L., Kwasiborski, S., Verma, V., Oflazer, A.M., Hirschschall, C., Sung, P., Liu, L., Everett, R., Bakker, M., Weisz, Wang, Y., Sampathkumar, V., Shaham, U., Xu, B., Altun, Y., Wang, M., Saeki, T., Chen, G., Taropa, E., Vasanth, S., Austin, S., Huang, L., Petrovic, G., Dou, Q., Golovin, D., Rozhdestvenskiy, G., Culp, A., Wu, W., Sano, M., Jain, D., Proskurnia, J., Cevey, S., Ruiz, A.C., Patil, P., Mirzazadeh, M., Ni, E., Snaider, J., Fan, L., Fréchette, A., Pierigiovanni, A., Iqbal, S., Lee, K., Fantacci, C., Xing, J., Wang, L., Irpan, A., Raposo, D., Luan, Y., Chen, Z., Ganapathy, H., Hui, K., Nie, J., Guyon, I., Ge, H., Vij, R., Zheng, H., Lee, D., Castaño, A., Baatarsukh, K., Ibagon, G., Chronopoulou, A., FitzGerald, N., Viswanadha, S., Huda, S., Moroshko, R., Stoyanov, G., Kolhar, P., Vaucher, A., Watts, I., Kuncoro, A., Michalewski, H., Kambala, S., Batsaikhan, B.-O., Andreev, A., Jurenka, I., Le, M., Chen, Q., Jishi, W.A., Chakera, S., Chen, Z., Kini, A., Yadav, V., Siddhant, A., Labzovsky, I., Lakshminarayanan, B., Bostock, C.G., Botadra, P., Anand, A., Bishop, C., Conway-Rahman, S., Agarwal, M., Donchev, Y., Singhal, A., Chaumont Quitry, F., Ponomareva, N., Agrawal, N., Ni, B., Krishna, K., Samsikova, M., Karro, J., Du, Y., Glehn, T., Lu, C., Choquette-Choo, C.A., Qin, Z., Zhang, T., Li, S., Tyam, D., Mishra, S., Lowe, W., Ji, C., Wang, W., Faruqui, M., Slone, A., Dalibard, V., Narayanaswamy, A., Lambert, J., Manzagol, P.-A., Karliner, D., Bolt, A., Lobov, I., Kusupati, A., Ye, C., Yang, X., Zen, H., George, N., Bhutani, M., Lacombe, O., Riachi, R., Bansal, G., Soh, R., Gao, Y., Yu, Y., Yu, A., Nottage, E., Rojas-Esponda, T., Noraky, J., Gupta, M.,

Kotikalapudi, R., Chang, J., Deur, S., Graur, D., Mossin, A., Farnese, E., Figueira, R., Moufarek, A., Huang, A., Zochbauer, P., Ingram, B., Chen, T., Wu, Z., Puigdomènech, A., Rechis, L., Yu, D., Padmanabhan, S.G.S., Zhu, R., Ko, C.-l., Banino, A., Daruki, S., Selvan, A., Bhaswar, D., Diaz, D.H., Su, C., Scellato, S., Brennan, J., Han, W., Chung, G., Agrawal, P., Khandelwal, U., Sim, K.C., Lustman, M., Ritter, S., Guu, K., Xia, J., Jain, P., Wang, E., Hill, T., Rossini, M., Kostelac, M., Misiunas, T., Sabne, A., Kim, K., Iscen, A., Wang, C., Leal, J., Sreevatsa, A., Evci, U., Warmuth, M., Joshi, S., Suo, D., Lottes, J., Honke, G., Jou, B., Karp, S., Hu, J., Sahni, H., Taïga, A.A., Kong, W., Ghosh, S., Wang, R., Pavagadhi, J., Axelsson, N., Grigorev, N., Siegler, P., Lin, R., Wang, G., Parisotto, E., Maddineni, S., Subudhi, K., Ben-David, E., Pochernina, E., Keller, O., Avrahami, T., Yuan, Z., Mehta, P., Liu, J., Yang, S., Kan, W., Lee, K., Funkhouser, T., Cheng, D., Shi, H., Sharma, A., Kelley, J., Eyal, M., Malkov, Y., Tallec, C., Bahat, Y., Yan, S., Xintian, Wu, Lindner, D., Wu, C., Caciularu, A., Luo, X., Jenatton, R., Zaman, T., Bi, Y., Kornakov, I., Mallya, G., Ikeda, D., Karo, I., Singh, A., Evans, C., Netrapalli, P., Nallatamby, V., Tian, I., Assael, Y., Raunak, V., Carbune, V., Bica, I., Madmoni, L., Cattle, D., Grover, S., Somandepalli, K., Lall, S., Vázquez-Reina, A., Patana, R., Mu, J., Talluri, P., Tran, M., Aggarwal, R., Skerry-Ryan, R., Xu, J., Burrows, M., Pan, X., Yvinec, E., Lu, D., Zhang, Z., Nguyen, D.D., Mu, H., Barcik, G., Ran, H., Beltrone, L., Choromanski, K., Kharrat, D., Albanie, S., Purser-haskell, S., Bieber, D., Zhang, C., Wang, J., Hudson, T., Zhang, Z., Fu, H., Mauerer, J., Bateni, M.H., Maschinot, A., Wang, B., Zhu, M., Pillai, A., Weyand, T., Liu, S., Akerlund, O., Bertsch, F., Premachandran, V., Jin, A., Roulet, V., Boursac, P., Mittal, S., Ndebele, N., Karadzhov, G., Ghalebikesabi, S., Liang, R., Wu, A., Cong, Y., Ghelani, N., Singh, S., Fatemi, B., Warren, Chen, Kwong, C., Kolganov, A., Li, S., Song, R., Kuang, C., Miryoosefi, S., Webster, D., Wendt, J., Socala, A., Su, G., Mendonça, A., Gupta, A., Li, X., Tsai, T., Qiong, Hu, Kang, K., Chen, A., Girgin, S., Xian, Y., Lee, A., Ramsden, N., Baker, L., Elish, M.C., Krayvanova, V., Joshi, R., Simsa, J., Yang, Y.