HoloDP: A Multilingual Benchmark Suite and Script-Aware Framework for Document Parsing

Mengxiao Zhu[1,2], Wencan Ying[1,2], Yingjie Cao[1,2], Jie Liu[1,2]*, Jiu Sha[3]

[1]School of Artificial Intelligence and Computer Science, North China University of Technology
[2]Beijing Key Laboratory of Key Technologies for AI+ Domain Applications
[3]School of Information Engineering and the Institute of National Security, Minzu University of China

Abstract

Millions of multilingual documents remain locked in image formats, but current document parsers are still largely evaluated and optimized for a narrow range of high-resource languages and regular writing layouts. As a result, their reliability drops sharply when facing underrepresented scripts, vertical or right-to-left writing, and mixed-script document structures. We present HoloDP, a multilingual benchmark suite and script-aware framework for document parsing. HoloDP first introduces a standardized evaluation suite covering 13 languages across seven script families, with tasks spanning page-level layout analysis, region-level parsing, and mixed-script structural composition. Built on this resource, HoloDP further develops a script-aware parser that combines a Swin Transformer encoder, a four-branch direction-adaptive vision-to-text module, and a multilingual decoder. The adaptive module explicitly captures horizontal, vertical, diagonal, and global spatial priors, enabling the model to better handle heterogeneous writing systems and layouts. Extensive experiments show that HoloDP delivers consistent gains on underrepresented and mixed-script benchmarks while remaining competitive on high-resource Chinese and English documents. Ablation studies confirm that progressive multilingual optimization, direction-aware adaptation, and script-conditioned modeling provide complementary benefits. Our model and benchmark will be available at https://github.com/NcutLLM/HoloDP.

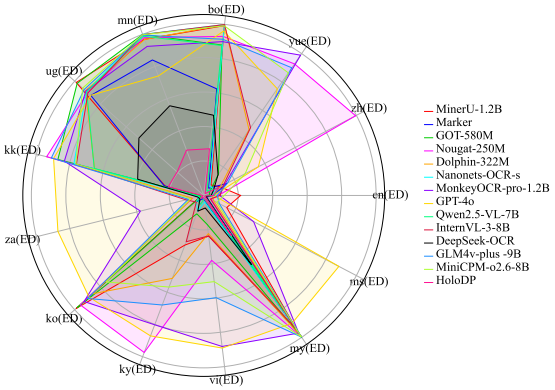


Figure 1: Comparison of HoloDP with representative baselines across multilingual document parsing benchmarks. A smaller enclosed area indicates stronger overall performance.

1 Introduction

Millions of multilingual documents in archives, educational materials, legal records, and cultural repositories remain accessible only as non-editable images. Document parsing aims to convert such page images into structured, machine-readable representations that preserve layout, reading order, and region-level content. Recent general-purpose vision-language models, such as GPT-4o (OpenAI, 2024) and Qwen2.5-VL (Bai et al., 2025), as well as specialized document parsing systems such as GoT (Wei et al., 2024) and dots.ocr (Li et al., 2025a), have achieved strong results on common document understanding tasks. However, their development and evaluation are concentrated on high-resource languages and regular layouts. As a result, document parsing for underrepresented scripts, heterogeneous writing directions, and mixed-script structures remains insufficiently studied.

This gap is difficult to close because two challenges interact with each other: supervision scarcity and structural script heterogeneity. Many underrepresented languages lack page-level annotations for layout structure, reading order, and fine-grained region content, limiting both model training and reliable evaluation. At the same time, their scripts often differ substantially in visual form and spatial organization. For example, Mongolian may appear in vertical text flows, Uyghur follows right-

*Corresponding author, liujie@ncut.edu.cn.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 32453–32470
October 24-29, 2026 ©2026 Association for Computational Linguistics

to-left writing conventions, and Brahmic-derived scripts involve complex grapheme composition. Such variation creates diverse alignment patterns between visual regions and textual sequences, making it difficult for parsers trained on visually homogeneous data to recover both content and structure.

To address these challenges, we introduce HoloDP, a multilingual benchmark suite and script-aware framework for document parsing. HoloDP is built on the premise that robust multilingual document parsing requires both reliable evaluation resources and models with explicit spatial-script inductive biases. On the resource side, we construct a large-scale multilingual corpus covering 13 languages across seven script families by combining public document resources, web-rendered document snapshots, and schema-driven synthetic documents. Based on this corpus, we build standardized benchmarks for page-level layout analysis, region-level content parsing, and mixed-script structural composition.

On the modeling side, HoloDP uses a Swin Transformer visual encoder, a four-branch adaptive vision-to-text module, and a multilingual decoder. The adaptive module models horizontal, vertical, diagonal, and global spatial priors, allowing the parser to better handle diverse writing directions and layout patterns. A script-conditioned projection further encourages parameter sharing across related writing systems, while a progressive three-stage training strategy stabilizes visual grounding and multilingual generation.

Experiments on public and newly constructed benchmarks show that HoloDP achieves strong gains on underrepresented and mixed-script document parsing tasks, while remaining competitive on high-resource Chinese and English settings. Further analyses show that multilingual data resources, direction-aware adaptation, and script-conditioned modeling provide complementary benefits.

Our contributions are summarized as follows:

(1) We introduce a multilingual document parsing resource suite, including a large-scale training corpus and standardized benchmarks covering 13 languages and seven script families, enabling page-level, region-level, and mixed-script evaluation under script-diverse settings.

(2) We propose a script-aware document parser that explicitly models direction-sensitive spatial priors through a four-branch adaptive vision-to-text module, script-conditioned projection, and progressive multilingual training.

(3) We conduct comprehensive experiments and analyses showing that HoloDP improves robustness on underrepresented and mixed-script scenarios while remaining competitive on high-resource Chinese and English benchmarks.

2 Related Work

2.1 Document Parsing Paradigms

Existing document parsing methods can be broadly grouped into pipeline tools, expert document VLMs, and general-purpose VLMs. Pipeline tools such as MinerU (Wang et al., 2024b), Dolphin (Feng et al., 2025), Marker[1], PP-StructureV3 (Cui et al., 2025), and Mamba-Yolo (Ma et al., 2026) decompose document parsing into layout detection, OCR, and post-processing, offering controllable intermediate outputs but suffering from error propagation under complex layouts and mixed writing directions. Expert document VLMs, including Nougat (Blecher et al., 2024), GOT-OCR (Wei et al., 2024), MonkeyOCR (Li et al., 2025b), dots.ocr (Li et al., 2025a), and DeepSeek-OCR (Wei et al., 2025), generate structured outputs in an end-to-end manner, but are mostly developed and evaluated on high-resource languages or visually regular documents. General-purpose VLMs such as GPT-4o (OpenAI, 2024), Qwen2.5-VL (Bai et al., 2025), GLM4v-plus (Zeng et al., 2024), InternVL-3 (Zhu et al., 2025), and MiniCPM-o (Yao et al., 2024) show strong multimodal generalization, yet they lack explicit mechanisms for direction-sensitive layout modeling and reading-order recovery in script-diverse documents.

2.2 Low-Resource and Script-Diverse Document Understanding

Low-resource document understanding is constrained by limited annotations and script-specific visual complexity. Prior work often focuses on character-level or line-level OCR, using transfer learning, few-shot adaptation, or synthetic data augmentation to reduce annotation costs (Wasfy et al., 2025; Sabbagh, 2025). However, these methods rarely address page-level layout structure, region-level parsing, and mixed-script reading order, which are essential for full document parsing.

[1]https://github.com/datalab-to/marker

2.3 Datasets and Multilingual Evaluation

Existing document understanding datasets support progress in layout analysis, formula recognition, table parsing, and document-level evaluation. Benchmarks such as OmniDocBench (Ouyang et al., 2025) and FoxPage (Liu et al., 2024a) provide document-level evaluation, while CPE (Wang et al., 2024a), PubTabNet (Zhong et al., 2020), and PubTab1M (Smock et al., 2022) focus on formulas and tables. Recent multilingual benchmarks have expanded this scope. XDocParse (Li et al., 2025a) covers broad OCR and layout parsing across 126 languages, SEA-Vision (Yue et al., 2026) evaluates document parsing and text-centric VQA across 11 Southeast Asian languages, and DocAtlas (Heakl et al., 2026) covers 82 languages and multiple OCR/document understanding tasks. HoloDP is complementary to these efforts: rather than claiming broad multilingual priority, we focus on jointly evaluating page-level layout analysis, region-level content parsing, and mixed-script structural composition across seven script families and heterogeneous writing directions, together with a direction- and script-aware parsing framework.

