Benchmarking Vision-Language Models on Chinese Ancient Documents: From OCR to Knowledge Reasoning

Haiyang Yu[1,2]*, Yuchuan Wu[1]*, Fan Shi[1]*, Jinghui Lu[2]*,
Ke Niu[1], Xiaodong Ge[1], Minghan Zhuo[1], Jingqun Tang [2], Bin Li[1]†

[1] Fudan University, [2] Bytedance Inc.

* Equal Contribution, †**Corresponding Author:** libin@fudan.edu.cn

Abstract

Chinese ancient documents, invaluable carriers of millennia of Chinese history and culture, hold rich knowledge across diverse fields but face challenges in digitization and understanding—traditional methods only scan images, while current Vision-Language Models (VLMs) struggle with their visual/linguistic complexity. Existing document benchmarks focus on English printed texts or simplified Chinese, leaving a gap for evaluating VLMs on ancient Chinese documents. To address this, we present AncientDoc, the first benchmark for Chinese ancient documents, designed to assess VLMs from OCR to knowledge reasoning. AncientDoc includes five tasks (page-level OCR, vernacular translation, reasoning-based QA, knowledge-based QA, linguistic variant QA) and covers 14 document types, over 100 books, and about 3,000 pages. Based on AncientDoc, we evaluate mainstream VLMs using multiple metrics, supplemented by a human-aligned large language model for scoring. The benchmark is available at https://bytedance.github.io/AncientDoc/.

1 Introduction

Chinese ancient documents, which carry thousands of years of Chinese history and culture, are treasure troves of knowledge spanning history, philosophy, medicine, astronomy, etc. They are invaluable cultural heritage for both China and the world. With the wave of digitization of Chinese ancient documents in libraries and museums, many captured Chinese ancient document images are produced. However, traditional digitization methods only stay at the level of image scanning, while many downstream applications (knowledge mining, historical exploration) urgently need the ability to deeply understand the content of ancient documents. At the same time, parsing and understanding Chinese ancient document images pose huge challenges, including visual complexity, linguistic complexity, and poor adaptability of current vision-language models (VLMs).

Existing general document understanding datasets and benchmarks (*e.g.*, DocVQA (Mathew et al., 2021)) are mainly based on printed documents and are predominantly in English. Even Chinese-related datasets (*e.g.*, CN-DocVQA) only involve simplified Chinese characters, which is totally different from Chinese ancient documents. In addition, with the development of large vision-language models (VLMs), an increasing number of VLMs have acquired the capabilities of document OCR and understanding. However, for VLMs, there is currently no benchmark that can systematically evaluate their OCR and understanding capabilities on Chinese ancient documents.

To address these issues, we construct the first Chinese ancient document benchmark called *AncientDoc*, which is used to comprehensively evaluate the capabilities of VLMs ranging from OCR to knowledge reasoning. AncientDoc includes five tasks: page-level OCR, vernacular translation, reasoning-based QA (question answering), knowledge-based QA, and linguistic variant QA. AncientDoc covers 14 types of ancient documents (such as collected works and Chuci-style poetry), approximately 100 books, and a total of 3000 document pages. In addition, we have adopted several evaluation metrics to evaluate most mainstream VLMs on the five tasks. To supplement these metrics, we additionally utilize a large language model to score the predictions of VLMs.

Our main contributions are as follows:

- We propose the first benchmark (AncientDoc) for Chinese ancient documents, aiming to comprehensively evaluate existing vision-language models from OCR to knowledge reasoning.
- AncientDoc contains five tasks for evaluating

Findings of the Association for Computational Linguistics: ACL 2026, pages 28793–28812

Task	DocVQA	TKH	MTH	OCRBench	OCRBench v2	AncientDoc
Page-level OCR	✗	✓	✓	✓	✓	✓
Vernacular Translation	✗	✗	✗	✗	✗	✓
Reasoning-based QA	✓	✗	✗	✓	✓	✓
Knowledge-based QA	✗	✗	✗	✗	✗	✓
Linguistic Variant QA	✗	✗	✗	✗	✗	✓

Table 1: Comparison of task type between different benchmarks.

VLMs: page-level OCR, vernacular translation, reasoning-based QA, knowledge-based QA, and linguistic variant QA. It covers 14 types of ancient documents, with 3,000 page images extracted from over 100 ancient books.

- We have conducted a comprehensive evaluation of existing mainstream vision-language models with various metrics. In addition, we also adopt a large language model that is the most consistent with human scoring to evaluate them.

2 Related Work

In recent years, document understanding tasks (Borchmann et al., 2021; Ma et al., 2024; Tanaka et al., 2024) have continuously expanded from traditional Optical Character Recognition (OCR) (Kang et al., 2022; Yin et al., 2017; Ingle et al., 2019) to higher-level semantic understanding tasks (Ding et al., 2023; Zhang et al., 2024), such as question answering (Mishra et al., 2019; Ding et al., 2024; Kang et al., 2024), translation (Zhang et al., 2018; Wang et al., 2023), and structured information extraction (Jaume et al., 2019; Huang et al., 2022). With the increasing complexity of tasks, multimodal document datasets have become increasingly abundant, providing an important basis for evaluating the capabilities of different models. However, most existing datasets still suffer from limitations in task dimensions, making it difficult to fully cover the multi-level cognitive and language transfer tasks involved in complex Chinese ancient documents.

Among existing datasets, DocVQA (Mathew et al., 2021) is an early representative multimodal dataset focusing on visual document question answering tasks, emphasizing linguistic reasoning on structured and unstructured text in document images. The question format of DocVQA usually relies on OCR results rather than the linguistic content, and its core task is text logical reasoning, which is suitable for evaluating the text understanding and document reading abilities of models. However, this dataset does not include page-level OCR task, so it cannot examine the recognition robustness of models when faced with complex visual inputs. Therefore, it is still insufficient in evaluating the comprehensive capabilities of models.

Additionally, TKH (Yang et al., 2018) and MTH (Yang et al., 2018) focus more on character-level recognition of Chinese historical documents, mainly used to evaluate the OCR capabilities of models on ancient books with low quality and numerous variant characters. Both are derived from real historical materials and have strong characteristics of ancient book images, such as vertical typesetting and cursive script, making them suitable for basic OCR evaluation datasets. However, neither of them involves tasks at the level of language understanding or language generation, so they cannot be used to evaluate the model's ability to understand the semantics, grammatical structure, or background knowledge of ancient Chinese. They also lack vernacular output or question-answer interaction forms, which limits their value in language transfer ability analysis.

To alleviate the above problem of single tasks, OCRBench (Liu et al., 2024b) and its enhanced version OCRBench v2 (Fu et al., 2024) have introduced more diverse settings, covering multiple subtasks, including character recognition, text localization, and some document-level question answering and logical reasoning problems. Among them, OCRBench v2 has improved in the task complexity compared to the first version, making it suitable for evaluating the comprehensive performance of multimodal models in real scenarios. However, these two datasets mainly focus on modern Chinese or English documents. At the same time, they do not cover tasks such as translation, knowledge question answering, or language style transfer, so they are difficult to use for evaluating the generalization performance of models in cross-lingual and

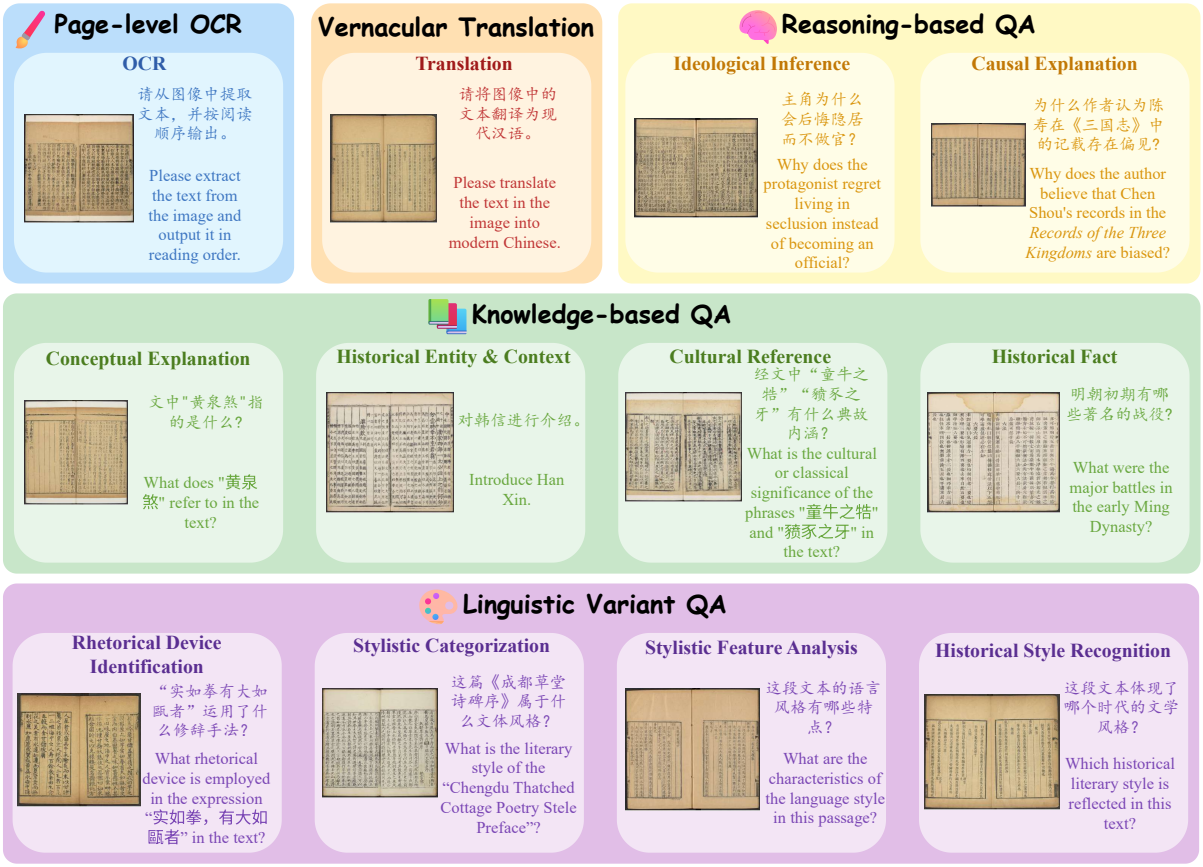


Figure 1: Some examples of each task in AncientDoc.

cross-style understanding abilities.

