CHRONICLES-OCR: A CROSS-TEMPORAL PERCEPTION BENCHMARK FOR THE EVOLUTIONARY TRAJECTORY OF CHINESE CHARACTERS

Anonymous authors
Paper under double-blind review

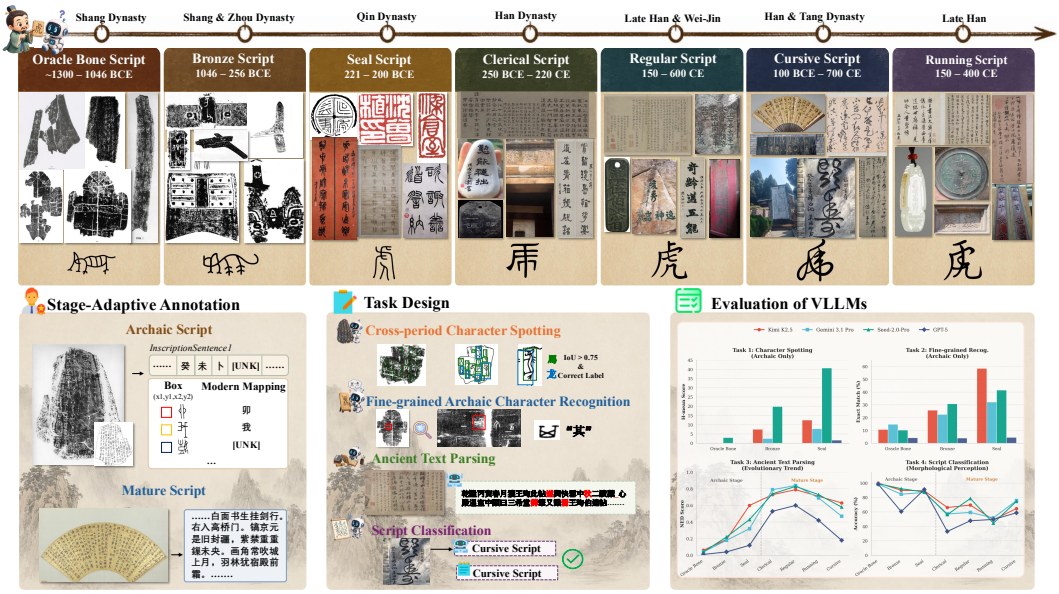

Figure 1: **Chronicles-OCR overview.** The top row showcases diverse physical artifact samples across seven script stages, alongside the morphological evolution of the modern Chinese character "虎" (Tiger). We introduce a stage-adaptive annotation paradigm and four progressive tasks.

ABSTRACT

Vision Large Language Models (VLLMs) have achieved remarkable success in modern text-rich visual understanding. However, their perceptual robustness in the face of the continuous morphological evolution of historical writing systems remains largely unexplored. Existing ancient text datasets typically focus on isolated historical periods, failing to capture the systematic visual distribution shifts spanning thousands of years. To bridge this gap and empower Digital Humanities, we introduce **Chronicles-OCR**, the first comprehensive benchmark specifically designed to evaluate the cross-temporal visual perception capabilities of VLLMs across the complete evolutionary trajectory of Chinese characters, known as the "Seven Chinese Scripts". Curated in collaboration with top-tier institutional domain experts, the dataset comprises 2,800 strictly balanced images encompassing highly diverse physical media, ranging from tortoise shells to paper-based calligraphy. To accommodate the drastic morphological and topological variations across different historical stages, we propose a novel *Stage-Adaptive Annotation Paradigm*. Based on this, Chronicles-OCR formulates four rigorous quantitative tasks: cross-period character spotting, fine-grained archaic character recognition via visual referring, ancient text parsing, and script classification. By isolating visual perception from semantic reasoning, Chronicles-OCR provides an authoritative platform to expose the limitations of current VLLMs, paving the way for robust, evolution-aware historical text perception. The benchmark and evaluation code will be made publicly available.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Chinese characters have evolved over millennia, leaving a written heritage that spans diverse scripts and physical media. Vision Large Language Models (VLLMs) have made substantial progress in modern text understanding Bai et al. (2025a;b), with strong performance on benchmarks such as OCRBench Fu et al. (2024), OmniDocBench Ouyang et al. (2025), and OCR-Reasoning Huang et al. (2025). Extending these capabilities to historical writing requires handling changes in character forms, spatial arrangements, and material conditions. Assessing this ability is an important step toward supporting historical transcription and research in Digital Humanities.

Across its long history, Chinese writing has developed a variety of scripts with distinct visual characteristics Guan et al. (2024); Jiao et al. (2025); Wang et al. (2022). Among them, Oracle Bone, Bronze, Seal, Clerical, Regular, Running, and Cursive represent seven major scripts. Recognizing the first three requires mapping substantially different historical forms to modern characters, while Clerical and Regular share more structural features with modern writing. Running and Cursive introduce simplification and connected strokes that can obscure character boundaries Qin et al. (2020); Liang et al. (2020). Material textures and degradation create further challenges across bones, bronze vessels, stone, and paper Philips & Tabrizi (2020). These variations place different demands on script identification, character localization and recognition, and reading order recovery.

Existing datasets typically focus on individual scripts or specific document collections, such as Oracle Bone inscriptions Chen et al. (2025); Li et al. (2020); Wang et al. (2024b) and historical books Shi et al. (2023; 2025). These resources establish baselines for particular settings, while systematic comparisons across scripts and among script identification, character recognition, and full-text transcription remain limited. Moreover, transcription performance reflects both character recognition and reading order, making it difficult to assess recognition alone. A unified framework is therefore needed to examine these related capabilities. *How well can VLLMs recognize historical characters when their locations are provided, and how does this ability compare with identifying scripts and transcribing complete texts?*

To bridge this evaluation gap and address these questions, we introduce **Chronicles-OCR** (illustrated in Fig. 1), a benchmark of 2,800 images curated with domain experts and balanced across the Seven Chinese Scripts. Our *Stage-Adaptive Annotation Paradigm* accommodates differences in character forms and textual organization and supports four complementary tasks. For Oracle Bone, Bronze, and Seal scripts, character-level annotations support cross-period character spotting and fine-grained archaic character recognition via visual referring. The latter highlights target characters to assess recognition when their locations are known. Across all seven scripts, sequence-level transcriptions and script labels support ancient text parsing and script classification, respectively. We evaluate mainstream open-source and closed-source VLLMs under this framework, comparing their script identification, recognition at specified locations, and full-text transcription capabilities across scripts.

In summary, the primary contributions of this paper are threefold:

- **Chronicles-OCR, the First Benchmark Covering the "Seven Chinese Scripts":** We bridge the evaluation gap in historical text perception with the first full-timespan coverage, from unstandardized archaic symbols to mature pre-modern scripts.
- **Stage-Adaptive Evaluation Tasks:** We design four tasks tailored to each historical stage, spanning character spotting and fine-grained recognition for archaic scripts, ancient text parsing and script classification across all seven scripts.
- **Comprehensive Benchmarking of Perceptual Deficiencies:** Through extensive evaluations of mainstream VLLMs, we establish performance baselines, quantify their perceptual bottlenecks across eras, and provide clear optimization directions for Digital Humanities.

2 RELATED WORK

2.1 EVOLUTION AND VISUAL CHARACTERISTICS OF CHINESE SCRIPTS

The morphological evolution of Chinese characters is conventionally categorized into the "Seven Chinese Scripts", developing continuously across successive dynasties. Originating in the Shang Dynasty, the Oracle Bone script represents the earliest mature writing system, carved on tortoise

shells and animal bones with thin, angular lines and strong pictographic characteristics, lacking standardized character sizes or spatial alignment Chen et al. (2025); Flad (2008); Keightley (1997). During the Shang and Zhou Dynasties, the Bronze script emerged on ceremonial vessels, exhibiting thicker strokes that gradually evolved into more regularized and aesthetic structures by the late Western Zhou period Guo (2020); Hua et al. (2025); Zhou et al. (2025). Following the unification of China, the Qin Dynasty standardized the Seal script, simplifying earlier variants into fixed structural patterns with pronounced curvilinear symmetry Fu (2026); Ou et al. (2024); Zhou et al. (2023). The Han Dynasty then witnessed the "Clerical Reformation" through the Clerical script Guoqing et al. (2022); Lei et al. (2025); Wu (2024). This script flattened characters and replaced curves with angular strokes, marking the transition of Chinese characters into symbolic forms and laying the foundation for modern typography. Emerging in the late Han and Wei-Jin periods, the Regular script established strict square-shaped structures and standardized strokes, remaining the dominant formal script to this day Peng (2017); Wang et al. (2022); Yang et al. (2018); Zhang (2021). While these first five scripts successively served as formal writing systems, the Cursive and Running scripts developed primarily for informal and rapid writing Wang et al. (2025b). Originating in the Han Dynasty and later evolving into unconstrained forms such as *Kuangcao*, the Cursive script uses rapid, continuous strokes, often eliminating independent character boundaries Qin et al. (2020); Liang et al. (2020); Wu et al. (2018); Chen & Lum (1995); Schlombs (1998); Gang et al. (2023). In contrast, the Running script functions as a fluid yet legible intermediate style. From a computer vision perspective, this historical progression introduces substantial visual distribution shifts, as texts evolve from pictographic drawings to structured geometric symbols and further into highly abstract continuous forms Guan et al. (2024); Jiao et al. (2025); Wang et al. (2022). These shifts create significant challenges for text localization, cross-period feature mapping, and reading order prediction, often causing modern VLMs to hallucinate by overfitting to contemporary glyph patterns rather than interpreting ancient structures Li et al. (2026a;b); Peng et al. (2025b).

2.2 VLLMs in Modern Text Perception

With the rapid advancement of LLMs Brown et al. (2020); Achiam et al. (2023); Singh et al. (2025), VLLMs have achieved strong cross-modal alignment between visual and textual representations Wang et al. (2024a); Bai et al. (2023; 2025b); Liu et al. (2023; 2024a;b); Dai et al. (2023); Peng et al. (2025a); OpenAI (2023); Zhu et al. (2023); Hou et al. (2026), driving a paradigm shift in text-rich visual understanding from cascaded OCR pipelines to end-to-end multimodal perception Cui et al. (2025); Team et al. (2025). Pre-trained on massive image-text corpora, state-of-the-art models (*e.g.,* GPT-4o Achiam et al. (2023), Qwen-VL Bai et al. (2025a;b); Qwen Team (2026)) demonstrate strong performance on modern benchmarks such as OCRBench Fu et al. (2024) and OmniDocBench Ouyang et al. (2025), effectively handling text localization, layout parsing, and cross-modal reasoning. However, this performance largely depends on the unified morphology, clear delimiters, and regular structures of modern writing systems, as their visual encoders are primarily optimized for standardized contemporary typography Liu et al. (2025). Consequently, when facing extreme historical distribution shifts, unconstrained layouts, and morphological diversity, the perceptual limitations of VLLMs across historical scripts remain largely unexplored Cao et al. (2025); Li et al. (2025a); Qiao et al. (2025); Yao et al. (2025).