Type	ISO	Language	Script Family	training data #pages/ #regions	benchmarks #pages/ #regions
Monolingual	zh	Standard Chinese	Han (CJK)	300000/700000	712/5000
	en	English	Latin	300000/700000	402/28825
	yue	Yue Chinese	Han (CJK)	40000/110000	2000/5000
	bo	Tibetan	Brahmic	60000/140000	2000/5000
	mn	Mongolian	Mongolic	30000/90000	2000/5000
	ug	Uyghur	Arabic-derived	35000/105000	2000/5000
	kk	Kazakh	Cyrillic	20000/90000	2000/5000
	za	Zhuang	Latin	15000/75000	2000/5000
	ko	Korean	Hangul	40000/140000	2000/5000
	ky	Kyrgyz	Cyrillic	10000/50000	2000/5000
	vi	Vietnamese	Latin	30000/120000	2000/5000
	my	Burmese	Brahmic	20000/100000	2000/5000
	ms	Malay	Latin	50000/130000	2000/5000
Cross-lingual			–	400000/-	11079/-
Total	–	–	–	1350000/2550000	34193/88825

Table 1: Statistics of the HoloDP dataset.

3 Data Construction and Benchmarks

Because document parsing involves multiple output granularities beyond plain OCR, we first define a unified task format to organize our corpus and benchmarks. Under this formulation, HoloDP provides both a multilingual training corpus and a standardized benchmark suite, covering 13 languages across seven script families, including high-resource anchor languages (Chinese and English) and 11 underrepresented languages. Table 1 summarizes the language coverage and statistics of both the training corpus and benchmark suite.

3.1 Task Definition

We define multilingual document parsing as the task of converting a document image $\mathbf{I}$ into a structured representation $\mathbf{Y}$ that preserves textual content, layout elements, region boundaries, and reading order. Unlike plain OCR, which mainly focuses on recognizing text strings, document parsing requires the model to recover both local content and global document structure.

Each training or evaluation instance is represented as $(\mathbf{I}, \mathbf{P}, \mathbf{Y})$, where $\mathbf{I}$ is the input document image, $\mathbf{P}$ is a task prompt specifying the language, task type, output format, and optional region coordinates, and $\mathbf{Y}$ is the target structured output. HoloDP considers three evaluation settings under this formulation. **Page-level layout analysis** predicts a document-level structural skeleton, including block categories, normalized bounding boxes, and reading order. **Region-level parsing** converts localized regions, such as paragraphs, tables, and formulas, into task-specific textual or markup-like outputs. **Mixed-script structural composition** evaluates whether a model can preserve document structure when multiple languages or scripts appear within the same layout. This unified formulation allows us to evaluate text fidelity, structural consistency, and cross-script robustness under a shared document parsing framework.

3.2 Training Corpus

Based on the task formulation, we construct a training corpus with both page-level and region-level supervision. The final corpus contains 1.35M page-level samples and 2.55M region-level samples, totaling 3.9M training instances. The corpus is constructed from three complementary sources.

High-resource public data. We collect Chinese and English document data from existing public resources (Feng et al., 2025; Liu et al., 2024b; Wei et al., 2024; Schuhmann et al., 2022; Gu et al., 2022; Hu et al., 2024). These samples serve as high-resource anchors that provide general layout priors and stabilize multilingual training, rather than replacing supervision for underrepresented scripts. We use PaddleOCR[2] to obtain pseudo labels and retain samples that pass OCR confidence and structural consistency checks. Samples with unclear licenses or terms are excluded. This source

[2] https://github.com/PaddlePaddle/PaddleOCR

contributes 2M Chinese/English instances.

Multilingual web-rendered data. To expand coverage to underrepresented languages, we collect public web pages from Wikipedia and official portals in 11 target languages. We render HTML and CSS pages into document images using `WeasyPrint` and `wkhtmltopdf`, and align text and structural labels from DOM trees. Controlled perturbations, including font substitution, spacing jitter, and mild background noise, are applied to increase layout diversity. This source contributes 0.7M multilingual instances.

Schema-driven synthetic data. We further construct synthetic documents from predefined schemas covering common formats such as news, forms, and transcripts. Each schema is populated with multilingual text fields and rendered using LaTeX, HTML/CSS, or SVG engines with script-specific fonts and writing directions. This source provides controllable supervision for scripts with limited naturally annotated data and contributes 1.2M image-text instances.

3.3 Benchmark Suite

We build a benchmark suite to evaluate document parsing under script-diverse settings.

Public benchmarks. We include OmniDocBench (981 pages; Chinese/English/code-mixed) (Ouyang et al., 2025) and FoxPage (212 bilingual pages) (Liu et al., 2024a) for document-level evaluation, and CPE (Wang et al., 2024a), PubTabNet (Zhong et al., 2020), and PubTab1M (Smock et al., 2022) for formula and table parsing.

Script-diverse page-level benchmarks. To evaluate layout structure, text content, and reading-order robustness under diverse scripts and writing directions, we construct two benchmarks: (1) OmniDocBench-HoloDP extends OmniDocBench with web-derived document pages in 11 underrepresented languages, totaling 11,000 samples. (2) Fox-Page-HoloDP extends Fox-Page from plain-text web pages in 11 underrepresented languages, totaling 11000 samples.

Mixed-script structural benchmark. We construct Mstruct-HoloDP, a synthetic benchmark where documents contain mixed-language content (e.g., Tibetan + Chinese) rendered with shared layout templates. Based on schema-driven reverse synthesis, this dataset with 11000 samples evaluates the model's ability to disentangle structure and language across scripts.

Region-level benchmark. To evaluate structure-aware region parsing, we construct RegL-HoloDP with 55000 regions by extracting and annotating paragraphs, formulas, and tables from real and synthetic documents in all 13 target languages.

3.4 Quality Control and Data Isolation

We apply the quality-control pipeline to the training corpus and the benchmark suite, covering automatic filtering, manual auditing, and train-test isolation. For all rendered and synthetic samples, automatic filters verify Unicode validity, font coverage, text overflow, bounding-box validity, reading-order consistency, and schema compliance. For web-rendered data, we additionally check DOM-to-image alignment and discard pages with broken rendering, missing fonts, element overlap, or incomplete structural labels. For pseudo-labeled public data, we retain only samples that pass OCR confidence and structural consistency thresholds.

We further conduct manual quality auditing. A team of 12 native or fluent annotators reviews a stratified 5% sample covering all languages, data sources, and task types. Each sampled item is independently checked by two annotators for text correctness, layout alignment, reading order, and output-format validity. Disagreements are resolved through discussion with a senior reviewer, and the final inter-annotator agreement reaches $\kappa > 0.85$. Detailed annotation and quality-control procedures are provided in Appendix A.

To reduce data leakage, all newly constructed benchmarks are built from disjoint source pools and independent rendering seeds. For web-rendered data, URLs and domains appearing in the training corpus are excluded from evaluation sets. For synthetic data, benchmark templates use separate layouts, fonts, field values, and rendering configurations from those used for training. We also apply document-level and schema-level deduplication to check for near-duplicate layouts or repeated content. In addition, we remove personally identifiable information and copyrighted media when constructing redistributable resources.

4 Method

4.1 Overview

We propose HoloDP, a script-aware document parsing framework that bridges layout-preserving visual perception and multilingual structured generation. As shown in Figure 2, HoloDP consists

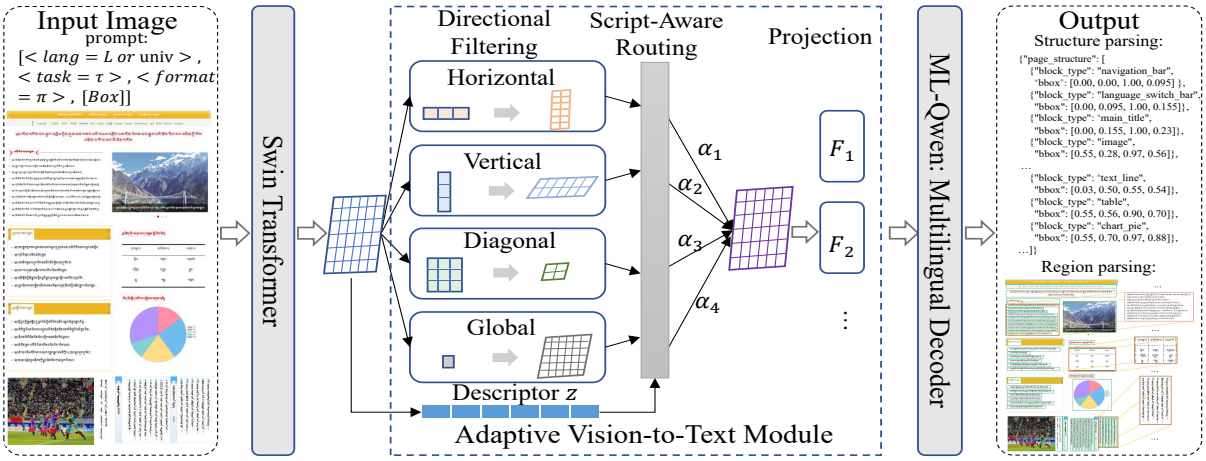


Figure 2: Overall architecture of the HoloDP framework.

of a Swin Transformer visual encoder, an Adaptive Vision-to-Text Module, and a multilingual decoder initialized from Qwen2.5-7B-Instruct. Given an image $\mathbf{I}$, the encoder extracts fixed-grid visual features $\mathbf{F}_{\text{vis}} \in \mathbb{R}^{H\times W\times D}$ and a page-level descriptor $\mathbf{z}$ obtained by global average pooling, linear projection, and layer normalization. Unlike direction-agnostic vision-to-language projectors, HoloDP performs script-aware directional routing: visual tokens are processed by multiple directional branches, fused using page/script context, projected into the decoder space, and decoded into task-specific structured outputs. Further formulations of visual-token construction, routing modes, prompt serialization, and macro-to-micro decoding are provided in Appendix B.