-Y., Ambroszczyk, P., Ghosh, D., Kar, A., Shanguan, Y., Yamamori, Y., Akulov, Y., Brock, A., Tang, H., Vashishtha, S., Munoz, R., Steiner, A., Andra, K., Eppens, D., Feng, Q., Kobayashi, H., Goldshtein, S., Mahdy, M.E., Wang, X., Jilei, Wang, Killam, R., Kwiatkowski, T., Koppurapu, K., Zhan, S., Jia, C., Bendebury, A., Luo, S., Recasens, A., Knight, T., Chen, J., Patel, M., Li, Y., Withbroe, B., Weesner, D., Bhatia, K., Ren, J., Eisenbud, D., Songhori, E., Sun, Y., Choma, T., Kementsietsidis, T., Manning, L., Roark, B., Farhan, W., Feng, J., Tatineni, S., Cobon-Kerr, J., Li, Y., Hendricks, L.A., Noble, I., Breaux, C., Kushman, N., Peng, L., Xue, F., Tobin, T., Rogers, J., Lipschultz, J., Alberti, C., Vlaskin, A., Dehghani, M., Sharma, R., Warkentin, T., Lee, C.-Y., Uria, B., Juan, D.-C., Chandorkar, A., Sheftel, H., Liu, R., Davoodi, E., Pigem, B.D.B., Dhamdhere, K., Ross, D., Hoech, J., Mahdieh, M., Liu, L., Li, Q., McCafferty, L., Liu, C., Mircea, M., Song, Y., Savant, O., Saade, A., Cherry, C., Hellendoorn, V., Goyal, S., Pucciarelli, P., Torres, D.V., Yahav, Z., Lee, H., Sjoesund, L.L., Kirov, C., Chang, B., Ghoshal, D., Li, L., Baechler, G., Pereira, S., Sainath, T., Boral, A., Grewe, D., Halumi, A., Phu, N.M., Shen, T., Ribeiro, M.T., Varma, D., Kaskasoli, A., Feinberg, V., Potti, N., Kahn, J., Wisniewski, M., Mohamed, S., Hrafnkelsson, A.M., Shahriari, B., Lespiau, J.-B., Patel, L., Yeung, L., Paine, T., Mei, L., Ramirez, A., Shivanna, R., Zhong, L., Woodward, J., Tubone, G., Khan, S., Chen, H., Nielsen, E., Ionescu, C., Prabhu, U., Gao, M., Wang, Q., Augenstein, S., Subramaniam, N., Chang, J., Iliopoulos, F., Luo, J., Khan, M., Kuo, W., Teplyashin, D., Perot, F., Kilpatrick, L., Globerson, A., Yu, H., Siddiqui, A., Sukhanov, N., Kandoor, A., Gupta, U., Andreetto, M., Ambar, M., Kim, D., Wośowski, P., Perrin, S., Limonchik, B., Fan, W., Stephan, J., Stewart-Binks, I., Kappedal, R., He, T., Cogan, S., Datta, R., Zhou, T., Ye, J., Kieliger, L., Ramalho, A., Kastner, K., Mentzer, F., Ko, W.-J., Suggala, A., Zhou, T., Butt, S., Strejček, H., Belenki, L., Venugopalan, S., Ling, M., Eltyshv, E., Deng, Y., Kovacs, G., Raghavachari, M., Dai, H., Schuster, T., Schwarcz, S., Nguyen, R., Nguyen, A., Buttimore, G., Mallick, S.B., Gandhe, S., Benjamin, S., Jastrzebski, M., Yan, L., Basu, S., Apps, C., Edkins, I., Allingham, J., Odisho, I., Kocisky, T., Zhao, J., Xue, L., Reddy, A., Anastasiou, C., Atias, A., Redmond, S., Milan, K., Heess, N., Schmit, H., Dafoe, A., Andor, D., Gangwani, T., Dragan, A., Zhang, S., Kachra, A., Wu, G., Xue, S., Aydin, K., Liu, S., Zhou, Y., Malihi, M., Wu, A., Gopal, S., Schumann, C., Stys,

P., Wang, A., Olšák, M., Liu, D., Schallhart, C., Mao, Y., Brady, D., Xu, H., Mery, T., Sitawarin, C., Velusamy, S., Cobley, T., Zhai, A., Walder, C., Katz, N., Jawahar, G., Kulkarni, C., Yang, A., Paszke, A., Wang, Y., Damoc, B., Borsos, Z., Smith, R., Li, J., Gupta, M., Kapishnikov, A., Prakash, S., Luisier, F., Agarwal, R., Grathwohl, W., Chen, K., Han, K., Mehta, N., Over, A., Azizi, S., Meng, L., Santo, N.D., Zheng, K., Shapiro, J., Petrovski, I., Hui, J., Ghafouri, A., Snoek, J., Qin, J., Jordan, M., Sikora, C., Malmaud, J., Kuang, Y., Świetlik, A., Sang, R., Shi, C., Li, L., Rosenberg, A., Zhao, S., Crawford, A., Peter, J.