Table 1: **Comparison of Ancient Chinese Script Benchmarks.**

Benchmark	Release Date	Script Types							Task Types			
		Oracle Bone	Bronze Script	Seal Script	Clerical Script	Regular Script	Cursive Script	Running Script	Character Spotting	Character Recognition	Text Parsing	Script Classification
TKH & MTH Yang et al. (2018)	2018.05					✓			✓	✓	✓	
HWOBC Li et al. (2020)	2020.11	✓							✓			
M5HisDoc Shi et al. (2023)	2023.11				✓	✓	✓	✓	✓	✓	✓	
HUST-OBC Wang et al. (2024b)	2024.01	✓							✓			
EVOBC Guan et al. (2024)	2024.01	✓	✓	✓	✓				✓			
OBI Component 20 Hu et al. (2024)	2024.06	✓							✓			
OBIMD Li et al. (2024)	2024.07	✓							✓	✓	✓	
OBI-Bench Chen et al. (2025)	2024.12	✓							✓	✓		
HisDoc1B Shi et al. (2025)	2025.01				✓	✓	✓	✓	✓	✓	✓	
RMOBS Li et al. (2025b)	2025.06	✓							✓			
PictOBI-20k Chen et al. (2026)	2025.09	✓							✓			
GEVO-Bench Song et al. (2026)	2026.04	✓	✓	✓	✓	✓			✓			✓
Chronicles-OCR	2026.09	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2.3 Evaluation for Ancient Script Perception

Existing research on ancient script perception suffers from two major limitations. First, most existing benchmarks focus only on specific historical periods or isolated script categories, resulting

in a highly fragmented evaluation landscape. For example, recent works such as HWOBC Li et al. (2020), HUST-OBC (Wang et al., 2024b), OBIMD (Li et al., 2024), and OBI-Bench Chen et al. (2025) are dedicated exclusively to Oracle Bone Inscriptions, while document benchmarks such as M5HisDoc Shi et al. (2023) and HisDoc1B (Shi et al., 2025) mainly target paragraph-level OCR in mature ancient Chinese books. Second, studies such as HWOBC (Li et al., 2020), HUST-OBC (Wang et al., 2024b), and GEVO-Bench (Song et al., 2026) primarily operate on isolated character glyphs, limiting the evaluation of document-level capabilities such as character localization, reading-order recovery, and full-text parsing. As VLLMs emerge as universal perception engines, it becomes increasingly important to evaluate their robustness against the systematic visual distribution shifts accumulated across thousands of years of script evolution Chen et al. (2025); Liu et al. (2025), yet the current fragmentation of datasets makes such a holistic evaluation fundamentally unattainable. To bridge this gap, **Chronicles-OCR** provides the first unified benchmark that establishes performance baselines and quantifies the perceptual bottlenecks of VLLMs across the complete evolutionary trajectory of Chinese scripts.

3 Chronicles-OCR Benchmark

In this section, we detail the construction of the Chronicles-OCR benchmark. As illustrated in Fig. 2, Chronicles-OCR is organized into three core components. We first introduce the expert-driven image sourcing and data curation process in Sec. 3.1. This is followed by a detailed exposition of the *Stage-Adaptive Annotation Paradigm* in Sec. 3.2, which is specifically tailored to address the morphological variations of scripts across different historical eras. Finally, we formulate four rigorous evaluation tasks and corresponding metrics to precisely quantify the perceptual bottlenecks of current VLLMs in Sec. 3.3. As summarized in Tab. 1, compared to existing datasets, Chronicles-OCR stands as the first unified benchmark to encompass the full evolutionary trajectory of the Seven Chinese Scripts while offering unprecedented task diversity.

3.1 Image Sourcing and Data Curation

Securing highly reliable historical image data is paramount, as it forms the foundational basis of the Chronicles-OCR benchmark. To ensure historical authenticity and structural diversity, our raw image sourcing process was conducted in close collaboration with domain experts and institutional partners. Specifically, the Oracle Bone image data was provided by domain experts specializing in oracle bone inscriptions. The Bronze and Seal script images were systematically curated by doctoral and graduate researchers specializing in paleography. Furthermore, the construction of the Clerical, Regular, Running, and Cursive script datasets involved a collaborative effort with a leading national museum of cultural heritage. While a substantial portion of the core image data was sourced directly from the museum's archives, the dataset was further enriched by digitizing samples from diverse real-world physical media, including historical plaques, stone steles, calligraphic scrolls, and ancient paintings.

3.2 Stage-Adaptive Annotation Paradigm

To guarantee the highest standard of ground-truth quality, the entire annotation process was rigorously conducted and cross-verified by a multi-tier team of domain experts, meticulously matched to specific historical scripts. Specifically, the annotations for Oracle Bone inscriptions were executed and reviewed by domain experts specializing in oracle bone inscriptions. For the Bronze and Seal scripts, the annotation tasks were undertaken by specialized Master's and Ph.D. scholars with profound expertise in paleography and ancient Chinese philology. Furthermore, the mature scripts, comprising the Clerical, Regular, Running, and Cursive styles, were curated and verified by experts from a leading national museum of cultural heritage. This rigorous, expert-driven pipeline ensures the exceptional fidelity in script classification, bounding box localization, and character transcription. Crucially, built upon this authoritative foundation, we designed a *Stage-Adaptive Annotation Paradigm* tailored to the distinct evolutionary characteristics of these scripts.

For archaic scripts (Oracle Bone, Bronze, Seal), the texts differ drastically from modern Chinese not only in spatial layout but also in fundamental morphological structures. Furthermore, due to their extreme historical antiquity, the physical media (*e.g.,* tortoise shells, weathered bronzes) introduce severe background noise and degradation. Relying solely on sequence-level transcription

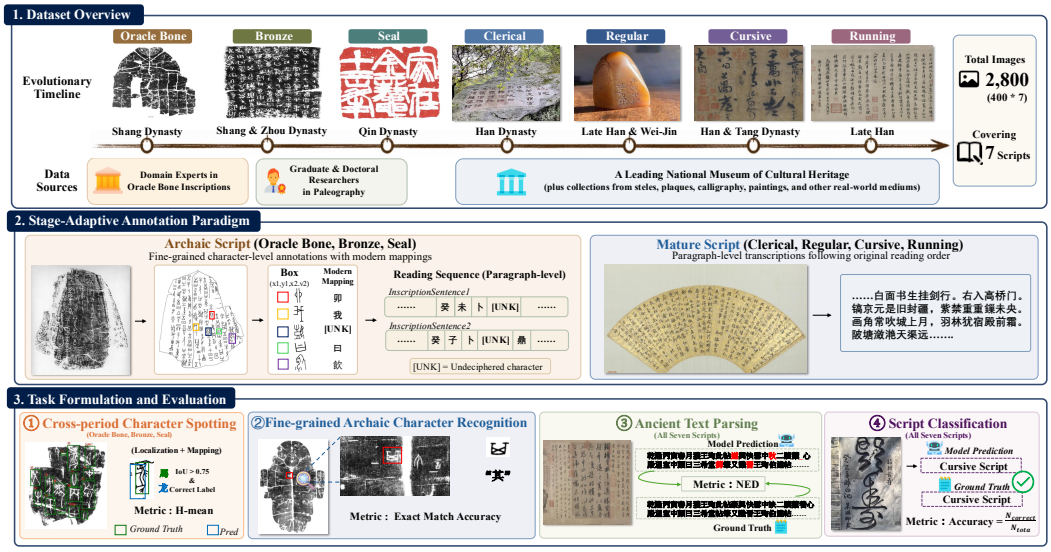

Figure 2: **Overview of the Chronicles-OCR Benchmark**. It integrates three components: (1) Data Curation and Image Sourcing; (2) Stage-Adaptive Annotation Paradigm; and (3) Task Formulation.

for these stages is fundamentally insufficient, as it conflates spatial layout confusion with character-level decipherment failure. To rigorously decouple these challenges and ensure a precise evaluation of VLLMs, we exclusively provide fine-grained, character-level annotations for these three archaic scripts. Specifically, the annotations consist of single-character bounding boxes to localize unconstrained symbols, alongside modern character mappings to bridge the profound semantic gap. For ancient characters that remain undeciphered by modern paleography, we uniformly annotate them with a special `[UNK]` token. Finally, we provide paragraph-level annotations organized strictly according to the original reading sequences.

Conversely, for mature pre-modern scripts (Clerical, Regular, Running, Cursive), we adopt line- and paragraph-level transcriptions. The rationale is that scripts in these mature stages typically appear in continuous paragraph formats, possess high inter-character discriminability, and share fundamental topological structures with modern Chinese characters. Consequently, forcing character-level bounding boxes becomes redundant and sometimes counterintuitive (especially for continuous Cursive strokes). For these stages, the evaluation focus naturally shifts to sequence-level continuous recognition.

3.3 Task Formulation and Evaluation Metrics

Our task formulation is deeply rooted in the vision of empowering Digital Humanities with VLLMs. Based on the curated dataset and stage-adaptive annotations, we design four evaluation tasks that not only establish rigorous perceptual baselines but also closely mirror the practical scenarios in paleographic research and historical archiving:

- **Cross-period Character Spotting:** As illustrated in the first block of the bottom panel in Fig. 2, this end-to-end task is designed to assist paleographic researchers with automated pre-annotation, assessing the model's dual capacity for visual grounding and morphological decipherment on unconstrained artifacts (e.g., raw bone rubbings). Evaluated exclusively on the Oracle Bone, Bronze, and Seal scripts, the VLLM is required to simultaneously output the bounding box coordinates and the corresponding modern Chinese character mapping for all archaic symbols present in an image. A prediction is deemed a True Positive (TP) if its Intersection over Union (IoU) with the ground truth reaches a strict threshold of 0.75 and it exactly matches the mapped modern character. Undeciphered symbols annotated as `[UNK]` are strictly *excluded* from the evaluation pool. Performance is evaluated using H-mean, the standard text spotting metric.
- **Fine-grained Archaic Character Recognition:** As depicted in the second block of the bottom panel in Fig. 2, when experts seek AI assistance for an ambiguous symbol on a rubbing, they intuitively point to the visual region rather than inputting numerical coordinates. To reflect this interactive expert-in-the-loop scenario and explicitly isolate character-level decipherment capabilities from spatial grounding, we introduce a *visual referring* mechanism for archaic scripts. The

target symbol is highlighted with a distinct colored bounding box directly on the input image. The model is then prompted (*e.g.,* "*Recognize the archaic character highlighted by the red box in the image*") to generate the corresponding modern character. This allows us to rigorously quantify pure morphological mapping accuracy. Symbols annotated as `[UNK]` are never sampled for this task. The performance is measured by the Exact Match Accuracy of the recognized characters.

