HisDoc-OCR: Restoring Visual Grounding in MLLMs for Chinese Historical Document OCR

Jiahuan Cao* **Yongxin Shi*** **Zeyu Shan** **Zhengyang Lu** **Lianwen Jin†**
South China University of Technology
jiahuanc@foxmail.com, yongxin_shi@foxmail.com, eezeyu_shan@mail.scut.edu.cn
eeluzhengyang@mail.scut.edu.cn, eelwjin@scut.edu.cn

Abstract

Chinese historical documents encode millennia of cultural heritage, yet remain largely inaccessible to computational analysis. While multimodal large language models (MLLMs) have achieved strong performance on modern document OCR, their application to historical Chinese texts suffers from severe hallucinations, character fabrication, uncontrolled repetition, and semantic drift. We identify the root cause as visual-textual misalignment: models prioritize linguistic priors over visual evidence, particularly problematic when archaic orthography and degraded image quality destabilize cross-modal correspondences. To address this, we propose **HisDoc-OCR**, which restores visual grounding through three synergistic strategies: (1) **Layout Injection**, which encodes two-dimensional layout structures into textual outputs using layout-aware delimiters; (2) **First-Occurrence Boost**, which emphasizes vision-dependent characters during training by reweighting first-occurrence characters; (3) **Self-Distilled Attention Focusing**, which guides the model's attention by distilling patterns from the most focused layer to the remaining layers. Extensive experiments demonstrate that HisDoc-OCR consistently outperforms general-purpose and OCR-specific MLLMs. The code will be publicly available.

1 Introduction

Chinese historical documents constitute an invaluable cultural heritage accumulated over millennia, preserving rich historical, cultural, and scholarly knowledge. However, a large portion of these documents remains inaccessible to computational analysis due to their reliance on scanned images rather than machine-readable text. Digitization is therefore a critical prerequisite for the preservation and utilization of historical documents, among which optical character recognition (OCR) serves as the core enabling technology. By transcribing document images into structured textual representations following the natural reading order, OCR provides the foundation for downstream applications such as information retrieval, content analysis, and knowledge mining.

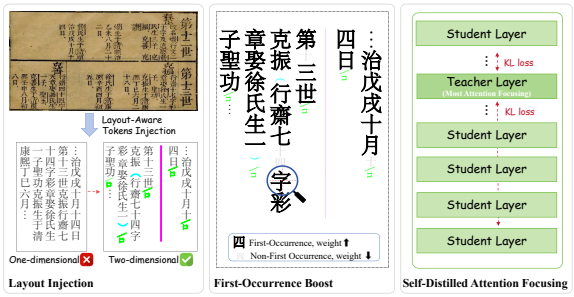


Figure 1: Our proposed HisDoc-OCR, which restores visual grounding through three synergistic strategies: (a) Layout Injection, (b) First-Occurrence Boost, and (c) Self-Distilled Attention Focusing.

Recently, with the rapid development of multimodal large language models (MLLMs) (Bai et al., 2025; Wang et al., 2025), document OCR has achieved remarkable progress. Both two-stage (Li et al., 2025; Cui et al., 2025; Zhang et al., 2025) and end-to-end (Mistral AI Team, 2025; Poznanski et al., 2025; Wei et al., 2025) methods have demonstrated strong performance. However, existing studies mainly focus on modern documents, while the application of MLLMs to Chinese historical document OCR remains largely unexplored. Compared with modern documents, Chinese historical documents typically exhibit more complex layouts, heterogeneous fonts, and severe noise interference, which make OCR for such documents considerably more challenging (Ma et al., 2020; Shi et al., 2023).

Directly applying general-purpose MLLMs to Chinese historical document OCR often leads to se-

* Equal contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 6053–6066
July 2-7, 2026

vere hallucination phenomena, including fabricated characters, uncontrolled repetition, and semantically drifted outputs. Prior studies (Bai et al., 2024; Zhou et al., 2024) have attributed hallucinations in MLLMs primarily to modality misalignment, namely the misalignment between visual evidence and textual generation. This issue becomes particularly pronounced in historical documents, where degraded visual quality and archaic orthography weaken cross-modal correspondences and cause the model to over-rely on linguistic priors rather than visual cues.

To address this, we propose **HisDoc-OCR**, which restores visual grounding through three synergistic strategies: **(1) Layout Injection.** Chinese historical documents contain rich two-dimensional layout structures, such as multi-block formats and Double-column Annotation, while conventional OCR outputs are typically one-dimensional text sequences. This structural mismatch exacerbates vision–text misalignment. Therefore, we inject explicit layout information into the textual output using layout-aware delimiters, thereby narrowing the gap between vision and text. **(2) First-Occurrence Boost.** During OCR transcription, characters that appear repeatedly can be inferred from contextual language priors, whereas characters appearing for the first time rely more heavily on visual evidence. However, standard training objectives treat all characters equally, causing vision-dependent alignment signals to be overwhelmed by semantic context. To mitigate this issue, we emphasize the learning of first-occurrence characters by increasing their loss weights, thereby strengthening vision–text alignment. **(3) Self-Distilled Attention Focusing.** During the human reading process, visual attention typically follows the reading order and focuses on the current recognition region. Inspired by this process, we introduce this strategy to guide the model's attention. Specifically, the layer exhibiting the most concentrated visual attention is selected as the teacher layer, and its attention patterns are distilled to the remaining layers via attention matching. This promotes consistent and locally focused attention across layers.

The effectiveness of HisDoc-OCR is extensively verified on several commonly used benchmarks. Experimental results show that HisDoc-OCR outperforms general and OCR-specific MLLMs. In addition, the effectiveness of the three proposed strategies is validated through ablation studies.

In summary, our main contributions include:

- We propose three effective and complementary visual-text alignment enhancement strategies, including Layout Injection, First-Occurrence Boost, and Self-Distilled Attention Focusing.
- We construct HisDoc-OCR, an outstanding MLLM-based OCR method for Chinese historical documents.
- Extensive experiments demonstrate the effectiveness of our method, which mitigates hallucination and improves OCR accuracy.

2 Related Work

2.1 MLLM for Document OCR

With the rapid development of multimodal large language models (MLLMs) (Bai et al., 2025; Wang et al., 2025), recent studies have explored the application of MLLMs to document OCR. Existing methods can be primarily categorized into Two-stage and end-to-end approaches. **Two-stage-based** methods first perform layout detection, followed by unified recognition of text, formulas, and tables within the detected regions. Representative approach including MinerU2.5 (Niu et al., 2025), PaddleOCR-VL (Cui et al., 2025), and Monkey-OCR v1.5 (Zhang et al., 2025). **End-to-end-based** methods directly generate structured text from document images. Representative approach including dots.ocr (Li et al., 2025), Mistral OCR (Mistral AI Team, 2025), olmOCR (Poznanski et al., 2025), SmolDocling (Nassar et al., 2025), and DeepSeek-OCR (Wei et al., 2025). Although these methods have demonstrated remarkable performance, they mainly focus on modern documents, leaving historical documents largely unexplored.