To this end, we have constructed a multi-task evaluation benchmark dataset *AncientDoc* for Chinese ancient documents. This dataset not only includes traditional OCR recognition tasks but also systematically integrates multiple high-level language understanding tasks. The comparison of task diversity is shown in Tab. 1.

3 Benchmark Construction

3.1 Task Definition

In this paper, we define five tasks for evaluating VLMs on Chinese ancient documents: page-level OCR, vernacular translation, reasoning-based QA (question answering), knowledge-based QA, and linguistic variant QA. Some examples of these tasks are shown in Fig. 1.

Page-level OCR: This task aims to directly extract complete and correctly ordered text content from an entire page of ancient documents, without relying on the character detection (Tian et al., 2016; Zhou et al., 2017), segmentation (He et al., 2016; Xu and Xiang, 2025), and recognition (Aksoushideh et al., 2025; Cui et al., 2025) processes in traditional OCR systems. This task presents the following challenges: 1) *Vertical texts*: Most Chinese ancient documents are vertically typeset from right to left, so the model needs to understand the correct reading order and the rules for line breaks within columns. 2) *Various annotations*: Chinese ancient documents sometimes contain interlinear notes, comments, small characters, postscripts, etc., requiring the model to have visual and semantic filtering capabilities. 3) *Traditional Chinese characters*: There are a large number of traditional Chinese characters or obsolete glyphs in Chinese ancient documents. Nevertheless, training samples of these characters are scarce.

Vernacular Translation: It aims to translate the Chinese texts in ancient documents into modern common vernacular expressions, enabling non-professional readers to understand the meaning of original texts and providing a clearer linguistic foundation for downstream tasks (such as question answering and summarization.). Unlike translation between languages, this task is intralingual translation. The difficulties of this task are: 1) *Polysemy*: Some ancient Chinese words often have

multiple meanings, and the model needs to understand the context or even the semantics of the entire paragraph to select the correct interpretation. 2) *Semantic punctuation*: There is a lack of punctuation in the text of Chinese ancient documents. Therefore, the model needs to insert punctuation based on semantic understanding.

Reasoning-based QA: Based on an image of a page from ancient documents, reasoning-based QA aims at extracting implicit information to answer questions that are not directly stated. Different from extractive question answering, reasoning-based QA requires the model to have the ability to understand and derive deep-level information such as facts, causality, and semantic relationships. The reasoning-based QA task is one of the advanced types of OCR-free document understanding. It integrates the understanding of graphic information, the interpretation of ancient Chinese language, and the connection and deduction of knowledge and logic, serving as an important scenario to test the deep language understanding and multi-step reasoning abilities of large models.

Knowledge-based QA: This task requires the model to answer questions related to objective knowledge in ancient documents, including time, place names, objects, medical terms, etc. Although this task belongs to factual QA, it still differs from encyclopedia QA. This task requires the model to understand the knowledge expression methods in ancient languages and have a certain reserve of historical and cultural background knowledge. Knowledge-based QA not only tests the model's ability to handle the explicit expression of knowledge but also deeply challenges its ability to summarize and infer knowledge under vague descriptions, putting forward new requirements for OCR-free document understanding and the knowledge transfer ability of VLMs in historical corpora.

Linguistic Variant QA: It aims to evaluate the model's ability to understand and reason about variant phenomena in ancient Chinese, such as language styles, rhetorical methods, and stylistic features. This task requires the model to generate or answer relevant questions around the aforementioned linguistic features. The linguistic variant question answering task is a key task connecting the understanding of linguistic artistry and generative language expression, representing a high-level language ability that moves from text recognition to the mastery of stylistic features. This task not only assesses whether the model "understands ancient Chinese" but also evaluates whether it understands how ancient Chinese is written, what its style is, and how rhetoric affects semantics.

3.2 Data Curation

3.2.1 Data Resources

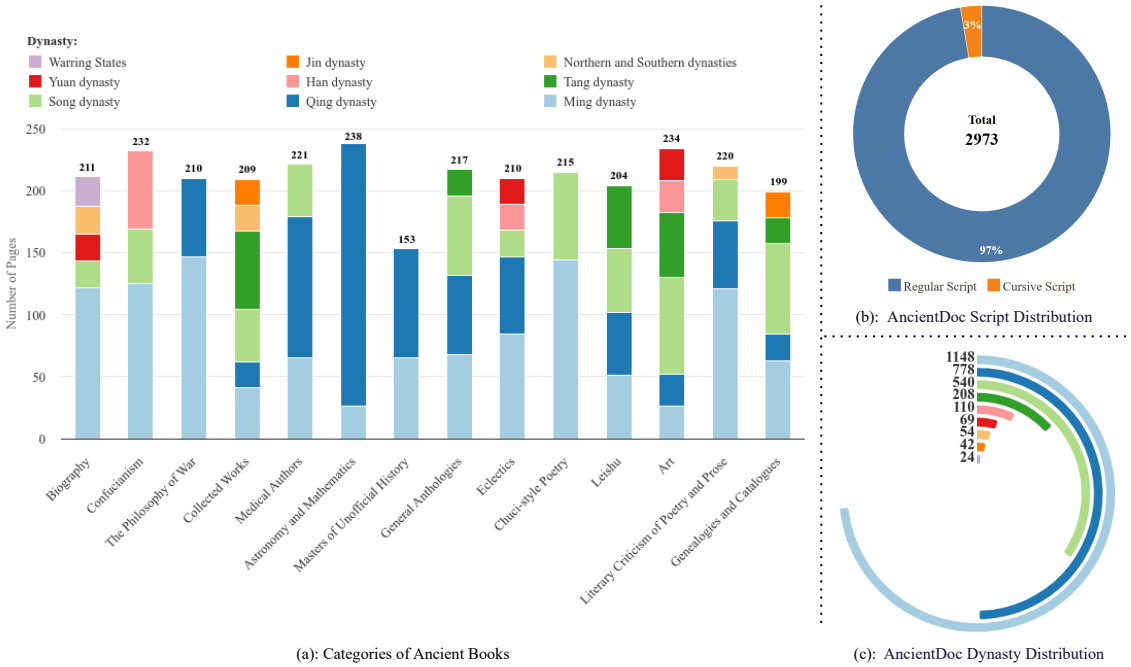
The original images of the proposed AncientDoc are mainly derived from the digitized resources of Chinese ancient documents held by the Harvard Library[1]. The digitized collection of ancient books in this library covers multiple dynasties and fields, featuring high scanning quality and rich typesetting styles, which provides a solid foundation for constructing a multi-task dataset of Chinese ancient documents. In combination with the requirements of semantic-related tasks, all collected ancient documents have undergone manual verification and classification. Finally, we divide the ancient documents into 14 semantic categories, including "collected works","Chuci-style Poetry","Literary Criticism of Poetry and Prose", "Eclectics", etc. All categories and corresponding explanations will be detailed in the supplementary materials. At the same time, we acknowledge that using a single institutional source may introduce collection-specific bias, including differences in preservation condition, edition preference, and document style distribution. Therefore, AncientDoc should be viewed as a realistic first-step benchmark built from currently accessible high-quality resources, rather than a fully source-balanced representation of all Chinese ancient documents.

3.2.2 Data Collection

In the process of data collection, we select representative Chinese printed ancient documents dating from the Qing Dynasty and earlier (The distribution of page counts across different dynasties is shown in Fig. 2(c)). To ensure the usability and challenge of the collected documents in terms of visual quality, linguistic content, and task adaptability, we establish the following priority criteria for collection:

1) *Vertical typesetting with traditional Chinese characters*: The selected pages conform to the typographic style of traditional ancient documents and feature structurally challenging information. Thus, they are suitable for evaluating the ability of OCR-free models to understand reading directions and inter-column structures.

[1]https://hollis.harvard.edu/



(a): Categories of Ancient Books

(b): AncientDoc Script Distribution

(c): AncientDoc Dynasty Distribution

Figure 2: (a) The page count distribution of different categories of ancient books. (b) The proportion of cursive script and regular script in AncientDoc. (c) The distribution of page counts across different dynasties in AncientDoc.

2) *Clear fonts with partial degradation or damage*: They should cover real-world scenarios such as ink blurring, simulating common degradation in the digitization of ancient documents. These selected samples can evaluate the robustness of models against low-quality inputs.

3) *Content with potential for linguistic, structural, knowledge-based, and reasoning tasks*: Poems, annotations, encyclopedias, historical biographies, and medical theories are widely chosen since they have high semantic density and are suitable for setting multi-level understanding tasks (*e.g.*, vernacular translation, reasoning-based QA, knowledge-based QA);

4) *High readability of images for convenient annotation*: We select versions with clear scans and complete page numbers to ensure effective OCR annotations, translation alignment, etc.