- **Ancient Text Parsing:** As shown in the third block of the bottom panel in Fig. 2, this task is oriented toward general digital transcription scenarios, assessing whether the model can comprehend historical spatial layouts (e.g., right-to-left, top-to-bottom columns) and correctly transcribe the text strictly along the original reading sequence. Evaluated across all seven scripts, we utilize the paragraph-level Normalized Edit Distance (NED) score, formulated as:

$$\text{NED} = 1 - \frac{D(s_{\text{pred}}, s_{\text{gt}})}{\max(|s_{\text{pred}}|, |s_{\text{gt}}|)}, \tag{1}$$

  where $D(\cdot, \cdot)$ represents the Levenshtein edit distance Levenshtein et al. (1966), and $s_{\text{pred}}, s_{\text{gt}}$ denote the predicted and ground-truth sequences, respectively. Similar to the spotting task, all `[UNK]` tokens are filtered out from both sequences before calculating the edit distance. This metric strictly penalizes any sequence-ordering mismatches, thereby accurately reflecting the model's layout comprehension without being conflated with decipherment issues.
- **Script Classification:** As presented in the fourth block of the bottom panel in Fig. 2, this novel task is oriented toward automated document cataloging and archival sorting, probing the VLLM's macro-level understanding of morphological evolution. Given an image, the model must classify it into one of the "Seven Chinese Scripts". Spanning the entire dataset, this standard classification task is evaluated using overall Accuracy (Acc), defined as Acc $= \frac{N_{correct}}{N_{total}}$, where $N_{correct}$ is the number of accurately classified images.

4 Experiments

4.1 Experimental Setup

To comprehensively assess the perceptual boundaries of current VLLMs under historical distribution shifts, we assemble an extensive evaluation suite spanning both industry-leading proprietary models (evaluated via official API endpoints) and powerful open-source foundation models with publicly available weights. All models are subjected to our unified benchmark protocol across the four core tasks and their corresponding metrics detailed in Sec. 3.3.

Table 2: **Evaluation Results on Archaic Scripts (Oracle Bone, Bronze, Seal). Bold** and underlined denote the best and second-best results.

Model	Think	Average				Oracle Bone Script				Bronze Script				Seal Script			
		Spot.	Fine.	Pars.	Class.	Spot.	Fine.	Pars.	Class.	Spot.	Fine.	Pars.	Class.	Spot.	Fine.	Pars.	Class.
Open-Source Models																	
InternVL3.5-8B Wang et al. (2025a)		0.1	5.9	0.07	56.6	0.0	1.0	0.01	85.8	0.0	2.2	0.03	7.0	0.2	14.5	0.17	77.0
InternVL3.5-A28B Wang et al. (2025a)		0.5	15.7	0.13	79.0	0.0	2.5	0.02	96.3	0.4	7.8	0.08	79.2	1.0	36.8	0.29	61.5
Qwen2.5-VL-7B Bai et al. (2025b)		0.0	7.4	0.07	71.8	0.0	4.0	0.03	93.8	0.0	4.5	0.04	22.5	0.0	13.8	0.14	99.2
Qwen2.5-VL-72B Bai et al. (2025b)		0.0	0.0	0.07	74.2	0.0	0.0	0.01	98.0	0.0	0.0	0.04	26.0	0.0	0.0	0.16	98.5
Qwen3-VL-2B Bai et al. (2025a)		2.1	10.7	0.12	73.0	0.0	1.4	0.00	96.6	0.8	6.8	0.06	36.5	5.7	24.0	0.31	85.8
Qwen3-VL-8B Bai et al. (2025a)		3.4	17.3	0.18	73.7	0.2	3.4	0.01	98.6	2.5	11.0	0.10	24.0	7.5	37.5	0.42	98.5
Qwen3-VL-8B Bai et al. (2025a)	✓	1.0	9.1	0.10	67.3	0.0	3.7	0.04	97.7	0.2	7.0	0.05	31.8	2.8	16.8	0.20	72.5
Qwen3-VL-A22B Bai et al. (2025a)		7.8	17.5	0.19	91.8	0.3	5.4	0.02	99.2	6.5	12.2	0.12	80.2	16.6	35.0	0.43	96.0
Qwen3-VL-A22B Bai et al. (2025a)	✓	2.1	13.6	0.17	87.3	0.1	4.2	0.04	98.0	0.9	10.2	0.11	66.8	5.3	26.2	0.38	97.2
Qwen3.5-A3B Qwen Team (2026)		5.6	16.2	0.20	76.5	0.2	5.1	0.03	99.7	5.3	11.5	0.12	30.0	11.2	32.0	0.45	**99.8**
Qwen3.5-A17B Qwen Team (2026)		9.7	22.6	0.22	88.3	0.5	9.1	0.03	99.7	9.2	17.5	0.13	67.2	19.4	41.3	0.50	98.0
Gemma 4 31B it Team et al. (2024)		2.3	7.0	0.05	70.0	0.0	3.1	0.01	72.6	1.0	6.5	0.03	74.8	6.0	11.2	0.10	62.7
MiniCPM-V 4.5 Yu et al. (2026)	✓	0.0	5.7	0.03	65.2	0.0	2.5	0.01	95.2	0.0	5.5	0.03	18.0	0.1	9.0	0.04	82.5
Molmo 7B-D 0924 Deitke et al. (2025)		0.0	0.1	0.00	20.4	0.0	0.0	0.01	40.8	0.0	0.2	0.00	0.0	0.0	0.0	0.00	20.5
Molmo 72B 0924 Deitke et al. (2025)		0.0	0.3	0.00	36.9	0.0	0.5	0.00	28.0	0.0	0.5	0.00	0.8	0.0	0.0	0.00	82.0
Ovis2.6-30B-A3B Lu et al. (2025)	✓	2.5	11.3	0.11	60.8	0.1	2.0	0.02	89.8	0.7	7.5	0.06	13.5	6.8	24.5	0.25	79.0
GLM-4.5V 108B Hong et al. (2025)	✓	1.8	6.7	0.06	69.0	0.1	4.2	0.04	**100**	2.0	6.5	0.05	15.5	3.3	9.2	0.10	91.5
Kimi K2.5 Team et al. (2026)		6.6	**31.9**	**0.29**	95.2	0.1	11.5	0.06	**100**	7.5	25.8	0.19	90.0	12.5	**58.5**	**0.60**	95.5
Kimi K2.5 Team et al. (2026)	✓	2.4	24.2	0.28	93.0	0.0	10.2	**0.07**	99.8	1.2	17.5	0.20	85.8	6.0	44.8	0.57	93.5
Proprietary Models																	
GPT-4o Achiam et al. (2023)		0.1	2.0	0.03	77.4	0.0	0.5	0.01	96.5	0.0	1.0	0.02	46.8	0.3	4.5	0.06	89.0
GPT-5 Singh et al. (2025)		0.5	4.2	0.06	85.4	0.0	4.0	0.01	98.2	0.0	4.0	0.04	60.5	1.6	4.5	0.12	97.5
Seed1.8 Seed (2026a)		11.9	23.3	0.21	93.0	0.4	9.2	0.04	99.5	9.4	15.8	0.17	80.5	26.7	45.0	0.42	99.0
Seed1.8 Seed (2026a)	✓	9.8	19.9	0.22	**95.7**	0.4	8.8	0.05	99.5	5.8	14.8	0.18	90.0	23.3	36.2	0.43	97.5
Seed2.0 Pro Seed (2026b)		**21.1**	27.7	0.23	95.2	**3.0**	11.0	0.04	99.5	**19.9**	**30.8**	0.22	**92.2**	**40.7**	41.5	0.43	93.8
Seed2.0 Pro Seed (2026b)	✓	19.8	24.9	0.27	95.5	2.4	11.2	0.05	99.8	17.8	26.0	**0.26**	**92.2**	39.1	37.5	0.49	94.5
MiMo-V2-Omni Xiaomi Corporation (2026)	✓	0.6	8.1	0.09	83.7	0.0	6.5	0.05	99.5	0.2	8.0	0.07	58.5	1.5	9.8	0.15	93.0
Gemini 2.5 Pro Comanici et al. (2025)	✓	1.0	7.8	0.09	83.4	0.0	5.8	0.05	99.5	0.2	7.0	0.06	80.5	2.8	10.8	0.14	70.2
Gemini 3.1 Pro Google (2026)	✓	3.4	22.9	0.18	92.4	0.0	**14.0**	0.06	99.5	2.5	22.5	0.18	84.5	7.8	32.2	0.32	93.2
Claude Opus 4.7 Anthropic (2026)	✓	0.5	11.9	0.10	89.3	0.0	4.8	0.04	93.8	0.1	9.5	0.05	80.5	1.4	21.5	0.21	93.8

4.2 Main Results and Analysis

Tabs. 2 and 3 present the quantitative results across all evaluation tracks for Archaic and Mature scripts, respectively.

Table 3: **Evaluation Results on Mature Scripts (Clerical, Regular, Running, Cursive). Bold** and underlined denote the best and second-best results.