2.2 LLM / MLLM for Historical Documents

With the rapid advancement of Large Language Models (LLMs), researchers have begun to explore their applications in historical documents. Some studies leverage LLMs to address specific tasks, such as ancient-to-modern translation (Cao et al., 2023) and entity recognition (Dang, 2025). SikuGPT (Liu et al., 2024) leverages massive classical Chinese corpora for generative pretraining. Xunzi (Xunzi-LLM-of-Chinese-classics, 2024), TongGu (Cao et al., 2024), and WenyanGPT (Yao et al., 2025) fine-tune general LLMs with classical Chinese–specific knowledge. For MLLMs, existing methods focus on fine-tuning general MLLMs

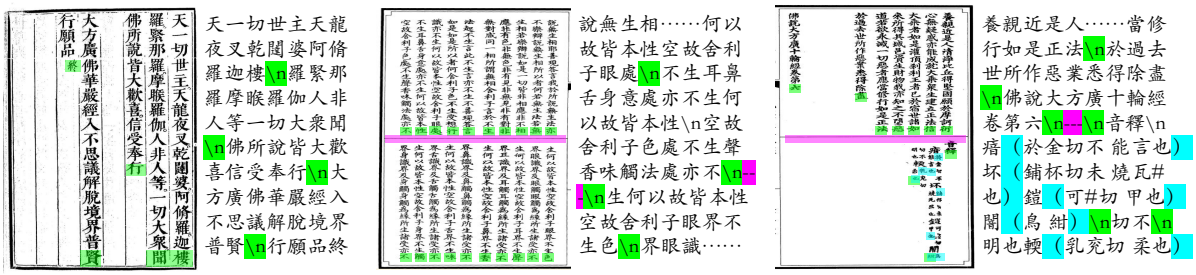


Figure 2: Several common layout elements in Chinese historical document images and the layout identifiers we introduced. Figure (a) shows a single-block structure, where the green positions indicate column breaks, and we insert newline tokens (\n) in the text to represent column break symbols; Figure (b) shows a multi-block structure, where the pink parts indicate column divisions, and in addition to column-break symbols, we introduce block-separation markers (——) to represent column switch symbols; the cyan parts in Figure (c) indicate the positions of double-column annotation notes, and we introduce parenthesis markers (()) to represent two-column interlinear note symbols.

to improve the understanding of historical documents. Representative approaches include XunZi-MLLM (Zhu et al., 2025), CalliReader (Luo et al., 2025), and TongGu-VL (Cao et al., 2025). These researchers primarily focus on the semantic understanding of historical documents, while the specific study of MLLM-based historical document OCR remains a research gap.

3 Methodology

Our study addresses a critical gap in the application of MLLMs to Chinese historical document OCR, a domain where existing general-purpose and OCR-specialized MLLMs encounter a fundamental challenge: severe hallucinations stemming from misalignment between visual perception and textual generation (Bai et al., 2024; Zhou et al., 2024). This misalignment manifests particularly acutely in historical documents, where archaic character forms, degraded image quality, and complex layouts conspire to destabilize the vision-language correspondence that underpins reliable OCR. To address this issue, we introduce a three-pronged approach that systematically strengthens different facets of visual-text alignment: Layout Injection, First-Occurrence Boost, and Self-Distilled Attention Focusing. The details of these strategies are elaborated in the following.

3.1 Layout Injection: Bridging the Dimensional Gap via Structural Tokenization

Historical document images inherently encode information in two modalities: a semantic modality (character identities) and a spatial modality (two-dimensional layout geometry). Standard autoregressive OCR models, however, operate on linearized token sequences $\mathbf{y} = [y_1, y_2, \ldots, y_N]$, creating a fundamental *dimensional mismatch*: while the visual encoder processes spatial features $\mathbf{V} \in \mathbb{R}^{H \times W \times D}$, the decoder generates text in a strictly one-dimensional manifold. This discrepancy induces a structural information bottleneck, in which spatial relationships (such as column adjacency and reading order) must be inferred implicitly during decoding. This mismatch becomes particularly severe under complex layouts with multiple columns or interleaved annotations, where ambiguous spatial cues make it difficult for the model to maintain stable visual–text correspondence.

To address this issue, we propose the Layout Injection strategy that incorporates two-dimensional layout information into the text recognition results, thereby narrowing the gap between vision and text. By encoding layout structures directly into the target sequence through structural tokenization, the OCR task is reformulated from unstructured sequence modeling to layout-aware generation. Specifically, as illustrated in Figure 2, these delimiters are employed for layout guidance as follows:

- **Column Separation**: A newline token (\n)

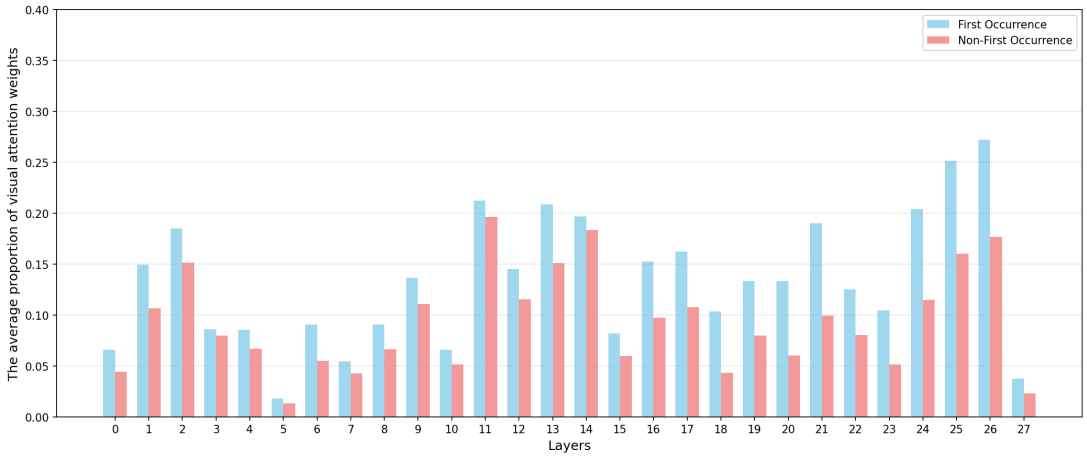


Figure 3: The visual attention weights of the first-occurring characters and non-first-occurring characters in Chinese historical documents across all layers of the MLLM.

is inserted at the end of each text column, transforming continuous text into a two-dimensional structure organized by columns.

