Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 3, No. 2, 2026

Multidimensional Information Extraction from Digitized Chinese Historical Documents via Relation-Aware Semantic Enhancement

Lei Zhang [1, *], Peng Ren [2], Rui Guo [1], Jinjiang Zhao [1, 2]

[1] Institute of Automation, Gansu Academy of Sciences, Lanzhou, 730000, Gansu, China

[2] College of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, Gansu, China

[*] **Corresponding author**: (Email: redeggs85@163.com)

Abstract: Complex layouts, rare characters, and irregular reading orders pose challenges to optical character recognition (OCR) and semantic information extraction from Chinese historical documents. To address these challenges, we propose Relation-Guided and Semantics-Enhanced TongGuOCR (RSE-TongGuOCR). A category-aware relation predictor integrates learned reading dependencies into geometry-based block construction, while dynamic programming selects recognition blocks to preserve both local context and character legibility. To promote consistency between layout analysis and text decoding, a cross-branch constraint aligns layout-based relation predictions with decoder-generated transition tokens in a shared spatial transition space. The framework retains rare-character vocabulary expansion and incorporates an ancient Chinese named entity recognition (NER) branch to identify six categories of historical entities in transcribed text. The OCR and NER components are trained in stages using distinct annotation sources and evaluated independently. RSE-TongGuOCR achieves an AR of 94.38 and an RO-ED of 2.86 on M5HisDoc, while its NER branch attains an F1 score of 85.84% on EvaHan2025 Dataset A. These task-specific results support the framework's effectiveness in document transcription, reading-order recovery, and structured historical information extraction.

Keywords: Chinese historical documents, optical character recognition, relation modeling, reading-order recovery, named entity recognition.

1. Introduction

Digitized Chinese historical documents require both reliable transcription and structured semantic annotation to support retrieval and archival indexing. Vertical writing, marginal notes, rare characters, and irregular reading paths make this task difficult. TongGuOCR [1] addresses historical OCR through layout-aware recognition blocks, rare-character vocabulary expansion, and line-transition tokens. Its block-level strategy preserves more context than isolated-line recognition while maintaining better character visibility than full-page processing.

Related document-parsing methods offer complementary perspectives. MonkeyOCR [2] decomposes parsing into structure detection, recognition, and relation prediction; MinerU2.5 [3] combines coarse layout analysis with native-resolution local recognition; and PaddleOCR-VL [4] integrates visual encoding and language generation in a compact model. Historical resources such as HisDoc1B [5] further expose the diversity of layouts and character categories that specialized recognition systems must address.

Two limitations motivate the present work. First, geometric penalties describe block size, overlap, aspect ratio, and local compatibility, but nearby text regions may belong to different reading sequences. Learned dependencies can therefore supplement geometric criteria when selecting recognition blocks. Second, layout processing and autoregressive decoding represent reading order differently: the former predicts relationships among regions, whereas the latter generates line-transition tokens. Without explicit agreement, these representations can provide inconsistent transition estimates.

Semantic extraction introduces an additional requirement. Ancient Chinese contains abbreviated expressions and ambiguous entity boundaries. Zhang et al. [6] combine GujiRoBERTa, BiLSTM, attention, and CRF; Lu and Lei [7] use domain-specific encoding with structured decoding; and Hua and Xu [8] constrain label transitions. EvaHan2025 [9] provides a dedicated ancient-Chinese NER evaluation. These approaches motivate a text-based semantic extension, but OCR-generated input must also preserve characters and reading sequence.

RSE-TongGuOCR connects relation-aware visual recognition with staged semantic extraction. Its contributions are organized around recognition granularity, reading-order agreement, and semantic utilization:

(1) Relation-informed recognition granularity. Category-aware dependencies are incorporated into the geometric segmentation objective, allowing dynamic programming to select blocks using both explicit layout constraints and learned reading relationships.

(2) Reading-order agreement across representations. Layout-level relation predictions and generation-level transition tokens are aligned in a shared spatial space, making consistency between the two branches an explicit training objective.

(3) Staged transcription-to-entity extraction. An established ancient-Chinese sequence-labeling architecture is connected to OCR output to identify six historical entity categories, using separate image and text annotations rather than requiring aligned image-level entity labels.