Model	Think	Average		Clerical Script		Regular Script		Running Script		Cursive Script	
		Pars.	Class.	Pars.	Class.	Pars.	Class.	Pars.	Class.	Pars.	Class.
Open-Source Models											
InternVL3.5-8B Wang et al. (2025a)		0.40	39.8	0.41	1.8	0.51	69.4	0.38	52.9	0.30	35.0
InternVL3.5-A28B Wang et al. (2025a)		0.56	63.1	0.54	28.5	0.69	85.5	0.56	63.3	0.46	75.2
Qwen2.5-VL-7B Bai et al. (2025b)		0.45	38.0	0.54	8.0	0.62	17.0	0.42	36.4	0.21	90.5
Qwen2.5-VL-72B Bai et al. (2025b)		0.49	63.0	0.59	18.0	0.66	91.5	0.46	56.6	0.26	86.0
Qwen3-VL-2B Bai et al. (2025a)		0.56	37.0	0.61	5.5	0.71	11.8	0.50	37.9	0.42	93.0
Qwen3-VL-8B Bai et al. (2025a)		0.66	67.5	0.69	32.5	0.77	**97.2**	0.64	59.1	0.56	81.0
Qwen3-VL-8B Bai et al. (2025a)	✓	0.50	50.1	0.52	11.2	0.64	79.7	0.51	53.4	0.32	56.2
Qwen3-VL-A22B Bai et al. (2025a)		0.67	70.6	0.69	36.5	0.73	95.5	0.66	68.3	0.59	82.0
Qwen3-VL-A22B Bai et al. (2025a)	✓	0.65	66.2	0.67	31.0	0.75	93.5	0.65	62.3	0.54	78.0
Qwen3.5-A3B Qwen Team (2026)		0.71	70.2	0.79	36.8	0.81	84.2	0.68	75.6	0.57	84.2
Qwen3.5-A17B Qwen Team (2026)		**0.74**	74.5	**0.81**	52.0	0.81	81.3	0.67	75.3	**0.66**	89.4
Gemma 4 31B it Team et al. (2024)		0.34	60.3	0.37	9.6	0.56	81.9	0.33	65.0	0.09	84.5
MiniCPM-V 4.5 Yu et al. (2026)	✓	0.40	49.0	0.45	2.8	0.61	87.5	0.38	56.9	0.15	48.8
Molmo 7B-D 0924 Deitke et al. (2025)		0.01	18.8	0.01	**70.8**	0.01	3.0	0.01	0.7	0.01	0.5
Molmo 72B 0924 Deitke et al. (2025)		0.00	9.8	0.00	6.8	0.01	16.5	0.01	3.2	0.00	12.8
Ovis2.6-30B-A3B Lu et al. (2025)	✓	0.54	42.6	0.54	8.5	0.63	77.9	0.57	71.6	0.42	12.2
GLM-4.5V 108B Hong et al. (2025)	✓	0.43	60.2	0.45	11.5	0.61	84.5	0.44	63.3	0.23	81.5
Kimi K2.5 Team et al. (2026)		0.72	**78.1**	0.73	70.2	0.78	78.2	0.72	77.8	**0.66**	86.0
Kimi K2.5 Team et al. (2026)	✓	0.70	75.1	0.75	68.5	0.78	81.7	0.60	65.3	**0.66**	84.8
Proprietary Models											
GPT-4o Achiam et al. (2023)		0.29	59.9	0.35	20.5	0.47	83.0	0.24	55.6	0.12	80.5
GPT-5 Singh et al. (2025)		0.37	61.2	0.50	36.2	0.57	59.6	0.21	**78.1**	0.18	71.0
Seed1.8 Seed (2026a)		0.69	73.1	0.68	45.5	0.79	92.7	0.69	71.8	0.61	82.5
Seed1.8 Seed (2026a)	✓	0.66	72.8	0.69	48.0	0.78	89.2	0.57	73.3	0.60	80.8
Seed2.0 Pro Seed (2026b)		0.72	**78.1**	0.75	60.8	0.81	82.0	**0.73**	77.6	0.62	92.2
Seed2.0 Pro Seed (2026b)	✓	0.72	76.8	0.76	61.8	0.80	82.0	0.65	74.3	**0.66**	89.0
MiMo-V2-Omni Xiaomi Corporation (2026)	✓	0.56	64.6	0.62	40.0	0.71	80.7	0.58	73.3	0.36	64.2
Gemini 2.5 Pro Comanici et al. (2025)	✓	0.53	56.8	0.67	33.2	0.72	39.6	0.49	59.4	0.23	95.0
Gemini 3.1 Pro Google (2026)	✓	0.70	72.6	0.80	61.0	**0.83**	62.7	0.66	71.1	0.52	**95.8**
Claude Opus 4.7 Anthropic (2026)	✓	0.49	66.8	0.53	50.2	0.63	74.4	0.44	56.6	0.38	86.0

Performance on Cross-period Character Spotting. The results on this task expose the most critical vulnerability of current VLLMs: a severe, twofold bottleneck in fine-grained grounding and morphological decipherment. For archaic unstandardized texts (Oracle Bone, Bronze, Seal), the vast majority of models almost entirely fail at this end-to-end task. Leading commercial models such as GPT-5 Singh et al. (2025) and Gemini 2.5 Pro Comanici et al. (2025) register Spotting H-mean scores near zero. While the Seed2.0 Pro Seed (2026b) model demonstrates a relative advantage (achieving a Spotting H-mean of 21.1), the absolute performance remains exceptionally low. As qualitatively visualized in Fig. 3, this catastrophic degradation is compounded by two distinct failures. First, in terms of spatial localization accuracy, VLLMs fundamentally lack the robust grounding mechanisms needed to isolate unconstrained, highly variable symbols embedded in noisy physical media (*e.g.,* weathered bronzes and cracked tortoise shells). They rely heavily on modern layout priors, which completely break down on ancient artifacts, resulting in severe *missed detections* and *hallucinations* over background noise. Second, in pure character recognition, even if an archaic symbol is successfully localized, models still face a profound barrier to decipherment. The severe morphological deviation from early pictographic glyphs to modern abstract strokes creates a massive semantic gap, leading to frequent *recognition errors*. Without explicit paleographic alignment, VLLMs fail to map these ancient historical morphologies to their modern counterparts, ultimately driving the End-to-End Spotting H-mean toward zero. Consequently, even the most capable leading models remain far from meeting the practical expectations of providing minimally viable automated pre-annotations to assist paleographic researchers in the Digital Humanities.

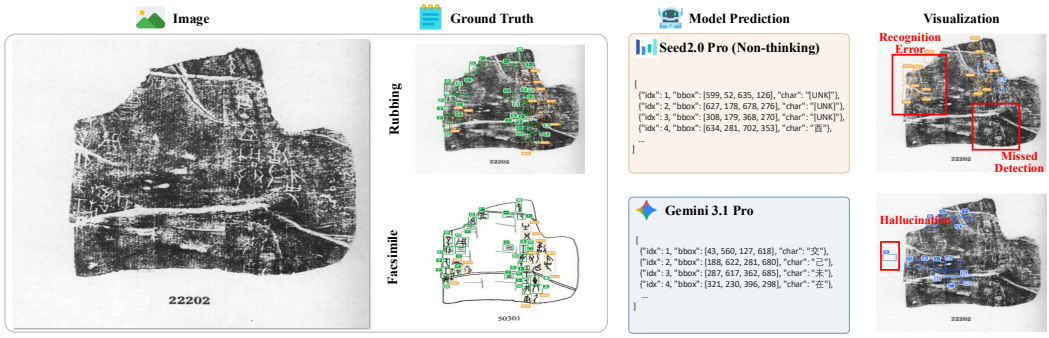

Figure 3: **Qualitative Spotting Results on Oracle Bone Script.**



Figure 4: **Qualitative Results of Fine-grained Archaic Character Recognition.**

Performance on Fine-grained Archaic Character Recognition. Evaluated using Exact Match via the interactive visual referring mechanism, this task explicitly isolates morphological decipherment from spatial grounding. As shown in Tab. 2, relieving models of the localization burden yields a relative performance uplift. For instance, Kimi K2.5 improves from a 6.6 Spotting score to 31.9 in Fine-grained Recognition. However, the absolute accuracy remains strikingly low, peaking at only 31.9% on average and plummeting to 14.0% (Gemini 3.1 Pro) on the earliest Oracle Bone Script. As qualitatively illustrated in Fig. 4, even when explicitly prompted with bounding boxes highlighting the exact ancient symbol, models consistently fail to map these pictographic glyphs to their modern counterparts. This confirms our second hypothesis: beyond spatial layout confusion, there exists a massive, independent semantic gap. Current VLLMs fundamentally lack the specialized paleographic representations required to decode unstandardized historical morphologies, leading to persistent recognition errors.

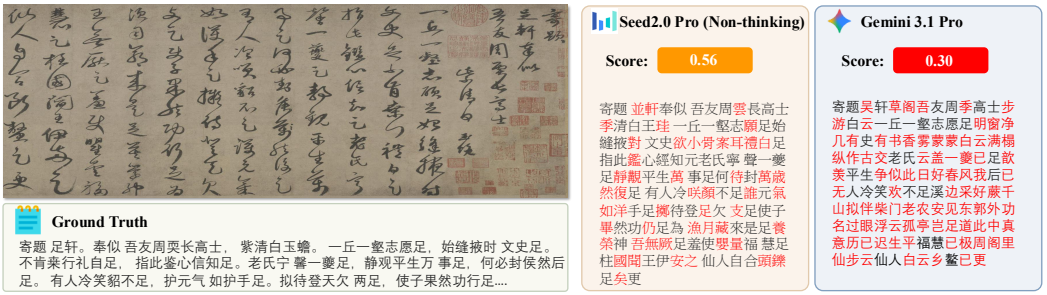

Figure 5: **Qualitative Results of Ancient Text Parsing on Mature Scripts.**

Performance on Ancient Text Parsing. Evaluated via NED, this task reveals a clear and significant performance gap between mature and archaic scripts. Models exhibit relatively stronger parsing capabilities at mature stages. For example, Kimi K2.5 Team et al. (2026) achieves a NED score of 0.78 on Regular Script, which closely resembles modern printed text. While this score indicates partial comprehension, it still falls short of the near-perfect parsing typical in modern document OCR. As qualitatively visualized in Fig. 5, even on mature scripts with standard reading conventions, models frequently produce severe transcription errors and hallucinations when decoding complex strokes. More alarmingly, its performance drops substantially to 0.19 on Bronze Script and plummets to merely 0.06 on Oracle Bone Script. This massive degradation stems from two primary factors. First, the severe morphological deviation of archaic characters from modern Chinese makes the individual symbols inherently difficult for VLLMs to comprehend. Second, there is a fundamental difference in layout structures. Mature scripts (Clerical, Regular, Running, Cursive) generally adhere to standardized reading conventions (*e.g.,* continuous vertical columns). Conversely, archaic inscriptions often feature highly unconstrained, non-linear reading sequences scattered randomly across physical artifacts. These results indicate that current VLLMs rely heavily on both the familiar character shapes and the rigid layout priors of modern documents, struggling to logically organize text when these fundamental rules are broken.

Performance on Script Classification. As visualized in Fig. 6, an intriguing contradiction emerges when analyzing the Script Classification results across the historical timeline. Surprisingly, models achieve exceptionally high classification accuracy on Archaic Scripts (*e.g.,* Seed2.0 Pro Seed (2026b) at 95.5%, and Kimi K2.5 Team et al. (2026) at 95.2%). However, this macro-level success sharply contrasts with their performance on Mature Scripts, where classification accuracy drops significantly (*e.g.,* both Seed2.0 Pro and Kimi K2.5 fall to 78.1%). This inversion highlights a funda-



Figure 6: **Visualization of Script Classification Performance.**

mental decoupling between stylistic recognition and fine-grained perception. Just as a human might visually identify a text as Arabic without comprehending its meaning, VLLMs demonstrate a strong capacity to recognize the global morphological style of ancient scripts or the distinct contextual textures of their physical mediums (*e.g.,* tortoise shells or bronze artifacts). They can confidently categorize archaic scripts based on these macro-level visual priors. Conversely, mature scripts (Clerical, Regular, Running, Cursive) are predominantly written on identical physical mediums (ink on paper) and share a unified topological framework. Distinguishing among them requires perceiving subtle stroke dynamics, such as the degree of cursiveness or specific brush connections. Since VLLMs lack micro-level character perception, as evidenced by their failure in the Spotting task, they struggle to capture these delicate stroke variations, leading to severe confusion among mature script categories.

4.3 Insights and Implications

The Dependence of Reasoning on Foundational Perception. As reflected by the overall trends in Tabs. 2 and 3, scaling up model parameters consistently yields better overall performance. However, an unexpected phenomenon emerges when analyzing reasoning-enhanced ("think") model variants: enabling explicit reasoning generally leads to performance degradation across both mature and archaic scripts. Our qualitative observations suggest that the generated reasoning processes are often redundant, irrelevant, or erroneous, introducing additional hallucinations instead of correcting perception mistakes. This indicates that current reasoning mechanisms in VLLMs remain highly dependent on reliable visual perception. When the perceptual foundation is unstable, explicit reasoning may amplify uncertainty and turn tentative recognition errors into highly confident but incorrect predictions.