- **Block Separation**: A block-separation marker (−) is inserted between adjacent blocks to explicitly indicate inter-block boundaries.
- **Double-column Annotation**: This is a special format (Shi et al., 2025) in Chinese historical documents, where two smaller text columns are placed below a larger one. These smaller columns typically serve to annotate or explain the larger column above. For this special layout, parenthesis markers are added. The format is as follows: *main text (right annotation left annotation)*.

3.2 Fisrt-Occurrence Boost

Let $P(c_i \mid v, c_{<i})$ denote the conditional probability of predicting character c_i given visual features v and preceding context $c_{<i}$. When a character or phrase appears repeatedly, the contextual information encoded in $c_{<i}$ provides a strong prior, allowing $P(c_i \mid v, c_{<i})$ to be largely determined by contextual regularities. In contrast, for characters that appear for the first time, the contextual prior is weak, and the prediction relies more heavily on the visual evidence v.

As illustrated in Figure 3, we quantify this phenomenon by analyzing cross-modal attention patterns across all Transformer layers. The Figure demonstrates that tokens corresponding to first-occurrence characters consistently exhibit higher text-to-visual attention ratios, particularly in middle and higher layers. This disparity confirms that the model implicitly distinguishes between vision-dependent and language-dependent tokens.

However, standard cross-entropy loss treats all tokens uniformly:

$$\mathcal{L}_{\text{standard}} = -\sum_{i=1}^{N} \log P(c_i \mid v, c_{<i}). \quad (1)$$

This homogeneous weighting creates an imbalanced gradient flow: language-driven gradients of repeated characters dominate updates to visual encoders, diluting the learning signal from truly vision-dependent tokens.

To restore the primacy of visual grounding, we propose an adaptive loss weighting scheme that amplifies the training signal for first-occurrence characters:

$$\mathcal{L}^{\text{char}}_{\text{weighted}} = -\sum_{i=1}^{N} w_i \cdot \log P(c_i \mid v, c_{<i}), \quad (2)$$

where the weight w_i is defined as:

$$w_i = 1 + \alpha \cdot \mathbb{I}_{\text{first}}(c_i). \quad (3)$$

Binary indicator $\mathbb{I}_{\text{first}}(c_i) \in \{0, 1\}$: Equals 1 if c_i appears for the first time in the sequence; 0 otherwise. For first-occurrence tokens with $w_i = 3$, the LLM decoder receives approximately $3\times$ stronger parameter updates compared to repeated tokens. This gradient rebalancing compels the model to prioritize visual feature extraction over linguistic pattern matching during early training phases, establishing a robust visual grounding foundation.

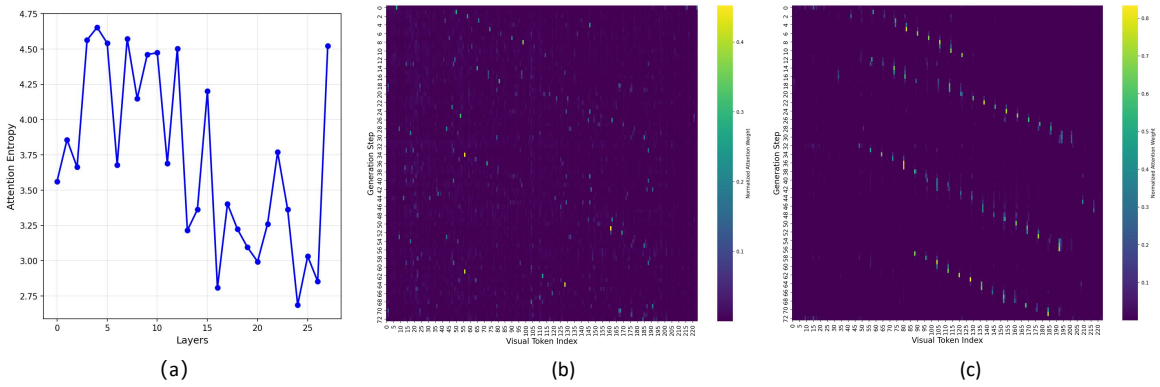


Figure 4: Illustration of"text-to-visual" attention. (a) The attention entropy across different layers. (b) The attention map of the layer with the highest attention entropy. (c) The attention map of the layer with the lowest attention entropy.

3.3 Self-Distilled Attention Focusing: Emergent Supervision from Internal Attention Dynamics

Human reading of historical documents exhibits a distinctive attentional pattern: visual focus progresses sequentially along the reading path, with explicit tracking mechanisms (e.g., finger-pointing) ensuring precise alignment between visual perception and character recognition. This spatially-locked attention prevents the cognitive system from hallucinating characters based solely on linguistic context; the visual evidence continuously grounds the recognition process. Inspired by this mechanism, we propose a strategy to explicitly guide MLLMs to focus on the region being recognized. A straightforward solution would be to constrain visual attention to the character's bounding box being recognized, but this requires labor-intensive character-level annotations.

To address this issue, we propose the **Self-Distilled Attention Focusing** strategy. As illustrated in Figure 4, we analyze the text-to-visual attention behavior of Qwen-3-VL-2B across layers. Figure 4(a) shows the attention entropy at different depths, where shallower layers exhibit higher entropy and deeper layers demonstrate progressively lower entropy, indicating increasingly concentrated attention. Figure 4(b) visualizes the attention map of a high-entropy layer, in which attention is broadly distributed across visual regions. In contrast, Figure 4(c) presents the attention map of the lowest-entropy layer, where attention is sharply focused on the visual regions corresponding to the currently recognized text.

These observations reveal that modern MLLMs spontaneously develop stratified attention patterns across layers without explicit supervision, and that a naturally focused layer can be identified based on attention entropy. Leveraging this emergent property, we adopt a self-distilled attention focusing approach, in which the layer l^* with the lowest attention entropy is treated as a pseudo-teacher that implicitly captures spatially grounded attention. Specifically, we first compute the text-to-visual attention entropy for each layer on a subset of training samples and select the most focused layer as the teacher. During training, we apply a KL-divergence loss to align the attention distributions of the remaining layers with that of the teacher layer, encouraging consistent and localized attention across layers. The distillation loss is as follows:

$$\mathcal{L}_{\text{distill}} = \sum_{l \in \mathcal{S}} \sum_{t} \sum_{v} \mathbf{A}_T^{(t,v)} \log \frac{\mathbf{A}_T^{(t,v)}}{\mathbf{A}_l^{(t,v)}}, \quad (4)$$

where $\mathbf{A}_l^{(t,v)}$ is the normalized attention from text token t to visual token v at student layer l, $\mathbf{A}_T^{(t,v)}$ is the corresponding teacher layer attention, and $\mathcal{S}$ is the set of student layers.

