MCS-Bench: A Comprehensive Benchmark for Evaluating Multimodal Large Language Models in Chinese Classical Studies

Yang Liu[†1,3], **Jiahuan Cao**[†1,3], **Hiuyi Cheng**[1,3], **Yongxin Shi**[1,3], **Kai Ding**[2,3], **Lianwen Jin**[*1,3,4]

[1]South China University of Technology [2]INTSIG Information Co., Ltd

[3]INTSIG-SCUT Joint Lab on Document Analysis and Recognition

[4]SCUT-Zhuhai Institute of Modern Industrial Innovation

ly10061105@gmail.com, jiahuanc@foxmail.com, eechenghiuyi1@mail.scut.edu.cn
yongxin_shi@foxmail.com, danny_ding@intsig.net, eelwjin@scut.edu.cn

Abstract

With the rapid development of Multimodal Large Language Models (MLLMs), their potential in Chinese Classical Studies (CCS), a field which plays a vital role in preserving and promoting China's rich cultural heritage, remains largely unexplored due to the absence of specialized benchmarks. To bridge this gap, we propose **MCS-Bench**, the first-of-its-kind multimodal benchmark specifically designed for CCS across multiple subdomains. MCS-Bench spans seven core subdomains (***Ancient Chinese Text***, ***Calligraphy***, ***Painting***, ***Oracle Bone Script***, ***Seal***, ***Cultural Relic***, and ***Illustration***), with a total of 45 meticulously designed tasks. Through extensive evaluation of 37 representative MLLMs, we observe that even the top-performing model (InternVL2.5-78B) achieves an average score below 50, indicating substantial room for improvement. Our analysis reveals significant performance variations across different tasks and identifies critical challenges in areas such as Optical Character Recognition (OCR) and cultural context interpretation. MCS-Bench not only establishes a standardized baseline for CCS-focused MLLM research but also provides valuable insights for advancing cultural heritage preservation and innovation in the Artificial General Intelligence (AGI) era. Data and code will be publicly available.

1 Introduction

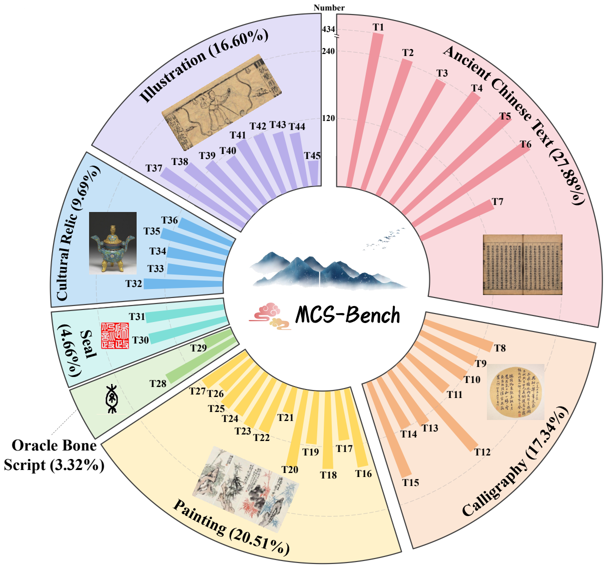
In recent years, the development of Multimodal Large Language Models (MLLMs) has significantly advanced visual and language understanding (Yin et al., 2023; Wu et al., 2024a), offering vast research and application potential in fields such as cultural heritage and cross-cultural communication (Li et al., 2024a; Zhang et al., 2024b). However, the capabilities of MLLMs in Chinese Classical Studies (CCS) remains largely underexplored. This field encompasses a wealth of cultural assets, including ancient Chinese texts, calligraphy, and painting, and plays a vital role in preserving and promoting China's rich cultural heritage. The convening of the World Classical Studies Conference in November 2024 further underscores the growing attention to this domain (Xu et al., 2024b). Due to the lack of targeted benchmarks, the evaluation of MLLMs' capabilities in CCS remains underdeveloped.

To fill this gap, we propose **MCS-Bench**, a comprehensive benchmark designed to evaluate MLLMs' performance across seven core domains in CCS: ***Ancient Chinese Text***, ***Calligraphy***, ***Painting***, ***Oracle Bone Script***, ***Seal***, ***Cultural Relic***, and ***Illustration***. Unlike existing benchmarks (Liu et al., 2021; Chiu et al., 2024; Vayani et al., 2024), which primarily target modern cultural contexts or a single specific subdomain within CCS, MCS-Bench is the first multimodal benchmark to evaluate MLLMs across diverse CCS subdomains. MCS-Bench features a diverse range of tasks that evaluate critical abilities such as cultural relic introduction, calligraphy and painting understanding and Optical Character Recognition (OCR) of ancient document and seal. Specifically, MCS-Bench offers three key advantages:

(a) Comprehensiveness and Diversity: As depicted in Figure 1 and Table 1, MCS-Bench focuses on seven core areas within CCS, featuring 45 fine-grained tasks that significantly exceed other datasets. These tasks encompass both Multiple-Choice Questions (MCQ) and open-ended Q&A, providing a comprehensive evaluation of the model's ability to understand complex cultural contexts. **(b) Complexity**: The CCS involves classical literature and artistic works with cultural and historical complexity. MCS-Bench includes high-difficulty tasks such as OCR for ancient texts and calligraphy appreciation, which can significantly challenge the limit capabilities of existing

[†]Equal contribution
[*]Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10435–10492
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



Ancient Chinese Text

T1	Text OCR
T2	Classical Chinese to Modern Chinese
T3	Punctuation
T4	Named Entity Recognition
T5	Word Explanation
T6	Reading Comprehension
T7	Source Attribution

Calligraphy

T8	Calligraphy OCR
T9	Calligraphy Appreciation
T10	Author Identification
T11	Dynasty Identification
T12	Title Identification
T13	Font Identification
T14	Genre Identification
T15	Work Introduction

Oracle Bone Script

T28	Oracle Bone Script OCR
T29	Pictographic Decoding

Cultural Relic

T32	Name Identification
T33	Dynasty Identification
T34	Collection Identification
T35	Cultural Relic Introduction
T36	Cultural Relic Classification

Painting

T16	Material Identification
T17	Painting Appreciation
T18	Author Identification
T19	Dynasty Identification
T20	Title Identification
T21	Image-based Q&A
T22	Background Introduction
T23	Painting Technique
T24	Source Identification
T25	Painting Description
T26	Painting OCR
T27	Work Introduction

Seal

T30	Seal OCR
T31	Owner Identification

Illustration

T37	Illustration OCR
T38	Illustration Description
T39	Entity Introduction
T40	Image-to-Poem Matching
T41	Source Identification
T42	Topic Classification
T43	Image-to-Person Matching
T44	Plot Introduction
T45	Image-based Q&A

Figure 1: Overview of MCS-Bench, covering 7 subdomains and 45 fine-grained tasks.

MLLMs. **(c) Motivating Research**: MCS-Bench offers detailed task design, extensive performance metrics, and in-depth result analysis, fostering research and application of MLLMs in CCS.

We evaluate 37 mainstream MLLMs, including closed-source models such as GPT-4o (OpenAI, 2024), Gemini-2.0-Flash (Google, 2024), and open-source models like InternVL series (Chen et al., 2024b; OpenGVLab, 2024) and QwenVL series (Wang et al., 2024a; Bai et al., 2023). The experimental results indicate that current MLLMs still have significant room for improvement in CCS. The top-performing InternVL2.5-78B achieves an average score below 50. We also provide an in-depth analysis of performance across different tasks (see Section 4 and Appendix E). Furthermore, through supplementary experiments, we find that explicitly decomposing recognition and understanding steps can improve performance in certain tasks related to ancient Chinese texts. However, in tasks requiring attention to fine details, very long OCR outputs may reduce the model's processing capability.

In summary, our contributions are as follows:

- We propose MCS-Bench, the first multimodal benchmark to evaluate MLLMs across diverse CCS subdomains, covering seven core subdomains and comprising 45 tasks.
- We evaluate 37 mainstream MLLMs, highlighting significant room for improvement in CCS.
- We conduct detailed analyses to reveal model performance across different tasks, providing valuable insights for future research in cultural heritage and innovation.

2 MCS-Bench

2.1 Task Definition

MCS-Bench covers seven core areas of CCS: Ancient Chinese Text, Calligraphy, Painting, Oracle Bone Script, Seal, Cultural Relic, and Illustration. Each area includes targeted evaluation tasks that comprehensively assess models' capabilities in recognition, understanding, knowledge, among other aspects within CCS. The task names represented by T1 to T45 are detailed on the right side of Figure 1.

Ancient Chinese Text refers to digitized image data of ancient documents obtained through photographic or scanning techniques. These images not only preserve the original textual content of ancient books but also retain their unique visual features, such as layout, typography, and paper texture. Seven tasks are designed around two main directions: **Text Recognition** (T1) and **Text Understanding** (T2, T3, T4, T5, T6, T7).

Calligraphy is an artistic expression that involves writing text using tools such as brushes and ink, adhering to specific elements like strokes, character forms, and layout to create visually appealing and expressive works. We focus on two main areas: **Calligraphy Recognition** (T8, T10, T11, T12, T13) and **Appreciation** (T9, T14, T15), defining a total of eight tasks.

Dataset	Domain	Modality	License	Scale	# Category	# Task	# LLM	Question Fromat		Method		
								MCQ	QA	HG	CI	MC
C-Eval	General	Text-only	CC BY-NC-SA-4.0	439	1	2	11	✓	✗	✓	✗	✓
Chinese SimpleQA	General	Text-only	CC BY-NC-SA-4.0	323	4	11	41	✗	✓	✓	✗	✓
CIF-Bench	General	Text-only	-	150	1	3	28	✗	✓	✓	✓	✗
CMMLU	General	Text-only	CC BY-NC-4.0	1,192	1	7	21	✓	✗	✓	✗	✗
GAOKAO-Bench	General	Text-only	Apache-2.0	145	1	2	12	✓	✓	✓	✗	✗
XiezhiBenchmark	General	Text-only	CC BY-NC-SA-4.0	2,060	2	3	47	✓	✗	✓	✗	✓
ACLUE	CCS	Text-only	CC BY-NC-4.0	4,967	5	15	8	✓	✗	✓	✓	✗
C-CLUE	CCS	Text-only	CC BY-SA-4.0	1,122	1	2	-	✗	✓	✓	✗	✗
CCLUE	CCS	Text-only	Apache-2.0	36,319	2	5	-	✓	✓	✓	✓	✗
CCPM	CCS	Text-only	-	2,720	1	1	-	✓	✗	✓	✗	✗
THUAIPoet	CCS	Text-only	-	5,173	1	3	-	✓	✓	✓	✗	✗
WenMind	CCS	Text-only	CC BY-NC-SA-4.0	4,875	3	42	31	✓	✓	✓	✓	✓
WYWEB	CCS	Text-only	-	69,700	5	9	-	✓	✓	✓	✓	✗
CII-Bench	General	Image-Text	Apache-2.0	137	1	1	21	✓	✗	✓	✗	✗
ALM-Bench	Culture	Image-Text	CC BY-NC-4.0	466	18	-	16	✓	✓	✓	✗	✓
CulturalBench	Culture	Image-Text	CC BY-4.0	117	3	-	18	✓	✗	✓	✗	✓
CVQA	Culture	Image-Text	-	311	10	-	16	✓	✗	✓	✗	✗
MaRVL	Culture	Image-Text	CC BY-4.0	1,012	11	1	-	✓	✗	✓	✗	✗
MCS-Bench (Ours)	CCS	Image-Text	CC BY-NC-SA-4.0	6,500	7	**45**	37	✓	✓	✓	✓	✓

Table 1: Comparison of existing datasets. "CCS" represents "Chinese Classical Studies"; "# LLM" represents "the number of LLMs evaluated in the paper"; "MCQ" represents "Multiple-Choice Questions"; "QA" represents "open-ended Q&A"; "HG" represents "Human Generated"; "CI" represents "Collection and Improvement of existing datasets"; and "MC" represents "Model Constructed." All datasets include only the portions related to CCS.

Painting is an artistic form that employs color, lines, and intricate composition to create works rich in historical culture and distinctive aesthetic meaning. For this domain, we provide twelve tasks from two perspectives: **Basic Information Recognition** (T16, T18, T19, T20, T24, T26) and **Content Understanding and Appreciation** (T17, T21, T22, T23, T25, T27).

Oracle Bone Script is an early form of Chinese characters inscribed on tortoise shells or animal bones, representing the early development of Chinese writing. This domain includes two tasks: **Assessing the model's recognition ability** (T28) and **Requiring the model to accurately understand the meanings of the pictographic characters** (T29).

Seal, as an important tool in ancient China used for verifying identity or signing documents, holds significant historical and cultural value through its unique artistic designs and inscriptions. This domain includes two tasks: **Evaluating the model's ability to recognize seal inscriptions** (T30) and **Requiring the model to accurately identify the owner based on the seal's font, style, and textual information** (T31).

Cultural relic is an artifact or site left behind from human history that holds artistic and scientific significance, serving as a bridge between the past and present. The cultural relic domain includes two evaluation dimensions: the first is **Information Recognition** (T32, T33, T34); the second is **Information Mining** (T35, T36).

Illustration supplements the content of ancient texts by visually representing scenes, characters, and emotions described in the text. We define tasks for the illustration domain from two dimensions: on one hand, **Recognition and Description** (T37, T38, T39, T41, T42); on the other hand, **Understanding and Association** (T40, T43, T44, T45).

2.2 Data Construction

As shown in Figure 2, the construction of MCS-Bench primarily consists of three components: Data Source, Generation Method, and Post-processing.

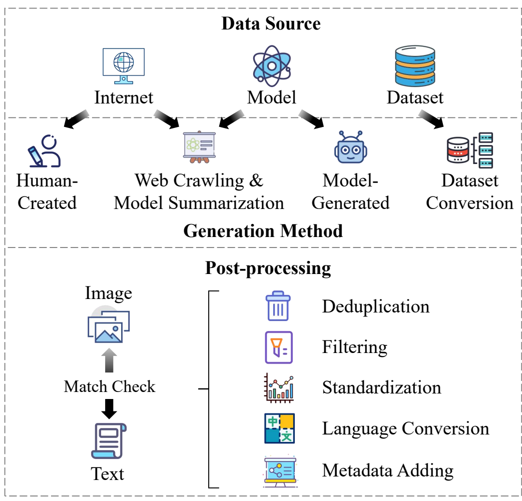


Figure 2: Construction pipeline of MCS-Bench.