4.2 Adaptive Vision-to-Text Module

Current document parsers often compress visual features through a single projector, which may oversmooth direction-specific spatial cues needed for vertical, right-to-left, or mixed-script documents. To address this issue, we introduce the Adaptive Vision-to-Text Module, a lightweight bridge between the visual encoder and the decoder.

4.2.1 Directional Filtering

We first decompose the visual feature map into four directional representations:

$$\mathbf{F}_b = f_b(\mathbf{F}_{\text{vis}}), \quad b \in \{\text{horiz}, \text{vert}, \text{diag}, \text{glob}\}. \tag{1}$$

The horizontal branch captures row-wise text continuity, the vertical branch captures column-wise dependencies, the diagonal branch captures slanted strokes and local structural boundaries, and the global branch preserves page-level context. In practice, the horizontal and vertical branches are implemented with depthwise $1\times k$ and $k\times 1$ convolutions, the diagonal branch uses a dilated 3×3 convolution, and the global branch applies global pooling followed by broadcasting. This decomposition provides explicit spatial priors without substantially increasing the number of parameters.

4.2.2 Script-Aware Routing

The importance of each spatial prior varies across scripts and regions. For example, horizontal features are more useful for Latin or CJK text, while vertical features are important for vertically typeset Mongolian. We therefore introduce a dual-conditioned routing mechanism. Given the page descriptor $\mathbf{z}$ and a language embedding $e_{\mathcal{L}}$, the routing weights are computed as:

$$\boldsymbol{\alpha}(\mathcal{L}, \mathbf{I}) = \text{Softmax}\big(W_a[e_{\mathcal{L}}; W_z\mathbf{z}] + b_a\big). \tag{2}$$

For page-level layout analysis, $e_{\mathcal{L}}$ is replaced by a universal embedding e_{univ}; for region-level parsing, it is conditioned on the target language or script. In benchmark experiments, $\mathcal{L}$ is provided by dataset annotations; in deployment, it can be estimated by a lightweight language/script identification module. The routed feature is:

$$\mathbf{F}_{\text{route}} = \sum_b \alpha_b(\mathcal{L}, \mathbf{I})\mathbf{F}_b. \tag{3}$$

4.2.3 Script-Conditional Projection

After routing, we project visual tokens into the decoder space using a script-conditioned projection:

$$\hat{\mathbf{F}} = \phi_{S(\mathcal{L})}(\mathbf{F}_{\text{route}}), \tag{4}$$

where $S(\mathcal{L})$ denotes either a universal projection for layout-level parsing or a script-family projection for fine-grained content parsing. In our implementation, $S(\mathcal{L})$ follows the script families defined in Table 1. Each $\phi_{S(\mathcal{L})}$ is implemented as a two-layer MLP. This design allows related scripts to share parameters while keeping separate projection paths for scripts with distinct visual structures.

4.3 Prompted Macro-to-Micro Parsing

HoloDP uses a unified prompt interface for different parsing granularities:

$$\texttt{Prompt} = [\texttt{lang} = \mathcal{L},\ \texttt{task} = \tau,\ \texttt{format} = \pi,\ \texttt{Box}]. \tag{5}$$

Here, τ specifies the task, such as `page_structure`, `line`, `para`, `table`, or `formula`; π specifies the output format; and `Box` is optionally used for region-level parsing. The routed visual features are augmented with 2D positional encodings and used as visual prefixes for the decoder. The page-level output is a JSON-style list of blocks, where each block contains a category, normalized bounding box, and reading-order index. Region-level outputs are serialized as plain text, Markdown, HTML tables, or LaTeX formulas depending on the task and output format.

At inference, HoloDP follows a macro-to-micro protocol. It first predicts a page-level structural skeleton, including block categories, bounding boxes, and reading order. It then parses each detected region using a language- or script-specific prompt. During training, page-level and region-level instances are mixed according to the task prompt, allowing the model to learn both macro-structural detection and localized content parsing. This protocol decouples global layout analysis from local content serialization, making the parsing process more robust for mixed-script documents.

4.4 Progressive Optimization

Since the directional module is randomly initialized while the decoder already contains multilingual linguistic priors, direct end-to-end training may lead the decoder to rely on language priors before the visual bridge is well aligned. To stabilize learning, we adopt a three-stage optimization strategy.

Stage 1: Language adaptation. We continue training the decoder on monolingual and code-switched corpora from underrepresented languages using LoRA, improving multilingual fluency and structural-token generation.

Stage 2: Visual alignment. We freeze the decoder and train the visual encoder together with the Adaptive Vision-to-Text Module. This stage encourages the directional branches and script-conditioned projections to align visual evidence with the decoder embedding space.

Stage 3: Joint tuning. We jointly fine-tune all modules with both multimodal parsing batches and text-only batches:

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{align}} + \lambda \mathcal{L}_{\text{LM}}. \tag{6}$$

Here, $\mathcal{L}_{\text{align}}$ is the autoregressive loss on multimodal parsing targets, and $\mathcal{L}_{\text{LM}}$ is the causal language modeling loss on text-only batches. This final stage balances structural grounding with multilingual generation.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate HoloDP on the benchmarks described in Section 3.3.

Baselines. We compare HoloDP with three groups of representative baselines: (1) pipeline tools, including MinerU-1.2B (Wang et al., 2024b), Dolphin-322M (Feng et al., 2025), Marker[3], and PP-StructureV3 (Cui et al., 2025); (2) expert document VLMs, including GOT-OCR-580M (Wei et al., 2024), Nougat-250M (Blecher et al., 2024), MonkeyOCR-pro-1.2B (Li et al., 2025b), Nanonets-OCR-s (Mandal et al., 2025), DeepSeek-OCR (Wei et al., 2025), and dots.ocr (Li et al., 2025a); and (3) general-purpose VLMs, including GPT-4o (OpenAI, 2024), Qwen2.5-VL-7B (Bai et al., 2025), GLM4v-plus-9B (Zeng et al., 2024), InternVL-3-8B (Zhu et al., 2025), and MiniCPM-o2.6-8B (Yao et al., 2024). All baselines are evaluated with their official checkpoints and recommended inference settings.

Metrics. For text recognition and region-level parsing, we report edit distance (ED, normalized version; lower is better) (Li et al., 2025b), Precision, Recall, F1, BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005). For structured outputs such as HTML, JSON, and Markdown, we use Tree Edit Distance Similarity (TEDS) (Zhong et al., 2020) to measure both content and structural consistency.

Implementation. HoloDP is implemented in PyTorch and trained on 16 A100 GPUs with mixed

[3] https://github.com/datalab-to/marker

Models	Omni	OmniDocBench-HoloDP											Avg.
	en&zh	yue	bo	mn	ug	kk	za	ko	ky	vi	my	ms	
Pipeline Tools													
MinerU-1.2B	0.2034	0.7134	0.6336	0.6526	0.7257	0.5788	0.6488	0.6485	0.5899	0.7270	0.6686	0.6714	0.6598
Dolphin-322M	0.3980	0.7164	0.6359	0.6549	0.7283	0.5812	0.6511	0.6512	0.5919	0.7299	0.6714	0.6743	0.6624
Marker	0.3965	0.2396	0.4615	0.9724	0.9627	0.9014	0.1346	0.2546	0.3981	0.1206	0.9362	0.2878	0.5154
PP-StructureV3	0.1875	0.1616	0.3869	0.3603	0.3213	0.3626	0.2637	0.2227	0.3409	0.1891	0.2649	0.2344	0.2826
Expert VLMs													
GOT-580M	0.3490	0.8967	0.9136	0.8763	0.8675	0.9078	0.1727	0.9462	0.9562	0.9255	0.9827	0.1019	0.7770
Nougat-250M	0.7125	0.9850	0.9418	0.9519	0.9914	0.9123	0.9498	0.9501	0.9183	0.9925	0.9606	0.9624	0.9560
Nanonets-OCR-s	0.4529	0.6849	0.6082	0.6261	0.6966	0.5556	0.6226	0.6224	0.5661	0.6980	0.6417	0.6443	0.6333
MonkeyOCR-pro-1.2B	0.2085	0.9724	0.9877	0.9923	0.9624	0.9758	0.9876	0.9686	0.9777	0.9602	0.9586	0.9528	0.9724
dots.ocr	0.1425	0.1265	0.3546	0.3214	0.2941	0.3309	0.2552	0.2139	0.3050	0.1620	0.2427	0.2068	0.2557
DeepSeek-OCR	**0.1400**	0.1736	0.3478	0.3477	0.3265	0.3428	0.1438	0.1146	0.3833	0.2880	0.4826	0.1591	0.2827
General VLMs													
GPT-4o	0.3160	0.9812	0.8712	0.8971	0.9978	0.7959	0.8919	0.8919	0.8109	0.9779	0.9194	0.9233	0.9053
Qwen2.5-VL-7B	0.2375	0.1378	0.8791	0.9392	0.9277	0.9019	0.3979	0.2067	0.2401	0.3879	0.9092	0.2397	0.5607
InternVL-3-8B	0.2570	0.9244	0.9765	0.9688	0.9052	0.9826	0.4199	0.2264	0.9912	0.9362	0.9471	0.1734	0.7683
GLM4v-plus-9B	0.3338	0.9990	0.8949	0.9496	0.9987	0.8538	0.9448	0.9445	0.8681	0.9867	0.9706	0.9742	0.9441
MiniCPM-o2.6-8B	0.4372	0.9236	0.9623	0.9241	0.9018	0.9169	0.1714	0.9288	0.9719	0.8925	0.8976	0.1882	0.7890
HoloDP	0.2248	**0.1149**	**0.3308**	**0.2731**	**0.2711**	**0.2746**	**0.1278**	**0.1082**	**0.2191**	**0.1129**	**0.2149**	**0.1017**	**0.1954**