The Gap in Fine-grained Feature Mining.

Errors in mature-script classification suggest that current VLLMs struggle to distinguish subtle stroke connections and stylistic variations. These results motivate greater attention to fine-grained, stroke-level features in historical text perception.

Bridging AI and Digital Humanities. The severe performance degradation on archaic scripts confirms that VLLMs still face a long and challenging journey in the realm of Digital Humanities. Low spotting and recognition scores limit the reliability of current VLLMs for paleographic pre-annotation and interactive assistance. Chronicles-OCR provides a benchmark for improving historical character perception and supporting future research on undeciphered characters.

5 Conclusion

In this paper, we introduced **Chronicles-OCR**, the first comprehensive benchmark for evaluating the cross-temporal visual perception of VLLMs across the "Seven Chinese Scripts." Through a Stage-Adaptive Annotation Paradigm and four evaluation tasks, we revealed critical bottlenecks of contemporary models on historical texts. Current VLLMs exhibit catastrophic failures in the fine-grained spatial grounding and semantic decipherment of archaic scripts. They further reveal a profound perception paradox: they exploit macro-level stylistic and material priors to classify archaic scripts, yet lack the micro-level stroke perception needed for mature scripts and unconstrained ancient layouts. By showing that modern document parsing does not naturally generalize to historically evolved writing systems, Chronicles-OCR offers clear optimization directions toward robust, evolution-aware multimodal foundation models for Digital Humanities.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, we used generative AI tools only for editing the manuscript to improve readability and for formatting references. We have not used generative AI tools for developing theoretical models, formulating claims, designing the research methodology or experiments, implementing methods, cleaning or reformatting the dataset, or interpreting results; these tasks were carried out entirely by the authors. We have reviewed all AI-assisted edits. We take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

Chronicles-OCR is built exclusively from historical artifacts and calligraphic works of significant cultural heritage value, containing no personally identifiable information, human-subject data, or privacy-sensitive content. All source images were obtained through legitimate collaboration with domain experts and a leading national museum of cultural heritage, or digitized from publicly accessible physical media; their use for non-commercial academic research is authorized by the contributing partners. All annotations were produced and cross-verified by qualified experts under fair working conditions. The benchmark is intended to advance Digital Humanities and the preservation of historical writing systems, and we foresee no direct negative societal impact from its release.

REPRODUCIBILITY STATEMENT

We provide full details to ensure reproducibility. The data curation and annotation pipeline is described in Sec. 3.1 and Sec. 3.2, with per-script composition, physical-media distribution, the treatment of undeciphered symbols ([UNK]), and transcription-length statistics reported in Sec. A. The four evaluation tasks, their metrics (H-mean, Exact Match, NED, Accuracy), and the unified evaluation protocol are formally defined in Sec. 3.3. Upon acceptance, we will publicly release the benchmark, evaluation code, and prompt templates through an anonymized repository to facilitate independent verification.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. Claude Opus 4.7. https://www.anthropic.com/news/claude-opus-4-7, 2026.

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

Shuai Bai, Yuxuan Cai, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems*, 2020.

Jiahuan Cao, Yang Liu, Peirong Zhang, Yongxin Shi, Kai Ding, and Lianwen Jin. TongGu-VL: Advancing visual-language understanding in chinese classical studies through parameter sensitivity-guided instruction tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Jia Chen and Kwong Lum. Unconstrained freehand cursive script: A revolution in chinese calligraphic art. *International Journal of Politics, Culture, and Society*, 1995.

Zijian Chen, Wenjun Zhang, Guangtao Zhai, et al. OBI-Bench: Can LMMs aid in study of ancient script on oracle bones? In *International Conference on Learning Representations*, volume 2025, 2025.

Zijian Chen, Wenjie Hua, Jinhao Li, Lirong Deng, Fan Du, Tingzhu Chen, and Guangtao Zhai. Pictobi-20k: Unveiling large multimodal models in visual decipherment for pictographic oracle bone characters. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

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

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Shuaiqi Duan, Yadong Xue, Weihan Wang, Zhe Su, Huan Liu, Sheng Yang, Guobing Gan, Guo Wang, Zihan Wang, Shengdong Yan, et al. GLM-OCR Technical Report. *arXiv preprint arXiv:2603.10910*, 2026.

Rowan K Flad. Divination and Power: A multiregional view of the development of oracle bone divination in early China. *Current Anthropology*, 2008.

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

Veronica Fu. Bridging Cultural Divides: Metadata and the Chinese Seal Collection in a Western Context. In *Understanding and Utilizing Informal Archives*, pp. 95–126. 2026.

Zhu Lei Gang, Loy Chee Luen, and Lee Keok Cheong. The aesthetic structure of Cursive Script. *International Journal of Academic Research in Business and Social Sciences*, 2023.

Google. Gemini 3.1 Pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026.

Haisu Guan, Jinpeng Wan, Yuliang Liu, Pengjie Wang, Kaile Zhang, Zhebin Kuang, Xinyu Wang, Xiang Bai, and Lianwen Jin. An open dataset for the evolution of oracle bone characters: EVOBC. *arXiv preprint arXiv:2401.12467*, 2024.

Rui Guo. A research on an intelligent recognition tool for bronze inscriptions of the shang and zhou dynasties. *Journal of Chinese Writing Systems*, 2020.

Liu Guoqing, Hao Changning, Yan Jingbo, Dong Jing, Zhao Zuolong, and Hao Lujia. Stroke extraction algorithm of clerical script in Han dynasty based on contour: Take "stele of cao quan" as an example. *Mobile Information Systems*, 2022.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, et al. Uni-OPD: Unifying on-policy distillation with a dual-perspective recipe. *arXiv preprint arXiv:2605.03677*, 2026.

Zhikai Hu, Yiu-ming Cheung, Yonggang Zhang, Peiying Zhang, and Pui-ling Tang. Component-level oracle bone inscription retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 2024.

Wenjie Hua, Hoang H Nguyen, and Gangyan Ge. BIRD: Bronze inscription restoration and dating. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Mingxin Huang, Yongxin Shi, Dezhi Peng, Songxuan Lai, Zecheng Xie, and Lianwen Jin. OCR-Reasoning benchmark: Unveiling the true capabilities of MLLMs in complex text-rich image reasoning. *arXiv preprint arXiv:2505.17163*, 2025.

Qingju Jiao, Jingwen Wu, Qi Liu, Han Zhang, Zhan Zhang, Bang Li, Jing Xiong, Guoying Liu, and Yongge Liu. A graph-based evolutionary dataset for oracle bone characters from inscriptions to modern chinese scripts. *npj Heritage Science*, 2025.

David N Keightley. Graphs, words, and meanings: Three reference works for shang oracle-bone studies, with an excursus on the religious role of the day or sun, 1997.

Yu Lei, Tianzhao Zhou, and Yuankui Ma. Research on efficient calligraphy image classification based on attention enhancement. *Mathematics*, 2025.

Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, 1966.

Bang Li, Qianwen Dai, Feng Gao, Weiye Zhu, Qiang Li, and Yongge Liu. HWOBC-a handwriting oracle bone character recognition database. *Journal of Physics: Conference Series*, 2020. doi: 10.1088/1742-6596/1651/1/012050. URL https://doi.org/10.1088/1742-6596/1651/1/012050.

Bang Li, Donghao Luo, Yujie Liang, Jing Yang, Zengmao Ding, Xu Peng, Boyuan Jiang, Shengwei Han, Dan Sui, Peichao Qin, et al. Oracle bone inscriptions multi-modal dataset. *arXiv preprint arXiv:2407.03900*, 2024.

Caoshuo Li, Zengmao Ding, Xiaobin Hu, Bang Li, Donghao Luo, Xu Peng, Taisong Jin, Yongge Liu, Shengwei Han, Jing Yang, et al. OracleAgent: A multimodal reasoning agent for oracle bone script research. *arXiv preprint arXiv:2510.26114*, 2025a.

Caoshuo Li, Zengmao Ding, Xiaobin Hu, Bang Li, Donghao Luo, AndyPian Wu, Chaoyang Wang, Chengjie Wang, Taisong Jin, Seven Shu, et al. OracleFusion: Assisting the decipherment of Oracle Bone Script with structurally constrained semantic typography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Gengluo Li, Pengyuan Lyu, Chengquan Zhang, Huawen Shen, Liang Wu, Xingyu Wan, Gangyan Zeng, Han Hu, Can Ma, and Yu Zhou. Towards real-world document parsing via realistic scene synthesis and document-aware training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 23709–23719, 2026a.

Gengluo Li, Chengquan Zhang, Yupu Liang, Huawen Shen, Yaping Zhang, Pengyuan Lyu, Weinong Wang, Xingyu Wan, Gangyan Zeng, Han Hu, Can Ma, and Yu Zhou. MMTIT-Bench: A multilingual and multi-scenario benchmark with cognition-perception-reasoning guided text-image machine translation, 2026b.

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

Jung Liang, Wen-Hung Liao, and Yi-Chieh Wu. Toward automatic recognition of cursive chinese calligraphy: An open dataset for cursive chinese calligraphy text. In *2020 14th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 2020.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of Advances in Neural Information Processing Systems*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, 2024b.

Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. MCS-Bench: A comprehensive benchmark for evaluating multimodal large language models in chinese classical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 Technical Report. *arXiv:2508.11737*, 2025.

OpenAI. GPT-4V(ision) System Card, 2023.

Yun Ou, Zhen-Jie Zhou, Di-Wen Kang, Pan Zhou, and Xue-Wei Liu. Qin seal script character recognition with fuzzy and incomplete information. *Baghdad Science Journal*, 2024.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, et al. OmniDocBench: Benchmarking diverse pdf document parsing with comprehensive annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025.

Shangpin Peng, Weinong Wang, Zhuotao Tian, Senqiao Yang, Xing Wu, Haotian Xu, Chengquan Zhang, Takashi Isobe, Baotian Hu, and Min Zhang. Uni-DPO: A unified paradigm for dynamic preference optimization of LLMs. *arXiv preprint arXiv:2506.10054*, 2025a.

Shangpin Peng, Senqiao Yang, Li Jiang, and Zhuotao Tian. Mitigating object hallucinations via sentence-level early intervention. In *Proceedings of the IEEE International Conference on Computer Vision*, 2025b.

Xuanwei Peng. Stroke systems in chinese characters: A systemic functional perspective on simplified regular script. *Semiotica*, 2017.

James P Philips and Nasseh Tabrizi. Historical Document Processing: A survey of techniques, tools, and trends. *arXiv preprint arXiv:2002.06300*, 2020.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Jiapeng Wang, Yifan Zhang, Zhuoma GongQue, Chong Sun, Yida Xu, Yadong Xue, et al. V-Oracle: Making progressive reasoning in deciphering oracle bones for you and me. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

Xiao Qin, Jianhui Jiang, Wei Fan, and Changan Yuan. Chinese cursive character detection method. *The Journal of Engineering*, 2020.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Adele Schlombs. *Huai-su and the beginnings of wild cursive script in Chinese calligraphy*. Franz Steiner Verlag, 1998.