Ultimately, we collect approximately 100 ancient books for AncientDoc, covering various themes and styles and containing a total of about 3,000 pages of image data. This distribution is not an intentional balancing choice, but a reflection of the current availability of digitized ancient documents that satisfy both image quality and annotation feasibility requirements.

3.2.3 Data Annotation

To construct a high-quality multi-task dataset for Chinese ancient documents, we adopt an annotation pipeline that combines large language model (LLM)-assisted pre-annotation with manual verification. Specifically, we use the Qwen2.5-VL-72B model (Bai et al., 2025b) to generate task-aligned initial annotations for each document page, including page-level OCR, vernacular translation, and multiple types of question–answer pairs. For each page, we construct one QA pair for page-level OCR and vernacular translation, and two QA pairs for each of the remaining three tasks.

All model-generated annotations are subsequently reviewed and refined by human annotators. Manual revisions focus on correcting OCR reading order and layout structure, improving translation accuracy and fluency, and ensuring semantic consistency across different tasks, rather than large-scale content rewriting. A small fraction of ambiguous or insufficiently grounded samples are conservatively filtered out during this process.

Instead of relying on numerical inter-annotator agreement metrics, which are less suitable for open-ended generation and high-level semantic tasks, annotation consistency is ensured through unified task definitions, detailed annotation guidelines, and iterative cross-checking. For edge cases such as

	Model	CER	Char Precision	Char Recall	Char F1	GPT-4o
Open-source MLLMs	**2B-4B MLLMs**					
	InternVL2.5-2B (2024)	120.72	1.76	1.69	1.72	2.03
	InternVL3-2B (2025)	95.59	2.46	1.79	2.07	2.52
	Qwen2.5-VL-3B (2025b)	50.36	9.44	9.02	9.23	5.56
	Qwen2.5-Omni-3B (2025)	63.42	5.4	4.78	5.07	4.56
	InternVL2.5-4B (2024)	79.75	5.27	4.05	4.58	3.65
	Qwen3-VL-4B (2025a)	66.05	8.62	10.22	9.35	5.71
	7B-8B MLLMs					
	Qwen2.5-VL-7B (2025b)	35.47	12.95	12.75	12.85	6.37
	Qwen2.5-Omni-7B (2025)	70.46	3.44	3.35	3.39	4.30
	LLaVA-1.5-7B (2024a)	129.71	0.04	0.03	0	0.05
	InternVL2.5-8B (2024)	96.88	2.87	2.59	2.72	3.00
	InternVL3-8B (2025)	51.8	7.3	6.67	6.97	5.06
	Qwen3-VL-8B (2025a)	**35.4**	**19.78**	**20.63**	**20.2**	**7.93**

Table 2: Evaluation on page-level OCR.

degraded pages or historically ambiguous texts, we follow an *evaluability-first* principle and exclude samples that cannot support clear and objective evaluation.

All annotations are produced by annotators with experience in reading classical Chinese. The goal of AncientDoc is not philological adjudication, but to provide a well-defined multi-task benchmark for systematically evaluating vision–language models on Chinese ancient document understanding.

3.3 Statistical Analysis

We have collected a total of 2,973 pages of Chinese ancient documents and conducted a systematic analysis from multiple dimensions. Firstly, in terms of chronological distribution, the dataset covers major dynasties from the Warring States, Qin, and Han dynasties to the Ming and Qing dynasties, showing a broad historical span (as shown in Fig. 2(c)). Among them, documents from the Ming Dynasty (1,148 pages) and the Qing Dynasty (778 pages) are the most abundant, together accounting for approximately 65% of the total pages. This reflects the relatively complete preservation and wide circulation of documents from these periods. Next are documents from the Song Dynasty (540 pages) and the Tang Dynasty (208 pages), while materials from the Han, Yuan, and Northern and Southern Dynasties are relatively scarce, largely due to historical loss and limited large-scale digitization.

In terms of genre distribution, the dataset covers 14 categories of mainstream traditional Chinese literature, including biographies, Confucianism, philosophy of war, collected works, medical authors, astronomy and mathematics, masters of unofficial history, general anthologies, eclectics, *Chuci*-style poetry, *Leishu*, art, literary criticism of poetry and prose, and genealogies and catalogues. Statistical results (see Fig. 2(a)) indicate that the three largest categories by page count are astronomy and mathematics (238 pages), Confucianism (232 pages), and art (234 pages), demonstrating a diverse coverage of disciplinary domains that supports multi-task evaluation such as OCR, translation, and question answering.

In terms of font style, the majority of pages in AncientDoc are written in regular script, with approximately 97% in regular script and 3% in cursive or semi-cursive forms (see Fig. 2(b)). This imbalance is not an arbitrary design choice, but rather reflects the historical transmission and preservation of Chinese ancient documents. Most large-scale digitized resources available today originate from recompiled or reprinted editions from the Ming and Qing dynasties, during which regular script gradually became the dominant and standardized writing form for both woodblock printing and manuscript copying. Earlier script forms or highly cursive variants are comparatively scarce, often fragmented across different collections, and difficult to digitize at scale under consistent quality standards. As a result, the observed script-style distribution largely mirrors the current state of accessible historical archives, rather than a bias introduced during dataset construction.

Overall, AncientDoc exhibits broad coverage across historical periods and literary genres, and reflects realistic visual and linguistic characteristics of extant Chinese ancient documents. This makes

it a practical and representative benchmark for evaluating vision–language models on a wide range of ancient document understanding tasks.

3.4 Evaluation Metric

To assess the accuracy of page-level OCR, we use the following four metrics: CER (Character Error Rate), Char Precision, Char Recall and Char F1. The detailed description of them are shown in the supplementary material. Differently, when evaluating the remaining four tasks (*i.e.*, vernacular translation, reasoning-based QA, knowledge-based QA and linguistic variant QA), we use CHRF++ (Popović, 2017) and BERTScore (BS-F1) as the key metrics (Zhang et al., 2019).

Considering that hard automatic metrics may not always align with human judgment, we additionally adopt a large language model as an evaluator to score model outputs on a 0–10 scale. Based on comparison with human ratings, GPT-4o exhibits the highest consistency with human judgment and is therefore adopted as the evaluator. Details are provided in the supplementary materials.

4 Results

4.1 Model Selection

For evaluation, we select models including GPT-4o (Hurst et al., 2024), Qwen series (Bai et al., 2025b, 2023; Xu et al., 2025), InternVL series (Chen et al., 2024; Zhu et al., 2025), as well as Doubao (Team, 2025), Gemini (Comanici et al., 2025), LLaVA (Liu et al., 2024a), DeepSeek (Wu et al., 2024), etc. Considering the limited pages, we only show part of results in the main text. More evaluation results are shown in the supplementary material.

4.2 Main Results

Results in Page-level OCR. Table 2 reports the performance of representative open-source MLLMs in the 2B–8B parameter range on page-level OCR. Overall, 7B–8B models consistently outperform 2B–4B models across character-level metrics. In particular, Qwen3-VL-8B achieves the best overall performance, obtaining the highest Char Precision (19.78), Char Recall (20.63), and Char F1 score (20.2), together with the highest GPT-4o score among the evaluated models. Compared with smaller variants, Qwen3-VL-8B exhibits substantially improved glyph discrimination and more stable transcription across long character sequences.

	Model	GPT-4o	BS-F1
Open-source MLLMs	**2B-4B MLLMs**		
	InternVL2.5-2B	0.03	53.92
	InternVL3-2B	0.10	51.83
	Qwen2.5-VL-3B	0.66	55.2
	Qwen2.5-Omni-3B	0.56	56.8
	InternVL2.5-4B	0.53	58.46
	Qwen3-VL-4B	3.01	70.75
	7B-8B MLLMs		
	Qwen2.5-VL-7B	2.30	65.59
	Qwen2.5-Omni-7B	0.46	53.99
	LLaVA-1.5-7B	0.01	50
	InternVL2.5-8B	0.58	59.24
	InternVL3-8B	0.29	53.16
	Qwen3-VL-8B	**3.52**	**71.66**

Table 3: Evaluation on vernacular translation.

	Model	GPT-4o	BS-F1
Open-source MLLMs	**2B-4B MLLMs**		
	InternVL2.5-2B	3.55	65.29
	InternVL3-2B	4.39	66.45
	Qwen2.5-VL-3B	4.67	65.83
	Qwen2.5-Omni-3B	4.90	66.96
	InternVL2.5-4B	4.75	66.12
	Qwen3-VL-4B	7.84	73.76
	7B-8B MLLMs		
	Qwen2.5-VL-7B	6.44	69.96
	Qwen2.5-Omni-7B	6.05	68.70
	LLaVA-1.5-7B	1.64	61.69
	InternVL2.5-8B	5.93	68.40
	InternVL3-8B	4.98	65.62
	Qwen3-VL-8B	**7.96**	**74.06**

Table 4: Evaluation on reasoning-based QA.

These results suggest that page-level OCR on ancient documents benefits from sufficient visual representation capacity and balanced multimodal modeling, rather than from aggressive language modeling alone. Detailed analyses on larger-scale open-source models and closed-source models are provided in the supplementary material.

Results in Vernacular Translation. Table 3 reveals a pronounced performance gap among open-source MLLMs below 8B parameters. Notably, the Qwen3-VL series exhibits a clear advantage over other models at comparable scales, with Qwen3-VL-4B and Qwen3-VL-8B achieving substantially higher BS-F1 scores than all other 2B–8B counterparts.