Table 2: Evaluation Results (ED, lower is better) on OmniDocBench and OmniDocBench-HoloDP. Avg. denotes the mean ED over the 11 newly introduced OmniDocBench-HoloDP language subsets. Best: **bold**, second: underline.

precision. Images are resized to 896×896, and visual features are aligned to a fixed spatial grid before entering the adaptive vision-to-text module. We use Adam (Kingma and Ba, 2015) with cosine learning-rate decay and train HoloDP with the three-stage strategy described in Section 4. Unless otherwise specified, we report the average results over three independent runs with different random seeds. Additional training, decoding, and evaluation settings are provided in Appendix C.

5.2 Main Results and Analysis

Mixed-content document parsing. Table 2 reports ED results on OmniDocBench and OmniDocBench-HoloDP. On high-resource Chinese/English documents, specialized OCR systems remain highly competitive. DeepSeek-OCR achieves the lowest ED on OmniDocBench, while dots.ocr and PP-StructureV3 also outperform HoloDP in this setting. This suggests that mature OCR-oriented systems still retain advantages on regular layouts and high-resource scripts.

The advantage of HoloDP becomes clearer in script-diverse and underrepresented settings. On the 11 newly introduced OmniDocBench-HoloDP language subsets, HoloDP achieves the lowest mean ED of 0.1954. It also obtains the lowest ED across the evaluated underrepresented languages, including Yue Chinese, Tibetan, Mongolian, Uyghur, Kazakh, Zhuang, Korean, Kyrgyz, Vietnamese, Burmese, and Malay. These results indicate that direction-aware visual routing and script-conditioned projection improve robustness under heterogeneous writing directions and script-specific layout patterns.

Models	ED↓	F1↑	Precision↑	Recall↑	BLEU↑	METEOR↑
Pipeline Tools						
MinerU-1.2B	0.4321	0.5262	0.6020	0.4981	0.3251	0.3886
Dolphin-322M	0.6121	0.5721	0.5781	0.6518	0.4312	0.5419
Marker	0.5639	0.5135	0.4220	0.4458	0.3120	0.3599
PP-StructureV3	0.4873	0.6247	0.7047	0.5686	0.3438	0.4254
Expert VLMs						
GOT-580M	0.7430	0.4274	0.4244	0.4409	0.3589	0.4123
Nougat-250M	0.7510	0.4123	0.4194	0.4348	0.3504	0.3315
Nanonets-OCR-s	0.7308	0.3072	0.3307	0.2961	0.3280	0.2530
MonkeyOCR-pro-1.2B	0.6381	0.4772	0.4316	0.3571	0.3331	0.3786
dots.ocr	0.4027	0.6881	0.7802	0.6514	0.4251	0.5082
DeepSeek-OCR	0.4102	0.4746	0.4997	0.4653	0.4079	0.4289
General VLMs						
GPT-4o	0.4299	0.6811	0.7792	0.6447	0.4208	0.5030
Qwen2.5-VL-7B	0.6250	0.4110	0.5106	0.5215	0.5142	0.5877
InternVL-3-8B	0.4525	0.6289	0.5132	0.5862	0.5872	0.6684
GLM4v-plus-9B	0.4423	0.5839	0.6392	0.5635	0.5172	0.5836
MiniCPM-o2.6-8B	0.4343	0.6381	0.7300	0.6040	0.3942	0.4712
HoloDP	**0.3422**	**0.6897**	**0.7852**	**0.6684**	**0.6210**	**0.6912**

Table 3: Evaluation Results on Multilingual Structural Composition Parsing. Best: **bold**, second: underline.

Mixed-script structural composition parsing. Table 3 reports results on Mstruct-HoloDP, which evaluates document parsing under mixed-script structural composition. HoloDP achieves the best performance across all reported metrics, with clear gains in ED, BLEU, and METEOR, and slight but consistent improvements in F1 and Precision over the strongest baselines. This indicates that HoloDP not only improves text-level fidelity, but also better preserves layout-level structural consistency in multilingual layouts.

Among the baselines, dots.ocr and GPT-4o achieve competitive F1 and Precision, suggesting that strong OCR-oriented or general-purpose VLMs can recover part of the local content and

block structure. However, their higher ED and lower generation metrics indicate remaining difficulties in preserving complete structured outputs when multiple scripts and layout patterns coexist. In contrast, HoloDP benefits from direction-aware visual routing and script-conditioned projection, which help align visual regions with serialized outputs under heterogeneous writing directions. These results support the importance of explicitly modeling script-aware spatial priors for multilingual structural composition parsing.

Module	Models	ED↓	F1↑	Precision↑	Recall↑	BLEU↑	METEOR↑
Decoder	w/o CLM	0.5378	0.4029	0.5022	0.4389	0.3965	0.5229
V2T	Cross Attention	0.4079	0.6290	0.7367	0.6384	0.5927	0.6722
	H-Reducer	0.3788	0.6633	0.7742	0.6397	0.6074	0.6833
	FC Layer	0.3541	0.6679	0.7735	0.6438	0.6189	0.6799
Joint FT	w/o Joint FT	0.4183	0.6023	0.7156	0.5927	0.4785	0.6137
Routing	Content-only	0.3752	0.6456	0.7411	0.6109	0.5876	0.6733
	Language-only	0.3951	0.6174	0.7344	0.5877	0.5886	0.6659
Direc-tional	1-branch	0.3976	0.6188	0.7249	0.5773	0.5878	0.6602
	2-branch	0.3619	0.6548	0.7603	0.6309	0.6028	0.6794
	3-branch	0.3518	0.6798	0.7798	0.6629	0.6138	0.6879
Proje-ction	Shared	0.3812	0.6332	0.7376	0.5902	0.5937	0.6708
	Per-language	0.3597	0.6789	0.7597	0.6248	0.6034	0.6863
Ours	HoloDP	**0.3422**	**0.6897**	**0.7852**	**0.6684**	**0.6210**	**0.6912**

Table 4: Ablation Study on Mixed-script Structural Composition Parsing.

5.3 Ablation Study

To quantify the individual contribution of each component, we conduct controlled ablations on Mstruct-HoloDP, as shown in Table 4. Overall, every removal or simplification degrades performance, confirming that HoloDP's improvements come jointly from the combination of progressive multilingual optimization, adaptive visual-textual alignment, direction-aware spatial filtering, script-aware routing, and script-conditioned projection.

Effect of progressive optimization. We first examine the training strategy. Removing continued language modeling causes the largest performance drop, with ED increasing from 0.3422 to 0.5378 and F1 decreasing from 0.6897 to 0.4029. This indicates that adapting the decoder to underrepresented language distributions is important for generating script-specific tokens and structured outputs. Removing joint fine-tuning also leads to clear degradation, increasing ED to 0.4183 and reducing F1 to 0.6023. These results show that language adaptation alone is insufficient; the model also needs final joint optimization to coordinate visual grounding and multilingual generation.

Effect of adaptive vision-to-text alignment. We then replace the proposed Adaptive Vision-to-Text Module with generic alignment alternatives. Cross-attention, H-Reducer (Hu et al., 2025), and a simple FC projector all underperform the full model, with ED increasing to 0.4079, 0.3788, and 0.3541, respectively. Although the FC projector is competitive among the simplified variants, it still lags behind HoloDP across all metrics. This suggests that a single direction-agnostic mapping is insufficient for multilingual structural composition, where visual tokens must be aligned with serialized outputs under heterogeneous writing directions and layout patterns.

Effect of script-aware routing. We further ablate the dual-conditioned routing mechanism. Content-only gating and language-only gating both degrade performance, increasing ED to 0.3752 and 0.3951, respectively. The content-only variant can still adapt to local visual layout, but lacks explicit script priors; the language-only variant preserves script-level information, but cannot sufficiently respond to local layout variability. The full routing mechanism performs best because it combines page-level visual context with language or script information, allowing the model to select suitable spatial branches for different document regions.