-T., Lei, Y., Garcia, X., Le, L., Wang, T., Amelot, J., Orr, D., Kacham, P., Alon, D., Tyen, G., Arora, A., Lyon, J., Kurakin, A., Ly, M., Guidroz, T., Yan, Z., Panigrahy, R., Xu, P., Kagohara, T., Cheng, Y., Noland, E., Lee, J., Lee, J., Yip, C., Wang, M., Nehoran, E., Bykovsky, A., Shan, Z., Bhagatwala, A., Yan, C., Tan, J., Garrido, G., Ethier, D., Hurley, N., Vesom, G., Chen, X., Qiao, S., Nayyar, A., Walker, J., Sandhu, P., Rosca, M., Swisher, D., Dekhtarev, M., Dillon, J., Muraru, G.-C., Tragut, M., Myaskovsky, A., Reid, D., Velic, M., Xiao, O., George, J., Brand, M., Li, J., Yu, W., Gu, S., Deng, X., Aubet, F.-X., Yeganeh, S.H., Alcober, F., Smith, C., Cohn, T., McKinney, K., Tschannen, M., Sampath, R., Cheon, G., Luo, L., Liu, L., Orbay, J., Peng, H., Botea, G., Zhang, X., Yoon, C., Magalhaes, C., Stradomski, P., Mackinnon, I., Hemingway, S., Venkatesan, K., May, R., Kim, J., Druinsky, A., Ye, J., Xu, Z., Huang, T., Abdallah, J.A., Dostmohamed, A., Fellinger, R., Munkhdalai, T., Maurya, A., Garst, P., Zhang, Y., Krikun, M., Bucher, S., Veerubhotla, A.S., Liu, Y., Li, S., Gupta, N., Adamek, J., Chen, H., Orlando, B., Zaks, A., Amersfoort, J., Camp, J., Wan, H., Choe, H., Wu, Z., Olszewska, K., Yu, W., Vadali, A., Scholz, M., Freitas, D.D., Lin, J., Hua, A., Liu, X., Ding, F., Zhou, Y., Severson, B., Tsihlias, K., Yang, S., Spalink, T., Yerram, V., Pankov, H., Blevins, R., Vargas, B., Jauhari, S., Miecznikowski, M., Zhang, M., Kumar, S., Farabet, C., Lan, C.L., Flennerhag, S., Bitton, Y., Ma, A., Bražinskas, A., Collins, E., Ahuja, N., Kudugunta, S., Bortsova, A., Giang, M., Zhu, W., Chi, E., Lundberg, S., Stern, A., Puttagunta, S., Xiong, J., Wu, X., Pande, Y., Jhindal, A., Murphy, D., Clark, J., Brockschmidt, M., Deines, M., McKee, K.R., Bahir, D., Shen, J., Truong, M., McDuff, D., Gesmundo, A., Rosseel, E., Liang, B., Caluwaerts, K., Hamrick, J., Kready, J., Cassin, M., Ingale, R., Lao, L., Pollom, S., Ding, Y., He, W., Bellot, L., Iljazi, J., Boppana, R.S., Han, S., Thompson, T., Khalifa, A., Bulanova, A., Mitrevski, B., Pang, B., Cooney, E., Shi, T., Coaguila, R., Yakar, T., Ranzato, M., Momchev, N., Rawles, C., Charles, Z., Maeng, Y., Zhang, Y., Bansal, R., Zhao, X., Albert, B., Yuan, Y., Vijayanarasimhan, S., Hirsch, R., Ramasesh, V., Vodrahalli, K., Wang, X., Gupta, A., Strouse, D., Ni, J., Patel, R., Taubman, G., Huo, Z., Gharibian, D., Monteiro, M., Lam, H., Vasudevan, S., Chaudhary, A., Albuquerque, I., Gupta, K., Riedel, S., Hegde, C., Ruderman, A., György, A., Wainwright, M., Chaugule, A., Ayan, B.K., Levinboim, T., Shleifer, S., Kalley, Y., Mirrokni, V., Rao, A., Radhakrishnan, P., Hartford, J., Wu, J., Zhu, Z., Bertolini, F., Xiong, H., Serrano, N., Tomlinson, H., Ott, M., Chang, Y., Graham, M., Li, J., Liang, M., Long, X., Borgeaud, S., Ahmad, Y., Grills, A., Mincu, D., Izzard, M., Liu, Y., Xie, J., O'Bryan, L., Ponda, S., Tong, S., Liu, M., Malkin, D., Salama, K., Chen, Y., Anil, R., Rao, A., Swavely, R., Bilenko, M., Anderson, N., Tan, T., Xie, J., Wu, X., Yu, L., Vinyals, O., Ryabtsev, A., Dangovski, R., Baumli, K., Keysers, D., Wright, C., Ashwood, Z., Chan, B., Shtefan, A., Guo, Y., Bapna, A., Soricut, R., Pecht, S., Ramos, S., Wang, R., Cai, J., Trinh, T., Barham, P., Friso, L., Stickgold, E., Ding, X., Shakeri, S., Ardila, D., Briakou, E., Culliton, P., Raveret, A., Cui, J., Saxton, D., Roy, S., Azizi, J., Yin, P., Loher, L., Bunner, A., Choi, M., Ahmed, F., Li, E., Li, Y., Dai, S., Elabd, M., Ganapathy, S., Agrawal, S., Hua, Y., Kunkle, P., Rajayogam, S., Ahuja, A., Conmy, A., Vasiloff, A., Beak, P., Yew, C., Mudigonda, J., Wydrowski, B., Blanton, J., Wang, Z., Dauphin, Y., Xu, Z., Polacek, M., Chen, X., Hu, H., Sho, P., Kunesch, M., Manshadi, M.H., Rutherford, E., Li, B., Hsiao, S., Barr, I., Tudor, A., Kecman, M., Nagrani, A., Pchelin, V., Sundermeyer, M., S, A.P., Karmarkar, A., Gao, Y., Chole, G., Bachem, O., Gao, I., BC, A., Dibb, M., Verzett, M., Hernandez-Campos, F., Lunts, Y., Johnson, M., Trapani, J.D., Koster, R., Brusilovsky, I., Xiong, B., Mohabey, M., Ke, H., Zou, J., Sabolić, T., Campos, V., Palowitch, J., Morris, A., Qiu, L., Ponnuramu, P., Li, F., Sharma, V., Sodhia, K., Tekelioglu, K., Chuklin, A., Yenugula, M., Gemzer, E., Strinopoulos, T., El-Husseini, S., Wang, H., Zhong, Y., Leurent, E., Natsev, P., Wang, W., Mahaarachchi, D., Zhu, T., Peng, S.,