This sharp improvement is likely related to the stronger language-centric pretraining and enhanced vision–language alignment in Qwen3-VL, which

	Model	GPT-4o	BS-F1
	2B-4B MLLMs		
Open-source MLLMs	InternVL2.5-2B	2.48	64.24
	InternVL3-2B	3.52	66.21
	Qwen2.5-VL-3B	3.77	62.86
	Qwen2.5-Omni-3B	4.11	66.10
	InternVL2.5-4B	3.92	64.81
	Qwen3-VL-4B	<u>7.68</u>	<u>73.63</u>
	7B-8B MLLMs		
	Qwen2.5-VL-7B	5.23	66.75
	Qwen2.5-Omni-7B	4.94	66.85
	LLaVA-1.5-7B	0.78	60.37
	InternVL2.5-8B	4.88	67.68
	InternVL3-8B	4.31	63.60
	Qwen3-VL-8B	**7.83**	**73.98**

Table 5: Evaluation on knowledge-based QA.

	Model	GPT-4o	BS-F1
	2B-4B MLLMs		
Open-source MLLMs	InternVL2.5-2B	2.25	**62.24**
	InternVL3-2B	2.83	58.95
	Qwen2.5-VL-3B	3.75	52.30
	Qwen2.5-Omni-3B	3.40	56.87
	InternVL2.5-4B	3.26	58.22
	Qwen3-VL-4B	3.36	56.02
	7B-8B MLLMs		
	Qwen2.5-VL-7B	**4.75**	57.48
	Qwen2.5-Omni-7B	<u>4.12</u>	58.62
	LLaVA-1.5-7B	0.87	56.56
	InternVL2.5-8B	3.63	<u>61.52</u>
	InternVL3-8B	3.53	55.96
	Qwen3-VL-8B	3.57	56.62

Table 6: Evaluation on linguistic variant QA.

are particularly beneficial for intralingual translation that requires accurate semantic grounding between ancient and modern Chinese.

Results in Reasoning-based QA. Table 4 shows a consistent performance advantage of the Qwen3-VL series on reasoning-based QA across different model scales.= In both the 2B–4B and 7B–8B ranges, Qwen3-VL models achieve the highest BS-F1 scores among all evaluated open-source MLLMs, indicating a stable superiority beyond parameter scale effects. This advantage is likely related to the stronger language-centric modeling and more robust long-context integration emphasized in Qwen3-VL, which are particularly beneficial for reasoning-based QA that requires multi-sentence understanding and stable semantic grounding over ancient document content.

Results in Knowledge-based QA. Table 5 illustrates the relationship between model scale and performance on knowledge-based QA. Overall, larger models in the 7B–8B range tend to achieve higher BS-F1 scores than 2B–4B models, indicating that increased model capacity generally benefits knowledge-intensive tasks.

However, performance does not scale monotonically with parameter size. Substantial performance gaps are observed among models of similar scales, suggesting that pretraining data composition and knowledge organization play a critical role beyond model size alone. Notably, Qwen3-VL-4B achieves performance comparable to 7B–8B models, highlighting that effective knowledge modeling can partially compensate for smaller scale.

Results in Linguistic Variant QA. Table 6 shows that performance on linguistic variant QA does not exhibit a clear correlation with model scale. Models in the 2B–4B range can achieve comparable or even better BS-F1 scores than their 7B–8B counterparts, indicating that larger parameter capacity does not necessarily translate into improved performance on this task.

Notably, the InternVL2.5 series consistently outperforms other models across different scales, suggesting that linguistic variant QA relies more on exposure to stylistically rich and form-diverse training data than on model size alone.

4.3 Comparison between BERTScore and GPT-4o Scoring

Through the results in Tab. 2- 6, BERTScore ratings largely align with those of GPT-4o scores, showing a positive correlation. This result demonstrates the rationality of selecting GPT-4o as a scoring tool for large models, which also validates our conclusion in Sec. 3.4. However, due to the existence of numerous possible outcomes in vernacular translation results, GPT-4o's scoring tends to be conservative. The top-performing model only achieves a score of 4.72 (out of 10) in GPT-4o's evaluation. Even more notably, when evaluating GPT-4o's own performance in vernacular translation, the score is merely 0.92, indicating that it does not exhibit biased scoring towards its own predictions.

4.4 Preliminary Diagnostic Analysis of Cross-task Bottlenecks

Beyond model ranking, the results in AncientDoc also reveal several diagnostic patterns about current VLM limitations on Chinese ancient docu-

ments. First, page-level OCR remains a major bottleneck. Even the strongest models only obtain relativel low character-level scores, indicating that vertically arranged text, annotation interference, degraded glyphs, and historical character forms are still difficult to transcribe faithfully. However, the gap between weak OCR performance and substantially stronger results on reasoning-based QA and knowledge-based QA suggests that exact character-level transcription is not the only prerequisite for downstream understanding. In many cases, models may still exploit partial lexical cues, local visual evidence, and semantic priors to answer questions without producing reliable full-page OCR. This pattern indicates that ancient-document understanding is constrained not only by recognition quality, but also by how models can ground higher-level reasoning on incomplete visual-textual evidence.

Second, different subtasks appear to expose different failure sources. The relatively weak performance on vernacular translation suggests that the challenge is not merely visual perception, but semantic transfer from classical Chinese to modern Chinese, which requires word-sense disambiguation, punctuation recovery, and context-sensitive paraphrasing. By contrast, reasoning-based QA more strongly depends on multi-sentence integration and stable answer grounding, while knowledge-based QA additionally exposes limitations in historical and cultural knowledge coverage. Finally, the fact that linguistic variant QA does not scale monotonically with model size suggests that stylistic understanding depends less on parameter count alone and more on exposure to stylistically rich pretraining data. Although these observations do not constitute controlled causal evidence, they provide a useful preliminary failure attribution for future ancient-document VLM development: improving OCR alone is unlikely to solve all downstream errors, and progress will likely require joint advances in visual grounding, classical-Chinese semantic modeling, and historically informed knowledge representation.

5 Conclusion

In this work, we present AncientDoc, the first benchmark for systematically evaluating vision-language models on Chinese ancient documents from OCR to higher-level understanding. AncientDoc covers five tasks—page-level OCR, vernacular translation, reasoning-based QA, knowledge-based QA, and linguistic variant QA—and spans 14 document categories, over 100 books, and about 3,000 page images. Extensive experiments show that Chinese ancient documents remain highly challenging for current VLMs, especially at the page-level OCR stage, while stronger performance on downstream QA tasks suggests that ancient-document understanding is constrained not only by visual transcription quality, but also by higher-level semantic grounding and reasoning ability.

Beyond benchmarking, our preliminary analysis suggests that different subtasks expose different bottlenecks: vernacular translation is closely tied to semantic transfer from classical Chinese to modern Chinese, reasoning-based QA depends more on multi-sentence integration and stable grounding, and knowledge-based QA further reveals limitations in historical and cultural knowledge coverage. In addition, the weak correlation between model scale and linguistic variant QA suggests that stylistic understanding may depend more on training data composition than on parameter count alone. At the same time, AncientDoc should be viewed as a realistic first-step benchmark built from currently accessible digitized resources, rather than a fully balanced representation of all Chinese ancient documents.

Limitations

This work still has several limitations. First, while AncientDoc covers diverse document types and periods, its data distribution is limited by digitized resource availability, overrepresenting Ming–Qing documents and regular script, with sparse coverage of earlier periods and highly cursive scripts. Second, the benchmark includes multiple tasks beyond OCR but remains page-level and text-centric, without explicit evaluation of cross-page reasoning, complex layout understanding, or document-level structural coherence. Third, high-level understanding tasks rely on automatic and LLM-based metrics, which may introduce biases and cannot fully replace expert human evaluation.

Acknowledgements

This work was supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

References

Alireza Akoushideh, Atefeh Ranjkesh Rashtehroudi, and Asadollah Shahbahrami. 2025. Persian/arabic scene text recognition with convolutional recurrent neural network. *IET Smart Cities*, 7(1):e70001.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

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. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Łukasz Borchmann, Michał Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Graliński. 2021. Due: End-to-end document understanding benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ran Cui, Aichun Zhu, and Zichen Ding. 2025. Enhanced chinese scene text recognition model base on cross-domain feature fusion. *Signal, Image and Video Processing*, 19(6):499.

Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023. Vqa: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 585–601. Springer.

Yihao Ding, Kaixuan Ren, Jiabin Huang, Siwen Luo, and Soyeon Caren Han. 2024. Mvqa: A dataset for multimodal information retrieval in pdf-based visual question answering. *arXiv preprint arXiv:2404.12720*.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, and 1 others. 2024. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *arXiv preprint arXiv:2501.00321*.

Pan He, Weilin Huang, Yu Qiao, Chen Loy, and Xiaoou Tang. 2016. Reading scene text in deep convolutional sequences. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

R Reeve Ingle, Yasuhisa Fujii, Thomas Deselaers, Jonathan Baccash, and Ashok C Popat. 2019. A scalable handwritten text recognition system. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 17–24. IEEE.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE.

Lei Kang, Pau Riba, Marçal Rusiñol, Alicia Fornés, and Mauricio Villegas. 2022. Pay attention to what you read: non-recurrent handwritten text-line recognition. *Pattern Recognition*, 129:108766.