From a regularization perspective, this loss imposes a cross-layer consistency constraint that stabilizes attention dynamics. Without distillation, different layers may attend to conflicting visual regions, leading to inconsistent representations that propagate through the decoder. By anchoring all layers to the focused pattern of l^*, we ensure that the entire network maintains coherent visual grounding, reducing the degrees of freedom that could lead to hallucinations.

Method	MTHv2						M^{5}HisDoc					
	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓
General MLLMs												
Qwen3-VL-2B	-207.22	22.69	88.23	32.31	5.36	39.25	-114.91	11.82	92.78	20.90	2.70	35.10
Qwen3-VL-4B	-51.49	59.56	51.48	61.85	29.12	16.63	-39.82	43.81	64.45	48.35	20.23	23.90
Qwen3-VL-8B	-56.23	60.52	51.37	62.78	29.48	19.38	-7.40	49.08	57.87	53.96	24.68	17.15
InternVL3.5-2B	26.54	55.05	47.91	61.19	30.93	3.38	2.28	44.56	58.74	50.51	22.93	12.00
InternVL3.5-4B	37.63	58.89	44.17	63.24	34.21	3.12	20.55	51.30	52.23	55.51	27.79	7.70
InternVL3.5-8B	48.14	57.94	44.15	62.81	33.77	2.13	34.23	53.72	49.32	57.73	29.09	5.50
Specialized MLLMs(two-stage)												
PaddleOCR-VL	-7.97	50.34	58.76	54.58	24.00	**0.00**	-5.64	31.79	72.84	38.56	14.31	**0.20**
MonkeyOCR-pro-3B	-137.52	48.74	60.07	57.71	24.54	14.37	-104.2	36.22	69.80	46.96	16.63	25.40
MinerU2.5	-166.64	45.24	63.83	51.49	21.61	5.37	-400.02	27.73	80.50	11.43	16.58	14.67
Specialized MLLMs(end2end)												
olmOCR-2-7B-1025	-22.15	54.30	53.26	60.98	28.34	14.37	-18.04	40.44	64.22	49.05	20.77	18.90
DeepSeek-OCR	-224.97	22.09	85.93	34.04	9.24	15.00	-202.51	6.83	95.18	11.85	2.83	21.60
dots.ocr	-96.03	62.47	54.63	64.30	26.87	18.12	-52.39	49.29	62.63	52.65	21.26	25.00
HisDoc-OCR-Qwen3-VL-2B (Ours)	69.36	70.67	29.97	69.74	42.37	0.13	52.13	60.70	41.14	61.20	33.61	3.40
HisDoc-OCR-InternVL3.5-2B (Ours)	65.79	70.55	30.09	70.02	42.37	0.25	56.32	61.13	40.44	62.12	34.36	2.10
HisDoc-OCR-InternVL3.5-4B-full (Ours)	**71.09**	**71.72**	**28.69**	**70.57**	**43.61**	**0.00**	**64.82**	**67.90**	**33.07**	**65.79**	**40.26**	0.50

Table 1: Performance comparison with various MLLMs on MTHv2 and M^{5}HisDoc. AR, CR, ED, and F1 indicate Accurate Rate, Correct Rate, Normalized Edit Distance, and F1-Score, respectively. **Bold** indicates the best score, and underline indicates the second best result.

3.4 Training Objective

The final training objective integrates the First-Occurrence Boost–weighted character loss and the Self-Distilled Attention Focusing loss. Specifically, the overall optimization objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{\text{char}}_{\text{weighted}} + \lambda \, \mathcal{L}_{\text{distill}}. \tag{5}$$

In our experiments, the weighting coefficient λ is set to 1.

4 Experiments

4.1 Implementation Details

To validate the generality of our strategies across different MLLMs, we build our models upon two mainstream MLLMs with high-resolution visual input: native-resolution-based (Qwen3-VL-2B (Bai et al., 2025)) and cropped-subimage-based (InternVL3.5-2B (Wang et al., 2025)). During training, the visual encoder and projector are frozen, and only the LLM component is fine-tuned. To constrain the sequence length, the maximum number of input visual tokens is set to 2048 for Qwen3-VL-2B, while the maximum number of cropped sub-images is set to 8 for InternVL3.5-2B. The training datasets include MTHv2 (Ma et al., 2020) and M^{5}HisDoc (Shi et al., 2023). To verify the scalability of our method, we further develop HisDoc-OCR-InternVL3.5-4B-full, which is built upon the InternVL3.5-4B with the number of cropped sub-images set to 12, and is trained with additional data (the validation set of M^{5}HisDoc). The models are trained with a batch size of 64 using the AdamW optimizer. The initial learning rate is set to 1×10^{-5} with a warm-up ratio of 0.03, followed by a cosine decay schedule. Training is conducted for 5 epochs. All experiments are conducted on 4 NVIDIA A6000 GPUs.

4.2 Evaluation Metrics

To evaluate the similarity between model predictions and ground truth text, we employ the following metrics: Accurate Rate (Wang et al., 2011), Correct Rate (Wang et al., 2011), Normalized Edit Distance (Lcvenshtcin, 1966), F1-Score (Fisher, 1936), and BLEU (Papineni et al., 2002). For the IC19HDRC (Saini et al., 2019) dataset, since the ground truth lacks reading-order annotations, we report only the character-level F1-Score (Fisher, 1936). In addition, the model may exhibit a repetition issue during sequence generation, where identical text segments are repeatedly produced at the end of the output. To quantify this behavior, we define the Repetition Rate (RR). An output is considered to enter repetition mode if: (1) the generated length reaches the maximum generation limit *max_new_tokens*; and (2) letting S be the last 100 characters of the output, the number of unique characters in S is less than 34. The Repetition Rate is then computed as the proportion of test samples whose outputs enter repetition mode.