Alabed, S., Lee, C.-C., Brohan, A., Szlam, A., Oh, G., Kovsharov, A., Lee, J., Wong, R., Barnes, M., Thornton, G., Gimeno, F., Levy, O., Sevenich, M., Johnson, M., Mallinson, J., Dadashi, R., Wang, Z., Ren, Q., Lahoti, P., Dhar, A., Feldman, J., Zheng, D., Ulrich, T., Panait, L., Blokzijl, M., Baetu, C., Matak, J., Harlalka, J., Shah, M., Marian, T., Dincklage, D., Du, C., Ley-Wild, R., Brownfield, B., Schumacher, M., Stuken, Y., Noghabi, S., Gupta, S., Ren, X., Malmi, E., Weissenberger, F., Huergo, B., Bauza, M., Lampe, T., Douillard, A., Seyedhosseini, M., Frostig, R., Ghahramani, Z., Nguyen, K., Krishnakumar, K., Ye, C., Gupta, R., Nazari, A., Geirhos, R., Shaw, P., Eleryan, A., Damen, D., Palomaki, J., Xiao, T., Wu, Q., Yuan, Q., Meadowlark, P., Bilotti, M., Lin, R., Sridhar, M., Schroecker, Y., Chung, D.-W., Luo, J., Strohman, T., Liu, T., Zheng, A., Emond, J., Wang, W., Lampinen, A., Fukuzawa, T., Campbell-Ajala, F., Roy, M., Lee-Thorp, J., Wang, L., Naim, I., Tony, N., Bensky, G., Gupta, A., Rogozińska, D., Fu, J., Pillai, T.S., Veličković, P., Drath, S., Neubeck, P., Tulsyan, V., Klimovskiy, A., Metzler, D., Stevens, S., Yeh, A., Yuan, J., Yu, T., Zhang, K., Go, A., Tsang, V., Xu, Y., Wan, A., Galatzer-Levy, I., Sobell, S., Toki, A., Salesky, E., Zhou, W., Antognini, D., Douglas, S., Wu, S., Lelkes, A., Kim, F., Cavallaro, P., Salazar, A., Liu, Y., Besley, J., Refice, T., Jia, Y., Li, Z., Sokolik, M., Kannan, A., Simon, J., Chick, J., Aharon, A., Gandhi, M., Daswani, M., Amiri, K., Birodkar, V., Ittycheriah, A., Grabowski, P., Chang, O., Sutton, C., Zhixin, Lai, Telang, U., Sargsyan, S., Jiang, T., Hoffmann, R., Brichtova, N., Hessel, M., Halcrow, J., Jerome, S., Brown, G., Tomala, A., Buchatskaya, E., Yu, D., Menon, S., Moreno, P., Liao, Y., Zayats, V., Tang, L., Mah, S., Shenoy, A., Siegman, A., Hadian, M., Kwon, O., Tu, T., Khajehnouri, N., Foley, R., Haghani, P., Wu, Z., Keshava, V., Gupta, K., Bruguier, T., Yao, R., Karmon, D., Zintgraf, L., Wang, Z., Piqueras, E., Jung, J., Brennan, J., Machado, D., Giustina, M., Tessler, M., Lee, K., Zhang, Q., Moore, J., Dagaard, K., Frömmgen, A., Beattie, J., Zhang, F., Kasenberg, D., Geri, T., Qin, D., Tomar, G.S., Ouyang, T., Yu, T., Zhou, L., Mathews, R., Davis, A., Li, Y., Gupta, J., Yates, D., Deng, L., Kemp, E., Joung, G.-Y., Vassilvitskii, S., Guo, M., LV, P., Dopson, D., Lachgar, S., McConnaughey, L., Choudhury, H., Dena, D., Cohen, A., Ainslie, J., Levi, S., Gopavarapu, P., Zablotzkaia, P., Vallet, H., Bahargam, S., Tang, X., Tomasev, N., Dyer, E., Balle, D., Lee, H., Bono, W., Mendez, J.G., Zubov, V., Yang, S., Rendulic, I., Zheng, Y., Hogue, A., Pundak, G., Leith, R., Bhoopchand, A., Han, M., Žanić, M., Schaul, T., Delakis, M., Iyer, T., Wang, G., Singh, H., Abdelhamed, A., Thomas, T., Brahma, S., Dib, H., Kumar, N., Zhou, W., Bai, L., Mishra, P., Sun, J., Anklin, V., Sukkerd, R., Agubuzu, L., Briukhov, A., Gulati, A., Sieb, M., Pardo, F., Nasso, S., Chen, J., Zhu, K., Sosea, T., Goldin, A., Rush, K., Hombaiah, S.A., Noever, A., Zhou, A., Haves, S., Phuong, M., Ades, J., Chen, Y.