2.2.1 Data Source

The data sources for MCS-Bench primarily fall into three categories: **(a) Internet**. This includes four components: "Websites and Platforms"(e.g., Saidajia Calligraphy (Sai, 2014) and Shuge Library (Wei, 2013)), "Baidu Baike" (Baidu, 2008), "Cloud Resources" (covering historical calligraphy, famous paintings, and ancient book resources), and "Internet Images". These sources provide a wealth of raw materials. **(b) Open-source Datasets**. Our research utilizes several open-source datasets, including ancient book datasets such as HisDoc1B (Shi et al., 2025) and M5HisDoc (Shi et al., 2024), as well as oracle bone script datasets like EVOBC (Guan et al., 2024) and HUST-OBC (Wang et al., 2024b). These datasets establish the foundation for evaluations in related domains. **(c) Models**. We leverage the assistance of LLMs to facilitate answer generation.

2.2.2 Generation Method

We categorize the generation methods into four types: **(a) Human-Created.** Some tasks are generated through manual writing, including creating MCQ based on existing materials, manually composing image-based questions, and OCR tasks for illustrations. These tasks are original, designed to meet specific evaluation needs, ensuring their uniqueness and relevance. **(b) Model-Generated**. For certain tasks, we use LLMs to generate questions and answers. For example, TongGu (Cao et al., 2024b) generates answers for translation and punctuation tasks, while the ERNIE-4 (Baidu, 2023) is used for named entity recognition, word explanation, and reading comprehension tasks. The generation of questions is entirely based on accurate OCR results. **(c) Web Crawling & Model Summarization**. For appreciation and introduction-type tasks, we gather related data through web crawling and use the Qwen2.5-32B (Yang et al., 2024b) to integrate key content, enhancing the accuracy and comprehensiveness of the answers. For description-type tasks, a combination of manual extraction and model extension is used to generate answers. **(d) Dataset Conversion**. We also utilize existing open-source datasets, converting their data into appropriate question formats. These datasets provide a stable foundation for questions, adjusted and formatted to meet the specific needs of this study's tasks.

These four generation methods complement each other, ensuring the diversity, accuracy, and effective support of questions for evaluation tasks. More details can be found in Appendix B.4.

2.2.3 Post-processing

To ensure data quality, we implement a series of post-processing steps: **(a) Deduplication**. Ensuring that images and question-answer pairs within the same task are free from duplicates. **(b) Filtering**. For images, we filter out non-compliant images, such as damaged or irrelevant ones. For question-answer pairs, we perform manual supplementation and verification across three dimensions: standardization and accuracy of questions, correctness and comprehensiveness of answers, and ensuring that the text content does not contain ethical concerns or unsafe content. **(c) Standardization**. Images are unified into JPG format, with pixel dimensions controlled within 1,365,984, while maintaining the original aspect ratio. We standardize the questioning instructions for the same task, and we all use Chinese punctuation marks instead of English ones. **(d) Language Conversion**. We conduct language conversion, providing both Simplified and Traditional Chinese versions. **(e) Metadata Adding**. Each data entry includes metadata, such as task ID, domain, and task name. **(f) Match Check**. Finally, a visual-text matching check is performed to ensure the correctness of image and question-answer pair pairing.

2.3 Data Statistics

We present the statistical data of MCS-Bench in Figure 1 and Table 2. The MCS-Bench consists of 6,500 carefully curated image-text pairs, covering 7 domains and 45 specific tasks. The Ancient Chinese Text domain accounts for the highest proportion at 27.88%, while the Oracle Bone Script domain accounts for the lowest at 3.32%, aligning with the typical distribution in CCS. In terms of task scale, Text OCR leads with 434 data points, while Illustration Image-based Q&A occupies the lowest position with 52 data points. The dataset features a balanced design of 23 MCQ and 22 open-ended Q&A, with image aspect ratios ranging from 0.16 to 44.92. The above reflects the diversity and richness of MCS-Bench across domains, task types, and image variations, while also providing a solid foundation for related research and evaluation.

Task & Text						
Domain	**# Task**	**# MCQ Task**	**# QA Task**	**# Q**	**Avg. Q**	**Avg. A**
Ancient Chinese Text	7	2	5	1,812	49.50	213.32
Calligraphy	8	5	3	1,127	44.15	102.62
Painting	12	7	5	1,333	38.61	91.17
Oracle Bone Script	2	0	2	216	33.89	20.77
Seal	2	1	1	303	47.11	2.43
Cultural Relic	5	4	1	630	50.46	24.69
Illustration	9	4	5	1,079	48.86	94.31
OverAll	45	23	22	6,500	45.69	114.81
Image						
Domain	**Avg. PV**	**Max. PV**	**Min. PV**	**Avg. AR**	**Max. AR**	**Min. AR**
Ancient Chinese Text	943,506	1,003,470	101,808	1.00	1.73	0.16
Calligraphy	645,218	1,365,984	25,600	1.22	20.29	0.23
Painting	889,748	1,003,500	69,984	2.78	44.92	0.22
Oracle Bone Script	33,984	151,321	1,248	0.86	1.18	0.29
Seal	658,660	1,003,275	131,010	0.97	1.30	0.22
Cultural Relic	647,728	1,003,353	18,600	1.26	14.50	0.25
Illustration	392,036	1,003,266	39,360	0.83	3.03	0.29
OverAll	717,050	1,365,984	1,248	1.39	44.92	0.16

Table 2: The statistics of MCS-Bench. "MCQ" represents "Multiple-Choice Questions"; "QA" represents "open-ended Q&A"; "Avg. Q" represents "the Average length of Questions"; "Avg. A" represents "the Average length of Answers"; "PV" represents "Pixel Value"; and "AR" represents "Aspect Ratio".

3 Experimental Details

3.1 Evaluated Models

We extensively evaluate 37 MLLMs and 3 proprietary OCR models. These include: **(a) Notable closed-source models** such as GPT-4o (OpenAI, 2024), Claude-3.5-Sonnet (Anthropic, 2024), and Gemini-2.0-Flash (Google, 2024). **(b) Open-source models**, including the InternVL series (Chen et al., 2024b; OpenGVLab, 2024) and QwenVL series (Wang et al., 2024a; Bai et al., 2023), which primarily focus on Chinese capabilities, and the LLaVA series along with its variants (Liu et al., 2024b,a; LinkSoul, 2024), which primarily focus on English capabilities. **(c) Proprietary OCR models** like general-purpose OCR models such as GOT (Wei et al., 2024) and PaddleOCR (Baidu, 2021), as well as specialized ancient text OCR models like KanDianGuJi (GuJi, 2023). Detailed information about all evaluated models can be found in Appendix C.1.

3.2 Evaluation Metrics

We use various evaluation metrics for different task types. Specifically, six evaluation metrics are employed (between 0 and 1, with higher values indicating better performance): **(a) Acc (Accuracy)**. Used for all MCQ and image-based Q&A tasks. **(b) F1-Score** (Fisher, 1936). Used for punctuation and named entity recognition tasks. **(c) BLEU** (Papineni et al., 2002). Used for translation task. **(d) The average of BERTScore and ANLS**: Used for open-ended Q&A, where BERTScore (Zhang et al., 2020) measures semantic similarity and ANLS (Biten et al., 2019) measures character similarity. **(e) CR** (Peng et al., 2023). Used for OCR tasks. The evaluation metrics and details for each task can be found in Table 14 and Appendix C.2. Additionally, other metrics such as AR (Peng et al., 2023), Edit Distance (Levenshtein, 1966a), are included in Appendix D for a comprehensive assessment.

3.3 Experiment Setup

To ensure the fairness of the evaluation, the evaluation settings for all assessed MLLMs are standardized. Specifically, we use bf16 half-precision inference with a maximum sequence length set to 2,048. The temperature parameter, Top-p sampling, and Top-k sampling are set to 0.8, 0.95, and 50, respectively. To evaluate the knowledge capacity of MLLMs, all models are prohibited from using external search engines. All experiments are conducted on NVIDIA A6000 GPUs.

4 Results and Analysis

4.1 Overall Performance Analysis

MLLMs still have significant room for improvement in CCS. As shown in Table 3, InternVL2.5-78B (Chen et al., 2024b) ranks first among all models, leading the second-place model by 2.77%, but scoring less than 50. The lowest-scoring model, Molmo-7B-O-0924 (Deitke et al., 2024), record a score of only 20.18. The average score across 37 models is 35.46, with 59% of the models scoring below 40. This underscores the low overall performance of MLLMs in the CCS tasks, highlighting the need for future investigation in this area.

Closed-source models perform strongly, while open-source models show varied performance. Figure 3 presents a heatmap of normalized scores for all models, leading to the following observations: **(a)** Closed-source models achieve an average score of 44.83 and an average rank of 5.75, demonstrating robust performance, as indicated by the predominantly red areas in the heatmap. Gemini-2.0-Flash (Google, 2024) performs the best among them. These models typically benefit from more powerful computing resources and richer datasets, contributing to their strong results. **(b)** Among open-source models, the InternVL2.5 series (Chen et al., 2024b) leads, followed by InternVL2 (OpenGVLab, 2024) and Qwen2-VL (Wang et al., 2024a), with GLM-4V (GLM et al., 2024) and MiniCPM-2.6 (Yao et al., 2024) also performing well. Models like DeepSeek (Wu et al.,

Model	OverAll	Rank	Domain							Question Format	
			Ancient Chinese Text	Calligraphy	Painting	Oracle Bone Script	Seal	Cultural Relic	Illustration	MCQ	QA
Closed-source Models											
Claude-3.5-Sonnet	44.55	7	38.53	41.48	57.77	24.12	21.45	68.21	60.30	60.82	34.67
Gemini-1.5-Pro	43.87	8	<u>50.64</u>	35.67	58.87	23.75	19.79	63.00	55.39	60.45	31.57
Gemini-2.0-Flash	46.19	3	**51.44**	39.66	59.04	21.25	21.90	**69.23**	60.82	**65.11**	32.02
GPT-4o	44.72	5	35.53	44.75	57.96	25.49	21.42	68.41	59.52	61.83	33.96
Open-source Models											
Chinese-LLaVA-CLLaMA2	21.00	36	13.47	15.57	27.50	21.70	13.17	31.45	24.17	21.74	21.81
DeepSeek-VL2-Tiny	34.19	21	22.43	31.42	46.76	21.32	17.91	53.25	46.24	43.84	28.57
DeepSeek-VL2-Small	26.33	29	15.86	27.86	33.45	17.48	16.53	37.63	35.53	32.05	23.77
DeepSeek-VL2	30.47	26	20.17	29.51	38.73	19.71	14.42	48.96	41.75	38.63	26.22
GLM-4V-9B	41.83	13	30.64	39.36	54.04	25.28	23.33	65.42	54.75	57.34	31.43
InternVL2-4B	34.78	20	26.14	37.59	43.96	20.95	17.14	50.71	46.94	42.24	31.48
InternVL2-8B	41.05	14	34.76	45.30	53.28	22.65	19.09	59.73	52.53	53.42	34.37
InternVL2-26B	42.70	11	32.61	42.67	55.33	22.93	25.37	61.48	58.49	57.00	34.23
InternVL2.5-1B	38.79	17	27.94	45.42	48.48	25.16	19.49	55.31	49.73	47.62	34.72
InternVL2.5-2B	38.77	18	28.80	39.82	51.09	24.86	18.79	55.02	53.04	48.98	33.00
InternVL2.5-4B	43.25	10	33.66	47.34	58.92	25.44	22.19	60.88	54.33	55.91	36.53
InternVL2.5-8B	43.61	9	36.85	47.61	57.10	25.22	20.25	61.17	57.04	56.68	36.70
InternVL2.5-26B	45.92	4	36.80	<u>48.45</u>	59.55	26.11	23.80	65.27	<u>61.44</u>	59.69	37.91
InternVL2.5-38B	<u>46.55</u>	2	39.52	47.11	<u>61.32</u>	<u>26.82</u>	<u>26.43</u>	63.95	60.72	60.61	<u>38.58</u>
InternVL2.5-78B	**49.32**	1	42.93	**51.65**	**62.72**	**27.48**	**28.96**	<u>68.79</u>	**62.70**	<u>64.46</u>	**40.43**
LLaVA-v1.5-7B	23.26	34	15.16	20.12	29.25	22.96	9.91	34.58	30.87	26.68	22.07
LLaVA-v1.5-13B	24.10	31	14.43	20.97	31.12	22.74	12.17	36.50	30.80	27.73	22.27
LLaVA-v1.6-Mistral-7B	23.43	33	15.12	19.83	28.19	24.04	17.49	33.06	26.26	26.82	21.63
LLaVA-v1.6-Vicuna-7B	24.46	30	13.24	24.02	31.68	23.19	14.78	35.09	29.23	28.28	22.27
LLaVA-v1.6-Vicuna-13B	23.62	32	14.02	19.13	29.87	22.55	16.20	32.45	31.11	27.21	22.30
MiniCPM-V	30.40	27	18.57	23.61	39.77	25.17	11.28	58.81	35.62	38.70	23.46
MiniCPM-V-2	32.46	24	19.76	26.38	41.96	25.22	16.74	57.71	39.46	43.11	24.28
MiniCPM-LLaMA3-V-2.5	32.41	25	20.94	31.08	37.72	25.81	13.60	54.12	43.60	41.97	25.54
MiniCPM-V-2.6	39.10	16	30.33	36.28	55.30	23.15	13.21	63.88	51.55	54.45	28.19
Molmo-7B-D-0924	21.19	35	18.07	22.31	25.01	17.49	16.91	25.33	23.20	29.93	16.59
Molmo-7B-O-0924	20.18	37	12.57	20.15	27.26	16.05	14.00	28.30	22.92	26.98	15.79
Ovis1.5-Gemma2-9B	30.17	28	17.91	22.15	40.17	23.69	12.61	53.53	41.09	39.57	23.38
Ovis1.6-Gemma2-9B	32.47	23	21.17	24.09	40.27	21.88	17.62	56.51	45.74	43.95	24.90
Qwen-VL-Chat	33.53	22	21.59	27.54	43.26	23.76	20.49	56.81	41.26	45.24	25.89
Qwen2-VL-2B-Instruct	35.58	19	26.43	33.80	45.76	24.69	15.58	58.01	44.75	45.26	29.70
Qwen2-VL-7B-Instruct	40.60	15	37.78	37.94	50.64	25.73	17.76	60.77	53.57	52.65	33.91
Qwen2-VL-72B-Instruct	44.56	6	42.81	43.79	58.75	24.73	17.36	65.60	58.91	59.97	34.91
QVQ-72B-Preview	42.54	12	34.97	45.44	54.44	20.17	22.80	63.29	56.65	58.79	33.12
Average	35.46	-	27.39	33.97	45.85	23.26	18.16	53.57	46.00	46.10	28.98

Table 3: Results of 37 MLLMs on MCS-Bench. "MCQ" represents "Multiple-Choice Questions"; "QA" represents "open-ended Q&A". Each score is the average of metrics across all corresponding tasks. **Bold** represents the best results, while <u>underline</u> represents the second best.