Lei Kang, Rubèn Tito, Ernest Valveny, and Dimosthenis Karatzas. 2024. Multi-page document visual question answering using self-attention scoring mechanism. In *International Conference on Document Analysis and Recognition*, pages 219–232. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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):220102.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2024. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19071–19079.

ByteDance Seed Team. 2025. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. 2016. Detecting text in natural image with connectionist text proposal network. In *European conference on computer vision*, pages 56–72. Springer.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Zhongguo Xu and Yang Xiang. 2025. Improving chinese word segmentation with character–lexicon class attention. *Neural Computing and Applications*, 37(5):3857–3867.

Hailin Yang, Lianwen Jin, Weiguo Huang, Zhaoyang Yang, Songxuan Lai, and Jifeng Sun. 2018. Dense and tight detection of chinese characters in historical documents: Datasets and a recognition guided detector. *IEEE Access*, 6:30174–30183.

Fei Yin, Yi-Chao Wu, Xu-Yao Zhang, and Cheng-Lin Liu. 2017. Scene text recognition with sliding convolutional character models. *arXiv preprint arXiv:1709.01727*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. *arXiv preprint arXiv:1810.03581*.

Jinxu Zhang, Yongqi Yu, and Yu Zhang. 2024. Cream: coarse-to-fine retrieval and multi-modal efficient tuning for document vqa. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 925–934.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Explanations of 14 Document Categories in AncientDoc

The 14 categories of Chinese ancient documents included in AncientDoc cover a wide range of genres, reflecting the diversity of knowledge and literary forms in traditional Chinese culture. Their specific explanations are as follows:

- Biography: These texts focus on recording the life stories, achievements, and moral virtues of historical figures, typically presenting chronological narratives that serve as moral examples for readers.
- Confucianism: Works centered on Confucian philosophy and ethics, including interpretations of the "Five Classics" (e.g., Book of Changes, Book of Songs) and writings by Confucian scholars that elaborate on concepts like benevolence, righteousness, and propriety.
- The Philosophy of War: Military strategy texts, such as Sun Tzu's Art of War, which explore tactics, command principles, and the

philosophy of warfare, emphasizing strategic thinking and battlefield wisdom.

- Collected Works: Comprehensive anthologies compiling poems, essays, or other writings by multiple authors (or a single author across genres), often organized thematically (e.g., nature, politics) or chronologically.

- Medical Authors: Ancient medical texts that discuss diagnostic methods, herbal remedies, acupuncture techniques, and classical medical theories (e.g., yin-yang and five elements), forming the foundation of traditional Chinese medicine.

- Astronomy and Mathematics: Scientific treatises on calendrical systems, astronomical observations (e.g., eclipses, planetary movements), and mathematical operations, reflecting ancient China's advancements in natural sciences.

- Masters of Unofficial History: Semi-historical or fictional works that offer alternative perspectives on historical events, anecdotes, or folklore, often filling gaps in official historical records with vivid narratives.

- General Anthologies: Anthologies dedicated to the literary works of a single author, distinct from "Collected Works" (which include multiple authors). These texts highlight the unique style and creative characteristics of individual writers.

- Eclectics: Works that integrate ideas from multiple philosophical schools, such as Confucianism, Taoism, and Legalism, aiming to synthesize diverse thoughts into a coherent worldview.

- Chuci-style Poetry: Poetry modeled after the Chuci (Songs of Chu), a classic collection of ancient Chinese poetry. This genre is known for its rich imagery, mythological allusions, and emotional intensity, often expressing patriotic or melancholic sentiments.

- Leishu: Traditional encyclopedic works organized by topic (e.g., astronomy, geography, literature), serving as reference tools for scholars and educators to access comprehensive knowledge efficiently.

- Art: Texts focusing on traditional Chinese arts, including painting, calligraphy, music, and handicrafts, discussing techniques, aesthetic principles, and the cultural significance of artistic creation.

- Literary Criticism of Poetry and Prose: Analytical writings that evaluate classical literature, focusing on literary form, rhetorical devices, and artistic merit. These works shape critical standards for poetry, essays, and other genres.

- Genealogies and Catalogues: Documents recording family lineages (genealogies), bibliographic records of texts, or catalogs of artifacts (e.g., antiques, books), playing a key role in preserving historical and cultural heritage.

Together, these 14 categories encompass historical records, philosophical treatises, literary works, scientific texts, and practical references, providing a comprehensive sample of Chinese ancient documents for evaluating VLMs' cross-domain understanding capabilities.

B Evaluation Metrics

To rigorously assess the performance of Vision-Language Models (VLMs) on the tasks in Ancient-Doc, we employ a set of metrics tailored to different task characteristics, as detailed below:

B.1 Metrics for Page-level OCR

Four character-level metrics are used to evaluate the accuracy of text extraction from ancient document pages:

Character Error Rate (CER): Measures the normalized number of character-level errors (substitutions, insertions, deletions) between the predicted sequence and the reference text:

$$\mathrm{CER} = \frac{S + I + D}{|T|} \tag{1}$$

where S, I, and D represent the counts of substituted, inserted, and deleted characters, respectively, and $|T|$ is the length of the reference sequence (Ground Truth).

Character Precision: Reflects the proportion of correctly predicted characters relative to the total number of characters in the model's output:

$$\mathrm{CharPrecision} = \frac{M}{|P|} \tag{2}$$

where M is the number of correctly predicted characters, and $|P|$ is the total number of characters in the predicted sequence.

Character Recall: Measures the proportion of correctly predicted characters relative to the total number of characters in the reference text:

$$\text{CharRecall} = \frac{M}{|T|} \tag{3}$$

where $|T|$ denotes the length of the reference sequence.

Character F1: A harmonic mean of Character Precision and Character Recall, balancing both metrics to provide a comprehensive evaluation of OCR accuracy:

$$\text{CharF1} = 2 \cdot \frac{\text{CharPrecision} \cdot \text{CharRecall}}{\text{CharPrecision} + \text{CharRecall}} \tag{4}$$

B.2 Metrics for Higher-Level Understanding Tasks

For vernacular translation, reasoning-based QA, knowledge-based QA, and linguistic variant QA, the following metrics are adopted:

CHRF++: Computes the F-score based on character and word n-grams, with $\beta = 2$ to emphasize recall:

$$\text{CHRF}^{++} = F_\beta(\text{Ngram}_{\text{char}}, \text{Ngram}_{\text{word}}) \tag{5}$$

BERTScore (BS-F1): A semantic-level metric that calculates precision and recall using cosine similarities between BERT embeddings of predicted and reference texts, then derives the F1 score:

$$\text{BS-F1}(P, T) = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

GPT-4o Score: A human-aligned metric where GPT-4o scores model predictions against references on a scale of 0–10. The final score is the average of all individual scores:

$$\text{GPT4o\_Score} = \frac{1}{N} \sum_{i=1}^{N} s_i \tag{7}$$

where s_i is the score for the i-th sample, and N is the total number of samples.

C Metrics for Evaluating Differences Between Human and LLM Scoring

To quantify the consistency and discrepancy between human scoring and large model scoring, we employ six statistical metrics, defined as follows:

Pearson correlation coefficient: Measures the linear correlation between two sets of scores (x for human scores, y for model scores):

$$\text{Pearson}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{8}$$

where $\bar{x}$ and $\bar{y}$ denote the means of x and y, respectively.

Spearman rank correlation coefficient: Evaluates the monotonic relationship by comparing score ranks:

$$\text{Spearman}(x, y) = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{9}$$

with $d_i = \text{rank}(x_i) - \text{rank}(y_i)$ representing the rank difference of the i-th sample, and n being the total number of samples.

Kendall tau coefficient: Assesses ordinal association by counting concordant and discordant pairs:

$$\text{Kendall}(x, y) = \frac{C - D}{\frac{1}{2}n(n-1)} \tag{10}$$

where C is the number of concordant pairs (consistent order) and D is the number of discordant pairs (inconsistent order).

Mean Squared Error (MSE): Quantifies the average squared difference between scores:

$$\text{MSE}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2 \tag{11}$$

Mean Absolute Error (MAE): Measures the average absolute difference between scores:

$$\text{MAE}(x, y) = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| \tag{12}$$

Bias: Represents the systematic deviation between the mean of human scores and model scores:

$$\text{Bias}(x, y) = \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{x} - \bar{y} \tag{13}$$

These metrics collectively capture linear/correlation trends, ordinal relationships, error magnitudes, and systematic deviations, providing a comprehensive assessment of scoring consistency.

D Analysis Between Human Scoring and LLM Scoring Across Five Subtasks

To ensure that the large language model used as an automatic evaluator reflects human judgment as faithfully as possible, we conduct a human-alignment analysis between model-based scores and human ratings across the five subtasks of AncientDoc. The goal of this analysis is to identify the evaluator whose scoring behavior is most consistent with human preferences.

Specifically, we compare the scoring results produced by several mainstream large models, including Qwen2.5-VL-72B, Gemini, Doubao, Qwen-Plus, and GPT-4o, against human ratings. For each subtask, we randomly sample 50 QA pairs, resulting in a total of 250 QA pairs. Each QA pair is independently scored by human annotators and by the candidate models on a scale from 0 to 10.

To quantify the alignment between model-based scores and human judgments, we compute six commonly used metrics, including Pearson, Spearman, and Kendall correlation coefficients, as well as mean squared error (MSE), mean absolute error (MAE), and Bias. These metrics jointly measure both rank-level consistency and absolute score deviation between the two sources of evaluation.

The results 7 demonstrate that GPT-4o consistently achieves the strongest alignment with human ratings across all five subtasks, exhibiting higher correlation and lower error compared to other candidate evaluators (as illustrated in Fig. 3–7). Based on these findings, GPT-4o is selected as the automatic evaluator in our main experiments.