-t., Yang, L., Pagadora, J., Bileschi, S., Cotruta, V., Saputro, R., Pramanik, A., Ammirati, S., Garrette, D., Villela, K., Blyth, T., Akbulut, C., Jha, N., Rustemi, A., Wongpanich, A., Nagpal, C., Wu, Y., Rivière, M., Kishchenko, S., Srinivasan, P., Chen, A., Sinha, A., Pham, T., Jia, B., Hennigan, T., Bakalov, A., Attaluri, N., Garmon, D., Rodriguez, D., Wegner, D., Jia, W., Senter, E., Fiedel, N., Petek, D., Liu, Y., Hardin, C., Lehri, H.T., Carreira, J., Smoot, S., Prasetya, M., Akazawa, N., Stefanoiu, A., Ho, C.-H., Angelova, A., Lin, K., Kim, M., Chen, C., Sieniek, M., Li, A., Guo, T., Baltateanu, S., Tafti, P., Wunder, M., Olmert, N., Shukla, D., Shen, J., Kovelamudi, N., Venkatraman, B., Neel, S., Thoppilan, R., Connor, J., Benzing, F., Stjerngren, A., Ghiasi, G., Polozov, A., Howland, J., Weber, T., Chiu, J., Girirajan, G.P., Terzis, A., Wang, P., Li, F., Shalom, Y.B., Tewari, D., Denton, M., Aharoni, R., Kalb, N., Zhao, H., Zhang, J., Filos, A., Rahtz, M., Jain, L., Fan, C., Rodrigues, V., Wang, R., Shin, R., Austin, J., Ring, R., Sanchez-Vargas, M., Hassen, M., Kessler, I., Alon, U., Zhang, G., Chen, W., Ma, Y., Si, X., Hou, L., Mirhoseini, A., Wilson, M., Bacon, G., Roelofs, B., Shu, L., Vasudevan, G., Adler, J., Dwornik, A., Terzi, T., Lawlor, M., Askham, H., Bernico, M., Dong, X., Hidey, C., Kilgour, K., Liu, G., Bhupatiraju, S., Leonhard, L., Zuo, S., Talukdar, P., Wei, Q., Severyn, A., Listík, V., Lee, J., Tripathi, A., Park, S., Matias, Y., Liu, H., Ruiz, A., Jayaram, R., Tolins, J., Marcenac, P., Wang, Y., Seybold, B., Prior, H., Sharma, D., Weber, J., Sirotenko, M., Sung, Y., Du, D., Pavlick, E., Zinke, S., Freitag, M., Dylla, M., Arenas, M.G., Potikha, N., Goldman, O., Tao, C., Chhaparia, R., Voitovich, M., Dogra, P., Ražnatović, A., Tsai, Z., You, C., Johnson, O., Tucker, G., Gu, C., Yoo, J., Majzoubi, M., Gabeur, V., Raad, B., Rhodes, R., Kolipaka,

K., Howard, H., Sampemane, G., Li, B., Asawaroengchai, C., Nguyen, D., Zhang, C., Cour, T., Yu, X., Fu, Z., Jiang, J., Huang, P.-S., Surita, G., Iturrate, I., Karov, Y., Collins, M., Baeuml, M., Fuchs, F., Shetty, S., Ramaswamy, S., Ebrahimi, S., Guo, Q., Shar, J., Barth-Marion, G., Addepalli, S., Richter, B., Cheng, C.-Y., Rives, E., Zheng, F., Griesser, J., Dikkala, N., Zeldes, Y., Safarli, I., Das, D., Srivastava, H., Khan, S.M., Li, X., Pandey, A., Markeeva, L., Belov, D., Yan, Q., Rybiński, M., Chen, T., Nawhal, M., Quinn, M., Govindaraj, V., York, S., Roberts, R., Garg, R., Godbole, N., Abernethy, J., Das, A., Thiet, L.N., Tompson, J., Nham, J., Vats, N., Caine, B., Helmholtz, W., Pongetti, F., Ko, Y., An, J., Hu, C.H., Ling, Y.-C., Pawar, J., Leland, R., Kinoshita, K., Khawaja, W., Selvi, M., Ie, E., Sinopalnikov, D., Proleev, L., Tripuraneni, N., Bevilacqua, M., Lee, S., Sanford, C., Suh, D., Tran, D., Dean, J., Baumgartner, S., Heitkaemper, J., Gubbi, S., Toutanova, K., Xu, Y., Thekkath, C., Rong, K., Jain, P., Xie, A., Virin, Y., Li, Y., Litchev, L., Powell, R., Bharti, T., Kraft, A., Hua, N., Ikonomidis, M., Hitron, A., Kumar, S., Matthey, L., Bridgers, S., Lax, L., Malhi, I., Skopek, O., Gupta, A., Cao, J., Rasquinha, M., Pöder, S., Stokowiec, W., Roth, N., Li, G., Sander, M., Kessinger, J., Jain, V., Loper, E., Park, W., Yarom, M., Cheng, L., Guruganesh, G., Rao, K., Li, Y., Barros, C., Sushkov, M., Ferng, C.-S., Shah, R., Aharoni, O., Kumar, R., McConnell, T., Li, P., Wang, C., Pereira, F., Swanson, C., Jamil, F., Xiong, Y., Vijayakumar, A., Shroff, P., Soparkar, K., Gu, J., Soares, L.B., Wang, E., Majmundar, K., Wei, A., Bailey, K., Kassner, N., Kawamoto, C., Žužić, G., Gomes, V., Gupta, A., Guzman, M., Dasgupta, I., Bai, X., Pan, Z., Piccinno, F., Vogel, H.N., Ponce, O., Hutter, A., Chang, P., Jiang, P.