4.3 Main Results

We evaluate the performance of HisDoc-OCR on various Chinese historical document OCR benchmarks, including MTHv2 (Ma et al., 2020), M^{5}HisDoc (Shi et al., 2023), and IC19HDRC (Saini et al., 2019). We compare

HisDoc-OCR with general-purpose MLLMs and OCR-specific MLLMs (Two-stage-based and End-to-end-based). The results are demonstrated in Table 1 and 2. Based on the results, we draw the following key conclusions. Firstly, our method demonstrates strong effectiveness in Chinese historical document OCR, achieving state-of-the-art performance across multiple benchmarks. Secondly, by integrating the proposed alignment-enhancement strategies, HisDoc-OCR maintains low repetition ratios across all benchmarks, demonstrating that our method effectively mitigates hallucination. In contrast, general MLLMs such as Qwen3-VL and InternVL3.5 exhibit limited performance on Chinese historical documents. Although larger models generally achieve better accuracy, they still suffer from severe repetition, indicating strong hallucination tendencies. Thirdly, HisDoc-OCR demonstrates strong generalization ability: even on the unseen IC19HDRC benchmark, our method still achieves strong performance.

Method	IC19HDRC	
	Char F1 ↑	RR ↓
General MLLMs		
Qwen3-VL-2B	14.94	31.31
Qwen3-VL-4B	20.78	20.82
Qwen3-VL-8B	30.00	12.29
InternVL3.5-2B	23.57	18.00
InternVL3.5-4B	28.14	9.30
InternVL3.5-8B	32.88	3.84
Specialized MLLMs (two-stage)		
PaddleOCR-VL	13.88	3.33
MonkeyOCR-pro-3B	26.93	23.04
MinerU2.5	21.88	10.24
Specialized MLLMs (end2end)		
olmOCR-2-7B-1025	32.54	6.57
DeepSeek-OCR	9.49	16.30
dots.ocr	17.93	29.52
HisDoc-OCR-Qwen3-VL-2B (Ours)	36.61	0.43
HisDoc-OCR-InternVL3.5-2B (Ours)	35.49	0.43
HisDoc-OCR-InternVL3.5-4B-full (Ours)	**36.66**	**0.09**

Table 2: Performance comparison with various MLLMs on IC19HDRC. Char F1 indicate character-level F1-Score.

4.4 Ablation Study

We conduct Ablation Study based on Qwen3-VL-2B. Additionally, the ablation results based on InternVL3.5-2B are provided in the Appendix A.

The Efficiency of Each Strategy. Table 3 presents an ablation study on the three proposed components: Layout Injection (LI), First-Occurrence Boost (FOB), and Self-distilled Attention Focusing (SDAF), evaluated on both MTHv2 and M^{5}HisDoc benchmarks. Starting from the base model without any alignment enhancement, all three components individually improve OCR performance to varying degrees. Among them, LI yields notable gains on both datasets, particularly on M^{5}HisDoc, indicating that explicit layout cues are crucial for handling complex historical document structures. FOB brings the most significant reduction in repetition ratio on MTHv2, demonstrating its effectiveness in alleviating hallucination by strengthening vision-dependent character learning. SDAF provides consistent improvements, especially in repetition reduction, suggesting that guiding visual attention helps stabilize the transcription process. When all three components are jointly applied, the model achieves the best overall performance across almost all metrics. These results indicate that the proposed components are complementary, together leading to robust OCR performance and effective hallucination suppression.

The Weight in First-Occurrence Boost. We investigate the impact of the hyperparameter α in our First Character Alignment strategy. As shown in Table 4, setting $\alpha = 2$ (corresponding to a total weight of 3) yields the best overall performance. Specifically, when $\alpha = 0$ (i.e., uniform weighting for all characters), the model exhibits relatively lower recognition accuracy and noticeably higher repetition ratios, particularly on the challenging M^{5}HisDoc benchmark. As α increases from 0 to 2, we observe consistent improvements in AR, CR, F1, and BLEU scores, accompanied by a significant reduction in the repetition ratio across both datasets. This indicates that appropriately emphasizing first-occurrence characters strengthens vision-dependent learning and effectively alleviates hallucination. Further increasing α beyond 2 does not lead to additional performance gains. Although $\alpha = 3$ and $\alpha = 4$ slightly improve AR on M^{5}HisDoc, overall recognition metrics become marginally worse or saturated, and no clear advantage is observed compared with $\alpha = 2$. This suggests that excessively large values of α may overemphasize first-occurrence characters, thereby disturbing the balance between visual grounding and contextual modeling. Overall, $\alpha = 2$ achieves the optimal trade-off between recognition accuracy and hallucination suppression, and is therefore adopted as the default setting in our experiments.

The Efficiency of Layout Injection. To deeply evaluate the efficiency of Layout Injection, we manually select 100 samples with highly complex

Method			MTHv2						M^5HisDoc					
LI	FOB	SDAF	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓
			64.14	69.66	31.38	69.32	41.50	1.00	33.59	57.06	45.57	58.41	30.97	10.50
✔			65.02	70.64	30.21	69.71	42.19	0.75	50.29	60.06	41.85	60.29	33.27	**2.80**
	✔		68.65	70.16	30.62	69.67	42.02	**0.13**	48.35	59.16	42.76	60.24	32.66	5.30
		✔	66.62	69.96	30.96	69.41	41.78	0.63	36.94	57.31	45.00	58.61	31.30	9.10
✔	✔	✔	**69.36**	**70.67**	**29.97**	**69.74**	**42.37**	**0.13**	**52.13**	**60.70**	**41.14**	**61.20**	**33.61**	3.40

Table 3: Ablation study on the proposed Layout Injection (LI), First-Occurrence Boost (FOB) and Self-Distilled Attention Focusing (SDAF).

α	MTHv2						M^5HisDoc					
	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓
0	64.14	69.66	31.38	69.32	41.50	1.00	33.59	57.06	45.57	58.41	30.97	10.50
1	65.82	70.03	30.90	69.50	41.82	0.50	43.84	**60.08**	**42.33**	**60.76**	**33.34**	6.79
2	68.65	**70.16**	**30.62**	**69.67**	**42.02**	**0.13**	48.35	59.16	42.76	60.24	32.66	5.30
3	**68.67**	70.04	30.77	69.58	41.88	**0.13**	**48.71**	58.82	43.17	60.10	32.40	**4.50**
4	66.86	69.95	31.01	69.53	41.75	0.37	47.96	58.71	43.13	60.13	32.29	4.60

Table 4: Ablation study on the impact of different α used in First-Occurrence Boost.