E Supplementary Analysis of Experimental Results

Results in Page-level OCR. The results in Tab. 2 and Tab. 8 demonstrate clear performance differences among MLLMs on page-level OCR for AncientDoc. Among all evaluated models, Qwen3-VL-8B achieves the best character-level performance, obtaining the highest Char Precision (19.78), Char Recall (20.63), and Char F1 score (20.2). This indicates strong glyph discrimination ability as well as stable and consistent transcription across long character sequences. Notably, Qwen3-VL-8B outperforms not only smaller open-source models but also several larger open- and closed-source MLLMs, highlighting the effectiveness of its vision-language modeling for OCR-centric tasks.

We attribute the superior performance of Qwen3-VL-8B to its balanced architectural scale and OCR-friendly inductive bias. Compared to smaller models, the 8B variant provides sufficient capacity to capture subtle stroke-level variations and complex spatial layouts in ancient documents. Meanwhile, unlike larger models with stronger reasoning or semantic correction tendencies, Qwen3-VL-8B appears to prioritize faithful visual transcription over semantic inference, which is crucial for ancient scripts that lack modern linguistic regularities and exhibit high visual similarity across characters.

Gemini-2.5-Pro achieves the lowest CER (32.03) and the second-best Char F1 (18.12), indicating strong overall stability at the page level and fewer edit operations. This suggests that Gemini models are particularly effective in maintaining global transcription consistency. In contrast, Doubao-V2 attains high recall but suffers from a substantially higher CER, implying frequent character substitutions or ordering errors despite detecting character-like regions, which reflects difficulty in fine-grained glyph differentiation.

The Qwen2.5 series also performs reliably across model scales. Interestingly, Qwen2.5-VL-7B consistently outperforms Qwen2.5-VL-72B on all OCR-related metrics and achieves the highest GPT-4o score, further suggesting that page-level OCR primarily depends on accurate visual perception and stable sequence generation rather than advanced reasoning capabilities. Excessive semantic priors in larger models may instead interfere with strict character-level transcription.

Overall, these results indicate that medium-scale vision-language models with strong visual inductive biases, such as Qwen3-VL-8B and Qwen2.5-VL-7B, together with Gemini-2.5-Pro, provide the most practical and reliable solutions for page-level OCR of ancient documents.

Results in Vernacular Translation. As shown in Tab. 3 and Tab. 10, Gemini-2.5-Pro achieves the best overall performance, obtaining the highest BS-F1 score (72.5) and the highest GPT-4o score (4.72), demonstrating its strong capability in vernacular translation of ancient texts. Its superior performance across both automatic and LLM-based evaluation metrics indicates robust semantic understanding and fluent modern Chinese generation, making it a strong baseline for vernacular translation tasks.

Among open-source models, the Qwen series consistently outperforms other alternatives. No-

	Pearson	Spearman	Kendall	MSE	MAE	Bias
Qwen2.5-VL-72B (Bai et al., 2025b)	0.833	0.824	**0.718**	3.415	1.364	1.044
Gemini (Comanici et al., 2025)	0.829	0.822	0.67	4.463	1.644	-1.436
Doubao (Team, 2025)	0.695	0.774	0.637	10.739	2.696	-2.616
Qwen-Plus (Qwen et al., 2024)	0.767	0.743	0.618	3.631	1.54	**0.724**
GPT-4o (Hurst et al., 2024)	**0.846**	**0.837**	0.689	**2.939**	**1.32**	-0.896

Table 7: Comparison of differences between large model scoring and human scoring.

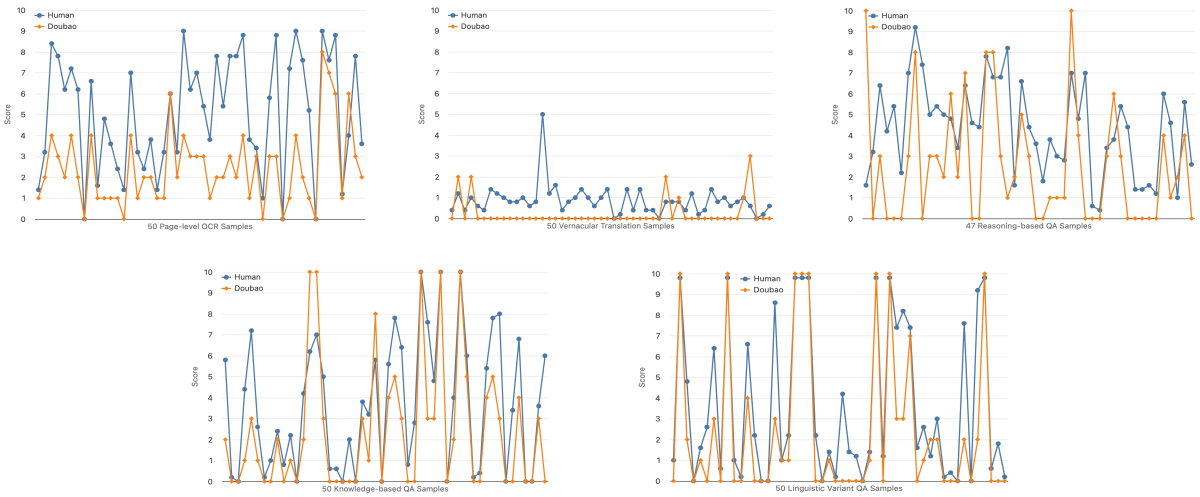


Figure 3: Comparison of consistency between Doubao scoring and human scoring.

tably, Qwen3-VL-8B achieves a BS-F1 score of 71.66, surpassing all other open-source models and approaching the performance of Gemini-2.5-Pro, despite its relatively moderate model size. This suggests that Qwen3-VL-8B possesses strong semantic alignment between ancient and modern Chinese, benefiting from effective language modeling and balanced multimodal representation learning. Compared to smaller variants, the 8B model provides sufficient capacity for capturing long-range semantic dependencies and implicit historical context, which are essential for faithful vernacular translation.

Within the Qwen family, Qwen-VL-Max (71.03) and Qwen2.5-VL-72B (69.87) also achieve strong results, while Qwen2.5-VL-7B (65.59) significantly outperforms most other open-source models, demonstrating that the Qwen series exhibits stable and scalable capabilities in ancient text understanding and modern Chinese expression. In contrast, the InternVL and LLaVA series generally perform worse on this task, suggesting that their multimodal modeling is more oriented toward general vision-language tasks and lacks specialized semantic modeling for ancient Chinese corpora.

It is also noteworthy that the closed-source model GPT-4o achieves a relatively low BS-F1 score (58.86), lagging behind several open-source models with fewer parameters, such as Qwen2.5-VL-7B and Qwen3-VL-4B. This further indicates that vernacular translation of ancient texts relies heavily on language-specific semantic alignment rather than general multimodal reasoning capabilities. Finally, due to the inherent diversity of valid vernacular translations, GPT-4o-based scores remain relatively low across all models: even the best-performing model only achieves an average score of 4.72 out of 10.

Results in Reasoning-based QA. As shown in Tab. 4 and Tab. 9, Qwen3-VL-8B achieves the best overall performance in BS-F1 (74.06), followed by Qwen3-VL-4B (73.76). This indicates that the Qwen3-VL series provides the strongest semantic matching to reference answers in reasoning-based QA over ancient texts. Notably, Qwen3-VL-8B outperforms both large-scale open-source models (e.g., Qwen2.5-VL-72B with 71.40) and several closed-source baselines, suggesting that model quality here is not purely scale-driven.

We attribute Qwen3-VL's advantage to two factors that are particularly critical for reasoning-based QA: (1) stronger context integration, i.e.,

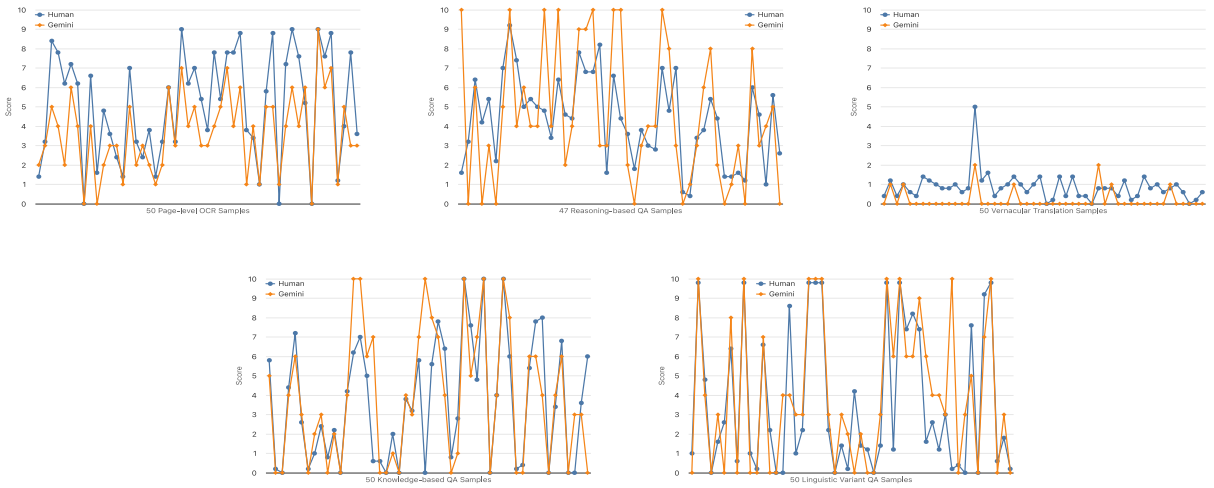


Figure 4: Comparison of consistency between Gemini scoring and human scoring.