Effect of directional spatial filtering. The branch-count ablation evaluates whether multiple directional priors are necessary. The 1-branch variant performs poorly, while adding more branches steadily improves both ED and F1. The 2-branch variant benefits from horizontal and vertical dependencies, and the 3-branch variant further improves performance by introducing diagonal spatial cues. The full four-branch design achieves the best results, indicating that horizontal, vertical, diagonal, and global spatial priors provide complementary signals for parsing mixed-script and structurally complex documents.

Effect of script-conditioned projection. Finally, we examine the projection granularity. A fully shared projection increases ED to 0.3812 and reduces F1 to 0.6332, suggesting that a single projection space cannot adequately represent diverse writing systems. A per-language projection performs better than the shared projection but remains below HoloDP, likely because it reduces parameter sharing and is less data-efficient for underrepresented languages. The script-conditioned projection achieves the best trade-off by sharing parameters within related script families while preserving script-specific visual patterns.

6 Conclusion

We presented HoloDP, a multilingual benchmark suite and script-aware framework for document parsing across diverse scripts and writing directions. On the resource side, HoloDP provides a large-scale multilingual corpus and standardized benchmarks covering 13 languages and seven script families, supporting page-level layout analysis, region-level parsing, and mixed-script structural composition. On the modeling side, HoloDP introduces an adaptive vision-to-text module that combines direction-aware spatial routing with script-conditioned projection, together with a progressive training strategy for multilingual visual-textual alignment. Experiments on public and newly constructed benchmarks show that HoloDP improves robustness in underrepresented and mixed-script scenarios while remaining competitive on high-resource Chinese and English settings. These results suggest that combining multilingual evaluation resources with script-aware modeling is a promising direction for more reliable document parsing under script diversity.

Limitations

HoloDP has several limitations.

First, although our benchmark covers 13 languages across seven script families, it still represents only a subset of multilingual writing systems. Many scripts, domains, and document traditions remain uncovered, so the reported results should be interpreted within the evaluated script-diverse settings rather than as evidence of universal multilingual document parsing.

Second, part of the newly constructed data comes from web-rendered snapshots and schema-driven synthetic documents. These sources provide controllable structural annotations and broad multilingual coverage, but may not fully match the noise distribution of real scanned documents, handwritten materials, camera-captured pages, or degraded archival documents. HoloDP may therefore remain less robust under extreme blur, compression artifacts, low resolution, handwriting, or irregular physical layouts.

Third, HoloDP uses script or language information during fine-grained region parsing. In our benchmark setting, this information is provided by annotations to ensure controlled evaluation; in deployment, it would need to be estimated by a language or script identification module. Errors in this step may affect routing and parsing quality, especially for mixed-script or visually similar scripts.

Finally, despite automatic filtering and manual auditing, the corpus may still inherit biases from source selection, rendering templates, OCR-based pseudo labels, and available fonts. These biases may affect performance on unseen document styles or communities with limited digital resources. Future work should expand coverage to more scripts and domains, evaluate stronger degradation scenarios, and explore lighter decoders and more script-agnostic representations for efficient deployment.

Acknowledgment

This work was supported by the Youth Research Special Project of NCUT (110051360025XN077-13), Provincial-Level Research Platform Operations Support-Beijing Key Laboratory of Key Technologies for AI+ Domain Applications, and the Beijing Natural Science Foundation Project(4252035).

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural optical understanding for academic documents. In *The Twelfth International Conference on Learning Representations*.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, Jingqun Tang, Hao Liu, and Can Huang. 2025. Dolphin: Document image parsing via heterogeneous anchor prompting. In *Findings of the Association for Computational Linguistics*, pages 21919–21936.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing Xu, and Hang Xu. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ahmed Heakl, Youssef Mohamed, Abdullah Sohail, Rania Elbadry, Ahmed Nassar, Peter WJ Staar, Fahad Shahbaz Khan, Imran Razzak, and Salman Khan. 2026. Docatlas: Multilingual document understanding across 80+ languages. *arXiv preprint arXiv:2605.12623*.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 3096–3120. Association for Computational Linguistics.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2025. mPLUG-DocOwl2: High-resolution compressing for OCR-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5817–5834.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Yumeng Li, Guang Yang, Hao Liu, Bowen Wang, and Colin Zhang. 2025a. dots. ocr: Multilingual document layout parsing in a single vision-language model. *arXiv preprint arXiv:2512.02498*.

Zhang Li, Yuliang Liu, Qiang Liu, Zhiyin Ma, Ziyang Zhang, Shuo Zhang, Zidun Guo, Jiarui Zhang, Xinyu Wang, and Xiang Bai. 2025b. Monkeyocr: Document parsing with a structure-recognition-relation triplet paradigm. *arXiv preprint arXiv:2506.05218*.

Chenglong Liu, Haoran Wei, Jinyue Chen, Lingyu Kong, Zheng Ge, Zining Zhu, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Focus anywhere for fine-grained multi-page document understanding. *arXiv preprint arXiv:2405.14295*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024b. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Wenkang Ma, Mingzhe Cao, Jinyue Ma, Zhenyang Dong, Chaozhi Yang, and Zongmin Li. 2026. Mamba-yolo: Multi-level adaptive rectangular convolution for document layout analysis. *Pattern Recognition*, 170:112031.

Souvik Mandal, Ashish Talewar, Paras Ahuja, and Prathamesh Juvatkar. 2025. Nanonets-ocr-s: A model for transforming documents into structured markdown with intelligent content recognition and semantic tagging.

OpenAI. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. 2025. Omnidocbench: Benchmarking diverse PDF document parsing with comprehensive annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24838–24848. Computer Vision Foundation / IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Christina Sabbagh. 2025. Enhanced htr accuracy for tibetan historical texts-optimising image preprocessing for improved transcription quality. *Revue d'Etudes Tibétaines*, 74:82–128.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642.

Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. 2024a. Unimernet: A universal network for real-world mathematical expression recognition. *arXiv preprint arXiv:2404.15254*.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and

Conghui He. 2024b. Mineru: An open-source solution for precise document content extraction. *CoRR*, abs/2409.18839.

Ahmed Wasfy, Omer Nacar, Abdelakreem Elkhateb, Mahmoud Reda, Omar Elshehy, Adel Ammar, and Wadii Boulila. 2025. Qari-ocr: High-fidelity arabic text recognition through multimodal large language model adaptation. *arXiv preprint arXiv:2506.02295*.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, Chunrui Han, and Xiangyu Zhang. 2024. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. *CoRR*, abs/2409.01704.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A GPT-4V level MLLM on your phone. *CoRR*, abs/2408.01800.

Pengfei Yue, Xingran Zhao, Juntao Chen, Peng Hou, Wang Longchao, Jianghang Lin, Shengchuan Zhang, Anxiang Zeng, and Liujuan Cao. 2026. Sea-vision: A multilingual benchmark for comprehensive document and scene text understanding in southeast asia. *arXiv preprint arXiv:2603.15409*.

Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, and 36 others. 2024. Chatglm: A family of large language models from GLM-130B to GLM-4 all tools. *CoRR*, abs/2406.12793.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479.

Check	Filtered cases
Unicode validity	Invalid encodings, broken characters
Font coverage	Missing glyphs, tofu boxes, unsupported scripts
Layout integrity	Text overflow, block overlap, empty regions
Bounding-box validity	Out-of-range boxes, zero-area regions, duplicated boxes
Reading-order consistency	Missing or inconsistent order annotations
Schema compliance	Invalid JSON/HTML, missing required fields
DOM-image alignment	Mismatched rendered regions and DOM nodes
OCR confidence	Low-confidence pseudo labels, unstable line grouping

Table 5: Automatic quality checks used in HoloDP data construction.

Item	Description
Annotators	12 native or fluent speakers
Sampling strategy	Stratified by language, source, and task type
Review mode	Two independent annotators per sample
Text check	Transcription correctness and Unicode validity
Layout check	Bounding-box coverage and region boundaries
Order check	Reading order under LTR, RTL, vertical, or mixed layouts
Format check	JSON/HTML/Markdown/text/schema validity
Resolution	Discussion and senior-review adjudication
Agreement	Inter-annotator agreement $\kappa > 0.85$

Table 6: Manual audit protocol for HoloDP quality control.

A Data Quality and Annotation Details

This appendix supplements Section 3.4 with operational details of the quality-control, annotation, and data-isolation procedures used in HoloDP.

A.1 Automatic Filtering Criteria

Table 5 summarizes the automatic checks applied during data construction. These checks are used for rendered documents, schema-driven synthetic samples, web-rendered pages, and pseudo-labeled public data, with source-specific filters applied when applicable.

For web-rendered pages, DOM nodes are matched with rendered image regions to detect broken rendering, missing fonts, collapsed layout elements, and incomplete DOM-to-image alignment. For pseudo-labeled public data, OCR confidence and structural consistency are used to reduce noisy supervision from automatically generated labels.

A.2 Manual Audit Protocol

We conduct manual auditing on a stratified subset sampled by language, data source, and task type. The audit covers page-level layout analysis, region-level parsing, and mixed-script structural composition. A total of 12 native or fluent annotators participate in the review process, collectively covering the 13 target languages.