Method	Score					
	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓
Qwen3-VL-4B	-39.82	43.81	64.45	48.35	20.23	23.90
	-103.77(-160.60%)	13.34(-69.55%)	92.96(-44.24%)	21.36(-55.82%)	3.35(-83.44%)	39.51(-65.31%)
InternVL3.5-4B	20.55	51.30	52.23	55.51	27.79	7.70
	-13.81(-167.19%)	35.83(-30.16%)	68.29(-30.75%)	42.33(-23.74%)	17.91(-35.55%)	13.58(-76.36%)
olmOCR-2-7B-1025	-18.04	40.44	64.22	49.05	20.77	18.90
	-43.79(-142.74%)	17.79(-56.01%)	87.53(-36.30%)	30.93(-36.94%)	7.02(-66.20%)	27.16(-43.70%)
MonkeyOCR-pro-3B	-104.2	36.22	69.80	46.96	16.63	25.40
	-191.31(-83.59%)	17.85(-50.72%)	88.33(-26.55%)	33.23(-29.24%)	5.46(-67.17%)	**55.56**(-118.74%)
dots.ocr	-52.39	49.29	62.63	52.65	21.26	25.00
	-80.89(-54.40%)	19.89(-59.65%)	89.12(-42.29%)	30.53(-42.01%)	6.74(-68.30%)	30.86(-23.44%)
HisDoc-OCR-InternVL3.5-4B-full (Ours)	64.82	67.90	33.07	65.79	40.26	0.50
	61.36(-5.34%)	63.61(-6.32%)	38.09(-15.18%)	60.65(-7.81%)	35.64(-11.48%)	0.00(+100.00%)

Table 5: The impact of layout complexity. For the performance of each model, the first row represents the score on the full M^5HisDoc dataset, while the second row represents the results on the 100 samples with the most complex layouts in M^5HisDoc.

layouts from the M^5HisDoc test set, forming a complex-layout subset. We report the performance gap between the original test set and the complex-layout subset in Table 5. The results demonstrate that most existing MLLM-based OCR models suffer from severe performance degradation when facing complex layouts. These indicate that without explicit layout modeling, the models are highly sensitive to layout complexity. In contrast, our HisDoc-OCR achieves the smallest performance drop across all metrics. These confirm that explicit Layout Injection effectively enhances the model's robustness to complex layout, enabling more stable recognition and accurate reading-order reasoning under challenging layout conditions.

5 Conclusion

This paper addresses a critical challenge in heritage document digitization: severe hallucinations when applying multimodal large language models to Chinese historical texts. We observe that the models over-rely on linguistic priors from modern corpora, failing to ground predictions in visual evidence when confronted with archaic orthography and degraded images. To address this challenge, we introduce HisDoc-OCR, which restores visual grounding through three synergistic strategies: Layout Injection preserves spatial topology via hierarchical structural tokenization; First-Occurrence Boost rebalances training dynamics by amplifying gradients for vision-dependent tokens; Self-Distilled Attention Focusing exploits emergent attention stratification to propagate spatially-localized patterns without character-level annotations. Beyond historical Chinese OCR, our findings reveal a broader insight: hallucinations in specialized domains stem from misaligned inductive biases rather than insufficient model capacity. Our approach demonstrates that lightweight, interpretable interventions can realign foundation models to specialized tasks, offering a scalable alternative to domain-specific pretraining for low-resource heritage languages.

Limitations

Despite the effectiveness of the proposed Layout Injection strategy, our method introduces additional annotation requirements. Specifically, the Layout Injection relies on textual transcriptions augmented with layout-aware delimiters, which assume the availability of layout information during training.

Ethical considerations

This study strictly adheres to academic ethical standards. All training and evaluation data consist of publicly available or legally authorized images of Chinese historical documents and their manually proofread transcriptions, and do not involve any personal privacy or sensitive information. Recognizing that multimodal large language models are prone to hallucinations in historical document OCR—such as generating fictitious characters, uncontrolled repetitions, or semantic drift—this work explicitly focuses on enhancing vision–text alignment as its core objective. Through layout-aware training and attention-focusing mechanisms, our approach constrains model generation to strictly rely on visual evidence, thereby preventing speculative outputs that deviate from the actual image content. All experiments follow the principle of faithful transcription; model outputs are used solely for academic research and are never employed to generate, alter, or disseminate unverified historical texts, thus ensuring the authenticity and reliability of the cultural heritage digitization process.

Acknowledgments

This research is supported in part by National Natural Science Foundation of China (Grant No.: 62476093, 62441604).

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Jiahuan Cao, Yang Liu, Peirong Zhang, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025. TongGu-VL: Advancing visual-language understanding in Chinese classical studies through parameter sensitivity-guided instruction tuning. In *Proc. ACM MM*, pages 11111–11120.

Jiahuan Cao, Dezhi Peng, Yongxin Shi, Zongyuan Jiang, and Lianwen Jin. 2023. Translating ancient Chinese to modern Chinese at scale: A large language model-based approach. In *Proc. ALT Workshop*, pages 61–69.

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024. TongGu: Mastering classical Chinese understanding with knowledge-grounded large language models. In *Proc. EMNLP*, pages 4196–4210.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, and 1 others. 2025. PaddleOCR-VL: Boosting multilingual document parsing via a 0.9B ultra-compact vision-language model. *arXiv preprint arXiv:2510.14528*.

Jianfei Dang. 2025. Entity recognition in Chinese classics via fine-tuned large language models. In *Proc. ISAEECE*, pages 630–634. IEEE.

Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

VI Lcvenshtcin. 1966. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Proc. Soviet physics-doklady*, volume 10.

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

Chang Liu, Dongbo Wang, Zhixiao Zhao, Die Hu, Mengcheng Wu, Litao Lin, Jiangfeng Liu, Hai Zhang, Si Shen, Bin Li, and 1 others. 2024. SikuGPT: a generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. *ACM Journal on Computing and Cultural Heritage*, 17(4):1–17.

Yuxuan Luo, Jiaqi Tang, Chenyi Huang, Feiyang Hao, and Zhouhui Lian. 2025. CalliReader: Contextualizing Chinese calligraphy via an embedding-aligned vision-language model. In *Proc. ICCV*, pages 23030–23040.

Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. 2020. Joint layout analysis, character detection and recognition for historical document digitization. In *Proc. ICFHR*, pages 31–36. IEEE.

Mistral AI Team. 2025. Mistral OCR: Introducing the world's best document understanding API.

Ahmed Nassar, Andres Marafioti, Matteo Omenetti, Maksym Lysak, Nikolaos Livathinos, Christoph Auer, Lucas Morin, Rafael Teixeira de Lima, Yusik Kim, A Said Gurbuz, and 1 others. 2025. SmolDocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. *arXiv preprint arXiv:2503.11576*.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, and 1 others. 2025. MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Jake Poznanski, Luca Soldaini, and Kyle Lo. 2025. olmOCR2: Unit test rewards for document OCR. *arXiv preprint arXiv:2510.19817*.