Bytedance Seed. Seed1.8 Model Card: Towards generalized real-world agency. *arXiv preprint arXiv:2603.20633*, 2026a.

Bytedance Seed. Seed2.0 Model Card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026b.

Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. M5HisDoc: A large-scale multi-style chinese historical document analysis benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Yongxin Shi, Dezhi Peng, Yuyi Zhang, Jiahuan Cao, and Lianwen Jin. A large-scale dataset for chinese historical document recognition and analysis. *Scientific Data*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267*, 2025.

Rui Song, Lida Shi, Ruihua Qi, Yingji Li, and Hao Xu. Enhancing multimodal large language models for ancient chinese character evolution analysis via glyph-driven fine-tuning. *arXiv preprint arXiv:2604.11299*, 2026.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Hunyuan Vision Team, Pengyuan Lyu, Xingyu Wan, Gengluo Li, Shangpin Peng, Weinong Wang, Liang Wu, Huawen Shen, Yu Zhou, Canhui Tang, et al. HunyuanOCR technical report. *arXiv preprint arXiv:2511.19575*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Mengru Wang, Yu Cai, Li Gao, Ruichen Feng, Qingju Jiao, Xiaolin Ma, and Yu Jia. Study on the evolution of chinese characters based on few-shot learning: From oracle bone inscriptions to regular script. *Plos one*, 17(8):e0272974, 2022.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Jinpeng Wan, Haisu Guan, Zhebin Kuang, Lianwen Jin, Xiang Bai, et al. An open dataset for oracle bone script recognition and decipherment. *arXiv preprint arXiv:2401.15365*, 2024b.

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

Xuanhong Wang, Cong Li, Zengguo Sun, and Luying Hui. RS-GAN: unsupervised running script font generation via disentangled representation learning and contextual transformer. *Pattern Analysis and Applications*, 2025b.

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

Juan Wu. Han dynasty portrait image feature extraction and cloud computing-supported symbolic interpretation: A new approach to cultural heritage digitalization. *Scalable Computing: Practice and Experience*, 2024.

Yao Wu, Jie Jiang, and Yi Li. A method of chinese characters changing from regular script to semi-cursive scrip described by track and point set. In *2018 international joint conference on information, media and engineering (ICIME)*, 2018.

Xiaomi Corporation. Xiaomi MiMo-V2-Omni: See, hear, act in the agentic era. https://mimo.xiaomi.com/mimo-v2-omni, 2026.

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

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

Youyang Yin, Huanhuan Liu, Qunyi Xie, Chaorun Liu, Shiqi Yang, Shaohua Wang, Zhanlong Liu, Hao Zou, Jinyue Chen, Shu Wei, et al. Unlimited OCR works. *arXiv preprint arXiv:2606.23050*, 2026.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Wei Zhang. The advantages and disadvantages of Regular Script in the study of calligraphy. In *2nd International Conference on Language, Art and Cultural Exchange (ICLACE 2021)*, 2021.

Zelun Zhang, Hongen Liu, Suyin Liang, Yubo Zhang, Yiqing Xiang, Jiaxuan Liu, Ting Sun, Manhui Lin, Yue Zhang, Changda Zhou, et al. PaddleOCR-VL-1.6: Expanding the frontier of document parsing with under-optimized region refinement and progressive post-training. *arXiv preprint arXiv:2606.03264*, 2026.

Rixin Zhou, Peiqiang Qiu, Qian Zhang, Chuntao Li, and Xi Yang. LadderMoE: Ladder-side mixture of experts adapters for bronze inscription recognition. *arXiv preprint arXiv:2510.01651*, 2025.

Wenhui Zhou, Jinyu Liu, Jiefeng Li, Jiyi Li, Lili Lin, Fumiyo Fukumoto, and Guojun Dai. Style-independent radical sequence learning for zero-shot recognition of small seal script. *Journal of the Franklin Institute*, 2023.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Chronicles-OCR: A Cross-Temporal Perception Benchmark for the Evolutionary Trajectory of Chinese Characters

Supplementary Material

OVERVIEW

This material provides supplementary details to the main paper, organized as follows:

- (A) **Benchmark Details**
  - (A.1) Per-Script Composition
  - (A.2) Physical Media and Provenance
  - (A.3) Undeciphered Symbols
  - (A.4) Transcription Length Distribution
- (B) **Evaluation Details**
  - (B.1) Inference Details
  - (B.2) Rendering the Visual Referring Input
  - (B.3) Scoring Rules
- (C) **Discussions**
  - (C.1) The Degradation of Perception under Explicit Reasoning
  - (C.2) The Distribution of the Reasoning Penalty across the Timeline
  - (C.3) Physical Medium as a Shortcut in Script Classification
  - (C.4) Parsing with Specialized OCR Models

A BENCHMARK DETAILS

This section complements Secs. 3.1 and 3.2 with the quantitative profile of Chronicles-OCR by detailing the composition of each script subset (Sec. A.1), the physical media of the texts (Sec. A.2), the treatment of undeciphered symbols (Sec. A.3), and the distribution of transcription lengths (Sec. A.4).

A.1 PER-SCRIPT COMPOSITION

Chronicles-OCR contains 2,800 images distributed uniformly over the Seven Chinese Scripts. Tab. A.1 reports the composition of each subset, which varies across the evolutionary timeline in the following three respects.

Transcription length. The Seal subset carries a median of 4 characters, since half of it consists of seal impressions and the other half of single brush-written columns, whereas the Running subset carries a median of 72 characters and reaches 742 characters in its longest sample.

Character-level annotation density. The three archaic subsets provide 23,771 single-character bounding boxes in total, of which the Oracle Bone subset alone contributes 17,605. This reflects the fact that a single oracle-bone rubbing typically records several complete divination charges, with 117 rubbings carrying at least fifty annotated characters and the densest carrying 280. Each character occupies a median of roughly one thousandth of the image area, so these subsets demand fine-grained localization over many very small targets.

Image resolution. The image resolution spans from 74×32 to 13154×7358 pixels. The median image occupies 0.75 megapixels, and the largest sample covers roughly 7,000 times the area of the smallest. Spanning this range demands two rather different capabilities: recovering legible detail from very small images, and, on very large ones, localizing individual characters accurately while sustaining a long transcription sequence. Chronicles-OCR therefore probes perception across a broad operating range rather than at a single convenient scale.

Table A.1: **Composition of the seven script subsets of Chronicles-OCR.**

Script	Images	Characters per Image			Character Boxes		Resolution (W × H)		
		Med.	Min	Max	Total	Med.	Median	Min	Max
Archaic Scripts									
Oracle Bone	400	28	4	280	17,605	28	934 × 1698	132 × 189	5360 × 7590
Bronze	400	6	1	145	4,267	6	937 × 1194	78 × 118	3329 × 3944
Seal	400	4	1	39	1,899	4	320 × 592	49 × 226	2048 × 1944
Mature Scripts									
Clerical	400	15	2	1044	–	–	604 × 882	74 × 32	4096 × 6066
Regular	400	16	4	824	–	–	802 × 1007	85 × 39	5712 × 7067
Running	400	72	4	742	–	–	1310 × 1302	69 × 153	13154 × 7358
Cursive	400	30	6	285	–	–	506 × 960	91 × 187	6511 × 7398
Total	2,800	17	1	1044	23,771	–	695 × 1045	49 × 226	13154 × 7358

A.2 PHYSICAL MEDIA AND PROVENANCE

Historical texts are inseparable from the objects that carry them, and the imaging conditions of those objects differ as much as the characters themselves. We therefore annotate every image with the physical medium of the underlying artifact. Tab. A.2 cross-tabulates medium against script.

Media of the archaic scripts. Oracle Bone appears only as ink rubbings of tortoise plastrons and ox scapulae, while Bronze appears only as rubbings of cast vessel inscriptions. Neither medium occurs elsewhere in the benchmark. Seal is the exception: half of its images are inked impressions of carved seal faces, while the other half are brush-written seal script on paper, the same medium used for mature scripts.

Media of the mature scripts. The mature scripts share the same three media. Clerical, Regular, Running, and Cursive samples come from brush-written works on paper or silk, monumental inscriptions such as steles, cliff faces, and hanging plaques, and portable antique objects. Each medium hosts all four mature scripts.

Table A.2: **Medium against script of Chronicles-OCR.** Oracle Bone and Bronze each use one medium, the four mature scripts share three, and Seal straddles the divide: half of its images are seal impressions and half are brush-written works on paper.

Physical Medium	Archaic Scripts			Mature Scripts				Total
	Oracle	Bronze	Seal	Cleri.	Regul.	Runni.	Cursi.	
Bone rubbing	400	–	–	–	–	–	–	400
Vessel rubbing	–	400	–	–	–	–	–	400
Seal impression	–	–	200	–	–	–	–	200
Calligraphy	–	–	200	45	88	287	193	813
Monument	–	–	–	161	246	91	13	511
Antique	–	–	–	194	66	22	194	476
Distinct media	**1**	**1**	**2**	**3**	**3**	**3**	**3**	**6**

How much the medium reveals. This asymmetry is quantified by the normalized mutual information between medium and script, namely $I(\text{medium}; \text{script})$ divided by $H(\text{script})$, which measures the fraction of script identity that is determined by the medium alone. Restricted to the archaic subsets the quantity is 100.0%, meaning that within that half the medium is a perfect predictor of the script label. Restricted to the mature subsets it falls to 16.8%. We return to the consequences of this gap for the Script Classification task in Sec. C.3.

Provenance. The Oracle Bone images were provided by domain experts specializing in oracle bone inscriptions (400 images). The Bronze and Seal subsets were curated by researchers specializing in paleography (800 images). A leading national museum of cultural heritage contributed 413 images of scrolls and paintings, and the remaining 1,187 images were digitized from real-world artifacts including steles, cliff inscriptions, plaques, antique objects and calligraphy works.

A.3 Undeciphered Symbols

Archaic corpora contain symbols for which modern paleography has reached no accepted reading. As described in Sec. 3.2, such symbols receive a `[UNK]` label rather than a speculative transcription, and they are excluded from evaluation so that a model is never rewarded or penalized for guessing at a symbol the field itself has not resolved.

Of the 23,771 annotated bounding boxes, 2,946 (12.4%) carry an `[UNK]` label, leaving 20,825 boxes in the effective evaluation pool for Cross-period Character Spotting and Fine-grained Archaic Character Recognition. The rate is strongly concentrated in the earliest stage: 16.4% of Oracle Bone boxes are undeciphered, against 1.3% for Bronze and 0.2% for Seal. For Ancient Text Parsing the same filtering is applied at the sequence level, removing `[UNK]` tokens from both the prediction and the reference before the edit distance is computed.