2. Proposed Method

2.1. Processing Pipeline

Given a historical document image, a structure–relation module detects text regions and estimates reading dependencies. A layout-aware optimizer then constructs recognition blocks by combining geometry and learned relations. The TongGuOCR backbone transcribes the resulting crops using rare-character and line-transition tokens. Finally, an ancient-Chinese semantic extractor assigns entity labels to the recognized text. The components are linked through the transcription, while OCR and NER retain separate training stages.

2.2. Relation-Guided Recognition-Block Construction

Following MonkeyOCR [2], a DETR-based detector predicts the bounding box and category of each document region. Its visual backbone and Transformer encoder extract page features; a lightweight foreground-scoring branch selects informative object queries; and the detector decoder predicts region positions and categories. These outputs supply the spatial information used by the relation branch rather than directly replacing the OCR transcriptions. For region i, the coordinate embedding encodes the two box corners, width, and height; a category embedding supplies region-type information. Their fusion is written as

$$P_i = \mathrm{Concat}(\mathrm{E}(x_i^1), \mathrm{E}(y_i^1), \mathrm{E}(x_i^2), \mathrm{E}(y_i^2), \mathrm{E}(w_i), \mathrm{E}(h_i)) \\ F_i = P_i + E_{\mathrm{c}}(c_i) \quad (1)$$

Here, E and E_c denote position and category embeddings, while c_i is the predicted region category. Combining them makes reading-dependency estimation sensitive to both region geometry and region type. A relation Transformer processes the fused features and produces contextual region representations h_i. The probability that region j follows region i is

$$q_{ij} = \frac{\exp(u_r^T \tanh(W_h h_i + W_k h_j + b_r))}{\sum_{k \neq i} \exp(u_r^T \tanh(W_h h_i + W_k h_k + b_r))} \quad (2)$$

The projections and scoring parameters in Eq. (2) are learned by the relation branch. The resulting score q_{ij} supplies reading-dependency evidence beyond spatial proximity, allowing neighboring regions to be distinguished when they belong to different reading sequences. Let L = [l_1, l_2, …] be the initially ordered text-line sequence and $B_{i,j}$ = {l_i, …, l_{j-1}} a candidate block of consecutive lines. For a partition S = (B_1, …, B_K), the geometric objective is extended by learned block-dependency scores:

$$P'(S) = \lambda_{\mathrm{blk}} \sum_{t=1}^{K} p_{\mathrm{blk}}(B_t) + \lambda_{\mathrm{adj}} \sum_{t=1}^{K-1} p_{\mathrm{adj}}(B_t, B_{t+1}) \\ + \lambda_{\mathrm{page}} P_{\mathrm{page}}(K) - \lambda_{\mathrm{rel}} \sum_{t=1}^{K-1} \log q(B_t, B_{t+1}) \quad (3)$$

The first three terms retain the block penalty p_{blk}, adjacent-block penalty p_{adj}, and page-level penalty P_{page} of TongGuOCR [1]. Together, they constrain block compactness, overlap, local layout compatibility, and segmentation complexity. The λ coefficients balance these criteria. The final term favors transitions supported by the learned dependency q(B_t, B_{t+1}), with λ_{rel} controlling its contribution. In this way, relational evidence supplements the geometric constraints while retaining their control over the number and shape of recognition blocks. Dynamic programming selects the sequence

$$S^* = \arg\min_{S \in \mathcal{S}(L)} P'(S) \quad (4)$$

Where $\mathcal{S}$(L) denotes candidate partitions of the ordered lines. Geometry and relations thus guide recognition granularity jointly rather than replacing one another. Boundary expansion, non-target masking, and square padding follow TongGuOCR [1] to obtain OCR-friendly crops while preserving local context and character legibility.

2.3. Token-Augmented Recognition and Reading-Order Agreement

Each selected crop is encoded visually and projected into the language-decoder representation space. TongGuOCR's vocabulary expansion is retained: historical characters that lack a single-token representation are assigned character-level tokens. Specifically, the added vocabulary contains single CJK characters absent from the original vocabulary whose original tokenization requires more than one token. Assigning these characters direct representations shortens decoding and strengthens the correspondence between a visual glyph and its transcription. This mechanism is inherited from TongGuOCR [1], whereas relation-guided block construction and cross-branch agreement provide the additional structural constraints.