2024b), LLaVA (Liu et al., 2024b), and Ovis (Lu et al., 2024) series perform poorly, likely due to limitations in data and language capabilities.

Data and training strategies significantly impact domain performance. Although both Qwen2-VL-7B and Molmo-7B-D-0924 are based on Qwen2-7B (Yang et al., 2024a), Qwen2-VL-7B achieves a score of 40.60, far surpassing Molmo-7B-D-0924's 21.19. This difference can be attributed to: **(a)** Qwen2-VL's use of dynamic resolution and M-RoPE, which better handle fine-grained images and complex text, such as in Ancient Chinese Texts. **(b)** Qwen2-VL's training data includes 1.4T tokens, significantly more than Molmo's 712K images in its general-purpose dataset. **(c)** Molmo emphasizes openness and dense visual descriptions, which enhance general performance but fall short in adapting to the specialized visual-textual demands of CCS.

Models exhibit variability across subdomains. In terms of subdomain, Cultural Relic achieves the highest average score (53.57), while Seal scores the lowest (18.16), indicating significant differences in the models' understanding across domains. This also reflects the general availability and scarcity of data. Additionally, Painting and Illustration show similar scores, likely due to their related presentation styles.

Slow-thinking models currently show limited suitability for CCS. As shown in Table 3 and 4, QVQ-72B-Preview (Qwen, 2024), an model which is derived from Qwen2-VL-72B (Wang et al., 2024a) through an o1-like training approach, does not demonstrate significant advantages over Qwen2-VL-72B. This may be because CCS tasks rely more on the model's domain-specific knowledge and understanding rather than complex reasoning processes. Despite taking significantly longer inference time (~12 times), its performance declines. Additionally, we observe that o1-like models have weaker instruction-following capabilities, often ignoring instructions to output their reasoning process directly, leading to overthinking (Chen et al., 2024a). Effectively leveraging o1-like models' potential in CCS tasks during reasoning remains a challenge. Appendix F.3 presents a more

Model	OverAll		T1		T8		T26		T28		T30		T37	
	CR↑	F1-Score↑	CR↑	F1-Score↑	CR↑	F1-Score↑	CR↑	F1-Score↑	CR↑	F1-Score↑	CR↑	F1-Score↑	CR↑	F1-Score↑
Closed-source Models														
Claude-3.5-Sonnet	32.15	41.17	39.64	59.00	40.25	58.56	45.63	53.82	5.44	5.44	8.22	9.13	53.72	61.09
Gemini-1.5-Pro	26.32	37.77	61.75	80.56	10.20	35.50	52.74	64.55	2.72	2.72	1.58	2.63	28.90	40.64
Gemini-2.0-Flash	41.02	50.94	72.22	81.53	27.64	64.73	66.12	69.43	5.44	5.44	6.46	10.76	**68.26**	73.72
GPT-4o	25.53	33.59	27.90	44.65	37.92	52.44	35.58	45.72	3.40	3.40	5.50	6.79	42.88	48.53
Open-source Models														
Chinese-LLaVA-CLLaMA2	2.84	2.42	1.49	7.54	1.50	1.34	1.96	2.43	0.00	0.00	6.33	1.61	5.75	1.57
DeepSeek-VL2-Tiny	19.33	39.51	18.33	47.81	18.23	58.79	25.66	53.34	2.04	2.04	4.48	7.74	47.21	67.36
DeepSeek-VL2-Small	16.17	28.60	5.30	16.93	20.74	48.25	15.16	31.15	3.40	3.40	5.06	8.52	47.33	63.37
DeepSeek-VL2	18.11	30.89	7.45	21.95	26.15	51.44	20.33	37.85	2.72	2.15	5.50	8.74	46.51	63.21
GLM-4V-9B	23.97	37.36	22.71	58.02	24.38	46.44	39.60	54.84	2.04	2.04	6.66	6.93	48.41	55.89
InternVL2-4B	36.33	46.06	54.36	66.64	47.97	67.52	57.03	61.74	3.40	2.84	6.27	11.92	48.93	65.72
InternVL2-8B	41.05	48.33	59.39	70.25	63.06	65.77	65.32	68.28	2.72	2.72	7.51	14.35	48.28	68.59
InternVL2-26B	39.63	49.24	56.69	69.92	52.97	67.68	61.98	65.62	4.76	4.76	9.41	16.74	51.95	70.70
InternVL2.5-1B	37.71	47.27	48.82	59.76	61.52	70.45	50.55	60.22	4.08	3.67	8.30	12.61	53.01	76.93
InternVL2.5-2B	34.50	43.58	53.00	64.06	41.62	66.52	56.65	61.86	2.72	2.72	4.91	8.11	48.07	58.22
InternVL2.5-4B	41.40	50.51	56.20	67.06	60.94	71.11	66.15	70.15	5.44	5.44	9.71	15.42	49.95	73.89
InternVL2.5-8B	42.02	51.29	60.56	70.24	60.42	71.80	68.07	72.78	3.40	3.40	7.83	12.65	51.81	76.86
InternVL2.5-26B	44.13	54.26	63.80	73.69	57.03	74.53	68.25	70.68	7.48	7.48	10.93	20.43	57.29	78.77
InternVL2.5-38B	45.36	55.71	68.05	76.04	57.37	75.59	70.46	73.17	6.80	6.80	11.53	20.56	57.97	**82.08**
InternVL2.5-78B	**49.54**	**57.73**	75.19	80.63	**68.73**	**80.97**	**72.42**	**74.18**	**8.16**	**8.16**	15.24	**21.31**	57.49	81.12
LLaVA-v1.5-7B	0.47	0.93	0.54	2.29	0.50	1.02	0.40	0.86	0.00	0.00	0.49	0.66	0.87	0.74
LLaVA-v1.5-13B	1.10	1.46	0.76	3.65	0.85	1.03	0.63	1.99	0.68	0.05	3.00	1.41	0.70	0.64
LLaVA-v1.6-Mistral-7B	2.84	2.88	2.05	8.65	2.06	2.05	3.23	4.52	3.40	0.20	3.65	0.93	2.62	0.94
LLaVA-v1.6-Vicuna-7B	2.48	2.57	1.72	6.86	2.60	1.77	1.08	3.27	2.04	0.31	4.22	1.65	3.21	1.58
LLaVA-v1.6-Vicuna-13B	2.64	2.82	1.71	8.26	3.15	1.95	1.32	4.00	1.36	0.11	4.39	1.47	3.88	1.11
MiniCPM-V	1.83	2.71	0.66	2.32	1.08	2.11	0.55	1.67	1.36	1.36	2.56	2.74	4.76	6.06
MiniCPM-V-2	5.75	7.22	1.49	6.61	5.61	18.70	6.21	4.98	3.40	3.40	4.80	2.32	12.98	7.31
MiniCPM-LLaMA3-V-2.5	10.37	21.62	4.31	14.84	8.36	35.66	9.03	24.58	5.44	5.44	1.87	2.19	33.22	47.00
MiniCPM-V-2.6	22.55	35.71	30.98	44.12	15.69	51.15	41.07	50.07	4.76	4.76	1.08	3.05	41.71	61.08
Molmo-7B-D-0924	0.56	0.26	0.12	0.29	0.05	0.12	0.03	0.08	0.00	0.00	3.15	1.09	0.00	0.00
Molmo-7B-O-0924	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ovis1.5-Gemma2-9B	1.72	3.21	2.16	8.02	1.12	2.09	1.37	4.23	0.68	0.01	1.22	1.17	3.78	3.75
Ovis1.6-Gemma2-9B	6.34	10.52	3.49	15.68	4.07	16.22	4.46	10.57	0.68	0.06	5.23	1.08	20.12	19.52
Qwen-VL-Chat	7.97	10.70	2.94	10.76	9.56	19.65	3.08	7.77	0.00	0.00	6.97	3.39	25.25	22.65
Qwen2-VL-2B-Instruct	22.70	32.77	20.99	33.60	27.25	50.98	35.49	43.59	3.40	2.27	4.49	7.11	44.55	59.04
Qwen2-VL-7B-Instruct	32.35	45.48	45.67	61.80	33.18	64.99	54.12	63.26	4.76	3.85	5.51	8.30	50.84	70.68
Qwen2-VL-72B-Instruct	36.81	49.34	59.51	77.15	32.89	66.77	59.70	68.75	3.40	3.40	5.38	7.25	59.95	72.73
QVQ-72B-Preview	35.34	28.01	46.60	38.51	45.67	47.95	41.67	38.31	3.40	1.38	**15.59**	6.22	59.11	35.71
OCR Models														
GOT	16.08	34.51	13.33	43.07	14.23	42.86	19.87	53.42	1.36	1.36	2.86	6.13	44.80	60.21
PaddleOCR	28.31	42.12	49.69	67.60	17.99	58.11	57.71	64.26	0.00	0.00	2.50	4.32	41.98	58.43
KanDianGuJi	37.24	50.42	**83.68**	**90.06**	39.43	73.88	66.11	71.78	0.00	0.00	0.99	1.66	33.24	65.16
Average	21.92	28.77	29.15	39.45	26.28	40.91	32.52	38.49	3.11	2.74	5.81	7.27	34.63	43.75

Table 4: Results of 37 MLLMs and 3 OCR models across six OCR tasks. Different colors represent different subdomains, with the color-domain mapping provided in Figure 1. In addition to CR, we also provide the F1-Score metric. "Average" represents the average metric of 37 MLLMs. **Bold** represents the best results, while underline represents the second best.

detailed analysis of slow-thinking models.

4.2 Subtasks Analysis

The analysis of subtask performance, as illustrated in Figure 4, reveals the following key observations: **(a)** Models perform notably well on tasks T20, T32, T36, and T42 because the answers are closely associated with the images, forming distinct image-caption pairs. **(b)** For OCR tasks, all models perform poorly on T28 and T30. This is mainly due to the high abstraction of Oracle Bone Script and Seal fonts, limited data availability, and the lack of specialized optimization in existing models, making accurate recognition and processing challenging. **(c)** In T2, T3, and T4, model performance is heavily concentrated in lower score ranges. Taking T2 as an example, most of MLLMs fail to achieve "one-step" solutions directly from images to desired results, struggling to implicitly handle transitions from images to OCR results and then translations. Even with OCR results, models often perform poorly due to inherent limitations (see Section 4.4). Additionally, some models fail to follow instructions and only output OCR results, suggesting overfitting issues.

4.3 OCR Tasks Analysis

Recently, the OCR performance of MLLMs has gained increasing attention (Fu et al., 2024; Shi et al., 2023). Given that six subdomains in the dataset involve OCR tasks, we specifically present model performance on these tasks in Table 4. Using the CR metric as an example, we obtain the following results:

MLLMs show room for improvement in OCR tasks within CCS. Among all models, InternVL2.5-78B (Chen et al., 2024b) performs the best, followed by InternVL2.5-38B, but neither achieves a CR score above 50. The average CR score across all models is only 21.92, indicating a relatively low performance level. This highlights challenges in recognizing the unique text, symbols, and image features specific to CCS, as well as the inherent complexity and diversity of these

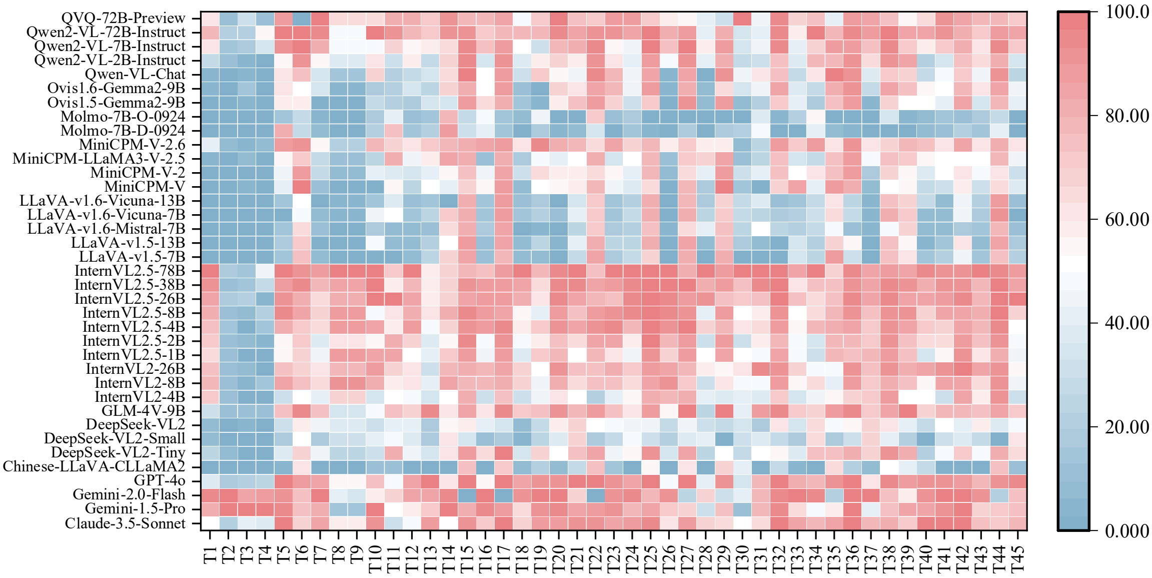


Figure 3: Normalized heatmap of model performance.

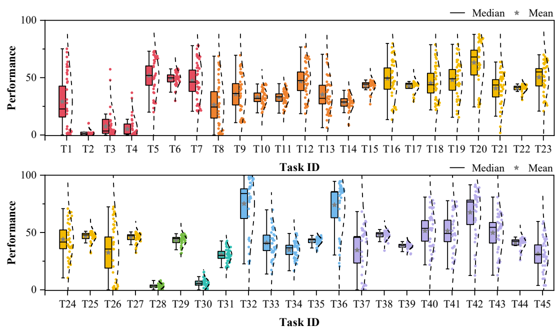


Figure 4: Performance distribution box plot of 37 MLLMs across 45 tasks.