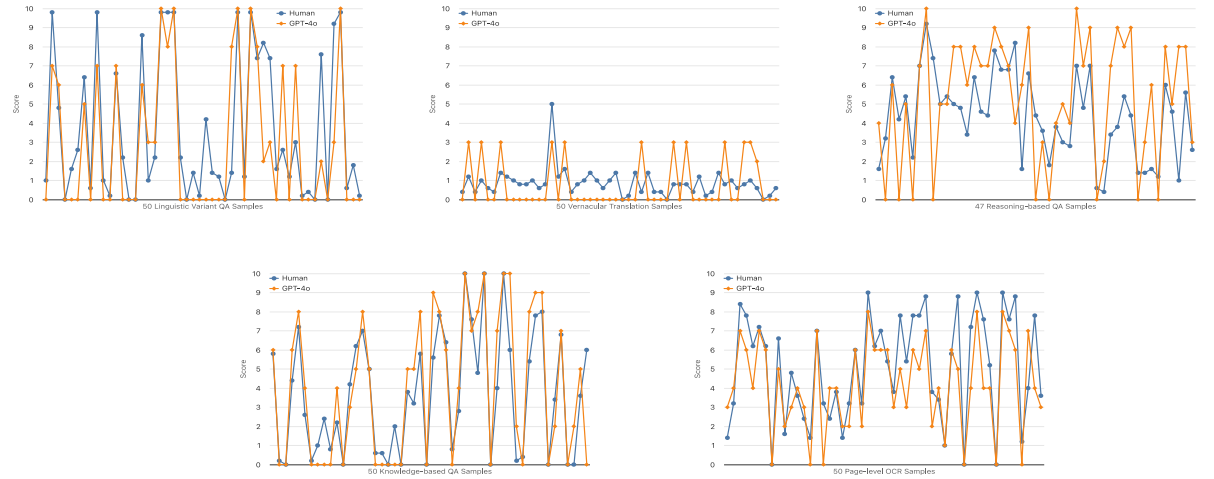


Figure 5: Comparison of consistency between GPT-4o scoring and human scoring.

better modeling of long-range dependencies and implicit relations in ancient passages; and (2) more stable answer grounding, where the model is less likely to drift into fluent but semantically misaligned responses. This is consistent with the observation that Qwen3-VL maintains high BS-F1 even at 4B scale, implying improved representation learning rather than simply increased parameters.

Within the Qwen2.5 family, Qwen2.5-VL-72B still achieves strong performance (BS-F1 71.40), and Qwen2.5-VL-7B remains competitive (69.96), indicating good reasoning efficiency at moderate scale. In contrast, the InternVL and LLaVA series generally obtain lower BS-F1 scores on this task, suggesting difficulties in establishing reliable contextual understanding and causal/implicit reasoning for ancient Chinese QA.

Finally, the GPT-4o-based scores are relatively higher than those in translation (with the best score reaching 7.76 from Gemini-2.5-Pro), which is reasonable because QA outputs are typically more constrained than free-form vernacular translation, reducing evaluation variance across different valid generations.

Results in Knowledge-based QA. From the experimental results in Tab. 5 and Tab. 11, Qwen3-VL-8B achieves the highest BS-F1 score (73.98), followed by Qwen3-VL-4B (73.63), indicating that the Qwen3-VL series exhibits the strongest overall performance in knowledge-based QA over ancient texts. These results suggest that Qwen3-VL benefits from effective knowledge grounding and strong semantic alignment, even at relatively moderate model scales.

Among closed-source models, GPT-4o achieves a competitive BS-F1 score (70.01), reflecting its strong general language modeling capability and broad coverage of ancient Chinese knowledge. In

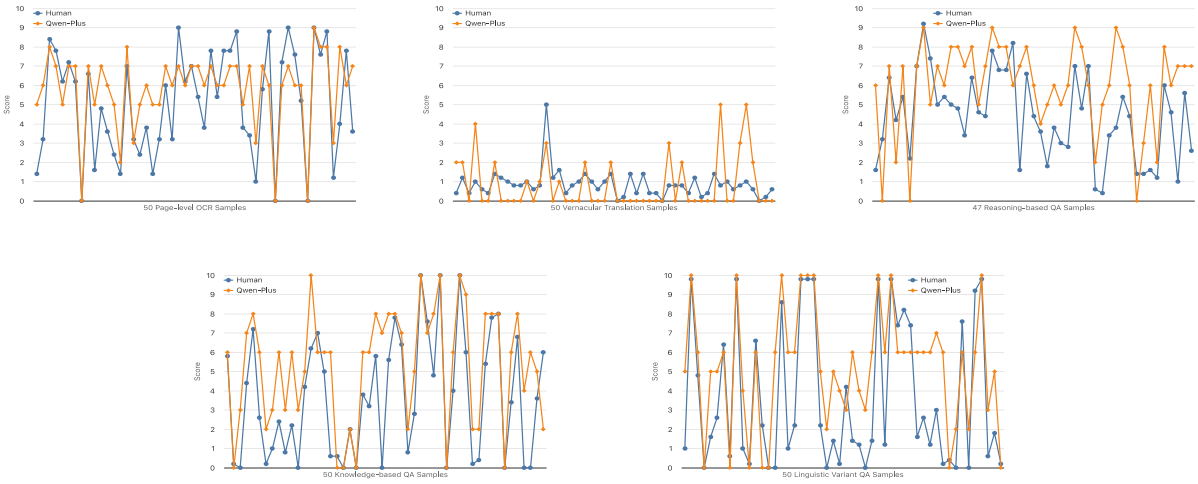


Figure 6: Comparison of consistency between Qwen-Plus scoring and human scoring.

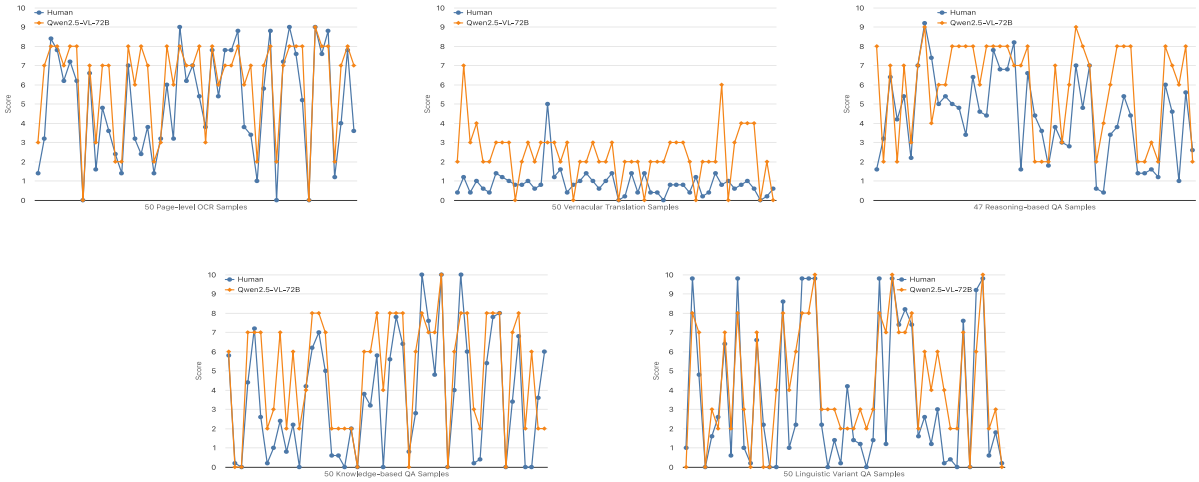


Figure 7: Comparison of consistency between Qwen-VL-72B scoring and human scoring.

terms of GPT-4o-based evaluation, Doubao-V2 and Gemini-2.5-Pro achieve the highest scores (7.36), indicating their advantage in producing fluent and knowledge-consistent answers under LLM-based evaluation criteria.

Within the Qwen2.5 family, Qwen2.5-VL-72B achieves a BS-F1 score (69.15) close to that of GPT-4o, suggesting that its training process incorporates a substantial amount of ancient-text-related data and provides relatively rich historical and cultural knowledge. In contrast, Qwen2.5-VL-7B does not demonstrate the same level of performance in knowledge-based QA as it does in page-level OCR, highlighting the increased demand for broad factual and cultural knowledge in this task.

These results indicate that, unlike OCR or reasoning-based QA, knowledge-based QA places stronger emphasis on the coverage and organization of pre-trained knowledge. While smaller and medium-sized models may perform well on perception- or reasoning-centric tasks, they can be limited by insufficient exposure to historical and cultural content during pre-training. The strong performance of Qwen3-VL suggests that improvements in data composition and knowledge-aware representation learning can partially compensate for model scale in knowledge-intensive settings.

Results in Linguistic Variant QA. From the evaluation results in Tab. 6 and Tab. 12, GPT-4o and Gemini-2.5-Pro once again demonstrate leading performance, achieving the highest BS-F1 scores of 64.58 and 62.06, respectively. These results indicate the reliability of large closed-source models in language style transformation and stylistic understanding, where fluent paraphrasing and preservation of linguistic style play a central role.