For adjacent lines, the horizontal and vertical displacement from the end of the current line to the start of the next is calculated and quantized into a predefined spatial transition category. This represents line-to-line movement explicitly within the generated sequence, including the irregular transitions encountered in historical layouts. The corresponding token r_i is inserted between line transcriptions:

$$Y_{\mathrm{aug}} = \mathrm{T}(l_1) \oplus r_1 \oplus \mathrm{T}(l_2) \oplus \cdots \oplus r_{n-1} \oplus \mathrm{T}(l_n) \quad (5)$$

T(l_i) denotes the transcription of line i, and ⊕ denotes concatenation. The autoregressive OCR loss is the mean negative log-likelihood of the target sequence conditioned on the visual representation and preceding tokens. It jointly supervises ordinary characters, expanded rare-character tokens, and transition tokens rather than treating reading-order markers as a separate post-processing output. At inference, transition tokens are converted into line boundaries, and block outputs are merged in the predicted reading sequence.

To connect layout analysis and generation, relation predictions are mapped to the same spatial categories used by transition tokens. Let p_{rel} and p_{token} denote the corresponding distributions, with m = (p_{rel} + p_{token})/2. Their agreement is encouraged by

$$\mathcal{L}_{\mathrm{cons}} = \frac{1}{2} D_{\mathrm{KL}}(p_{\mathrm{rel}} \parallel m) + \frac{1}{2} D_{\mathrm{KL}}(p_{\mathrm{token}} \parallel m) \quad (6)$$

The constraint operates at the distribution level: the relation branch estimates transitions from layout features, while the decoder estimates them during transcription. Aligning these complementary predictions makes reading-order consistency explicit instead of relying on independently learned representations.

2.4. Semantic Extraction and Staged Training

The recognized text is processed by the GujiRoBERTa–BiLSTM–Attention–CRF architecture adopted from ancient-Chinese sequence labeling [6]. GujiRoBERTa supplies

domain-oriented contextual representations; BiLSTM models dependencies in both directions; attention emphasizes entity-related features; and CRF jointly decodes labels using emission scores and transition constraints. This combination addresses the abbreviated expressions, flexible grammatical structures, and ambiguous entity boundaries of ancient Chinese by using contextual features together with sequence-level label constraints. The output covers person, location, time, official title, dynasty, and book title. Here, multidimensional extraction refers to these six entity categories, not to additional visual modalities. The semantic output provides structured annotations alongside the searchable transcription.

The loss terms of the framework are summarized as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{OCR}} + \lambda_1 \mathcal{L}_{\text{REL}} + \lambda_2 \mathcal{L}_{\text{NER}} + \lambda_3 \mathcal{L}_{\text{cons}} \quad (7)$$

$\mathcal{L}_{\text{OCR}}$ is the autoregressive transcription loss, $\mathcal{L}_{\text{REL}}$ the relation-prediction loss, and $\mathcal{L}_{\text{NER}}$ the CRF-based entity-recognition loss. Because the image and text datasets provide different annotation types, Eq. (7) organizes the objectives across training stages rather than implying fully end-to-end optimization. The structure–relation module and OCR backbone are trained using document-image annotations; the semantic branch is subsequently trained on entity-annotated historical text. During application, OCR transcriptions provide the interface between the stages.

3. Experimental Results and Analysis

3.1. Experimental Setup

OCR was evaluated on M5HisDoc, which contains 4,000 images, 403,824 text lines, 4,367,392 characters, and 16,151 character categories. Following TongGuOCR [1], 2,000 images were used for training, 1,000 for validation, and 1,000 for testing. EvaHan2025 Dataset A [9] was used to evaluate the semantic branch across the six entity categories. AR measures character recognition, while lower NED and RO-ED indicate better transcription fidelity and reading-order recovery. NER was assessed using Precision, Recall, and F1.