Rajkumar Saini, Derek Dobson, Jon Morrey, Marcus Liwicki, and Foteini Simistira Liwicki. 2019. ICDAR 2019 historical document reading challenge on large structured Chinese family records. In *Proc. ICDAR*, pages 1499–1504. IEEE.

Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. 2023. M5HisDoc: A large-scale multi-style Chinese historical document analysis benchmark. In *Proc. NeurIPS*, volume 36, pages 78483–78495.

Yongxin Shi, Dezhi Peng, Yuyi Zhang, Jiahuan Cao, and Lianwen Jin. 2025. A large-scale dataset for Chinese historical document recognition and analysis. *Scientific Data*, 12(1):169.

Qiu-Feng Wang, Fei Yin, and Cheng-Lin Liu. 2011. Handwritten Chinese text recognition by integrating multiple contexts. *IEEE transactions on pattern analysis and machine intelligence*, 34(8):1469–1481.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

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

Xunzi-LLM-of-Chinese-classics. 2024. Xunziallm.

Xinyu Yao, Mengdi Wang, Bo Chen, and Xiaobing Zhao. 2025. WenyanGPT: A large language model for classical Chinese tasks. *arXiv preprint arXiv:2504.20609*.

Jiarui Zhang, Yuliang Liu, Zijun Wu, Guosheng Pang, Zhili Ye, Yupei Zhong, Junteng Ma, Tao Wei, Haiyang Xu, Weikai Chen, and 1 others. 2025. MonkeyOCR v1.5 technical report: Unlocking robust document parsing for complex patterns. *arXiv preprint arXiv:2511.10390*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. In *Proc. ICLR Workshop*.

Dongmei Zhu, Chang Liu, Xue Zhao, Zhixiao Zhao, Si Shen, and Dongbo Wang. 2025. XunZi-MLLM: a multimodal large language model for ancient text and image recognition. *Digital Scholarship in the Humanities*, 40(2):709–722.

A Ablation Studies

Qwen3-VL and InternVL3.5 represent two mainstream high-resolution visual methods. The former achieves high resolution through a Vision Transformer (ViT) with native resolution, while the latter increases the upper limit of resolution by splitting the entire image into sub-images, feeding them into a ViT with fixed resolution, and then stitching the sub-images back together. Table 7 presents the ablation results of our proposed Layout Injection (LI), First-Occurrence Boost (FOB), and Self-Distilled Attention Focusing (SDAF) on InternVL3.5-VL-2B. The results demonstrate that our proposed methods are effective on MLLMs with mainstream architectures.

B Qualitative results

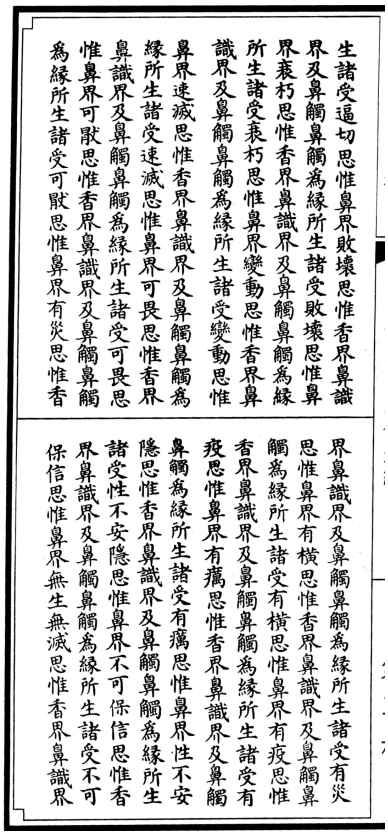
Figure 5-8 demonstrate the outstanding performance of HisDoc-OCR in tackling challenging Chinese historical document text recognition tasks, including printed fonts, complex layouts, cursive handwritten text, and dense text.

α	MTHv2						M^5HisDoc					
	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓
0	61.11	**69.75**	31.67	69.71	41.58	1.25	36.76	57.42	45.24	59.28	31.69	9.20
1	62.85	**69.75**	31.62	**69.86**	41.70	1.12	41.02	58.39	43.93	60.59	32.42	6.20
2	65.64	69.62	31.60	69.76	41.74	1.25	47.97	**58.56**	**43.47**	60.75	**32.56**	5.10
3	**66.08**	69.61	**31.51**	69.75	**41.76**	**0.75**	48.27	58.52	43.58	60.82	32.49	5.80
4	62.33	69.28	31.88	69.79	41.49	1.12	**49.67**	58.13	43.77	**60.88**	32.37	**4.00**

Table 6: Ablation study on the impact of different α used in First-Occurrence Boost based on InternVL3.5-2B.

Method			MTHv2						M^5HisDoc					
LI	FOB	SDAF	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓	AR ↑	CR ↑	ED ↓	F1 ↑	BLEU ↑	RR ↓
			61.11	69.75	31.67	69.71	41.58	1.25	36.76	57.42	45.24	59.28	31.69	9.20
✔			65.51	**70.91**	**29.96**	**70.02**	**42.46**	**0.25**	50.59	60.49	41.50	61.53	33.95	2.70
	✔		65.64	69.62	31.60	69.76	41.74	1.25	47.97	58.56	43.47	60.75	32.56	5.10
		✔	61.94	69.91	31.43	69.71	41.74	1.50	35.21	57.21	45.57	59.29	31.57	8.70
✔	✔	✔	**65.79**	70.55	30.09	**70.02**	42.37	**0.25**	**56.32**	**61.13**	**40.44**	**62.12**	**34.36**	**2.10**

Table 7: Ablation study on the proposed Layout Injection (LI), First-Occurrence Boost (FOB) and Self-Distilled Attention Focusing (SDAF) based on InternVL3.5-VL-2B.