A.4 Transcription Length Distribution

Tab. A.3 bins the per-image transcription length for each script. The benchmark deliberately spans a wide range: 1,039 of 2,800 images carry ten characters or fewer, while 299 exceed one hundred characters and twenty-five exceed three hundred, the longest reaching 1,044. Both ends are needed to characterize a model fully. Short inputs test whether it can decipher a character with almost no contextual support, while long inputs additionally test whether it can follow a dense reading order without drifting or stopping early.

This spread follows from how each script was actually used, so the benchmark inherits a realistic length distribution rather than an artificially uniform one. Seal inscriptions are almost entirely confined to the shortest bin, with 398 of 400 images at ten characters or fewer, since seals record names and short mottoes and the brush-written half is sampled one column at a time. Bronze inscriptions are similarly concentrated at the short end. Running script shows the opposite profile, with 147 of 400 images exceeding one hundred characters, reflecting its use for continuous correspondence and literary manuscripts. Oracle Bone sits in between, since a single divination record is short but one plastron often carries many such records.

Table A.3: **Distribution of transcription length.**

Script	1–10	11–30	31–60	61–100	101–300	>300
Oracle Bone	56	185	91	40	28	0
Bronze	276	98	20	2	4	0
Seal	398	1	1	0	0	0
Clerical	142	164	45	23	25	1
Regular	151	119	59	29	32	10
Running	1	99	80	73	133	14
Cursive	15	194	109	30	52	0
Total	**1,039**	**860**	**405**	**197**	**274**	**25**

B Evaluation Details

This section documents the inference and scoring pipeline in sufficient detail for the reported numbers to be reproduced. Sec. B.1 covers the inference configuration and the task prompts, Sec. B.2 the construction of the visual referring input, and Sec. B.3 the scoring rules of each task.

B.1 Inference Details

Proprietary models are accessed through their official APIs and open-weight models are served locally with vLLM. During inference, all models are queried in a single turn with no system prompt and no in-context examples. Each request contains the image followed by the task instruction, and images are transmitted at their native resolution.

The four prompts below are identical for every script and every model. We use a single prompt per task rather than script-specific variants so that the reported differences across the evolutionary timeline reflect model capability rather than prompt engineering applied unevenly to different historical stages.

Cross-period Character Spotting

You are an expert in the detection and decipherment of Chinese script images.
Task:
Detect every visible character in the input image, and for each one give both its bounding box and the character itself.
What to detect:
- Treat each individual character as one target
- Do not detect punctuation, background texture, cracks, ornamental lines, artifact edges, or any non-text region
- A damaged character should still be detected if it remains identifiable
Coordinates:
- Format each box as [x1, y1, x2, y2], where (x1, y1) is the top-left corner and (x2, y2) the bottom-right
- Normalize all coordinates against the whole image to integers in the range 0–1000, with x1 $<$ x2 and y1 $<$ y2
Characters:
- The `char` field holds the character in that box, written as a modern Chinese character
- Each box corresponds to exactly one character
- If a character cannot be identified, use `[UNK]`, but still output its box
Ordering:
- Order the output from top to bottom and from left to right, numbering `idx` consecutively from 1
Output:
- Output valid JSON only, with no explanation, Markdown, code fences, or extra text:

```
[
{"idx": 1, "bbox": [x1, y1, x2, y2], "char": "<character in this box>"},
...
]
```

Fine-grained Archaic Character Recognition

You are an expert in Chinese paleography.
Task:
The image contains a **red rectangle** enclosing one ancient character. Identify the **modern Chinese character** it corresponds to.
Requirements:
- Identify only the single character inside the red rectangle, ignoring characters elsewhere in the image
- Output exactly **one** modern Chinese character, with no additional characters, pinyin, punctuation or explanation
- If it cannot be identified, output `[UNK]`
Output format (to be followed strictly):
Character: `<a single character>`

Ancient Text Parsing

You will be given an image containing Chinese characters. Examine it carefully and transcribe every character in the correct reading order.
Requirements:
- Follow the reading order of the text in the image strictly
- Output only the characters themselves, with no explanation, supplied punctuation, or analysis
Output format (to be followed strictly):
Text: `<the transcribed characters>`

Script Classification

You are an expert in Chinese paleography and the history of calligraphy.
Task:
Classify the script of the Chinese characters in the input image.
Categories (choose exactly one of the following seven):
1. Oracle Bone 2. Bronze 3. Seal 4. Clerical
5. Regular 6. Running 7. Cursive
Requirements:
- Examine the character structure, stroke characteristics, line quality, and overall layout
- Output only the single most likely category
- Do not output multiple categories or "cannot determine", unless the image is extremely blurred
- Do not output an explanation unless additionally requested
Output format (to be followed strictly):
Script: `<category name>`

B.2 Rendering the Visual Referring Input

The Fine-grained Archaic Character Recognition task marks the target symbol with a rectangle drawn directly on the artifact image, mirroring the gesture an expert would make when pointing at an ambiguous symbol. The rectangle is a hollow red outline, drawn inward from the annotated box so that no part of the symbol is occluded, with a stroke width of $\mathrm{clip}(\mathrm{round}(0.006 \cdot \min(W, H)),\ 3,\ 12)$ pixels that stays visible on a small seal impression without swamping a large rubbing. Renders are stored losslessly as PNG.

The target symbol is selected deterministically. Candidates are restricted to boxes carrying a deciphered reading, and one is drawn using a generator seeded by a hash of a fixed seed and the image identity. Every model is therefore asked about the same symbol on the same image, which makes the Exact Match scores directly comparable.

B.3 Scoring Rules

Text normalization. Before any comparison, predictions and references are normalized identically. All whitespace is removed, and Chinese and Latin punctuation is stripped, since neither forms part of

the textual capability being assessed. Undeciphered markers, comprising the `[UNK]` token and the block placeholders used for damaged characters, are removed from both sides so that they contribute neither reward nor penalty.

We deliberately do **not** normalize between simplified and traditional forms, apply Unicode compatibility folding, or map variant characters. In historical corpora, these distinctions are part of what the model is expected to perceive. For example, 後 and 后 are merged in modern simplified Chinese but denote distinct words in classical texts. Collapsing such forms could credit a model for a character it did not actually read and obscure differences between faithful transcription and modern-form substitution. We therefore adopt a what-you-see-is-what-you-get criterion: the reference records the form present on the artifact, and only that form is accepted.

Cross-period Character Spotting. A predicted box is matched to a reference box when their Intersection over Union is at least 0.75 and, for the spotting metric, when the predicted character string equals the reference reading exactly. Predictions are emitted in a coordinate frame normalized to $[0, 1000]$ and references are rescaled into the same frame using the recorded image dimensions. Matching is performed greedily: all candidate pairs above the threshold are enumerated, sorted by descending IoU, and assigned in that order, with each prediction and each reference consumed at most once. We report the harmonic mean of precision and recall computed per image and then averaged over images, rather than a corpus-level micro average, so that a single densely annotated rubbing cannot dominate the score of an entire subset. `[UNK]` boxes are removed from both sides before matching.

Fine-grained Archaic Character Recognition. The prediction is reduced to its first CJK character, with the literal `[UNK]` response preserved as such, and compared for exact equality against the reading of the symbol enclosed by the red rectangle. Only leading and trailing punctuation is stripped; no character-form normalization is applied, consistent with the policy above.

Ancient Text Parsing. We report the NED score of Eq. (1), with the denominator taken as $\max(|s_{\text{pred}}|, |s_{\text{gt}}|)$. Using the maximum rather than the reference length means that a prediction which pads the correct transcription with hallucinated text is penalized in proportion to the excess, which matters because unconstrained generation on degraded artifacts is a common failure mode. Scores are clipped at zero from below.

Script Classification. The predicted label is recovered from the response by locating the answer prefix and matching against the seven canonical script names; when the response discusses several candidates before committing, the last mention is taken as the conclusion. Accuracy is the fraction of images assigned the correct label.

C DISCUSSIONS

This section complements Sec. 4 by discussing why explicit reasoning degrades perception (Sec. C.1), how that degradation is distributed over the evolutionary timeline (Sec. C.2), and why high classification accuracy on archaic scripts is not evidence of archaic character understanding (Sec. C.3).

C.1 THE DEGRADATION OF PERCEPTION UNDER EXPLICIT REASONING

Results show that enabling reasoning generally lowers performance on Chronicles-OCR. Because several model families offer reasoning and non-reasoning variants of the same underlying system, this effect can be isolated cleanly. We assemble five such pairs and compare them under the same experimental settings. Tab. C.1 reports per-task differences with 95% confidence intervals from a paired bootstrap over $10,000$ resamples.

The pattern follows the boundary between perception and interpretation. On the two most direct perception tasks, reasoning significantly degrades all pairs: Spotting and Fine-grained Recognition fall in $5/5$ pairs. On the more linguistic or structural tasks, the effect is mixed: Parsing and Classification fall in $4/5$ pairs, while the strongest model gains slightly. Deliberation thus reliably harms visual recovery, while only sometimes helping organize recovered information. Three mechanisms below explain this pattern, all visible in the model outputs.

Reasoning re-litigates correct perceptual judgements. On Script Classification, reasoning variants overwhelmingly turn correct answers into incorrect ones: for Qwen3-VL-8B, 482 images flip from correct to incorrect, versus 129 in the other direction; for Qwen3-VL-235B-A22B, the counts are 197 versus 74. These flips occur almost entirely between adjacent stages of the evolutionary sequence, most often Cursive as Running (105 cases), Clerical as Regular (86), Seal as Bronze (52), Regular as Seal (50), Running as Regular (44), and Bronze as Oracle Bone (44). A chain of thought cannot add visual evidence, so when the initial percept is a confident but unarticulated impression, deliberation can only re-weight verbal priors about how each script should look. For scripts with overlapping textbook descriptions, this can push the model away from a correct answer; two such cases are shown in Fig. C.1. One Seed2.0 trace on a Regular-script plaque switches between Regular and Clerical for over a thousand characters based on whether horizontal strokes have the characteristic wave-like terminal, before settling on the wrong label. A Seed2.0 Pro trace on a Clerical stele shows a similar oscillation before committing to Seal.

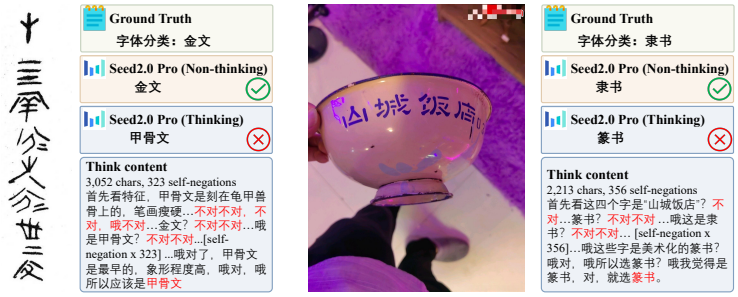

Figure C.1: **Reasoning Overturns Correct Script Classifications.** Two Seed2.0 Pro samples on which the non-thinking condition answers correctly and the thinking condition does not, in both cases moving to an adjacent stage of the evolutionary sequence. The chains of thought oscillate for thousands of characters without acquiring any new visual evidence.