Each sampled item is independently reviewed by two annotators. They verify transcription correctness, bounding-box coverage, reading-order consistency, and output-schema validity. Disagreements are resolved through discussion, and difficult cases are adjudicated by a senior reviewer. The final inter-annotator agreement reaches $\kappa > 0.85$.

A.3 Annotator Recruitment and Payment

Annotators were recruited based on their native or fluent proficiency in the target languages and their familiarity with document reading conventions, including left-to-right, right-to-left, vertical, and mixed-layout documents. Before annotation, all annotators received task guidelines and example cases covering transcription, bounding-box verification, reading-order checking, and schema validation. Participation was voluntary, and annotators were informed of the research purpose and annotation requirements before starting the task.

Annotators were compensated for their work according to the amount and difficulty of completed annotations, with payment rates agreed upon before annotation began. Only qualified annotations that passed consistency checks were included in the final audited set, while rejected or uncertain cases were returned for correction or senior review.

A.4 Data Isolation and Deduplication

We enforce source-level, document-level, and template-level isolation between the training corpus and benchmark suite. For web-rendered data, benchmark pages are collected from source pools disjoint from the training corpus; URLs and domains appearing in training are excluded from evaluation. For schema-driven synthetic data, benchmark templates are created separately from training templates and use different layout structures, font configurations, field values, and rendering seeds.

We also remove repeated or near-duplicate pages using source metadata, normalized text content, and rendered layout signatures. At the schema level, we check whether benchmark samples share the same template, field arrangement, or rendering configuration with training samples. Potential overlaps are manually inspected and removed before finalizing the benchmark split.

A.5 Licensing and Privacy

We retain only public resources with permissive research licenses or clearly stated usage conditions. Samples with unclear licenses, restricted redistribution terms, personally identifiable information, proprietary logos, copyrighted media, or sensitive visual content are excluded from redistributable data. For synthetic data, controlled templates and

curated multilingual text fields are used to avoid introducing private or restricted content.

B Supplementary Method Details

This appendix supplements Section 4 with additional methodological details that are omitted from the main text for brevity. We focus on the construction of fixed-grid visual tokens, branch-level fusion, script-conditioned routing, visual-prefix serialization, and the unified training objective.

B.1 Fixed-Grid Visual Token Construction

The visual encoder produces a spatial feature map from the input document image. To make the subsequent directional routing independent of the original image size, we convert the encoder output into a fixed-grid representation:

$$\mathbf{F}_{\text{vis}} = \text{Resize}\big(\psi(\mathbf{F}_{\text{enc}})\big) \in \mathbb{R}^{H\times W\times D}, \quad (7)$$

where $\mathbf{F}_{\text{enc}}$ denotes the final visual feature map produced by the Swin encoder, $\psi(\cdot)$ is a channel projection, and $\text{Resize}(\cdot)$ aligns the feature map to a fixed spatial grid. The fixed grid provides a shared coordinate system for directional filtering, routing, and 2D positional encoding.

The page-level descriptor used by the routing module is computed from the same fixed-grid representation:

$$\mathbf{z} = \text{LN}\big(W_p\,\text{GAP}(\mathbf{F}_{\text{vis}})\big), \quad (8)$$

where $\text{GAP}(\cdot)$ denotes global average pooling over spatial locations, W_p is a learnable linear projection, and $\text{LN}(\cdot)$ denotes layer normalization. Thus, $\mathbf{z}$ summarizes the page-level visual context while keeping the routing module independent of variable image sizes.

B.2 Directional Branch Fusion

The four directional branches operate on the same input feature map $\mathbf{F}_{\text{vis}}$ and preserve its spatial resolution and channel dimension:

$$\mathbf{F}_b = f_b(\mathbf{F}_{\text{vis}}) \in \mathbb{R}^{H\times W\times D}. \quad (9)$$

Where $b \in \{\text{horiz}, \text{vert}, \text{diag}, \text{glob}\}$. This shape consistency allows all branch outputs to be fused by scalar routing weights.

For the global branch, the pooled page representation is projected and broadcast back to the full spatial grid:

$$\mathbf{F}_{\text{glob}} = \text{Broadcast}\big(g(\text{GAP}(\mathbf{F}_{\text{vis}}))\big) \in \mathbb{R}^{H\times W\times D}. \quad (10)$$

Here, $\text{Broadcast}(\cdot)$ repeats the projected page-level vector at all spatial locations. This allows the global branch to provide document-level context while remaining compatible with the local directional branches.

The routing weights are branch-level coefficients rather than token-wise attention maps:

$$\boldsymbol{\alpha}(\mathcal{L}, \mathbf{I}) = [\alpha_{\text{horiz}}, \alpha_{\text{vert}}, \alpha_{\text{diag}}, \alpha_{\text{glob}}]. \quad (11)$$

The routed feature is therefore:

$$\mathbf{F}_{\text{route}} = \sum_b \alpha_b(\mathcal{L}, \mathbf{I})\mathbf{F}_b. \quad (12)$$

This design keeps the adaptive module lightweight while allowing each page or region to emphasize different directional priors.

B.3 Routing Modes for Page-Level and Region-Level Parsing

HoloDP uses different routing conditions for macro-level and micro-level parsing. For page-level structural prediction, the language input is replaced by a universal embedding:

$$e_{\mathcal{L}} = e_{\text{univ}}. \quad (13)$$

This setting encourages the model to focus on script-agnostic layout cues such as block boundaries, table structures, and reading-order relations.

For region-level parsing, the routing module uses the language or script embedding associated with the target region:

$$e_{\mathcal{L}} = e_{\text{lang/script}}. \quad (14)$$

For mixed-script pages, HoloDP first performs page-level parsing with the universal embedding and then applies region-specific routing during localized content parsing. In benchmark evaluation, the region language or script label is taken from annotations. In deployment, it can be obtained from a lightweight language or script identification module.

B.4 Script-Conditional Projection Paths

After routing, visual features are mapped into the decoder embedding space through a projection path selected by $S(\mathcal{L})$:

$$\hat{\mathbf{F}} = \phi_{S(\mathcal{L})}(\mathbf{F}_{\text{route}}). \quad (15)$$

The function $S(\mathcal{L})$ maps each language to either the universal layout projection or a script-family

projection. For page-level layout analysis, $S(\mathcal{L}) = \mathrm{Univ}$. For region-level parsing, $S(\mathcal{L})$ follows the script family of the target language. This design avoids assigning a separate projection to every language and instead shares parameters among visually related scripts.

B.5 Visual Prefix Serialization

The projected visual features are augmented with 2D positional encodings:

$$\mathbf{V} = \mathrm{Flatten}\big(\hat{\mathbf{F}} + \mathbf{E}_{\mathrm{2D}}\big), \quad \mathbf{V} \in \mathbb{R}^{N \times d_m}, \tag{16}$$

where $N = H \times W$ and d_m is the decoder embedding dimension. The 2D positional encoding preserves row and column information after flattening.

The decoder input is formed by concatenating the visual token sequence with the textual prompt embedding:

$$\mathbf{X}_0 = [\mathbf{V}; \mathrm{Emb}(\texttt{Prompt})]. \tag{17}$$

The decoder then autoregressively generates the serialized target:

$$p(\mathbf{y} \mid \mathbf{I}, \texttt{Prompt}) = \prod_{t=1}^{T} p(y_t \mid y_{<t}, \mathbf{X}_0). \tag{18}$$

This formulation makes page-level layout prediction, region-level text recognition, table parsing, and formula parsing all compatible with the same decoder interface.

B.6 Prompt and Target Format

The prompt contains four fields:

$$\texttt{Prompt} = [\texttt{lang} = \mathcal{L}, \texttt{task} = \tau, \texttt{format} = \pi, \texttt{Box}]. \tag{19}$$

The Box field is optional and is used only for region-level parsing. When present, it is represented as normalized coordinates:

$$\texttt{Box} = (x_1, y_1, x_2, y_2), \quad x_1, y_1, x_2, y_2 \in [0, 1]. \tag{20}$$

For page-level parsing, the target sequence is a serialized list:

$$\mathcal{B} = \{(c_i, \mathbf{b}_i, o_i)\}_{i=1}^{M}, \tag{21}$$

where c_i denotes the block category, $\mathbf{b}_i$ denotes the normalized bounding box, and o_i denotes the reading-order index. For region-level parsing, the target sequence is serialized according to the requested output format, such as plain text, Markdown, HTML, or LaTeX.

B.7 Unified Training Formulation

HoloDP treats page-level and region-level tasks as conditional generation problems. For a multimodal training instance $(\mathbf{I}, \texttt{Prompt}, \mathbf{y})$, the visual alignment objective is:

$$\mathcal{L}_{\mathrm{align}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, \mathbf{I}, \texttt{Prompt}). \tag{22}$$

This objective is used for both page-level structural outputs and region-level content outputs. The task type is specified by the prompt rather than by separate task-specific heads.

For text-only language adaptation, the decoder is trained with the causal language modeling objective:

$$\mathcal{L}_{\mathrm{LM}} = -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}). \tag{23}$$

The final joint objective combines multimodal parsing and text-only language preservation:

$$\mathcal{L}_{\mathrm{joint}} = \mathcal{L}_{\mathrm{align}} + \lambda \mathcal{L}_{\mathrm{LM}}. \tag{24}$$

The first term encourages visual grounding and structural serialization, while the second term helps preserve multilingual generation ability during joint tuning.