OCR tasks. Future efforts could focus on enhancing models' learning of CCS-specific features and improving dataset quality and annotation to boost performance.

Traditional OCR models show a performance gap compared to the upper limits achieved by MLLMs. The best-performing MLLM achieves a CR score of 49.54, surpassing GOT (Wei et al., 2024), PaddleOCR (Baidu, 2021), and KanDianGuJi (GuJi, 2023) by 33.46%, 21.23%, and 12.30%. This indicates that: **(a)** MLLMs demonstrate higher potential in handling complex or domain-specific OCR tasks, likely due to their advantages in processing multimodal information, understanding contextual semantics, and generalization. **(b)** The KanDianGuJi model excels in the Ancient Chinese Text domain, breaking the 80-point threshold, but falls significantly behind in other domains, highlighting its limited generalization. As technology advances and models continue to improve, general-purpose MLLMs are expected to achieve broader applicability and higher performance in OCR tasks, particularly in specialized areas like CCS.

4.4 Supplementary Experiment Analysis

We observe that model performance in the Ancient Chinese Text domain heavily depends on OCR capabilities. These tasks can essentially be divided into two steps: recognition and understanding. To investigate further, we design supplementary experiments with three settings: **Setting 1** inputs only the image and question; **Setting 2** inputs the image, question, and the model's own OCR results to assess whether explicitly separating the two steps improves performance; **Setting 3** inputs the image, question, and accurate OCR results, eliminating OCR performance differences to focus on the model's understanding ability. We conduct experiments on five tasks (T2 to T6) using four representative models. Results are shown in Figure 5.

For T2, T3, T4, and the T2-T6 average, we observe the following: **(a)** Explicitly separating the steps significantly improves performance. **(b)** Providing more accurate OCR results further enhances performance. **(c)** When OCR performance differences are eliminated, GPT-4o (OpenAI, 2024) demonstrates a substantial improvement and ranks first among the four models, indicating its strong capability in handling ancient texts. In contrast, other models remain at relatively low performance levels.

For T5 and T6, however, providing OCR results leads to a performance decline. These tasks require

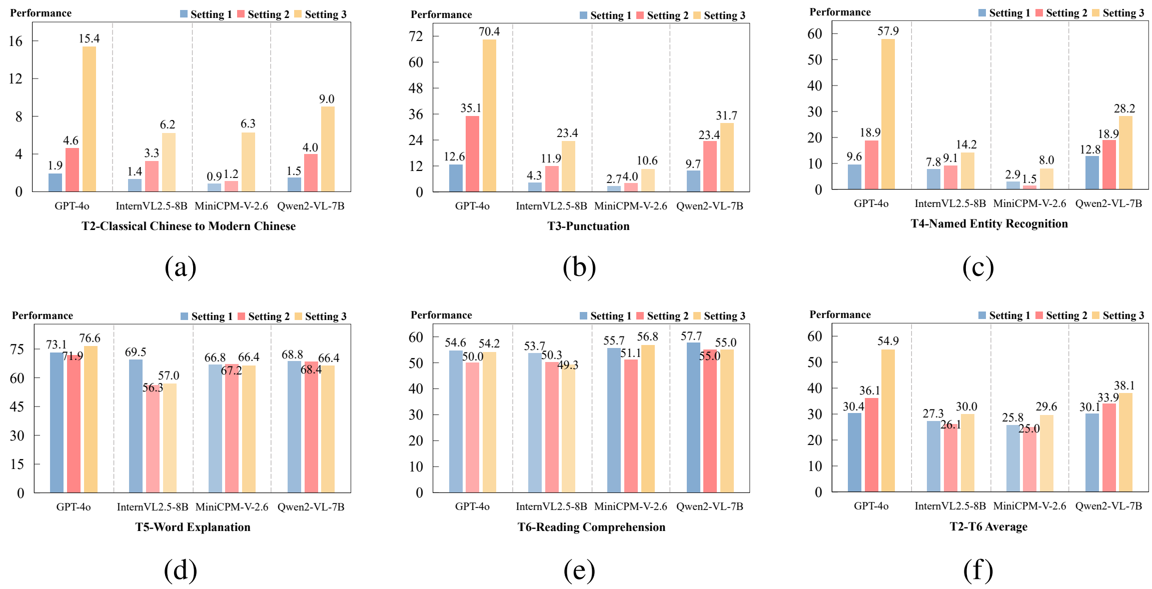


Figure 5: Bar chart of OCR supplementary experiment results. Different settings are provided in Section 4.4.

models to focus on capturing local details. When the steps are explicitly separated and OCR results are provided, the resulting lengthy text distracts the models from processing local details effectively, causing a slight decrease in performance.

5 Conclusion

This paper introduces MCS-Bench, a multimodal benchmark specifically designed for the Chinese Classical Studies (CCS) domain. It encompasses seven core subdomains — ***Ancient Chinese Text***, ***Calligraphy***, ***Painting***, ***Oracle Bone Script***, ***Seal***, ***Cultural Relic***, and ***Illustration*** — and features 45 meticulously designed tasks to comprehensively evaluate the capabilities of Multimodal Large Language Models (MLLMs) in complex cultural contexts. Evaluations of 37 mainstream MLLMs reveal significant room for improvement in CCS-related tasks. Additionally, we provide an in-depth analysis of performance variations across tasks and their influencing factors, offering valuable insights for future research. We believe that MCS-Bench can serve as a foundational benchmark to drive the development of MLLMs fostering innovation and progress in the understanding and processing of cultural heritage.

6 Limitations

Our primary evaluation limitation lies in the inability of certain metrics to fully capture model performance in complex ancient text scenarios. Nevertheless, we provide multiple evaluation metrics for OCR and open-ended Q&A tasks to quantify the models' true capabilities in this domain as thoroughly and accurately as possible.

7 Ethical Statement

When using the MCS-Bench dataset, special attention must be given to potential historical biases and cultural misunderstandings to avoid reinforcing incorrect or outdated societal perspectives in processing content related to CCS. It is essential to respect the profound cultural significance of CCS, interpret images and question-answering content with care, and prevent cultural misinterpretations or offenses arising from improper use. The MCS-Bench dataset is intended solely for academic research to promote the digital preservation and innovative transmission of Chinese classical culture. We strictly adhere to copyright requirements for data sources and prohibit its use for commercial purposes or any unethical applications, ensuring that the dataset is employed for legitimate goals aligned with cultural dissemination and technological advancement.

Acknowledgements

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

References

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, et al. 2024. Maya: An instruction finetuned multilingual multimodal model. *arXiv preprint arXiv:2412.07112*.

Anthropic. 2024. Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

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*, 1(2):3.

Baidu. 2008. Baidu Baike. https://baike.baidu.com/.

Baidu. 2021. PaddleOCR. https://github.com/PaddlePaddle/PaddleOCR.

Baidu. 2023. ERNIE-4.0. https://yiyan.baidu.com/.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Jiahuan Cao, Yang Liu, Yongxin Shi, Kai Ding, and Lianwen Jin. 2024a. WenMind: A comprehensive benchmark for evaluating large language models in Chinese classical literature and language arts. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024b. TongGu: Mastering classical Chinese understanding with knowledge-grounded large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4196–4210, Miami, Florida, USA. Association for Computational Linguistics.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. 2024a. Do not think that much for 2+ 3=? On the overthinking of o1-like LLMs. *arXiv preprint arXiv:2412.21187*.

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

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. 2024. CulturalBench: A robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of LLMs. *arXiv preprint arXiv:2410.02677*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and Pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Ethan. 2021. CCLUE: Classical Chinese language understanding evaluation benchmark: Datasets, baselines, pre-trained models, corpus and leaderboard. https://github.com/Ethan-yt/CCLUE.

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

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

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools.

Google. 2024. Introducing Gemini 2.0: Our new AI model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

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

KanDian GuJi. 2023. KanDianGuJi OCR. https://kandianguji.com/ocr.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, et al. 2024. Chinese SimpleQA: A Chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

InternLM. 2024. InternLM3-8B-Instruct. https://huggingface.co/internlm/internlm3-8b-instruct.

Zijing Ji, Yuxin Shen, Yining Sun, Tian Yu, and Xin Wang. 2021. C-CLUE: A benchmark of classical Chinese based on a crowdsourcing system for knowledge graph construction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6*, pages 295–301. Springer.

Joint Laboratory of HIT and iFLYTEK Research. 2024. LLaMA3-Chinese-8B-Instruct. https://huggingface.co/hfl/llama-3-chinese-8b-instruct.

VI Levenshtein. 1966a. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

VI Levenshtein. 1966b. Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. CultureLLM: Incorporating cultural differences into large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024b. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

LinkSoul. 2024. Chinese-LLaVA-CLLaMA2. https://huggingface.co/LinkSoul/Chinese-LLaVA-Cllama2.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu. 2024. LLaVA-V1.6-34B. https://huggingface.co/liuhaotian/llava-v1.6-34b.

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.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

OpenGVLab. 2024. InternVL2: Better than the best expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

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

Dezhi Peng, Lianwen Jin, Weihong Ma, Canyu Xie, Hesuo Zhang, Shenggao Zhu, and Jing Li. 2023. Recognition of handwritten Chinese text by segmentation: A segment-annotation-free approach. *IEEE Transactions on Multimedia*, 25:2368–2381.

Qwen. 2024. QVQ: To see the world with wisdom. https://qwenlm.github.io/blog/qvq-72b-preview/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Dajia Sai. 2014. Saidajia calligraphy. http://www.51-sdj.com/.

Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. 2024. M5HisDoc: A large-scale multi-style Chinese historical document analysis benchmark. *Advances in Neural Information Processing Systems*, 36.

Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. Exploring OCR capabilities of GPT-4V (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*.

Yongxin Shi, Dezhi Peng, Yuyi Zhang, Jiahuan Cao, and Lianwen Jin. 2025. HisDoc1B: A large-scale dataset for Chinese historical document recognition and analysis. *Scientific Data*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, et al. 2024. All languages matter: Evaluating LMMs on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*.

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

Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Jinpeng Wan, Haisu Guan, Zhebin Kuang, Lianwen Jin, Xiang Bai, and Yuliang Liu. 2024b. An open dataset for oracle bone character recognition and decipherment. *Scientific Data*, 11:976.

Ceng Wei. 2013. Shuge library. https://www.shuge.org/.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. 2024. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, Ping Luo, Yu Qiao, and Jifeng Dai. 2024a. VisionLLM v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024b. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024a. LLaVA-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

Liuliu Xu, Langchen Sun, and Jie Wu. 2024b. World conference of classics creates new platform for mutual learning. https://www.globaltimes.cn/page/202411/1322664.shtml. Published on 2024-11-07.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xinrun Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, et al. 2024a. Can MLLMs understand the deep implication behind Chinese images? *arXiv preprint arXiv:2410.13854*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Wei Zhang, Wong Kam-Kwai, Biying Xu, Yiwen Ren, Yuhuai Li, Minfeng Zhu, Yingchaojie Feng, and Wei Chen. 2024b. CultiVerse: Towards cross-cultural understanding for paintings with large language model. *arXiv preprint arXiv:2405.00435*.

Yixuan Zhang and Haonan Li. 2023. Can large langauge model comprehend ancient Chinese? A preliminary test on ACLUE. In *Proceedings of the Ancient Language Processing Workshop*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. WYWEB: A NLP evaluation benchmark for classical Chinese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319, Toronto, Canada. Association for Computational Linguistics.

A Related Work

A.1 Text-only Benchmarks

In the text-only domain, researchers have developed several benchmarks to evaluate Chinese Large Language Models (LLMs) from various dimensions. C-Eval (Huang et al., 2024) provides a comprehensive evaluation suite consisting of 13,948 Multiple-Choice Questions (MCQ) covering 52 disciplines, focusing on fundamental knowledge and reasoning capabilities. Similarly, CMMLU (Li et al., 2024b) assesses knowledge and reasoning abilities across 67 disciplines, ranging from basic to advanced levels, while Chinese SimpleQA (He et al., 2024) focuses on factual evaluation across six topics. In the domain of Classical Chinese Studies (CCS), ACLUE (Zhang and Li, 2023) offers specially designed MCQ to test models' understanding of Ancient Chinese Texts, whereas C-CLUE (Ji et al., 2021) targets tasks such as Named Entity Recognition and Relation Extraction. CCLUE (Ethan, 2021) expands the evaluation scope by providing five distinct tasks, including Sequence Labeling and Sentence Classification. WYWEB (Zhou et al., 2023) contributes nine evaluation tasks, ranging from Text Classification to Punctuation and Machine Translation. WenMind (Cao et al., 2024a) adopts a more holistic approach by combining various task formats and evaluation methods, aiming to align more closely with human intuition for a more accurate assessment of Chinese LLMs' capabilities.

However, most existing CCS datasets remain text-only and relatively small in scale. Except for CCLUE and WYWEB, the majority contain only a few thousand examples each, which are comparatively easier to collect but do not cover visual modalities. Our MCS-Bench is the first benchmark in the CCS domain to use images to assess models' related capabilities.

A.2 Multimodal Benchmarks

In the visual-text multimodal domain, multilingual and cross-cultural evaluation benchmarks have recently garnered widespread attention. The MaRVL benchmark (Liu et al., 2021) systematically assesses the performance of models in visual cultural understanding through tasks such as title classification, pairwise title matching, and cultural label selection. ALM-Bench (Vayani et al., 2024), a multilingual multimodal VQA benchmark comprising 100 languages and 22.7K question-answer pairs, provides an important reference for evaluating models' cross-linguistic and cross-cultural understanding capabilities. Similarly focusing on cultural understanding, CulturalBench (Chiu et al., 2024) encompasses 17 cultural themes, ranging from cuisine to etiquette. In the domain of Chinese visual understanding, CII-Bench (Zhang et al., 2024a) evaluates models' higher-order perception and reasoning abilities through complex images with implicit visual meanings, such as abstract art, comics, and posters. Collectively, these benchmarks drive the advancement of multilingual multimodal models in the domain of cross-cultural understanding.

However, a reliable evaluation benchmark remains lacking in the multimodal CCS domain. To address this gap, we propose MCS-Bench, which fills the void in this area. Compared to existing multimodal cultural benchmarks, it offers significant advantages in terms of dataset size, task variety, question formats, and construction methodology. Moreover, we have minimized evaluation errors by employing diverse metrics and evaluating a comprehensive set of models. In addition, MCS-Bench covers a broader spectrum of task types, with an emphasis on diversity and high quality. We have implemented strict quality-control measures in data annotation and question design to ensure robust evaluation, thereby more fully revealing the true capabilities of LLMs in the CCS field.