-P., Gog, I., Ionescu, V., Manyika, J., Pedregosa, F., Ragan, H., Behrman, Z., Mullins, R., Devin, C., Pyne, A., Gawde, S., Chadwick, M., Gu, Y., Tavakkol, S., Twigg, A., Goyal, N., Elue, N., Goldie, A., Venkatachary, S., Fei, H., Feng, Z., Ritter, M., Leal, I., Dasari, S., Sun, P., Rochman, A.R., O'Donoghue, B., Liu, Y., Sproch, J., Chen, K., Clay, N., Petrov, S., Sidhwani, S., Mihailescu, I., Panagopoulos, A., Piergiovanni, A., Bai, Y., Powell, G., Karkhanis, D., Yacovone, T., Mitrichev, P., Kovac, J., Uthus, D., Yazdanbakhsh, A., Amos, D., Zheng, S., Zhang, B., Miao, J., Ramabhadran, B., Radpour, S., Thakoor, S., Newlan, J., Lang, O., Jankowski, O., Bharadwaj, S., Sarr, J.-M., Ashraf, S., Mondal, S., Yan, J., Rawat, A.S., Velury, S., Kochanski, G., Eccles, T., Och, F., Sharma, A., Mahintorabi, E., Gurney, A., Muir, C., Cohen, V., Thakur, S., Bloniarz, A., Mujika, A., Pritzler, A., Caron, P., Rahman, A., Lang, F., Onoe, Y., Sirkovic, P., Hoover, J., Jian, Y., Duque, P., Narayanan, A., Soergel, D., Haig, A., Maggiore, L., Buch, S., Dean, J., Figotin, I., Karpov, I., Gupta, S., Zhou, D., Huang, M., Vaswani, A., Semturs, C., Shivakumar, K., Watanabe, Y., Rajendran, V.K., Lu, E., Hou, Y., Ye, W., Vashishth, S., Nti, N., Sakenas, V., Ni, D., DeCarlo, D., Bendersky, M., Bagri, S., Cano, N., Peake, E., Tokumine, S., Godbole, V., Guía, C., Lando, T., Selo, V., Ellis, S., Tarlow, D., Gillick, D., Epasto, A., Jonnalagadda, S.R., Wei, M., Xie, M., Taly, A., Paganini, M., Sundararajan, M., Toyama, D., Yu, T., Petrova, D., Pappu, A., Agrawal, R., Buthpitiya, S., Frye, J., Buschmann, T., Crocker, R., Tagliasacchi, M., Wang, M., Huang, D., Perel, S., Wieder, B., Kazawa, H., Wang, W., Cole, J., Gupta, H., Golan, B., Bang, S., Kulkarni, N., Franko, K., Liu, C., Reid, D., Dalmia, S., Whang, J., Cen, K., Sundaram, P., Ferret, J., Isik, B., Ionita, L., Sun, G., Shekhawat, A., Mohammad, M., Pham, P., Huang, R., Raman, K., Zhou, X., Mcilroy, R., Myers, A., Peng, S., Scott, J., Covington, P., Erell, S., Joshi, P., Oliveira, J.G., Noy, N., Nasir, T., Walker, J., Axelrod, V., Dozat, T., Han, P., Chu, C.-T., Weinstein, E., Shukla, A., Chandrakaladharan, S., Poklukur, P., Li, B., Jin, Y., Eruvbetine, P., Hansen, S., Dabush, A., Jacovi, A., Phatale, S., Zhu, C., Baker, S., Shomrat, M., Xiao, Y., Pouget-Abadie, J., Zhang, M., Wei, F., Song, Y., King, H., Huang, Y., Zhu, Y., Sun, R., Franco, J.V., Lin, C.-C., Arora, S., Hui, Li, Xia, V., Vilnis, L., Schain, M., Alarakya, K., Prince, L., Phillips, A., Habtegebriel, C., Xu, L., Gui, H., Ontanon, S., Aroyo, L., Gill, K., Lu, P., Katariya, Y., Madeka, D., Krishnan, S., Raghvendra, S.S., Freedman, J., Tay, Y., Menghani, G., Choy, P., Shetty, N., Abolafia, D., Kukliansky, D., Chou, E., Lichtarge, J., Burke, K., Coleman, B., Guo, D., Jin, L., Bhattacharya, I., Langston, V., Li, Y., Kotecha, S., Yakubovich, A., Chen, X., Petrov, P., Powell, T., He, Y., Quick, C., Garg, K., Hwang, D., Lu, Y., Bhojanapalli, S., Kijms, K., Mehran, R., Archer, A., Hasselt, H., Balakrishna, A., Kearns, J., Guo, M., Riesa, J., Sazanovich, M., Gao, X., Sauer, C., Yang, C., Sheng, X., Jimma, T., Gansbeke, W.V., Nikolaev,

V., Wei, W., Millican, K., Zhao, R., Snyder, J., Bolelli, L., O'Brien, M., Xu, S., Xia, F., Yuan, W., Neelakantan, A., Barker, D., Yadav, S., Kirkwood, H., Ahmad, F., Wee, J., Grimstad, J., Wang, B., Wiethoff, M., Settle, S., Wang, M., Blundell, C., Chen, J., Duvarney, C., Hu, G., Ronneberger, O., Lee, A., Li, Y., Chakladar, A., Butryna, A., Evangelopoulos, G., Desjardins, G., Kanerva, J., Wang, H., Nowak, A., Li, N., Loo, A., Khurshudov, A., Shafey, L.E., Baddi, N., Lenc, K., Razeghi, Y., Lieber, T., Sinha, A., Ma, X., Su, Y., Huang, J., Ushio, A., Klimczak-Plucińska, H., Mohamed, K., Chen, J., Osindero, S., Ginzburg, S., Lamprou, L., Bashlovkina, V., Tran, D.