B.8 Macro-to-Micro Decoding

At inference, HoloDP first generates a page-level structural skeleton:

$$\hat{\mathcal{B}} = \{(\hat{c}_i, \hat{\mathbf{b}}_i, \hat{o}_i)\}_{i=1}^{M}. \tag{25}$$

The predicted blocks are sorted by the reading-order index $\hat{o}_i$. Each block is then parsed with a region-level prompt:

$$\begin{aligned} \texttt{Prompt}_i = [&\texttt{lang} = \hat{\mathcal{L}}_i,\ \texttt{task} = \tau_i, \\ &\texttt{format} = \pi_i,\ \texttt{Box} = \hat{\mathbf{b}}_i]. \end{aligned} \tag{26}$$

This produces localized content predictions for lines, paragraphs, tables, or formulas. The final document output is obtained by composing the page-level skeleton and region-level serialized outputs according to the predicted reading order.

C Implementation Details

Table 7 summarizes the implementation and training settings used for HoloDP. All main experiments and ablation studies follow the same configuration unless otherwise specified; for each ablation, only

Item	Setting
Basic implementation	
Framework	PyTorch
Hardware	16×NVIDIA A100 GPUs
Precision	mixed precision
Image resolution	896×896
Image normalization	normalized to $[0, 1]$
Patch size	$P = 4$
Maximum decoding length	4096 tokens
Random seeds	3 independent runs
Visual encoder	
Backbone	hierarchical Swin Transformer
Number of stages	4
Layers per stage	$[2, 2, 14, 2]$
Attention heads per stage	$[4, 8, 16, 32]$
Window size	7
Optimization	
Optimizer	Adam (Kingma and Ba, 2015)
Learning-rate schedule	cosine decay
Warmup ratio	0.03
Weight decay	0.01
Gradient clipping	1.0
Stage 1: language adaptation	
Trainable modules	decoder LoRA parameters
Frozen modules	visual encoder and Adaptive Vision-to-Text Module
Learning rate	2×10^{-5}
Effective batch size	128
Training epochs	3
LoRA rank	8
LoRA α	16
Stage 2: visual alignment	
Trainable modules	visual encoder and Adaptive Vision-to-Text Module
Frozen modules	decoder
Learning rate	5×10^{-5}
Effective batch size	128
Training epochs	2
Stage 3: joint tuning	
Trainable modules	all modules
Learning rate	1×10^{-5}
Effective batch size	64
Training epochs	1

Table 7: Implementation and training settings of HoloDP.

the target component is modified while the remaining settings are kept unchanged. Each setting is run with three independent random seeds, and the main tables report the averaged results.

The number of training epochs is gradually reduced across the three stages as more modules are unfrozen. This schedule first adapts the decoder to multilingual structural generation, then aligns visual features with the frozen decoder space, and finally performs joint tuning with a smaller learning rate and batch size to stabilize multimodal generation.

D Supplementary Experimental Results and Analysis

D.1 Additional Main Results and Analysis

Pure-text document OCR. Table 8 reports edit distance results on Fox-Page and Fox-Page-HoloDP. HoloDP achieves the lowest overall average ED of 0.1138, substantially improving over the strongest baseline dots.ocr (0.2150) and DeepSeek-OCR (0.2241). This shows that HoloDP generalizes better across script-diverse OCR settings, rather than only performing well on high-resource scripts.

At the language level, the results reveal a clear contrast between high-resource and underrepresented settings. On English and Chinese, specialized OCR-oriented systems remain stronger: dots.ocr obtains the lowest ED on both languages, while DeepSeek-OCR and Nanonets-OCR-s are also highly competitive. HoloDP is competitive but not the best on these two high-resource languages, suggesting that mature OCR systems still benefit from dense supervision and broad coverage of common fonts and glyph distributions.

The advantage of HoloDP becomes more evident on Fox-Page-HoloDP. It achieves the best ED on all 11 underrepresented or script-diverse languages, including Yue Chinese, Tibetan, Mongolian, Uyghur, Kazakh, Zhuang, Korean, Kyrgyz, Vietnamese, Burmese, and Malay. In contrast, several pipeline tools, expert OCR/VLM systems, and general VLMs show large performance fluctuations across scripts, especially on Tibetan, Mongolian, Uyghur, Burmese, and other visually distinctive scripts. These results indicate that HoloDP better captures script-level visual regularities and is more robust to rare glyph forms, script-specific typography, and font variation.

Region-level evaluation. Table 9 reports region-level parsing results on three public benchmarks and RegL-HoloDP. HoloDP achieves the best results on most metrics, including ED and F1 on CPE and PubTabNet, F1 on PubTab1M, and both TEDS and F1 on RegL-HoloDP. The only exception is PubTab1M ED, where dots.ocr obtains the lowest error, while HoloDP ranks second. These results show that HoloDP is competitive on established region-level benchmarks and particularly effective on the more challenging RegL-HoloDP setting.

On RegL-HoloDP, HoloDP achieves a TEDS score of 0.9024 and an F1 score of 0.9493, clearly outperforming the strongest baseline dots.ocr by 0.0662 and 0.0792 points, respectively. This indicates that HoloDP better preserves fine-grained correspondence between visual regions and serialized textual structures under complex layouts and heterogeneous region types.

Compared with pipeline tools, expert OCR/VLM systems, and general VLMs, HoloDP more faithfully reconstructs multi-row tables and nested text regions, reducing errors such as mis-grouped cells, duplicated lines, and fragmented regions. The

Models	Fox-Page		Fox-Page-HoloDP											Avg.
	en	zh	yue	bo	mn	ug	kk	za	ko	ky	vi	my	ms	
Pipeline Tools														
MinerU-1.2B	0.2136	0.1082	0.0736	0.9922	0.9690	0.9883	0.2313	0.0190	0.9884	0.3076	0.2376	0.9572	0.1506	0.4797
Dolphin-322M	0.0479	0.1090	0.4627	0.9987	0.9678	0.9169	0.6524	0.0679	0.8601	0.5146	0.2230	0.9958	0.0648	0.5294
Marker	0.0558	0.1506	0.0510	0.6231	0.8406	0.8754	0.2367	0.0283	0.0540	0.0207	0.0258	0.9924	0.0187	0.3056
PP-StructureV3	0.0893	0.0793	0.2257	0.5267	0.4311	0.2486	0.4079	0.1025	0.1187	0.1029	0.1484	0.4868	0.1390	0.2389
Expert VLMs														
GOT-580M	0.0950	0.1370	0.1501	0.8791	0.9989	0.9810	0.8716	0.0155	0.9820	0.1138	0.1699	0.9998	0.0090	0.4925
Nougat-250M	0.0846	0.9951	0.9273	0.9296	0.9789	0.8994	0.9389	0.0681	0.9570	0.9742	0.3792	0.9963	0.0488	0.7060
Nanonets-OCR-s	0.0480	0.0787	0.0613	0.9982	0.9994	0.9089	0.7755	0.0281	0.0582	0.0985	0.0378	0.9321	0.0147	0.3876
MonkeyOCR-pro-1.2B	0.0922	0.1385	0.9889	0.8990	0.9249	0.9138	0.8328	0.3785	0.9276	0.8023	0.8824	0.9698	0.3267	0.6983
dots.ocr	**0.0310**	**0.0470**	0.2023	0.4970	0.4105	0.2444	0.4051	0.0763	0.0883	0.0899	0.1420	0.4492	0.1124	0.2150
DeepSeek-OCR	0.0433	0.0635	0.1456	0.4688	0.5564	0.5025	0.3949	0.0365	0.0291	0.0973	0.0731	0.4874	0.0150	0.2241
General VLMs														
GPT-4o	0.0585	0.3580	0.7516	0.9612	0.7415	0.8761	0.8971	0.8713	0.8891	0.8712	0.8917	0.9012	0.8854	0.7657
Qwen2.5-VL-7B	0.0610	0.0982	0.0311	0.8762	0.9879	0.9071	0.6540	0.0221	0.0468	0.0551	0.0198	0.8131	0.0090	0.3524
InternVL-3-8B	0.0934	0.1289	0.4778	0.9987	0.9981	0.8977	0.7607	0.1063	0.0573	0.2871	0.2361	0.9332	0.0439	0.4630
GLM4v-plus-9B	0.0787	0.1561	0.8999	0.9129	0.9945	0.9268	0.9105	0.0918	0.9010	0.6795	0.5972	0.9982	0.1017	0.6345
MiniCPM-o2.6-8B	0.1191	0.2711	0.8642	0.9912	0.9974	0.9390	0.7642	0.0772	0.7566	0.5681	0.5034	0.9964	0.0852	0.6102
HoloDP	0.0545	0.0971	**0.0137**	**0.2730**	**0.2814**	**0.2061**	**0.2119**	**0.0034**	**0.0220**	**0.0129**	**0.0147**	**0.2823**	**0.0067**	**0.1138**

Table 8: Evaluation Results (ED, lower is better) on Fox-Page and Fox-Page-HoloDP. Avg. denotes the mean ED across all 13 language subsets. Best: **bold**, second: underline.