B Data

B.1 Data Examples

Figure 6 showcases examples from seven subdomains in MCS-Bench.

B.2 Tasks Statistics

Table 5 provides detailed statistical data for 45 tasks, including the number of questions, average length of questions, average length of answers, average values, maximum values, and minimum values for pixel values and aspect ratios.

B.3 Detailed Source of Data

Table 13 presents detailed information on the data sources and generation methods for 45 tasks, along with the corresponding license information for each source. We strictly adhere to the licenses of the original data, ensuring compliance with all relevant regulations during data processing and usage. The data generation methods fall into four categories:

Figure 6: Data examples from MCS-Bench.

M1: Human-Created; M2: Web Crawling & Model Summarization; M3: Model-Generated; M4: Dataset Conversion.

The detailed data construction process is provided in Appendix B.4.

B.4 Detailed Construction of Data

The process of constructing the MCS-Bench can be divided into two stages: **the data collection and generation stage**, and **the post-processing stage**.

B.4.1 The Data Collection and Generation Stage

Human-Created:

Responsible parties: Graduate Student A and Volunteer B.

Selection Criteria: Requires strong literary skills (college entrance examination Chinese score of 120 or above) and meticulous attention to detail.

Process: (a) Graduate Student A is responsible for manually constructing multiple-choice tasks. First, data sources and scope are determined based on the task type, with a standardized question format. Then, the number of options is established, typically 3 to 5 options. Subsequently, correct and distractor options are designed according to the task requirements, ensuring that options are moderately misleading but logically consistent, avoiding overly extreme choices. Finally, the options are randomly ordered by the responsible party to prevent response bias, ensuring the scientific nature and fairness of the multiple-choice questions. (b) Volunteer B is responsible for manually constructing image-based Q&A tasks and OCR tasks in the fields of painting, illustration, and calligraphy. After manually collecting data, suitable paintings or illustrations are selected for image-based Q&A tasks according to task requirements, generating questions from different perspectives. For OCR tasks, textual content on images is annotated word by word.

Additional Notes: (a) For multiple-choice tasks, the model's output requirement is set in the prompt (i.e., the model can only output the letters of selected options). To ensure fairness, the final check of multiple-choice outputs from MLLMs will be conducted uniformly. (b) In image-based Q&A tasks, the correct answer must be retrievable directly from the image or deducible through inference. Questions should avoid ambiguity and can be designed from aspects such as counting, color, position, etc.

Model-Generated:

Responsible party: Graduate Student C.

Selection Criteria: Requires extensive familiarity with LLMs and at least one publication in the field of LLMs.

Process: Graduate Student C first collects images of ancient texts and OCR results, then manually selects the original text to be processed by LLMs. Finally, LLMs are called upon to generate answer pairs for tasks such as classical Chinese

ID	Task Name	# Q	Avg. Q	Avg. A	Avg. PV	Max. PV	Min. PV	Avg. AR	Max. AR	Min. AR
T1	Text OCR	434	40.90	358.83	946,768	1,003,470	436,293	1.02	1.73	0.26
T2	Classical Chinese to Modern Chinese	245	46.00	467.70	955,175	1,003,470	510,340	1.03	1.61	0.53
T3	Punctuation	236	45.00	376.58	939,505	1,003,470	157,471	1.03	1.61	0.16
T4	Named Entity Recognition	247	57.00	52.28	951,012	1,003,470	501,650	1.04	1.61	0.53
T5	Word Explanation	256	70.03	1.00	938,153	1,003,356	194,910	1.04	1.60	0.55
T6	Reading Comprehension	249	33.66	56.37	949,591	1,003,356	101,808	1.04	1.61	0.16
T7	Source Attribution	145	66.63	1.00	906,760	1,003,470	524,560	0.67	0.88	0.53
T8	Calligraphy OCR	130	40.00	17.47	275,139	728,200	65,340	1.02	2.40	0.40
T9	Calligraphy Appreciation	121	15.00	514.55	405,939	1,003,014	94,620	1.52	13.16	0.39
T10	Author Identification	119	56.03	1.00	822,189	1,003,392	246,016	1.29	20.29	0.31
T11	Dynasty Identification	105	56.18	1.00	844,613	1,003,392	246,016	1.34	20.29	0.31
T12	Title Identification	215	65.98	1.00	610,650	1,003,260	94,620	1.40	19.40	0.39
T13	Font Identification	125	57.00	1.00	900,581	1,003,395	241,434	0.74	5.05	0.23
T14	Genre Identification	118	57.00	1.00	839,438	1,003,181	274,000	0.78	1.69	0.36
T15	Work Introduction	194	11.00	259.98	581,616	1,365,984	25,600	1.42	13.16	0.39
T16	Material Identification	149	49.00	1.00	1,002,184	1,003,460	998,538	4.63	39.69	0.24
T17	Painting Appreciation	97	8.00	445.54	988,981	1,003,426	353,312	3.93	37.58	0.24
T18	Author Identification	146	53.07	1.00	993,606	1,003,392	353,312	2.46	23.59	0.24
T19	Dynasty Identification	110	50.87	1.00	900,517	1,003,400	69,984	2.28	37.58	0.29
T20	Title Identification	147	61.90	1.00	935,017	1,003,440	138,600	2.69	31.62	0.24
T21	Image-based Q&A	55	19.60	1.55	648,523	1,002,996	214,442	0.98	2.31	0.42
T22	Background Introduction	111	13.00	227.88	990,756	1,003,500	353,312	4.31	25.07	0.24
T23	Painting Technique	120	52.00	1.00	729,710	1,003,236	137,280	0.68	1.82	0.22
T24	Source Identification	96	66.78	1.00	554,748	1,003,107	232,944	0.88	2.27	0.49
T25	Painting Description	109	15.00	215.22	985,837	1,003,500	353,312	4.51	44.92	0.24
T26	Painting OCR	81	40.00	84.05	668,913	1,003,080	284,490	1.11	2.19	0.49
T27	Work Introduction	112	8.00	195.50	992,010	1,003,500	353,312	3.07	37.58	0.25
T28	Oracle Bone Script OCR	147	39.00	1.00	30,994	151,321	1,248	0.83	1.18	0.29
T29	Pictographic Decoding	69	23.00	62.90	40,354	97,969	2,916	0.93	1.05	0.44
T30	Seal OCR	153	41.00	3.83	662,505	1,003,230	131,010	0.96	1.26	0.22
T31	Owner Identification	150	53.33	1.00	654,739	1,003,275	144,724	0.97	1.30	0.43
T32	Name Identification	150	73.87	1.00	686,372	1,003,113	49,000	1.20	2.52	0.44
T33	Dynasty Identification	116	51.33	1.00	494,078	1,003,286	18,600	1.36	14.50	0.34
T34	Collection Identification	120	66.49	1.00	673,135	1,003,353	57,750	1.27	2.52	0.26
T35	Cultural Relic Introduction	132	8.00	114.07	714,948	1,003,276	53,550	1.23	7.36	0.25
T36	Cultural Relic Classification	112	51.10	1.00	648,668	1,002,960	60,000	1.29	2.13	0.36
T37	Illustration OCR	182	39.00	4.05	275,976	583,156	73,185	0.93	2.61	0.29
T38	Illustration Description	151	9.58	207.60	410,245	1,003,266	133,285	0.63	1.29	0.34
T39	Entity Introduction	128	12.00	283.38	373,593	1,003,200	39,360	0.65	1.44	0.37
T40	Image-to-Poem Matching	110	156.94	1.00	997,013	1,002,960	430,137	0.69	0.70	0.67
T41	Source Identification	126	60.28	1.00	263,497	666,852	48,884	0.61	1.24	0.29
T42	Topic Classification	120	50.05	1.00	354,121	493,500	177,471	0.68	1.70	0.52
T43	Image-to-Person Matching	110	75.13	1.00	347,995	495,175	176,337	0.64	1.50	0.54
T44	Plot Introduction	100	26.30	328.49	246,172	506,319	133,285	2.02	3.03	0.70
T45	Image-based Q&A	52	16.88	1.77	283,643	496,125	72,819	0.75	1.50	0.42

Table 5: Statistics of 45 tasks. "Avg. Q" represents "the Average length of Questions"; "Avg. A" represents "the Average length of Answers"; "PV" represents "Pixel Value"; and "AR" represents "Aspect Ratio".

translation, punctuation, named entity recognition, word explanation, and reading comprehension.

Additional Notes: (a) The responsible party tests different prompt words and guides LLMs in a few-shot manner to generate well-structured and accurate answer pairs. (b) TongGu (Cao et al., 2024b) is used for generating translations and punctuation based on accurate OCR results. TongGu demonstrates strong performance in translation and punctuation tasks due to its large-scale incremental pretraining and fine-tuning on ancient texts. (c) ERNIE-4.0 (Baidu, 2023) is used for generating named entity recognition, word explanation, and reading comprehension answer pairs. According to the WenMind (an ancient literature benchmark) (Cao et al., 2024a), ERNIE-4.0 excels in ancient text processing and instruction-following capabilities, making it the chosen model.

Web Crawling & Model Summarization:

Responsible parties: Graduate Student A and Graduate Student C.

Selection Criteria: Graduate Student A is skilled in web crawling, and Graduate Student C has extensive familiarity with LLMs and at least one publication in the field of LLMs.

Process: (a) For appreciation tasks (e.g., painting appreciation, calligraphy appreciation) and introduction tasks (e.g., entity introduction, cultural relic introduction), Graduate Student A collects official and accurate content from the Internet. Graduate Student C then uses Qwen2.5-32B (Yang et al., 2024b) to integrate key information. (b) For descriptive tasks (e.g., painting description, illustration description), simple descriptions are first manually generated, and relevant descriptive keywords are extracted. Qwen2.5-32B is then used to refine and expand the descriptions, enriching the text. The responsible party selects and verifies the final descriptions to ensure accuracy.

Additional Notes: The use of Qwen2.5-32B is chosen due to its low cost, moderate memory usage, and excellent instruction-following and long-text summarization capabilities.

Dataset Conversion:

Responsible party: Graduate Student D.

Selection Criteria: Requires a certain level of proficiency in classical Chinese and prior experience in data research, processing, and related tasks.

Process: (a) The responsible party conducts research and collection of existing open-source datasets, selecting classic tasks and high-quality data within CCS. The scarcity and construction difficulty of task data are evaluated to ensure the selection of valuable and currently scarce data, thereby supplementing and improving the evaluation benchmarks. (b) The processing involves text-image filtering and format conversion, guiding the conversion of raw data into question-answer pairs tailored to different tasks.

B.4.2 The Post-processing Stage

The post-processing stage consists of six steps: deduplication, filtering, standardization, language conversion, metadata adding, and match check. We focus on the "deduplication" and "filtering" steps, while the other steps are described in Section 2.2.3 of the main text.

Description of the "deduplication" step:

(a) For image deduplication, we use the CLIP model (Radford et al., 2021) to extract feature vectors of the images and compute the cosine similarity between images within the same task. A threshold of 0.95 is set to determine whether an image is considered duplicate.

(b) For text deduplication, we apply the MinHash algorithm to estimate the Jaccard similarity between question-answer pairs. By mapping each question-answer pair into a low-dimensional hash space and setting a similarity threshold of 0.95, we effectively identify and remove duplicate question-answer pairs.

Description of the "filtering" step:

(a) Responsible parties: Graduate Student A, Graduate Student C, and Volunteer E, where Volunteer E has expertise in CCS.

(b) Approximately 10,000 data points are collected. On one hand, irrelevant, blurry, damaged, or incorrect images are excluded. On the other hand, manual verification of the question-answer pairs is conducted across three dimensions: the correctness and standardization of questions, the accuracy and comprehensiveness of answers, and whether the textual content raises ethical or safety concerns.

(c) Data with inconsistent or unsafe answers is either removed or revised. The remaining data is balanced and filtered for quality through secondary review, resulting in a final dataset of 6,500 entries.

(d) Efforts are made to ensure data security, and it is emphasized that the dataset is solely intended for evaluating model performance in CCS, not for supporting biased viewpoints or inappropriate uses.

B.5 More Examples

Figure 7 to 13 provide additional image examples from different subdomains. Figure 17 to 25 present detailed examples of 45 tasks.

Figure 7: Image examples in the subdomain of Ancient Chinese Text.

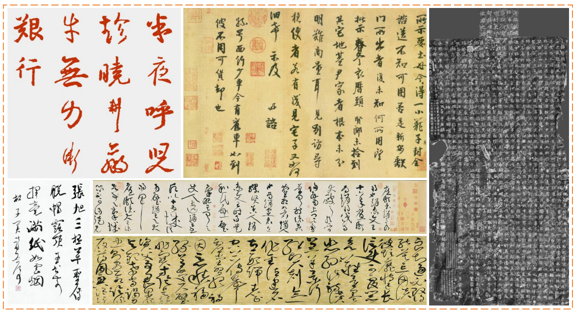


Figure 8: Image examples in the subdomain of Calligraphy.

Figure 9: Image examples in the subdomain of Painting.

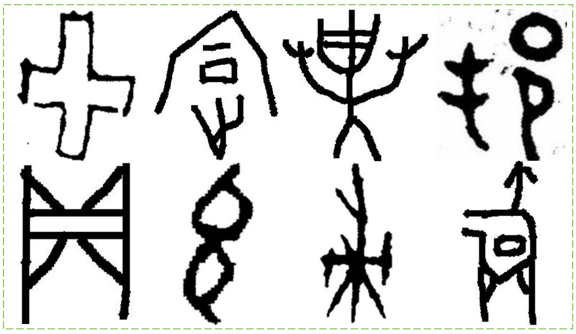


Figure 10: Image examples in the subdomain of Oracle Bone Script.

Figure 11: Image examples in the subdomain of Seal.

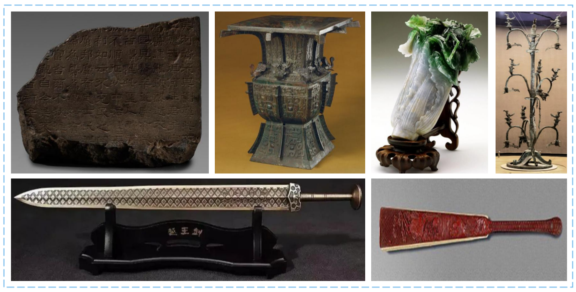


Figure 12: Image examples in the subdomain of Cultural Relic.