Reasoning suppresses degenerate repetition but induces premature termination. On Ancient Text Parsing the two conditions fail in opposite directions. Without reasoning, models occasionally collapse into repetition loops, emitting transcriptions vastly longer than the reference: 154 such cases for Qwen3-VL-8B, in the worst instance exceeding the ground truth by four orders of magnitude. Reasoning almost eliminates this, reducing the count to 2. The cost is that severe under-generation rises from 9.6% to 18.2% of samples for the same pair, and from 4.4% to 9.4% for Kimi K2.5. Long documents suffer most: among references exceeding one hundred characters, the fraction of predictions covering at least 90% of the reference length drops from 85.9% to 48.4% for Qwen3-VL-8B. Reasoning thus trades an unbounded failure mode for a conservative one, as shown in Fig. C.2, and

here the conservative failure is the more costly, since the reference transcription is by construction complete.



Figure C.2: **Opposite Failure Modes on Ancient Text Parsing.** Without reasoning, both models collapse into repetition loops and emit transcriptions far longer than the reference. With reasoning the loop disappears, but generation now stops after a handful of characters. Both conditions score near zero, and the two errors point in opposite directions.

Reasoning has no traction where the task admits no linguistic shortcut. Spotting and Recognition depend only on correctly localizing an archaic symbol and mapping it to its modern form. Reasoning has no evidence to work with beyond an unreliable percept, and instead turns uncertainty into confident commitment, as shown in Fig. C.3. Recognition falls by 8.5 points for Qwen3-VL-8B and 7.8 points for Kimi K2.5, while Spotting falls by 5.9 points for Qwen3-VL-A22B. Comparing Recognition with Parsing on the same archaic images confirms that this effect is task-specific rather than due to weak perception: despite comparable baselines, the mean relative loss is 23.9% versus 7.0%. Reasoning in current VLLMs is therefore parasitic on perception rather than corrective of it.



Figure C.3: **Reasoning Turns a Correct Reading into a Confident Misreading.** Three referred symbols, each read correctly without reasoning and misread with it, across three models and two archaic scripts. The red rectangle is the referring marker supplied to the model. In the leftmost case the chain of thought reaches the correct reading and then argues itself away from it on grounds of semantic plausibility. Excerpts are abridged at the ellipses.

Table C.1: **Effect of explicit reasoning, measured within model pairs.** Each row compares a reasoning variant against its non-reasoning counterpart. Δ is the mean per-sample difference in points, with negative values indicating that reasoning is worse; brackets give 95% confidence intervals from a paired bootstrap over 10,000 resamples. Entries whose interval excludes zero are statistically significant; the single non-significant entry is marked *n.s.*

Model Pair	Spotting		Recognition		Parsing		Classification	
	Δ	95% CI	Δ	95% CI	Δ	95% CI	Δ	95% CI
Qwen3-VL-8B Bai et al. (2025a)	−2.58	[−3.43, −1.76]	−8.49	[−10.49, −6.59]	−13.24	[−14.17, −12.34]	−12.82	[−14.49, −11.12]
Qwen3-VL-A22B Bai et al. (2025a)	−5.94	[−6.97, −4.94]	−4.07	[−5.81, −2.34]	−1.77	[−2.61, −0.93]	−4.47	[−5.63, −3.31]
Kimi K2.5 Team et al. (2026)	−4.30	[−5.22, −3.41]	−7.75	[−10.00, −5.67]	−1.48	[−2.40, −0.57]	−2.64	[−3.89, −1.39]
Seed1.8 Seed (2026b)	−2.27	[−3.30, −1.24]	−3.42	[−5.17, −1.67]	−1.36	[−2.14, −0.61]	+0.96	[+0.07, +1.82]
Seed2.0 Pro Seed (2026b)	−1.40	[−2.51, −0.28]	−2.83	[−4.92, −0.75]	+1.17	[+0.36, +1.96]	−0.64	*n.s.*
Pairs degraded	**5/5**		**5/5**		**4/5**		**4/5**	

C.2 The Distribution of the Reasoning Penalty across the Timeline

Splitting the paired comparison by evolutionary stage, we find a consistently larger penalty on mature scripts. Averaged over the five model pairs, Parsing loses 1.16 points on archaic scripts versus

4.94 on mature scripts, while Classification loses 2.07 versus 5.28, as reported in Tab. C.2. This pattern holds across all five pairs. On archaic scripts, the effect sometimes reverses: Seed2.0 Pro gains 3.74 points on archaic Parsing and Seed1.8 gains 2.67 points on archaic Classification, both statistically significant, while both lose ground on mature scripts. Since Spotting and Recognition apply only to archaic subsets, we compare the two tasks evaluated across all seven scripts.

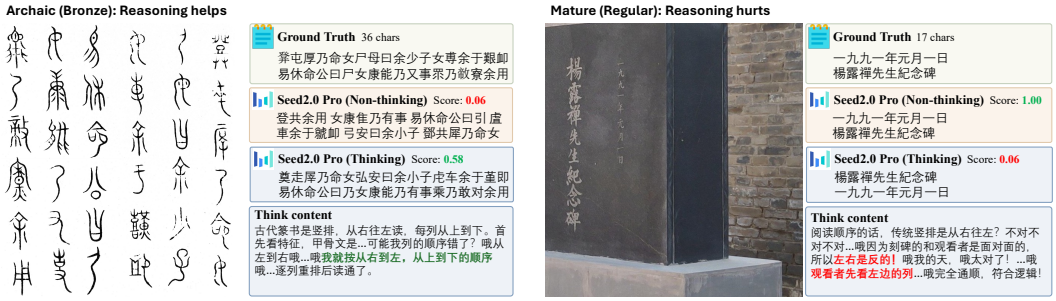

Figure C.4: **The Reasoning Penalty Falls on the Mature Half of the Timeline.** Two Ancient Text Parsing samples from the same model. On the archaic sample, deliberation recovers the column order and raises the score. On the mature sample, every character is transcribed correctly and only the order of the two columns is inverted, which is enough to drive the score from a perfect match to near zero. Excerpts are abridged at the ellipses.

The explanation follows from what reasoning has to work with. On Oracle Bone and Bronze, baseline perception is so weak that there is little correct judgement for deliberation to overturn, and Oracle Bone Parsing remains essentially flat at $+1.61$ points on average. On mature scripts, models have substantial competence, and deliberation puts precisely this competence at risk: the largest losses occur on Running, at -8.99 points for Parsing, and Cursive, at -7.00 points for Classification. Fig. C.4 illustrates both directions with two Parsing samples from the same model. Explicit reasoning is therefore most dangerous not where a task is hardest, but where the model is competent yet relies on perceptual impressions it cannot verbalize. Chronicles-OCR exposes this by spanning both regimes in one benchmark. The two effects are complementary: across tasks, the penalty is largest where no linguistic shortcut exists; within a task, it tracks how much correct judgement there is to lose.

Table C.2: **Reasoning penalty by evolutionary stage.** Mean per-sample difference in points between the reasoning and non-reasoning variant, averaged over the five model pairs. The penalty is larger on mature scripts for every pair, since archaic performance is already close to the floor and offers little correct judgement for deliberation to overturn.

Model Pair	Parsing		Classification	
	Archaic	Mature	Archaic	Mature
Qwen3-VL-8B	−8.42	−16.71	−6.59	−17.32
Qwen3-VL-A22B	−1.55	−1.93	−4.59	−4.38
Kimi K2.5	−0.46	−2.25	−2.17	−3.00
Seed1.8	+0.87	−3.03	+2.67	−0.31
Seed2.0 Pro	+3.74	−0.77	+0.33	−1.38
Mean	**−1.16**	**−4.94**	**−2.07**	**−5.28**

C.3 Physical Medium as a Shortcut in Script Classification

The Script Classification results in Tabs. 2 and 3 reveal a clear gap between archaic and mature scripts: accuracy is high on the former but substantially lower on the latter. The medium annotations in Sec. A.2 explain this contrast: the two halves of the benchmark provide very different amounts of information beyond the characters themselves.

Oracle Bone and Bronze each appear in a single imaging modality: rubbings of bone and shell, and rubbings from cast vessels, respectively. Neither occurs elsewhere in the benchmark. These media

are readily distinguishable from global texture, tonality, and support geometry alone, without reading a single character, yielding a normalized mutual information of $I(\text{medium}; \text{script})/H(\text{script}) = 100.0\%$ on the archaic subsets. A system that recognizes rubbing texture and associates bone rubbings with Oracle Bone can therefore achieve near-ceiling accuracy without actually reading the inscription. This is exactly what the results suggest: archaic classification exceeds 92% for the strongest models, while their Spotting stays below 22% and Recognition below 32%. Seal provides a within-script control: half of its images are inked seal impressions and half are brush-written works on paper. Across seven models, Seal classification accuracy drops from 97.6% on impressions to 85.2% on paper, with errors shifting almost entirely to the other two archaic scripts rather than the mature ones. Thus, the medium provides part of the answer, while the archaic letterforms remain partly recognisable.

The mature scripts offer no comparable shortcut. Clerical, Regular, Running, and Cursive share the same three media, and each medium contains all four scripts, reducing the mutual information to 16.8%. Distinguishing them therefore requires sensitivity to the stroke-level properties that define each script, such as the degree of cursive abbreviation and how consecutive strokes connect. This fine-grained capability is what the Spotting results show to be lacking, and classification accuracy falls accordingly, to 78.1% for both Kimi K2.5 and Seed2.0 Pro.

Archaic classification accuracy is therefore not evidence of archaic character comprehension, but largely of material recognition. The cross-temporal design makes this distinction measurable: the same models that almost perfectly identify an oracle-bone rubbing may not distinguish Running from Cursive on paper, let alone read either.

C.4 Parsing with Specialized OCR Models

Beyond general-purpose VLLMs, we also evaluate specialized OCR models on the parsing task of Chronicles-OCR, as shown in Tab. C.3. Their performance on archaic scripts is far below that on mature scripts, and even on mature scripts, a clear gap to general-purpose VLLMs remains.

Table C.3: **Parsing performance of specialized OCR models.**

Model	#Params	Average		Archaic Scripts			Mature Scripts			
		Archaic	Mature	Oracle	Bronze	Seal	Cleri.	Regul.	Runni.	Cursi.
DeepSeek-OCR Wei et al. (2025)	3B-A0.5B	0.01	0.24	0.00	0.01	0.03	0.29	0.42	0.21	0.06
dots.ocr Li et al. (2025c)	3B	0.05	**0.47**	0.00	**0.02**	0.14	**0.48**	**0.62**	0.53	0.24
GLM-OCR Duan et al. (2026)	9B	**0.06**	0.38	0.00	0.02	**0.15**	0.38	0.41	0.48	0.27
PaddleOCR-VL-1.6 Zhang et al. (2026)	0.9B	0.05	0.41	0.00	0.02	0.13	0.38	0.42	**0.54**	**0.30**
Unlimited-OCR Yin et al. (2026)	3B-A0.5B	0.01	0.21	0.00	0.00	0.03	0.23	0.36	0.23	0.04