A notable observation is that the InternVL2.5 series performs significantly better in this task than in previous OCR, translation, and QA tasks. In particular, InternVL2.5-2B achieves the highest BS-F1 score (62.24) among all open-source models, even surpassing Gemini-2.5-Pro. This suggests that InternVL2.5 exhibits strong capability in modeling stylistic patterns and surface-level linguistic transformations, despite its relatively small parameter size.

Moreover, the overall performance of the InternVL2.5 series is consistently higher than that of the InternVL3 series, indicating a regression rather than improvement for this specific task. A plausible explanation is that InternVL2.5 incorporates more ancient-style or classical-language-related data during training, which benefits the modeling of classical expressions, rhetorical structures, and stylistic allusions required in this task. In contrast, newer model versions may emphasize general multimodal capabilities at the expense of style-specific linguistic sensitivity.

By comparison, the Qwen2.5 and Qwen3-VL series do not exhibit the same level of advantage in language-style transformation as they do in OCR or reasoning-based tasks. This further suggests that style transformation relies more heavily on language-centric pretraining and exposure to stylistically rich corpora, rather than on visual grounding or advanced reasoning abilities.

F All evaluation results

Since the limited length of the main text, we only show some results in the main text. In the following table, we have displayed all evaluation results (shown in Tab. 8- 12).

Table 8: Evaluation Results on Page-level OCR of AncientDoc

Model	**CER**	**CharPrecision**	**CharRecall**	**CharF1**	**GPT-4o**
Qwen-VL-Max	66.39	9.62	9.93	9.77	5.7
DeepSeek-VL2	147.28	0.25	0.22	0.11	0.6
InternVL2.5-1B	97.77	2.38	1.92	2.13	2.63
InternVL2.5-2B	120.72	1.76	1.69	1.72	2.03
InternVL2.5-4B	79.75	5.27	4.05	4.58	3.65
InternVL2.5-8B	96.88	2.87	2.59	2.72	3
Qwen2-VL-2B	81.75	4	2.66	3.2	3.79
Qwen2-VL-7B	71.61	7.99	6.8	7.35	4.66
Qwen2.5-VL-3B	50.36	9.44	9.02	9.23	5.56
Qwen2.5-VL-7B	35.47	12.95	12.75	12.85	6.37
Qwen2.5-VL-32B	59.65	14.72	14.65	14.68	5.67
Qwen2.5-VL-72B	58.83	10.62	9.56	10.06	5.73
Qwen2.5-Omni-3B	63.42	5.4	4.78	5.07	4.56
Qwen2.5-Omni-7B	70.46	3.44	3.35	3.39	4.3
Doubao-V2	71.95	13.14	20.44	16	6.27
Gemini2.5-Pro	32.03	17.73	18.53	18.12	6.08
GPT-4o	75.1	4.83	2.72	3.48	2.97
InternVL3-1B	97.49	3.57	1.83	2.42	1.83
InternVL3-2B	95.59	2.46	1.79	2.07	2.52
InternVL3-8B	51.8	7.3	6.67	6.97	5.06
InternVL3-38B	103.02	3.26	3.75	3.49	2.98
InternVL3-78B	78.67	4.7	4.89	4.79	4.04
LLaVA-Onevision-5B	215.31	0.05	0.07	0.01	0
LLaVA-Onevision-72B	190.67	0.29	0.43	0.25	0.01
LLaVA1.5-7B	129.71	0.04	0.03	0	0.05

Table 9: Evaluation Results on Reasoning-based QA of AncientDoc

Model	CHRF++	BS-F1	GPT-4o
Qwen-VL-Max	8.6	71.3	7.46
DeepSeek-VL2	3.42	59.09	1.01
InternVL2.5-1B	4.58	63.8	2.38
InternVL2.5-2B	5.01	65.29	3.55
InternVL2.5-4B	5.77	66.12	4.75
InternVL2.5-8B	7.47	68.4	5.93
Qwen2-VL-2B	6.12	67.8	4.45
Qwen2-VL-7B	6.91	69.03	5.79
Qwen2.5-VL-3B	4.8	65.83	4.67
Qwen2.5-VL-7B	7.34	69.96	6.44
Qwen2.5-VL-32B	9.69	70.9	7.49
Qwen2.5-VL-72B	9.04	71.4	7.43
Qwen2.5-Omni-3B	5.63	66.96	4.9
Qwen2.5-Omni-7B	6.52	68.7	6.05
Doubao-V2	7.15	68.78	7.4
Gemini2.5-Pro	8.68	69.33	7.76
GPT-4o	7.99	70.52	6.92
InternVL3-1B	5.12	64.24	2.85
InternVL3-2B	5.66	66.45	4.39
InternVL3-8B	5.21	65.62	4.98
InternVL3-38B	5.36	67.29	5.78
InternVL3-78B	4.95	65.99	5.18
LLaVA-Onevision-5B	2.21	55.79	0.89
LLaVA-Onevision-72B	5.96	67.2	5.48
LLaVA1.5-7B	3.99	61.69	1.64

Table 10: Evaluation Results on Vernacular Translation of AncientDoc

Model	CHRF++	BS-F1	GPT-4o
Qwen-VL-Max	12.3	71.03	3.22
DeepSeek-VL2	0.49	50.27	0
InternVL2.5-1B	1.35	52.73	0.02
InternVL2.5-2B	1.33	53.92	0.03
InternVL2.5-4B	2.84	58.46	0.53
InternVL2.5-8B	3.2	59.24	0.58
Qwen2-VL-2B	1.9	52.31	0.46
Qwen2-VL-7B	4.34	60.74	1.17
Qwen2.5-VL-3B	1.63	55.2	0.66
Qwen2.5-VL-7B	7.04	65.59	2.3
Qwen2.5-VL-32B	9.9	67.9	2.87
Qwen2.5-VL-72B	9.77	69.87	2.98
Qwen2.5-Omni-3B	2.06	56.8	0.56
Qwen2.5-Omni-7B	1.13	53.99	0.46
Doubao-V2	0.56	52.39	2.12
Gemini2.5-Pro	11.41	72.5	4.72
GPT-4o	3.02	58.86	0.92
InternVL3-1B	1.86	54.92	0.06
InternVL3-2B	0.97	51.83	0.1
InternVL3-8B	0.85	53.16	0.29
InternVL3-38B	1.99	56.93	0.56
InternVL3-78B	4.45	62.4	1.21
LLaVA-Onevision-5B	0.76	50.85	0
LLaVA-Onevision-72B	0.44	48.83	0.03
LLaVA1.5-7B	0.43	50	0.01

Table 11: Evaluation Results on Knowledge-based QA of AncientDoc

Model	CHRF++	BS-F1	GPT-4o
Qwen-VL-Max	7.58	68.67	6.78
DeepSeek-VL2	3.77	59.83	0.94
InternVL2.5-1B	3.73	61.59	1.68
InternVL2.5-2B	4.47	64.24	2.48
InternVL2.5-4B	4.95	64.81	3.92
InternVL2.5-8B	6.93	67.68	4.88
Qwen2-VL-2B	5.27	66.6	3.37
Qwen2-VL-7B	5.75	66.82	4.83
Qwen2.5-VL-3B	4.47	62.86	3.77
Qwen2.5-VL-7B	5.87	66.75	5.23
Qwen2.5-VL-32B	9.44	69.35	6.84
Qwen2.5-VL-72B	7.82	69.15	6.84
Qwen2.5-Omni-3B	5.33	66.1	4.11
Qwen2.5-Omni-7B	5.67	66.85	4.94
Doubao-V2	8.75	69.15	7.36
Gemini2.5-Pro	8.88	68.94	7.36
GPT-4o	8.02	70.01	6.53
InternVL3-1B	4.46	63.05	1.82
InternVL3-2B	5.43	66.21	3.52
InternVL3-8B	4.75	63.6	4.31
InternVL3-38B	4.84	65.45	5.11
InternVL3-78B	5.25	65.79	5.18
LLaVA-Onevision-5B	2.32	56.67	0.51
LLaVA-Onevision-72B	5.56	66.31	5.04
LLaVA1.5-7B	3.23	60.37	0.78

Table 12: Evaluation Results on Linguistic Variant QA of AncientDoc

Model	CHRF++	BS-F1	GPT-4o
Qwen-VL-Max	3.31	58.77	5.67
DeepSeek-VL2-Tiny	3.7	60.05	0.91
InternVL2.5-1B	2.61	59.27	1.59
InternVL2.5-2B	3.42	62.24	2.25
InternVL2.5-4B	2.85	58.22	3.26
InternVL2.5-8B	3.65	61.52	3.63
Qwen2-VL-2B	3.49	58.97	3.1
Qwen2-VL-7B	2.3	56.73	3.67
Qwen2.5-VL-3B	1.03	52.3	3.75
Qwen2.5-VL-7B	2.65	57.48	4.75
Qwen2.5-VL-32B	4.62	61.18	5.61
Qwen2.5-VL-72B	3.65	59.34	5.63
Qwen2.5-Omni-3B	2.12	56.87	3.4
Qwen2.5-Omni-7B	2.66	58.62	4.12
Doubao-V2	3.32	57.7	5.79
Gemini2.5-Pro	5.22	62.06	5.92
GPT-4o	4.16	64.58	5.03
InternVL3-1B	3.69	61.92	1.51
InternVL3-2B	2.9	58.95	2.83
InternVL3-8B	1.93	55.96	3.53
InternVL3-38B	1.96	56.87	3.99
InternVL3-78B	2.38	57.78	4.13
LLaVA-Onevision-5B	1.72	56.12	0.48
LLaVA-Onevision-72B	2.51	57.57	3.43
LLaVA1.5-7B	2.27	56.56	0.87