C Models and Metrics

C.1 The Evaluated Models

Details of all evaluated MLLMs are shown in Table 6.

C.2 Metrics

Accuracy (Acc). The accuracy is used for all multiple-choice and image-based Q&A tasks. The accuracy can be calculated as follows:

$$Accuracy(Acc) = \frac{N_{\text{correct}}}{N} \tag{1}$$

where N_{correct} and N denote the number of correct answers and the total number of answers. Only when the correct answer is present in the model's response will it be considered correct. When the model answers multiple-choice questions, it is required to output only the letter of the selected option. To ensure fair assessment and avoid the inclusion of multiple letters in the model's responses, we will perform a manual review of the model's answers.

F1-Score. The F1-Score is used to evaluate punctuation, named entity recognition and OCR tasks. Given the predicted and ground truth Key-Value pairs, the F1-Score is formulated as follows:

$$Precision = \frac{N_3}{N_2}, \tag{2}$$

$$Recall = \frac{N_3}{N_1}, \tag{3}$$

$$F1\text{-}Score = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{4}$$

where N_1, N_2, and N_3 denote the number of ground-truth Key-Value pairs, predicted Key-Value pairs, and correctly matched Key-Value pairs, respectively.

BLEU. BLEU (Papineni et al., 2002) is used for translation task. BLEU evaluates prediction quality by comparing n-gram match rates between prediction and ground truth sequences. For each n-gram type, precision is calculated as the ratio of matching n-grams to total prediction n-grams. The final BLEU score is the geometric mean of these precision values multiplied by a penalty BP, which is defined as:

$$BLEU = BP * exp(\sum_{n=1}^{N} w_n \log p_n), \tag{5}$$

$$BP = \begin{cases} 1 & L_p \geq L_g \\ e^{(1-\frac{L_p}{L_g})} & L_p < L_g \end{cases}, \tag{6}$$

Model	Open-Source	# Params	Vision Encoder	Base LLM	Institution	Deployment	Domain
Claude-3.5-Sonnet (Anthropic, 2024)	No	-	-	-	Anthropic	Official API	General
Gemini-1.5-Pro (Team et al., 2024)	No	-	-	-	Google	Official API	General
Gemini-2.0-Flash (Google, 2024)	No	-	-	-	Google	Official API	General
GPT-4o (OpenAI, 2024)	No	-	-	-	OpenAI	Official API	General
Chinese-LLaVA-CLLaMA2 (LinkSoul, 2024)	Yes	7B	CLIP ViT-L	Chinese-Llama-2-7B	LinkSoul	Locally Load	General
DeepSeek-VL2-Tiny (Wu et al., 2024b)	Yes	3B (MoE)	SigLIP-SO400M	DeepSeekMoE-3B	Deepseek-AI	Locally Load	General
DeepSeek-VL2-Small (Wu et al., 2024b)	Yes	16B (MoE)	SigLIP-SO400M	DeepSeekMoE-16B	Deepseek-AI	Locally Load	General
DeepSeek-VL2 (Wu et al., 2024b)	Yes	27B (MOE)	SigLIP-SO400M	DeepSeekMoE-27B	Deepseek-AI	Locally Load	General
GLM-4V-9B (GLM et al., 2024)	Yes	9B	EVA-CLIP-L	GLM-4-9B	Tsinghua	Locally Load	General
InternVL2-4B (OpenGVLab, 2024)	Yes	4B	InternViT-300M-448px	Phi-3-mini-128k-instruct	Shanghai AI Lab	Locally Load	General
InternVL2-8B (OpenGVLab, 2024)	Yes	8B	InternViT-300M-448px	InternLM2_5-7b-chat	Shanghai AI Lab	Locally Load	General
InternVL2-26B (OpenGVLab, 2024)	Yes	26B	InternViT-6B-448px-V1-5	InternLM2_5-20b-chat	Shanghai AI Lab	Locally Load	General
InternVL2.5-1B (Chen et al., 2024b)	Yes	1B	InternViT-300M-448px-V2_5	Qwen2.5-0.5B-Instruct	Shanghai AI Lab	Locally Load	General
InternVL2.5-2B (Chen et al., 2024b)	Yes	2B	InternViT-300M-448px-V2_5	InternLM2_5-1_8b-chat	Shanghai AI Lab	Locally Load	General
InternVL2.5-4B (Chen et al., 2024b)	Yes	4B	InternViT-300M-448px-V2_5	Qwen2.5-3B-Instruct	Shanghai AI Lab	Locally Load	General
InternVL2.5-8B (Chen et al., 2024b)	Yes	8B	InternViT-300M-448px-V2_5	InternLM2_5-7b-chat	Shanghai AI Lab	Locally Load	General
InternVL2.5-26B (Chen et al., 2024b)	Yes	26B	InternViT-6B-448px-V2_5	InternLM2_5-20b-chat	Shanghai AI Lab	Locally Load	General
InternVL2.5-38B (Chen et al., 2024b)	Yes	38B	InternViT-6B-448px-V2_5	Qwen2.5-32B-Instruct	Shanghai AI Lab	Locally Load	General
InternVL2.5-78B (Chen et al., 2024b)	Yes	78B	InternViT-6B-448px-V2_5	Qwen2.5-72B-Instruct	Shanghai AI Lab	Locally Load	General
LLaVA-v1.5-7B (Liu et al., 2024a)	Yes	7B	CLIP ViT-L	Vicuna-7B	UW_Madison	Locally Load	General
LLaVA-v1.5-13B (Liu et al., 2024a)	Yes	13B	CLIP ViT-L	Vicuna-13B	UW_Madison	Locally Load	General
LLaVA-v1.6-Mistral-7B (Liu et al., 2024b)	Yes	7B	CLIP ViT-L	Mistral-7B	UW_Madison	Locally Load	General
LLaVA-v1.6-Vicuna-7B (Liu et al., 2024b)	Yes	7B	CLIP ViT-L	Vicuna-7B	UW_Madison	Locally Load	General
LLaVA-v1.6-Vicuna-13B (Liu et al., 2024b)	Yes	13B	CLIP ViT-L	Vicuna-13B	UW_Madison	Locally Load	General
MiniCPM-V (Yao et al., 2024)	Yes	3B	SigLip-400M	MiniCPM-2.4B	OpenBMB	Locally Load	General
MiniCPM-V-2 (Yao et al., 2024)	Yes	3B	SigLip-400M	MiniCPM-2.4B	OpenBMB	Locally Load	General
MiniCPM-LLaMA3-V-2.5 (Yao et al., 2024)	Yes	8B	SigLip-400M	Llama3-8B-Instruct	OpenBMB	Locally Load	General
MiniCPM-V-2.6 (Yao et al., 2024)	Yes	8B	SigLip-400M	Qwen2-7B	OpenBMB	Locally Load	General
Molmo-7B-D-0924 (Deitke et al., 2024)	Yes	7B	CLIP ViT-L	Qwen2-7B	AllenAI	Locally Load	General
Molmo-7B-O-0924 (Deitke et al., 2024)	Yes	7B	CLIP ViT-L	OLMo-7B-1024	AllenAI	Locally Load	General
Ovis1.5-Gemma2-9B (Lu et al., 2024)	Yes	9B	SigLip-400M	Gemma2-9B-It	AIDC-AI	Locally Load	General
Ovis1.6-Gemma2-9B (Lu et al., 2024)	Yes	9B	SigLip-400M	Gemma2-9B-It	AIDC-AI	Locally Load	General
Qwen-VL-Chat (Bai et al., 2023)	Yes	7B	CLIP ViT-bigG	Qwen-7B-Chat	Alibaba	Locally Load	General
Qwen2-VL-2B-Instruct (Wang et al., 2024a)	Yes	2B	DFN CLIP ViT-L	Qwen2-1.5B-Instruct	Alibaba	Locally Load	General
Qwen2-VL-7B-Instruct (Wang et al., 2024a)	Yes	7B	DFN CLIP ViT-L	Qwen2-7B-Instruct	Alibaba	Locally Load	General
Qwen2-VL-72B-Instruct (Wang et al., 2024a)	Yes	72B	DFN CLIP ViT-L	Qwen2-72B-Instruct	Alibaba	Locally Load	General
QVQ-72B-Preview (Qwen, 2024)	Yes	72B	DFN CLIP ViT-L	Qwen2-72B-Instruct	Alibaba	Locally Load	General

Table 6: Details of all evaluated MLLMs. Zoom in for better view.

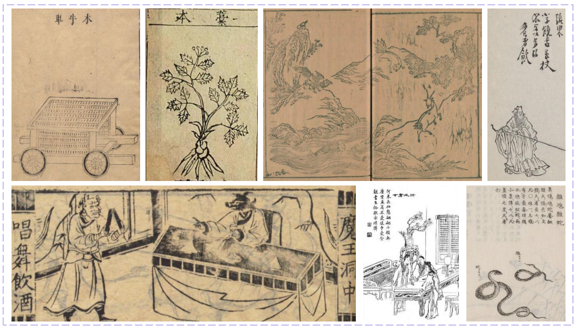


Figure 13: Image examples in the subdomain of Illustration.

where p_n represents the precision of n-grams, L_p represents the length of prediction sequence, L_g represents the length of ground truth sequence, w_n is weight factor, usually evenly distributed ($w_n = \frac{1}{N}$). Typically, N is set to 4.

BERTScore and ANLS. We used the average of BERTScore (Zhang et al., 2020) and ANLS (Biten et al., 2019) for open-ended QA.

$$ANLS = \frac{1}{N}\sum_{i=1}^{N}\max(0, 1 - \frac{\text{NLD}(y_i, \hat{y}_i)}{|y_i|}) \tag{7}$$

$$NLD(y_i, \hat{y}_i) = \frac{\text{LD}(y_i, \hat{y}_i)}{\max(|y_i|, |\hat{y}_i|)} \tag{8}$$

where N represents the number of samples, y_i represents the ground truth text for the i-th sample, $\hat{y}_i$ represents the predicted text for the i-th sample, NLD represents the Normalized Levenshtein Distance, $|y_i|$ represents the length of the ground truth text and LD represents the Levenshtein Distance.

CR and AR. AR and CR (Peng et al., 2023) are used for OCR task. They can be calculated as follows:

$$AR = \frac{N_t - D_e - S_e - I_e}{N_t} \tag{9}$$

$$CR = \frac{N_t - D_e - S_e}{N_t} \tag{10}$$

where D_e, S_e, and I_e represent the total number of deletion, substitution, and insertion errors, respectively, and N_t is the total number of characters in the annotations.

ROUGE. ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004) are used for translation task. They can be calculated as follows:

$$ROUGE\text{-}1 = \frac{\sum_{g\in\text{ref}}\sum_{u\in\{\text{1-gram}\}}\text{Count}_{\text{match}}(u)}{\sum_{g\in\text{ref}}\sum_{u\in\{\text{1-gram}\}}\text{Count}(u)} \quad (11)$$

$$ROUGE\text{-}2 = \frac{\sum_{g\in\text{ref}}\sum_{b\in\{\text{2-gram}\}}\text{Count}_{\text{match}}(b)}{\sum_{g\in\text{ref}}\sum_{b\in\{\text{2-gram}\}}\text{Count}(b)} \quad (12)$$

$$ROUGE\text{-}L = \frac{(1+\beta^2)\text{P}_{\text{lcs}}\text{R}_{\text{lcs}}}{\text{R}_{\text{lcs}}+\beta^2\text{P}_{\text{lcs}}} \quad (13)$$

$$P_{\text{lcs}} = \frac{\text{LCS}(X,Y)}{|X|} \quad (14)$$

$$R_{\text{lcs}} = \frac{\text{LCS}(X,Y)}{|Y|} \quad (15)$$

where $\text{Count}_{\text{match}}(u)$ represents the number of matched n-grams between reference and candidate, $\text{Count}(u)$ represents the total number of n-grams in reference, $\text{LCS}(X,Y)$ represents the length of Longest Common Subsequence between X and Y, X represents the candidate text, Y represents the reference text, $|X|$ represents the length of candidate text, $|Y|$ represents the length of reference text and β represents a parameter that determines the importance of precision and recall (usually $\beta = 1.2$).

Normalized Edit Distance (NED). Normalized Edit Distance (NED) (Levenshtein, 1966b) is used for OCR task, which measures string similarity by computing the minimum number of operations needed to transform one string into another. The calculation is formulated as follows:

$$NED(S_1,S_2) = \frac{ED(S_1,S_2)}{\max(len(S_1),len(S_2))} \quad (16)$$

where $ED(S_1,S_2)$ represents the edit distance between the prediction string S_1 and the ground truth S_2. The NED value of 0 indicates identical strings, while 1 indicates completely different strings.

The evaluation metrics and question formats corresponding to 45 tasks are shown in Table 14.

D Detailed Metrics

Table 15 to 21 present the metrics for 37 MLLMs across all tasks. The metrics corresponding to each task are from Table 14 (Metrics-Main). Table 22 to 28 provide the additional metrics for 37 MLLMs across all tasks. The metrics for each task are from Table 14 (Metrics-Others).

E Error Analysis

We conduct an error analysis for all tasks below.

Figure 14: Challenging text OCR samples.

E.1 T1-Text OCR

As shown in Figure 14, MLLMs primarily achieve lower scores on the following types of ancient text images: Text-intensive, Image Distortion, Complex Layout, and Difficult-to-Read Fonts. Additionally, some MLLMs recognize text in an incorrect reading order, leading to lower scores (the correct reading order for ancient texts is right-to-left and top-to-bottom). Figure 26 presents sample model responses for T1.

E.2 T2-Classical Chinese to Modern Chinese, T3-Punctuation, T4-Named Entity Recognition

Since the challenges of T2, T3, and T4 are similar, we use T2 as an example for illustration. Through sample analysis, we identify the main challenges of the task as follows: (a) Compared to T1, the difficulty of images in other tasks is slightly reduced; however, most models still struggle to recognize complete content, which affects subsequent translation. (b) The translation capabilities of the models themselves are relatively poor. (c) Error-prone images tend to include those with dense text or darker tones. (d) Some models exhibit overfitting behavior, fail to follow translation instructions, and output only the OCR results. (e) During translation, models often produce incomplete outputs, translating only the initial content or generating repetitive outputs. Figure 27 presents sample model responses for T4.