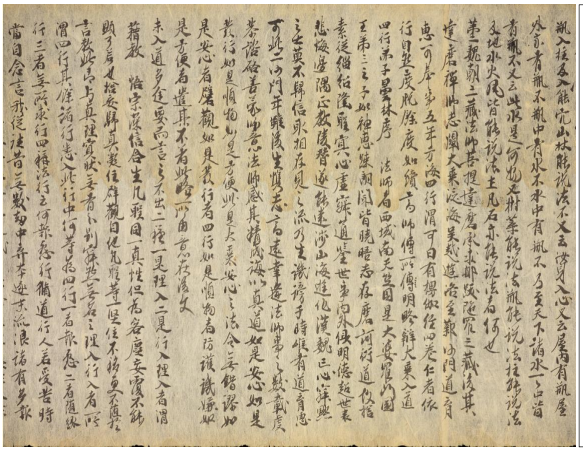


生諸受逼切思惟鼻界敗壞思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受敗壞思惟鼻界衰朽思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受衰朽思惟鼻界變動思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受變動思惟鼻界速滅思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受速滅思惟鼻界可畏思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受可畏思惟鼻界可猒思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受可猒思惟鼻界有災思惟香

界鼻識界及鼻觸鼻觸爲緣所生諸受有災思惟鼻界有橫思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受有橫思惟鼻界有疫思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受有疫思惟鼻界有癘思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受有癘思惟鼻界性不安隱思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受性不安隱思惟鼻界不可保信思惟香界鼻識界及鼻觸鼻觸爲緣所生諸受不可保信思惟鼻界無生無滅思惟香界鼻識界

Input image OCR output

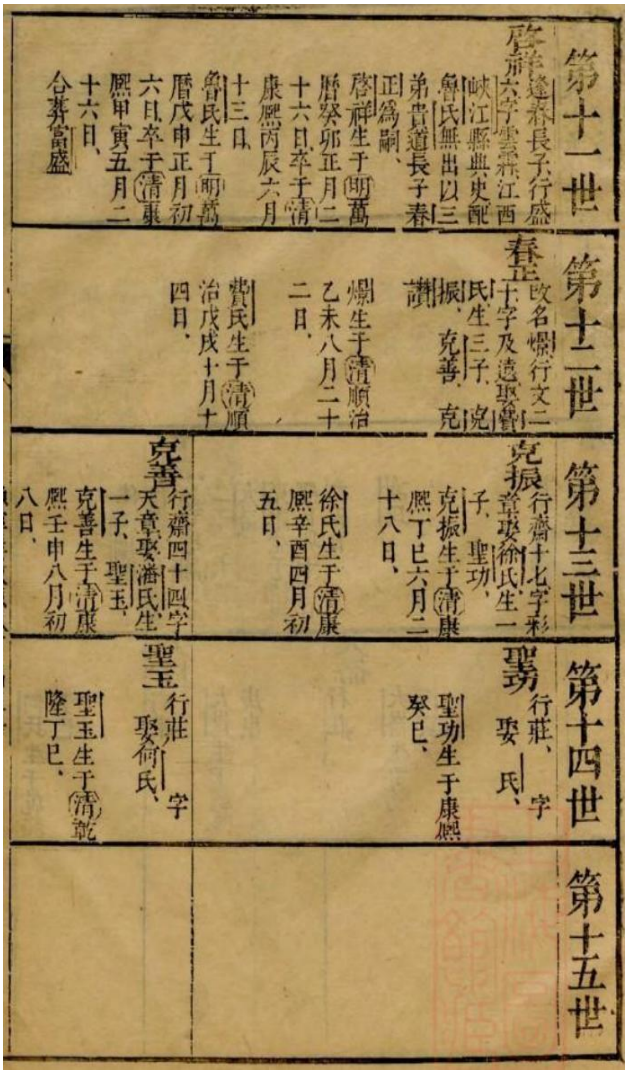
Figure 5: Robust text recognition results of HisDoc-OCR on a Chinese historical document with printed text.



瓶入柱及入能穴山林缽說法不又云汝身入心又云屋內有瓶屋
外亦有瓶不瓶中有水不水中有瓶不有至天下諸水一合皆
有瓶不又云此水是何物又樹葉缽說法瓶缽說法柱缽說法
及地水火風皆缽說法土凡石尔缽說法者何也
堂巍朝三藏法師菩提達磨承求鄔跋陁羅之藏後其
達磨禪師懃闍大乘宗海吳越遊洛至敦西門道育
恵可奉事五年於海四行渭可白有標學従四卷仁者依
行自然度脫餘度如續高师傳所傳明經辩大乘入道
四行弟子曇林學法师者西域南天竺國是大婆羅山國
王第三之子如神恵踰朗聞皆曉晤志存廣訶衍道仏捨
素従緇紹隆聡空心虚寂通鑒世事内外俱明德超世表
悲悔違隅正教陵替遂缽述涉山海遊化漢魏正心寂黜
之志莫不歸信念相存見之流乃生識涛于時唯有道育恵
可此一四門年雖後生懼志高遠辜遊法师事之數載度
恭諮耻善蒙师意法师感其精誠誨以真道如是安心如是
發行如是順物如是方便此是大乘安心之法令無錯謬如
是安心者磨觀如是發行者四行如是順物者防護識嫌如
是方便者遣其不者此略所由意疰後攵
未入道多達要而言之不出二種一是理入二是行入理入者渭
藉教悟宗深信合生凡聖同一真性但為客塵妄覆不缽
顯了差也捨妄歸真凝住辟觀自他凡聖等一境住不移更不墮於
言教此即与真理真状無有分別寂然無名之理入行入者所
渭四行其餘諸行悉入此行中何等為四行一者報怨二者隨緣
行三者無所求行四補法行云何報怨行脩道行人若受苦時
當自念言我從徒昔無數劫中弃拳逐末流浪諸有多報

Input image OCR output

Figure 6: Robust text recognition results of HisDoc-OCR on a Chinese historical document with cursive handwritten text.



第十一世
#達春長子行盛
六字雲霖江西
峽江縣典史配
魯氏無出以三
弟貴道長子春
正爲嗣
啓祥生于朗萬
曆癸卯正月二
十六日卒于清
康熙丙辰六月
十三日
魯氏生于朗萬
曆戊申正月初
六日卒于清康
熙甲寅五月二
十六日
合葬富盛

第十二世
春（改名煩行文二 十字及遠取爵）
民生三子
振克善克
讚
燥生于清順治
乙未八月二十
二日
費氏生于清順
治戊戌十月十
四日

第十三世
克振（行齋七十四字彩 章婁徐氏
生一）
子聖功
克振生于清康
熙丁巳六月二
十八日
徐氏生于清康
熙辛酉四月初
五日
克善（行齋四十四字 天章襄潘氏
生）
一子聖玉
克善生于清康
熙壬申八月初
八日

第十四世
聖（行莊字 娶氏）
聖功生于康熙
癸巳

第十五世

Input image OCR output

Figure 7: Robust text recognition results of HisDoc-OCR on a Chinese historical document with complex layout.

Figure 8: Robust text recognition results of HisDoc-OCR on a Chinese historical document with dense text.