The implementation used PyTorch, AdamW, and bfloat16 mixed precision. The OCR backbone was initialized from TongGuOCR with a learning rate of 1×10^{-5}, cosine decay, and a 0.1 warm-up ratio. The NER branch used a 256-dimensional BiLSTM, dropout of 0.4, maximum sequence length of 256, and learning rate of 2×10^{-5}. The relation and consistency loss coefficients were initially $\lambda_1 = 0.5$ and $\lambda_3 = 0.3$, respectively.

The OCR comparison includes HunyuanOCR [10], GLM-OCR [11], PaddleOCR-VL-1.6 [12], MinerU2.5-Pro [13], and TongGuOCR [1]. NER comparisons include the official SikuRoBERTa baseline [9], attention-enhanced GujiRoBERTa [6], Simple RoBERTa-NER [14], and Multi-Strategy NER [15]. Published results on the corresponding official test sets were used when available. OCR and NER are evaluated separately because their annotation sources are not aligned.

3.2. Main Results

Table 1. Main results on M5HisDoc and EvaHan2025 Dataset A.

(a) Historical OCR on M5HisDoc			
Method	AR ↑	NED ↓	RO-ED ↓
HunyuanOCR [10]	48.87	12.12	7.53
GLM-OCR [11]	46.73	48.76	34.06
PaddleOCR-VL-1.6 [12]	53.59	29.93	16.91
MinerU2.5-Pro [13]	38.48	45.75	35.39
TongGuOCR [1]	93.76	6.15	3.49
RSE-TongGuOCR (Ours)	94.38	5.61	2.86
(b) Ancient Chinese NER on EvaHan2025 Dataset A			
Method	Precision ↑	Recall ↑	F1 ↑
SikuRoBERTa-BiLSTM-CRF [9]	85.90	77.50	81.48
GujiRoBERTa-BiLSTM-Attention-CRF [6]	86.30	78.78	82.37
Simple RoBERTa-NER [14]	90.77	76.75	83.17
Multi-Strategy NER [15]	89.13	79.32	83.94
RSE-TongGuOCR (Ours)	90.80	81.40	85.84

Note: ↑ indicates higher is better; ↓ indicates lower is better. Bold marks the best value within each task. NER scores are percentages.

On M5HisDoc, RSE-TongGuOCR reaches 94.38 AR, 5.61 NED, and 2.86 RO-ED. Relative to TongGuOCR, AR increases by 0.62 percentage points, while NED and RO-ED decrease by 8.78% and 18.05%, respectively. The larger relative reduction in RO-ED is consistent with the method's emphasis on reading dependencies and cross-branch agreement. The proposed method also obtains better values than the other listed OCR systems on all three metrics.

On EvaHan2025 Dataset A, the semantic branch achieves 90.80% Precision, 81.40% Recall, and 85.84% F1. F1 exceeds Multi-Strategy NER by 1.90 percentage points and the listed attention-enhanced GujiRoBERTa model by 3.47 percentage points. Relative to Multi-Strategy NER, Recall increases from 79.32% to 81.40%, while Precision rises from 89.13% to 90.80%. The reported gain therefore involves both the proportion of recovered entities and the precision of their identification. These scores describe entity recognition on the NER benchmark; they do not measure entity extraction from the M5HisDoc OCR outputs or establish an end-to-end gain from improved transcription.

3.3. Component Analysis and Limitations

The component results on M5HisDoc further distinguish the two OCR additions. Adding relation modeling to TongGuOCR increases AR from 93.76 to 94.05 and reduces

NED/RO-ED from 6.15/3.49 to 5.89/3.13. Adding cross-branch consistency then yields 94.38 AR, 5.61 NED, and 2.86 RO-ED. The incremental changes support using learned relations for block construction and aligning them with decoder transitions.

The main limitation is the separation of image-level OCR annotations from text-level entity annotations, which prevents the present evaluation from establishing fully end-to-end image-to-entity performance. Unseen rare characters and highly irregular layouts remain potential sources of transcription errors. Aligned historical image–entity annotations are needed to evaluate their downstream semantic effects directly.