E.3 T5-Word Explanation

Through error analysis, we identify the following issues: (a) Models struggle to accurately locate the positions of relevant terms in ancient Chinese texts. Due to the multiple meanings of the same term in ancient contexts, models often fail to determine the correct meaning based on the surrounding context. (b) Even when models locate the correct position of a term, they are easily confused by other distractions. For example, the term “Qu” in Classical Chinese can mean “approach,” “small, quick steps,” “interest,” or “urge.” Based on the article’s content, “Qu” should mean “to take small steps toward the door,” but many models choose distracting options such as “swiftly” or “urge,” even when provided with correct OCR results. This indicates that models do not have a solid understanding of the multiple meanings of terms and their relation to the article’s content.

E.4 T6-Reading Comprehension

The errors made by the models are concentrated in the following types of questions: (a) Questions requiring the synthesis of complex details, such as summarizing sacrificial rituals and character behaviors, where models struggle to extract and integrate all key details. (b) Historical questions with high background knowledge requirements, such as the context of events or the changes in place names (e.g., the evolution of “Ancheng”), where models make mistakes due to insufficient knowledge. (c) Questions involving abstract evaluation and sentiment analysis, such as analyzing social phenomena or the author’s attitude, where models perform poorly in identifying emotional tendencies and abstract understanding. (d) Questions requiring the reconstruction of scene details, such as depictions of nighttime or battle scenes, where models fail to accurately reproduce the intricate details of complex scenarios.

The main causes of errors include: (a) Incomplete detail extraction, where models tend to overlook or simplify complex details. (b) Insufficient background knowledge, which hinders their ability to handle questions requiring extensive historical and cultural understanding. (c) Weak reasoning abilities, resulting in poor performance on questions requiring logical inference or contextual connections. (d) Limited multimodal integration capabilities, making it difficult for models to accurately interpret content in ancient Chinese texts.

E.5 T7-Source Attribution

The books in which models are prone to errors in source attribution tasks include Song Shu, Han Shu, Huainan Honglie Jie, Baopuzi, Sanguozhi, Daode Zhenjing Jiyi, and Chunqiu Fanlu. The models demonstrate a poor understanding of the content associated with these texts.

Font: Cursive script

OCR: 失去了也没关系对自己我只想问心无愧

Font: Seal script

OCR: 风流不见秦淮海寂寞人间五百年

Figure 15: Challenging calligraphy OCR samples.

E.6 T8-Calligraphy OCR

We find that models exhibit poor OCR recognition capabilities for Cursive Script and Seal Script (as shown in Figure 15). The primary reasons are as follows:

First, the structural complexity of Cursive Script and Seal Script is high, with deformations in character shapes and strokes, as well as frequent ligatures, making accurate recognition challenging. When models lack sufficient ability to extract and identify character details, errors are more likely to occur. Second, the character shapes of Cursive Script and Seal Script differ significantly from Regular Script and have an indirect correspondence with modern Chinese characters. The limited training data for these scripts during the pretraining process leads to weak recognition performance for these calligraphic styles. The low recognition scores for Cursive Script and Seal Script reflect the models’ limitations in handling complex character structures and insufficient training data. Additionally, some MLLMs are unaware of the correct reading order for calligraphy, resulting in further errors. Figure 28 presents sample model responses for T8.

E.7 T9-Calligraphy Appreciation

The calligraphy works with lower scores in appreciation tasks primarily include Sangluan Tie, Zheng Wen Gong Bei, Yi He Ming, Cuan Longyan Bei, Dongfang Shuo Hua Zan Bei, and Hanqie Tie. It is evident that MLLMs exhibit weaker appreciation capabilities for “inscription-style” calligraphy works and are not sufficiently familiar with this category.

The low scores of MLLMs are primarily attributed to several factors: (a) The models fail to effectively recognize the calligraphy works, resulting in an inability to provide detailed analysis of specific artistic features. (b) The models do not accurately express the technical and artistic aspects of the calligraphy, lacking in-depth analysis of brushwork, character structure, composition, and emotional expression. (c) The models sometimes fail to accurately grasp the historical context and cultural significance of the calligraphy works, leading to an oversight of their impact in the context of their time and culture.

E.8 T10-Author Identification

We find that: (a) Models often confuse the works of famous calligraphers such as Zhao Mengfu, Huang Tingjian, and Wang Xizhi. (b) For some calligraphers with fewer works or lesser-known reputations, models tend to favor more prominent calligraphers' options.

E.9 T11-Dynasty Identification

We find that: (a) Models are most prone to making errors in judgment for calligraphy works from the Yuan Dynasty, followed by the Ming and Sui-Tang periods. (b) Calligraphy works from different dynasties often share similar styles, making it difficult for models to differentiate them. Additionally, many models struggle to accurately identify the specific calligraphy work from an image, which limits their ability to reason about the corresponding author and dynasty, due to a lack of calligraphy image knowledge and the limited capability for multi-hop reasoning.

E.10 T12-Title Identification

The calligraphy works with a high error rate include Huang Ying Qu, Lian Po Lin Xiangru Liezhuan, Chen Shuyong Old Collection, Province Exam After Winter Solstice Wangsong Palace, Shenji Tie, Zhuzi Jiaxun, Shu Su Tie, Ni Kuanzan, Junren Tie, and others. Surprisingly, models exhibit lower accuracy on renowned calligraphy works such as Shu Su Tie, Ni Kuanzan, and Junren Tie, indicating that some models struggle to directly identify these works from images, often resorting to random guessing for multiple-choice questions. Figure 29 presents sample model responses for T12.

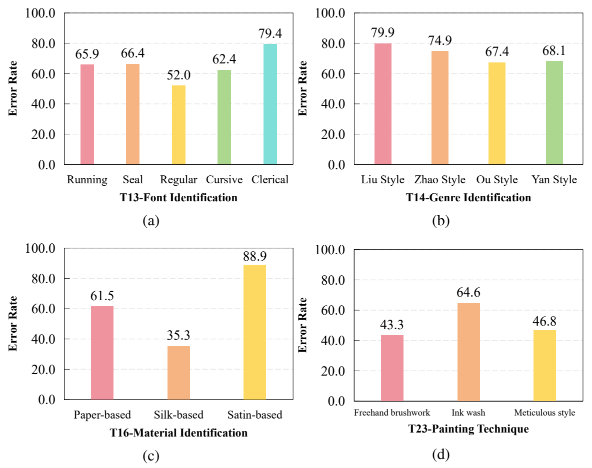


Figure 16: The error rates for each option in T13, T14, T16, and T23.

E.11 T13-Font Identification

As shown in Figure 16 (a), the font with the highest error rate in the model is Clerical Script, followed by Seal Script, Running Script, Cursive Script, and Regular Script. The model finds Regular Script easier to differentiate. Compared to other fonts, Regular Script has a more regular stroke structure with clear and distinct strokes. It is the most commonly used font in everyday life, making it more straightforward for the model to recognize. The model has a higher error rate when recognizing Clerical Script, primarily due to its complex character structure, curved and varying strokes. Additionally, the training data may be insufficient, and Clerical Script significantly differs from modern, commonly used fonts in terms of strokes, shapes, and styles, making accurate recognition challenging.

E.12 T14-Genre Identification

Although the model performs well in differentiating Regular Script from other fonts, its ability to distinguish between different branches within Regular Script is weak. As shown in Figure 16 (b), the most frequently confused branches are Liu style, followed by Zhao, Yan, and Ou styles. Among all calligraphy subdomain recognition tasks, the poorest performance is in genre identification, followed by dynasty identification tasks.

E.13 T15-Work Introduction

The works with lower scores primarily include: Stele of King Gwanggaeto, Poems of General Pei, Dazifujing by Chu Suiliang, Heart Sutra, Stele of Military Strategy, Songs from Songfeng Pavilion, Cuanglongyan Stele, Ascension of the Immortal Prince, Yan Qinli Stele, and others. Similar to T9, the model demonstrates weaker familiarity with "inscription-style" works. Additionally, when providing work descriptions, we observe:

(a) Repetitiveness and redundancy. Responses may contain excessive repetition or have multiple answers to similar questions that are overly similar, lacking focus and depth. (b) Insufficient details. The descriptions of some works may be overly brief, failing to thoroughly explore the background, style, historical significance, etc.

E.14 T16-Material Identification

In calligraphy works, the model's ability to differentiate between "paper" and "silk" materials is relatively weak. As shown in Figure 16 (c), the error rates, in descending order, are as follows: Satin-based, Paper-based, and Silk-based.

E.15 T17-Painting Appreciation

The model's responses often lack in the following aspects: (a) Insufficient detail description. The model tends to overlook or simplify details such as texture, composition, and spatial hierarchy. (b) Absence of specialized terminology. The model struggles to accurately describe certain art terms (e.g., "Pi Ma Cun," "Tie Xian Miao"). (c) Weak overall context. When integrating the scene and narrative logic, the model may fragment or distort the context. (d) Insufficient understanding of cultural background. Some artworks' historical and cultural context (e.g., Emperor Huizong's prayers or the openness of Tang dynasty attire) are crucial for appreciation, but the model may lack a deep understanding or fail to effectively connect these aspects.

E.16 T18-Author Identification

The identification rate for painters such as Dai Jin, Sheng Mao, Wang Zhong, Fu Ru, Ye Fanglin, and Zhou Kun is relatively low. Among them, works by Dai Jin, Zhou Kun, and others consistently receive lower scores.

E.17 T19-Dynasty Identification

The identification rate for dynasties such as the Ming and Qing is relatively low. Similar to T11, many models struggle to accurately identify the specific painting works corresponding to the images, leading to difficulties in obtaining the associated authorship and dynasty information.

E.18 T20-Title Identification

In paintings, works with lower recognition rates mainly include: Wan Jing Chunxi, Chibi Tu, Duju Tu, Huangzhong Changyue Tu scroll, Wanjia Chunyu, Jiaoshu Tu, Taibai Mountain Tu, etc. Among them, we find that the model is less familiar with the paintings in the "Wan Chun Jiqing" album.

E.19 T21-Image-based Q&A

In this task, we mainly set questions including counting, character profession and behavior, color recognition, specific item search, pattern recognition, and other types of questions. Among these, the model is least proficient in counting questions, particularly those requiring precise answers about specific objects (e.g., animals, seals, leaves, etc.). Figure 30 presents sample model responses for T21.

E.20 T22-Background Introduction

Paintings with lower recognition rates mainly include: Zhao Mengfu's "Rider Painting," Song Huizong's "Snowy River Returning Boat," Shen Zhou's "Two Rivers Scenic Views," Qiu Ying's "Peach Blossom Spring," Li Song's "Peddler Painting," and Fu Baoshi's "Erxiang Tu," among others. These works have a certain level of fame, and the model has a basic understanding of the paintings' titles and key information. However, the model is less familiar with the contextual background of these works, especially the stories and imagery associated with them.

E.21 T23-Painting Technique

As shown in Figure 16 (d), for the three painting techniques, the highest error rate is for the option "Ink wash," followed by meticulous style painting, and finally freehand brushwork painting.

E.22 T24-Source Identification

We select representative albums as questions, mainly including: Xingxing Erbashi Shensheng Tu, Mofa Jiyao Tu Juan, Jiuxia Anhe Tu Ce, Shier Jinyi Jing Tu Scroll, and others. Among these, the model has lower recognition rates for works such as: Wan Chun Jiqing, Momeiao Zhulin, Bafeng Tiaoyu Tu Ce, and Porcelain Zhi Yun Tu.

In all identification tasks within the painting subdomain, the model performs poorly mainly in author identification and source identification, followed by dynasty identification.

E.23 T25-Painting Description

In painting description, works with lower scores mainly include: Four Scenery Mountain and Water Scroll, Chushi Beijiang Scroll, Han Xizai Night Banquet Scroll, You Ran Jian Nanshan, and Zhong Kui Rainy Night Outing Scroll, among others.

The model's low score in answering these painting description questions has several main reasons: (a) Inadequate details. These questions require the model to describe specific details and content of the artwork. Often, only general answers are provided, lacking in-depth analysis of the composition, character activities, background details, and other aspects. If descriptions are too simplistic, they may fail to fully capture the richness and depth of the work, affecting the score. (b) Language expression issues. While the model's responses can effectively convey the theme of the painting, the language may sometimes lack fluency or vividness, missing the nuanced visual depiction. An excellent painting description should enable readers to "see" the artwork rather than merely summarize its content. Additionally, some model responses are overly brief and vague, applying the same phrasing to most artworks, which fails to provide specificity.

E.24 T26-Painting OCR

In painting OCR, works with lower scores mainly include: Porcelain Zhi Yun Tu and Momeiao Zhulin. Among these, the painting section in Porcelain Zhi Yun Tu occupies a large portion, while text accounts for a very small part. As a result, the model is easily influenced by the visual content, making it difficult to capture the finer textual details.

E.25 T27-Work Introduction

Works with lower scores mainly include: Fu Zui Tu, Nü Le Tu, Song Wang Kui Ying Shan Tu, Hu Tinghui Chunshan Fanzhou Tu Scroll, Bie Yuan Gualan Tu, and Qingxi Yinma Tu, among others. The model performs well on more well-known paintings (e.g., Nüshi Zhen Tu, Erjun Tu, Xuejing Hanlin Tu). However, for works with lower scores, the model has difficulty identifying the titles of the works, and responses are often too broad. Figure 31 presents sample model responses for T27.

E.26 T28-Oracle Bone Script OCR

The oracle bone OCR task is highly challenging, with the overall score rate for the model approaching zero. Only a few simple oracle bone characters have an average score rate above zero, such as: 王, 听, 平, 可, 舌, 铜, 工, 鸟, 子, 甲, 井, among others, totaling 36 characters. The remaining 111 characters have an OCR score of zero. Figure 32 presents sample model responses for T28.

E.27 T29-Pictographic Decoding

The model performs poorly in interpreting certain characters, including: 未, 已, 辛, 戊, etc.

The poor performance can be attributed to the following reasons: (a) Lack of sufficient knowledge of pictographic characters. The model struggles with understanding the historical and cultural context of oracle bone shapes and meanings, making it difficult to accurately reconstruct the origins and evolution of pictographic characters. (b) Limited semantic reasoning ability. Interpreting pictographic characters requires in-depth reasoning that combines shape features with abstract meanings, an area where the model performs poorly. (c) Some oracle bone structures are complex, making it difficult for the model to accurately interpret their pictographic meanings.

E.28 T30-Seal OCR

The model performs poorly on many questions. The main reasons include: (a) Complex shapes. The characters in seals often have unique seal carving styles, with intricate brushstrokes and diverse forms, making them difficult to accurately interpret. (b) Deformation and irregularity. Seal characters may be distorted, have broken strokes, or exaggerated brushstrokes due to carving techniques, which increases the difficulty of recognition. (c) High-density arrangement. Seal text is often densely packed, with small character spacing, leading to segmentation errors or confusion.