Model	Open-source Benchmark						RegL-HoloDP	
	CPE		PubTabNet		PubTab1M			
	ED↓	F1↑	ED↓	F1↑	ED↓	F1↑	TEDS↑	F1↑
Pipeline Tools								
Dolphin-322M	0.2945	0.7480	0.4538	0.5791	0.3982	0.6381	0.6258	0.6615
PP-StructureV3	0.1954	0.8178	0.5483	0.6686	0.2892	0.7831	0.7547	0.7635
Expert VLMs								
DeepSeek-OCR	0.1803	0.8631	0.2316	0.8906	0.2731	0.7466	0.7787	0.6076
dots.ocr	0.1261	0.9212	0.0485	0.9911	**0.0375**	0.9795	0.8362	0.8701
General VLMs								
Qwen2.5-VL-7B	0.1691	0.8810	0.3831	0.6541	0.3538	0.6852	0.7054	0.7412
GPT-4o	0.2278	0.8188	0.3308	0.7096	0.2757	0.7778	0.7301	0.7734
HoloDP	**0.1199**	**0.9266**	**0.0419**	**0.9921**	0.0468	**0.9797**	**0.9024**	**0.9493**

Table 9: Region-level evaluation results. Best: **bold**, second: underline.

script-aware routing mechanism adaptively weights directional visual features, which improves boundary recognition and region alignment in irregular or mixed-layout documents. These findings suggest that HoloDP better bridges local visual patterns with document-level structural organization, leading to stronger layout consistency across diverse document types.

D.2 Case Study

Figure 3 presents representative examples from web document parsing. The examples include horizontal text, vertical text, and table-rich layouts, illustrating HoloDP's ability to perform page-level layout detection, fine-grained element parsing, and structured output reconstruction within a unified macro-to-micro protocol.

The horizontal and vertical text examples show that HoloDP can localize text regions and preserve reading-order information under different writing directions. The table example further shows that HoloDP can maintain row-column alignment and generate coherent structured output, reducing common errors such as fragmented regions or misaligned cells. These cases qualitatively support the quantitative findings: the main advantage of HoloDP comes from better alignment between script-aware visual features and serialized document structures, especially for documents involving heterogeneous layouts and underrepresented

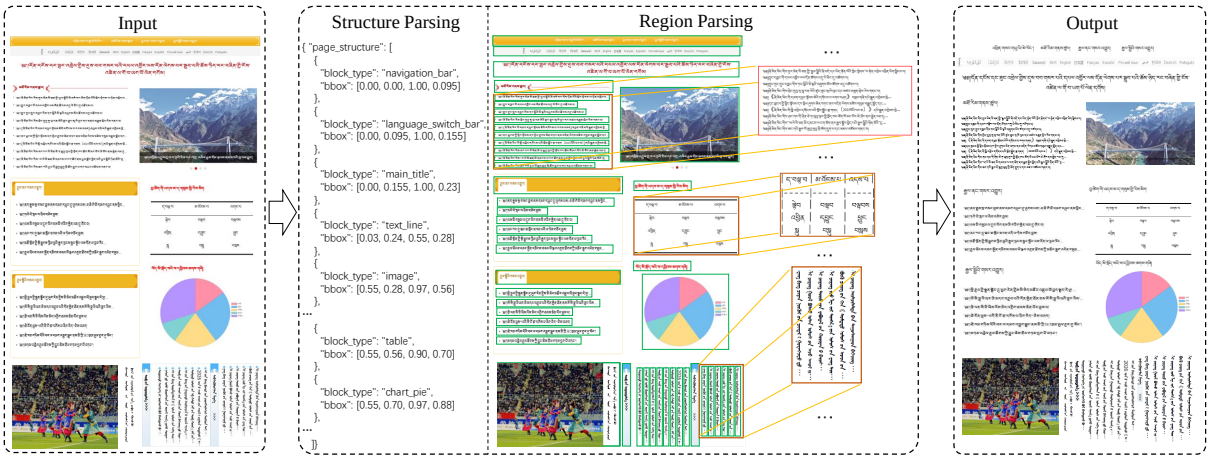


Figure 3: Representative examples from web document parsing, demonstrating HoloDP's ability to handle both horizontal and vertical text layouts, as well as table parsing, highlighting HoloDP's cross-script generalization and robustness in low-resource scenarios.

scripts.

D.3 Efficiency and Scalability

Although HoloDP uses a 7B multilingual decoder and is larger than OCR-specific baselines such as Dolphin-322M and MinerU, its inference cost remains practical for document parsing. On A100 GPUs, HoloDP takes 1.9 s on average to process a full A4 web-rendered document, including page-level structural prediction and region-level content parsing. This latency is higher than lightweight OCR pipelines but remains within the range of recent VLM-based document parsers.

More importantly, HoloDP brings larger accuracy gains on complex multilingual and mixed-script layouts, suggesting a favorable accuracy-efficiency trade-off for script-diverse document understanding. The main computational cost comes from autoregressive decoding with the multilingual decoder, while the adaptive vision-to-text module adds only lightweight directional filtering and routing operations. Future work may explore smaller multilingual decoders, decoder distillation, or more efficient vision adapters for deployment-oriented settings.

E Discussion

The results across benchmarks suggest that HoloDP is not only effective on individual tasks, but also more stable under varying document conditions. As shown in OmniDocBench, Fox-Page, Mstruct-HoloDP, and RegL-HoloDP, many baselines perform well on high-resource or visually regular subsets but degrade more noticeably when layout com-

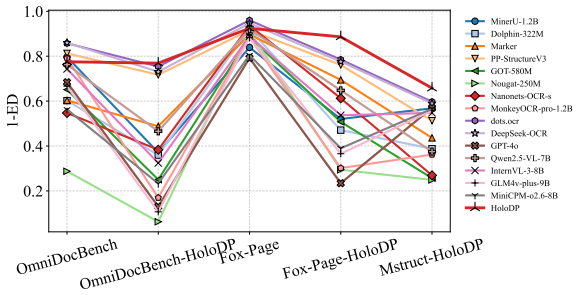


Figure 4: **Cross-dataset performance comparison.** HoloDP exhibits relatively stable performance across the evaluated benchmarks.

plexity, script diversity, or structural composition increases. In contrast, HoloDP shows smaller performance variation across these settings, suggesting that its improvements are less tied to a single benchmark or parsing granularity and more related to robustness under heterogeneous document layouts, as illustrated in Figure 4.

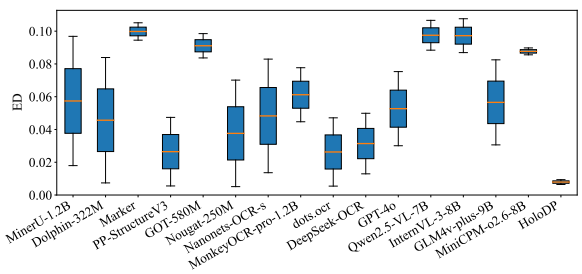


Figure 5: **Cross-lingual stability on a single dataset.** Lower variance indicates more consistent performance across languages. HoloDP exhibits the lowest cross-language variance among the evaluated models.

A similar trend appears within multilingual

benchmarks. While several OCR-oriented systems and general VLMs achieve competitive results on high-resource or visually regular languages, their errors fluctuate substantially across scripts. HoloDP maintains lower cross-language variance, indicating that the model is less sensitive to script-specific changes in glyph shape, writing direction, and layout convention. This supports the design choice of combining script-aware routing with shared multilingual generation, rather than relying only on language-specific recognition patterns, as shown in Figure 5.

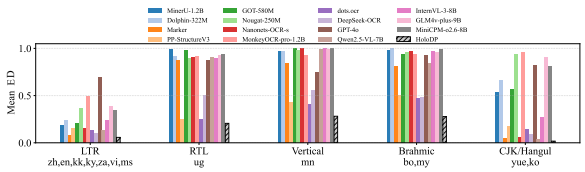


Figure 6: **Performance across representative script and writing-direction groups.**

The stability of HoloDP is closely related to its explicit modeling of script and writing-direction diversity. Conventional OCR pipelines and VLM-based parsers often perform best under dominant left-to-right layouts, whereas multilingual documents may involve vertical, right-to-left, or mixed-direction structures. HoloDP addresses this issue through directional visual branches and page- or region-level routing, allowing the model to emphasize different spatial priors for different scripts and layout regions. The results in Figure 6 further show that this design helps preserve reading order and structural boundaries under heterogeneous writing directions.

Finally, the gains on Mstruct-HoloDP and RegL-HoloDP indicate that HoloDP remains effective when documents combine multiple scripts, nested layouts, and fine-grained region structures. These settings require more than local text recognition: the model must align visual regions with serialized outputs while preserving reading order and structural hierarchy. Overall, the findings suggest that HoloDP's advantage comes from reducing performance variance across scripts, layouts, and task granularities, rather than from isolated improvements on a small set of high-resource languages.