4. Conclusions

RSE-TongGuOCR combines relation-guided block construction, cross-branch reading-order agreement, and staged ancient-Chinese semantic extraction. Learned dependencies complement geometric constraints, while consistency aligns layout and generation representations. The reported M5HisDoc results support improved transcription and reading-order recovery, and the separate EvaHan2025 evaluation supports six-category entity extraction. The framework connects digitization with structured historical-text annotation without assuming fully aligned image–entity supervision. Future work will address image-level entity annotations, unseen rare characters, and extensions toward historical relation extraction and knowledge organization.

Acknowledgements

This research was funded by the Key R&D Program of Gansu Province under Grant 24YFFA029, in part by the Key R&D Project of Gansu Provincial Academy of Sciences under Grant 2025PTXM-03, in part by the Youth Fund Project of Gansu Academy of Sciences under Grant 2025QN-01, and in part by the Achievement Transformation and Technical Application Service Project of Gansu Academy of Sciences 2026CY-02.

Data availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflict of interest: There is no conflict of interest.

References

[1] Zhou, Z., Sun, Y., He, H., et al. (2026). TongGuOCR: A layout-aware and token-augmented OCR MLLM for Chinese historical documents. arXiv preprint arXiv:2608.07917v2.

[2] Li, Z., Liu, Y., Liu, Q., et al. (2026). MonkeyOCR: Document parsing with a structure-recognition-relation triplet paradigm. arXiv preprint arXiv:2506.05218v2.

[3] Niu, J., Liu, Z., Gu, Z., et al. (2025). MinerU2.5: A decoupled vision-language model for efficient high-resolution document parsing. arXiv preprint arXiv:2509.22186.

[4] Cui, C., Sun, T., Liang, S., et al. (2025). PaddleOCR-VL: Boosting multilingual document parsing via a 0.9B ultra-compact vision-language model. arXiv preprint arXiv:2510.14528.

[5] Shi, Y., Peng, D., Zhang, Y., Cao, J., & Jin, L. (2025). A large-scale dataset for Chinese historical document recognition and analysis. Scientific Data, 12, Article 169. https://doi.org/10.1038/s41597-025-04897-8

[6] Zhang, Q., Duan, Z., Ma, S., Liu, S., Yuan, Z., & Ma, R. (2025). Multi-domain ancient Chinese named entity recognition based on attention-enhanced pre-trained language model. In Proceedings of the Second Workshop on Ancient Language Processing (pp. 237–241).

[7] Lu, Y., & Lei, M. (2025). Construction of NER model in ancient Chinese: Solution of EvaHan 2025 challenge. In Proceedings of the Second Workshop on Ancient Language Processing (pp. 165–169).

[8] Hua, W., & Xu, S. (2025). When less is more: Logits-constrained framework with RoBERTa for ancient Chinese NER. In Proceedings of the Second Workshop on Ancient Language Processing (pp. 192–196).

[9] Li, B., Chang, B., Liu, R., et al. (2025). Overview of EvaHan2025: The first international evaluation on ancient Chinese named entity recognition. In Proceedings of the Second Workshop on Ancient Language Processing (pp. 156–164).

[10] Hunyuan Vision Team, Lyu, P., Wan, X., et al. (2025). HunyuanOCR Technical Report. arXiv preprint arXiv:2511.19575.

[11] Duan, S., Xue, Y., Wang, W., et al. (2026). GLM-OCR Technical Report. arXiv preprint arXiv:2603.10910.

[12] Zhang, Z., Liu, H., Liang, S., et al. (2026). PaddleOCR-VL-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. arXiv preprint arXiv:2606.03264.

[13] Wang, B., He, T., Ouyang, L., et al. (2026). MinerU2.5-Pro: Pushing the Limits of Data-Centric Document Parsing at Scale. arXiv preprint arXiv:2604.04771.

[14] Zhang, Y., Liu, M., Tang, H., Lu, S., & Xue, L. (2025). Simple Named Entity Recognition (NER) System with RoBERTa for Ancient Chinese. In Proceedings of the Second Workshop on Ancient Language Processing (pp. 206–212). https://doi.org/10.18653/v1/2025.alp-1.27

[15] Dong, W., & Liu, M. (2025). Multi-strategy named entity recognition system for ancient Chinese. In Proceedings of the Second Workshop on Ancient Language Processing (pp. 213–220).