-H., Khodaei, A., Anand, A., Di, Y., Eskander, R., Vuyyuru, M.R., Liu, J., Kamath, A., Goldenberg, R., Bellaiche, M., Pluto, J., Rosgen, B., Mansoor, H., Wong, W., Ganesh, S., Bailey, E., Baird, S., Deutsch, D., Baek, J., Jia, X., Lee, C., Friesen, A., Braun, N., Lee, K., Panda, A., Hernandez, S.M., Williams, D., Liu, J., Liang, E., Autef, A., Pitler, E., Jain, D., Kirk, P., Bunyan, O., Elias, J.S., Yin, T., Reid, M., Pope, A., Putikhin, N., Samanta, B., Guadarrama, S., Kim, D., Rowe, S., Valentine, M., Yan, G., Salcianu, A., Silver, D., Song, G., Singh, R., Ye, S., DeBalsi, H., Merey, M.A., Ofek, E., Webson, A., Mourad, S., Kakarla, A., Lattanzi, S., Roy, N., Sluzhaev, E., Butterfield, C., Tonioni, A., Waters, N., Kopalle, S., Chase, J., Cohan, J., Rao, G.R., Berry, R., Voznesensky, M., Hu, S., Chiafullo, K., Chikkerur, S., Scrivener, G., Zheng, I., Wiesner, J., Macherey, W., Lillicrap, T., Liu, F., Walker, B., Welling, D., Davies, E., Huang, Y., Ren, L., Shabat, N., Agostini, A., Iinuma, M., Zelle, D., Sathyanarayana, R., D'olimpio, A., Redshaw, M., Ginsberg, M., Murthy, A., Geller, M., Matejovicova, T., Chakrabarti, A., Julian, R., Chan, C., Hu, Q., Jarrett, D., Agarwal, M., Challagundla, J., Li, T., Tata, S., Ding, W., Meng, M., Dai, Z., Vezzani, G., Garg, S., Bulian, J., Jasarevic, M., Cai, H., Rajamani, H., Santoro, A., Hartmann, F., Liang, C., Perz, B., Jindal, A., Bu, F., Seo, S., Poplin, R., Goedeckemeyer, A., Ghazi, B., Khadke, N., Liu, L., Mather, K., Zhang, M., Shah, A., Chen, A., Wei, J., Shivam, K., Cao, Y., Cho, D., Scarpati, A.S., Moffitt, M., Barbu, C., Jurin, I., Chang, M.-W., Liu, H., Zheng, H., Dave, S., Kaeser-Chen, C., Yu, X., Abdagic, A., Gonzalez, L., Huang, Y., Zhong, P., Schmid, C., Petrini, B., Wertheim, A., Zhu, J., Nguyen, H., Ji, K., Zhou, Y., Zhou, T., Feng, F., Cohen, R., Rim, D., Phal, S.M., Georgiev, P., Brand, A., Ma, Y., Li, W., Gupta, S., Wang, C., Dubov, P., Tarbouriech, J., Majumder, K., Li, H., Rink, N., Suman, A., Guo, Y., Sun, Y., Nair, A., Xu, X., Elhawaty, M., Cabrera, R., Han, G., Eisenschlos, J., Bai, J., Li, Y., Bansal, Y., Sellam, T., Khan, M., Nguyen, H., Mao-Jones, J., Parotsidis, N., Marcus, J., Fan, C., Zimmermann, R., Kochinski, Y., Graesser, L., Behbahani, F., Caceres, A., Riley, M., Kane, P., Lefdal, S., Willoughby, R., Vicol, P., Wang, L., Zhang, S., Gill, A., Liang, Y., Prasad, G., Mariooryad, S., Kazemi, M., Wang, Z., Muralidharan, K., Voigtlaender, P., Zhao, J., Zhou, H., D'Souza, N., Mavalankar, A., Arnold, S., Young, N., Sarvana, O., Lee, C., Nasr, M., Zou, T., Kim, S., Haas, L., Patel, K., Bulut, N., Parkinson, D., Biles, C., Kalashnikov, D., To, C.M., Kumar, A., Austin, J., Greve, A., Zhang, L., Goel, M., Li, Y., Yaroshenko, S., Chang, M., Jindal, A., Clark, G., Taitelbaum, H., Johnson, D., Roval, O., Ko, J., Mohananey, A., Schuler, C., Dodhia, S., Li, R., Osawa, K., Cui, C., Xu, P., Shah, R., Huang, T., Gruzewska, E., Clement, N., Verma, M., Sercinoglu, O., Qian, H., Shah, V., Yamaguchi, M., Modi, A., Kosakai, T., Strohmman, T., Zeng, J., Gunel, B., Qian, J., Tarango, A., Jastrzebski, K., David, R., Shan, J., Schuh, P., Lad, K., Gierke, W., Madhavan, M., Chen, X., Kurzeja, M., Santamaria-Fernandez, R., Chen, D., Cordell, A., Chervonyi, Y., Garcia, F., Kannen, N., Perot, V., Ding, N., Cohen-Ganor, S., Lavrenko, V., Wu, J., Evans, G., Santos, C.N., Sewak, M., Brown, A., Hard, A., Puigcerver, J., Zheng, Z., Liang, Y., Gladchenko, E., Ingle, R., First, U., Sermanet, P., Magister, C., Velimirović, M., Reddi, S., Ricco, S., Agustsson, E., Adam, H., Levine, N., Gaddy, D., Holtmann-Rice, D., Wang, X., Sathe, A., Roy, A.G., Bratanić, B., Carin, A., Mehta, H., Bonacina, S., Cao, N.D., Finkelstein, M., Rieser, V., Wu, X., Altché, F., Scandinaro, D., Li, L., Vieillard, N., Sethi, N., Tanzer, G., Xing, Z., Wang, S., Bhatia, P., Citovsky, G., Anthony, T., Lin, S., Shi, T., Jakobovits, S., Gibson, G., Apte, R., Lee, L., Chen, M., Byravan, A., Maniatis, P., Webster, K., Dai, A., Chen, P.-C., Pan, J., Fadeeva, A., Gleicher, Z., Luong, T., Bhumihar, N.K.: Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities (2025). <https://arxiv.org/abs/2507.06261>

- [121] OpenAI: Gpt-5 system card. Technical report, OpenAI (August 2025). Accessed: September 20, 2025. <https://cdn.openai.com/gpt-5-system-card.pdf>
- [122] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.*: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022)
- [123] Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980* (2020)
- [124] Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V (2023). <https://arxiv.org/abs/2310.11441>
- [125] Qwen Team: QVQ: To See the World with Wisdom. <https://qwenlm.github.io/blog/qvq-72b-preview/>. Accessed: 2025-08-22 (2024)
- [126] Telea, A.: An image inpainting technique based on the fast marching method. *Journal of Graphics Tools* **9**(1), 23–34 (2004)
- [127] Bertalmio, M., Bertozzi, A.L., Sapiro, G.: Navier–stokes, fluid dynamics, and image and video inpainting. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 355–362. IEEE, ??? (2001)
- [128] Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing* **13**(9), 1200–1212 (2004)
- [129] Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pp. 417–424. ACM Press/Addison-Wesley Publishing Co., ??? (2000)
- [130] Pawlik, M., Augsten, N.: Tree edit distance: Robust and memory-efficient. *Information Systems* **56**, 157–173 (2016) <https://doi.org/10.1016/j.is.2015.08.004>
- [131] Wang, B., Wu, F., Ouyang, L., Gu, Z., Zhang, R., Xia, R., Zhang, B., He, C.: Cdm: A reliable metric for fair and accurate formula recognition evaluation. *arXiv e-prints*, 2409 (2024)
- [132] Bai, L., Cai, Z., Cao, Y., Cao, M., Cao, W., Chen, C., Chen, H., Chen, K., Chen, P., Chen, Y., Chen, Y., Cheng, Y., Chu, P., Chu, T., Cui, E., Cui, G., Cui, L., Cui, Z., Deng, N., Ding, N., Dong, N., Dong, P., Dou, S., Du, S., Duan, H., Fan, C., Gao, B., Gao, C., Gao, J., Gao, S., Gao, Y., Gao, Z., Ge, J., Ge, Q., Gu, L., Gu, Y., Guo, A., Guo, Q., Guo, X., He, C., He, J., Hong, Y., Hou, S., Hu, C., Hu, H., Hu, J., Hu, M., Hua, Z., Huang, H., Huang, J., Huang, X., Huang, Z., Jiang, Z., Kong, L., Li, L., Li, P., Li, P., Li, S., Li, T., Li, W., Li, Y., Lin, D., Lin, J., Lin, T., Lin, Z., Liu, H., Liu, J., Liu, J., Liu, J., Liu, K., Liu, K., Liu, K., Liu, S., Liu, S., Liu, W., Liu, X., Liu, Y., Liu, Z., Lu, Y., Lv, H., Lv, H., Lv, H., Lv, Q., Lv, Y., Lyu, C., Ma, C., Ma, J., Ma, R., Ma, R., Ma, R., Ma, X., Ma, Y., Ma, Z., Mi, S., Ning, J., Ning, W., Pang, X., Peng, J., Peng, R., Qiao, Y., Qiu, J., Qu, X., Qu, Y., Ren, Y., Shang, F., Shao, W., Shen, J., Shen, S., Song, C., Song, D., Song, D., Su, C., Su, W., Sun, W., Sun, Y., Tan, Q., Tang, C., Tang, H., Tang, K., Tang, S., Tong, J., Wang, A., Wang, B., Wang, D., Wang, L., Wang, R., Wang, W., Wang, W., Wang, J., Wang, Y., Wang, Z., Wu, L.-L., Wu, W., Wu, Y., Wu, Z., Xiao, L., Xing, S., Xu, C., Xu, H., Xu, J., Xu, R., Xu, W., Yang, G., Yang, Y., Ye, H., Ye, J., Ye, S., Yu, J., Yu, J., Yu, J., Yuan, F., Zang, Y., Zhang, B., Zhang, C., Zhang, C., Zhang, H., Zhang, J., Zhang, Q., Zhang, Q., Zhang, S., Zhang, T., Zhang, W., Zhang, W., Zhang, Y., Zhang, Z., Zhao, H., Zhao, Q., Zhao, X., Zhao, X., Zhou, B., Zhou,

D., Zhou, P., Zhou, Y., Zhou, Y., Zhu, D., Zhu, L., Zou, Y.: Intern-S1: A Scientific Multimodal Foundation Model (2025). <https://arxiv.org/abs/2508.15763>

- [133] Nguyen, L., Scialom, T., Piwowarski, B., Staiano, J.: Loralay: A multilingual and multimodal dataset for long range and layout-aware summarization. arXiv preprint arXiv:2301.11312 (2023)