Figure 33 presents sample model responses for T30.

E.29 T31-Owner Identification

The characters with lower identification rates mainly include: Li Shihang, Huang Zhou, Jin Yue, Ning Fu Cheng, Jin Cheng, He Zuopeng, and others.

The reasons include: (a) Difficulties in shape recognition. To identify the owner of a seal, it is necessary to first recognize the textual content associated with the seal. The complex and diverse styles of seal fonts make it challenging for the model to accurately interpret the text. (b) Ambiguous feature matching. Identifying the identity of a person requires matching seal text with options, but seals may lack obvious visual indicators. (c) Limited information. Isolated seal texts often provide insufficient information, requiring a robust knowledge base to accurately interpret them.

E.30 T32-Name Identification

The task of identifying cultural relic names is relatively simple, with the model achieving high overall performance, even reaching 100% accuracy with GPT-4o (OpenAI, 2024). The cultural relic names with higher error rates mainly include: Guandi Sitting Statue, Anyang Pingjian Cloth, Ding Chou Jinshi, Qingxu Family Mountain and Water Fan, and Qing Qianlong Tiantao Jiuru Red and Black Ink, among others.

The model performs well due to the following reasons: (a) The format and structure of cultural relic names are relatively standardized. Naming cultural relics often follows a fixed descriptive pattern, such as dynasty, material, and purpose, which facilitates model identification. (b) High semantic clarity. Cultural relic names have unique semantic features that distinctly differentiate them from language patterns in other fields, reducing ambiguity. (c) Clear task options. Options provide clear distinctions, offering a well-defined benchmark for comparison. (d) Cultural relic data is more prevalent compared to data in other fields, making it easier for the model to learn relevant knowledge from training data.

E.31 T33-Dynasty Identification

The model has a high error rate for cultural relics from the Ming dynasty, followed by the Qing dynasty, Eastern Han, and others.

The model performs poorly due to the following data types: (a) Lack of distinct features. Some cultural relics may lack clear dynasty features, such as material, style, or craftsmanship, making it difficult to distinguish between different dynasties. (b) Similar dynasties. For certain dynasties, such as Ming and Qing, there may be many similarities in the appearance of cultural relics, making it challenging for the model to make accurate distinctions.

E.32 T34-Collection Identification

The model performs moderately in this task, with no significant difference in identification accuracy across different museums.

The main reasons include: (a) Insufficient visual information. Features of the collection are not adequately represented in the artifact images, making it difficult for the model to make accurate distinctions based on these details, requiring a strong knowledge base. (b) Repetitive artifact features. Many artifacts have similar appearances, making it challenging for the model to differentiate their origins.

E.33 T35-Cultural Relic Introduction

Cultural relics with lower scores mainly include: Terracotta Warriors, Wu Ling Clay Figurines, Modern Liang Qichao Letter Roll, Ink Bamboo Stone Scroll, White Script Stone Seal “Wu Junqing”, among others.

We find that the model performs well for very famous cultural relics (e.g., Galloping Horse Statue, King Goujian Sword, Zeng Houyi Bell, Four Sheep Square Zun, etc.), but performs worse for less well-known relics. In fact, for cultural relics with lower scores, the model may even incorrectly identify the type of relic.

Figure 34 presents sample model responses for T35.

E.34 T36-Cultural Relic Classification

The task of classifying cultural relics is relatively straightforward, and the model generally performs well. The reasons for this include: (a) Clear classification. Cultural relic types and categories typically follow standardized classification criteria, allowing the model to identify them based on these clear standards. (b) Distinct category differences. Different categories of cultural relics exhibit significant differences in form, material, and use, making it easier for the model to distinguish them based on these visual features. (c) Rich visual features. Cultural relics have unique appearance characteristics in different classifications, making it effective for the model to use these features for identification.

In the identification tasks within the cultural relic subdomain, name identification performs the best on average, followed by cultural relic classification. Collection identification performs the worst on average.

E.35 T37-Illustration OCR

The lower-scoring questions mainly come from Dream of the Red Chamber, Ben Cao Gang Mu, and Wu Bei Zhi.

The main reasons include: (a) The model is easily disrupted by illustrations, making it difficult to focus on the textual portions, especially illustrations from Dream of the Red Chamber. (b) Some of the characters in illustrations are in traditional Chinese, making the shapes more complex or blurry, leading to difficulties in recognition. (c) The reading order of text in illustrations is inconsistent with modern conventions.

Figure 35 presents sample model responses for T37.

E.36 T38-Illustration Description

The lower-scoring questions mainly come from Shan Hai Jing and Tiangong Kaiwu.

The main reasons include: (a) Unique illustration style. The illustrations in Shan Hai Jing are bold and exaggerated, with abstract details, while Tiangong Kaiwu features densely detailed and intricate linework. The model may struggle to accurately understand and describe these illustrations. (b) High background knowledge requirement. Understanding the illustrations in Shan Hai Jing and Tiangong Kaiwu requires a certain level of historical and cultural knowledge, which the model may not fully grasp, leading to imprecise descriptions. (c) Complex subjects. Illustrations often include mythological, natural, and abstract elements, making it challenging for the model to describe these intricate and symbolic images accurately.

E.37 T39-Entity Introduction

The lower-scoring questions mainly involve the identification of person entities (e.g., Liexian Wine Plaque) and plant identification in Xinbian Leiyao Tu Zhu Ben Cao.

The reasons include: (a) Difficulties in identifying person entities. The figures in Liexian Wine Plaque are often abstract or mythologized, requiring a certain level of historical knowledge. (b) Unclear plant features. In Xinbian Leiyao Tu Zhu Ben Cao, plant representations are often simplified, and there are many similar-looking plants, making it challenging for the model to accurately differentiate and identify them. (c) High domain knowledge requirement. These images involve specific historical backgrounds, cultural symbols, and botanical knowledge, necessitating not only entity recognition but also the knowledge of related information.

E.38 T40-Image-to-Poem Matching

The model's overall performance is moderate, with four data points showing significantly lower accuracy.

The reasons include: (a) The natural elements and emotional descriptions in the poetry are rich, involving concepts like flowers, wind, and mountains, which require higher understanding. (b) The images include multiple elements or layers, making them complex, and the model struggles to accurately match the images with the corresponding poetry. (c) There is strong ambiguity between the options for these four data points.

E.39 T41-Source Identification

The identification rate for illustrations from Yin Shan Zheng Yao is relatively low.

The reasons include: (a) The illustrations in Yin Shan Zheng Yao primarily depict food-related content, with relatively simple visual features and no significant visual differences, making them susceptible to misidentification with other similar themes, such as those in Ben Cao Gang Mu or Mao Shi Pin Wu Tu Kao, which also cover plants and food-related topics, leading to higher distractor rates. (b) Other types of illustrations require the model to have a certain level of background knowledge, leading to additional errors.

E.40 T42-Topic Classification

In this task, the model often confuses the categories of "story" and "character." The reason is that the model finds it challenging to distinguish whether the illustration focuses on "portraying a specific story" or "depicting a specific character." The task has a high overall performance, with the model demonstrating a relatively good level of performance, largely because the classification is not overly difficult.

E.41 T43-Image-to-Person Matching

The high error rate in this task may be attributed to: (a) The historical or legendary figures involved in these questions are complex and may appear in different cultural contexts and historical events. The features in the images may not be sufficiently intuitive, making it challenging for the model to accurately match the characters. (b) Some illustrations of figures are quite ambiguous, especially for characters like "Yang Guifei," "Yizhu," or "Xiang Zhong." (c) Certain options in the task require the model to identify typical visual clues, such as clothing, posture, etc., which increases the difficulty of making accurate judgments.

E.42 T44-Plot Introduction

We select Journey to the West and Strange Stories from a Chinese Studio as representative topics. Compared to Journey to the West, the overall accuracy rate is lower.

The reasons include: (a) Journey to the West is a richly detailed and complex classical novel with numerous storylines, many of which may carry strong cultural backgrounds or symbolic meanings, making it challenging for the model to accurately match illustrations with specific content. In contrast, Strange Stories from a Chinese Studio consists of individual stories, making the task less difficult. (b) Compared to Strange Stories from a Chinese Studio, the illustrations in Journey to the West are more abstract, which hinders the model from accurately capturing and describing specific storylines.

E.43 T45-Image-based Q&A

The questions with lower accuracy rates include: (a) Topics that involve abstract content or excessive detail, such as "What is the person in the image doing?" These types of questions require the model to accurately extract and understand details from images, which presents a challenge. (b) Some questions, like "How many layers are there of objects in the image?" or "How many nails are there on the large pry in the image?" require precise counting and spatial recognition abilities, which are complex tasks for image understanding. (c) Certain questions involve specific cultural backgrounds or symbols that are not easily interpretable (e.g., "Ziwengui," "Tianxiang"), potentially making it difficult for the model to understand the context and provide accurate answers.

Figure 36 presents sample model responses for T45.

F Supplementary Experiments and Analysis

F.1 Human Performance Baseline

In the evaluation results of MCS-Bench, the human performance baseline reaches 68.24, which is significantly higher than the scores of all current mainstream MLLMs, outperforming the best-performing model by 18.92 points. This indicates that there remains substantial room for improvement in the CCS domain. Although some models perform comparably on a few individual tasks, none surpass the 50-point threshold overall. This suggests notable limitations in their ability to comprehend ancient texts, recognize calligraphic styles, interpret visual details, and grasp cultural contexts. The performance of models fluctuates particularly in tasks requiring cultural background knowledge, visual-semantic integration, and complex reasoning.

Model	**Overall**
Qwen2-VL-72B-Instruct	44.56
GPT-4o	44.72
InternVL2.5-26B	45.92
Gemini-2.0-Flash	46.19
InternVL2.5-38B	46.55
InternVL2.5-78B	49.32
Human Baseline	**68.24**

Table 7: Comparison of the human baseline and the top six MLLMs on MCS-Bench.

The human baseline is derived from responses to the entire set of benchmark tasks, completed by two graduate students majoring in electronic information, each scoring 120 or above in the Chinese language section of the national college entrance examination (Gaokao). With solid literary competence and multimodal comprehension abilities, they represent a general population with basic humanities literacy. We deliberately select evaluators without specialized backgrounds in classical studies to ensure a more representative human benchmark while avoiding inflated scores due to expert knowledge. This setup allows for a more accurate reflection of the current performance gap between MLLMs and human-level capabilities in the CCS domain.

F.2 OCR+LLM Performance Baseline

We conduct experiments to validate the effectiveness of the OCR+LLM paradigm as a baseline for our benchmark, which emphasizes OCR-heavy tasks.

First, we evaluate the OCR performance of two tools: PaddleOCR (Baidu, 2021), a general-purpose OCR system, and KanDianGuJi (GuJi, 2023), which is specifically designed for ancient Chinese texts. In the "ancient texts" subdomain, PaddleOCR achieves an OCR accuracy of 49.69, while KanDianGuJi significantly outperforms it with a score of 83.68. This substantial gap highlights the necessity of domain-specific OCR solutions for handling historical documents.

Table 8 presents the BLEU scores of three LLMs

Model	GT	KanDianGuJi	PaddleOCR
Qwen2.5-7B-Instruct	**16.84**	**14.77**	**6.76**
InternLM3-8B-Instruct	13.13	11.26	4.29
LLaMA3-Chinese-8B-Instruct	1.69	1.61	0.79

Table 8: Performance of the OCR+LLM paradigm on the T2 task (Metric: BLEU).

Model	BLEU
Qwen2.5-VL-7B-Instruct	2.50
Qwen2-VL-7B-Instruct	1.51
InternVL2.5-8B	1.36
MiniCPM-LLaMA3-V-2.5	0.05
Gemini-2.0-Flash	**10.36**
Gemini-1.5-Pro	10.13

Table 9: Performance of direct use of MLLMs on the T2 task.

Model	OverAll
Gemini-2.0-Flash	46.19
Qwen2-VL-72B-Instruct	44.56
Gemini-2.0-Flash-Thinking-Exp	**48.14**
QVQ-72B-Preview	42.54
LLaMA-3.2V-11B-CoT	30.11

Table 10: Comparison of slow-thinking and baseline models on MCS-Bench.

given the outputs from the two OCR tools as well as the ground truth (GT). The results show that KanDianGuJi consistently leads to better translation performance than PaddleOCR across all models, confirming the strong impact of OCR quality on downstream tasks.

Among the evaluated LLMs, Qwen2.5-7B-Instruct (Yang et al., 2024b) consistently achieves the highest BLEU scores, demonstrating its superior capacity for ancient text translation compared to InternLM3-8B-Instruct (InternLM, 2024) and LLaMA3-Chinese-8B-Instruct (Joint Laboratory of HIT and iFLYTEK Research, 2024).

Table 9 compares these OCR+LLM results with direct image-based translation performed by several MLLMs. The OCR+LLM paradigm yields substantially higher BLEU scores, highlighting the advantage of decoupling text recognition and language understanding for this task.

In summary, the combination of high-quality OCR with powerful LLMs not only improves translation performance but also establishes a strong and interpretable baseline for OCR-related tasks within the benchmark.

F.3 Slow-Thinking Model Performance

To investigate the underperformance of slow-thinking models, we conduct a series of comparative experiments and qualitative analyses.

We select three representative multimodal slow-thinking models and compare them with their respective baseline counterparts.

As shown in the table 10, Gemini-2.0-Flash-Thinking-Exp (Google, 2024) slightly outperforms Gemini-2.0-Flash (Google, 2024), suggesting that stronger slow-thinking models can yield marginal gains through additional reasoning, particularly in multiple-choice and appreciation-related tasks. In contrast, QVQ-72B-Preview (Qwen, 2024) performs worse than Qwen2-VL-72B-Instruct (Wang et al., 2024a), and LLaMA-3.2V-11B-CoT (Xu et al., 2024a) performs significantly worse overall. These results indicate that weaker slow-thinking models tend to "overthink," leading to answer fluctuations and degraded performance. Manual inspection reveals that such models often fail to follow instructions precisely and frequently oscillate between correct and incorrect options.

Regarding instruction-following capability, we observe substantial deficiencies in weaker slow-thinking models, such as irrelevant analysis of options and failure to follow explicit prompt constraints (particularly evident in the QVQ series). In contrast, stronger slow-thinking models (e.g., Gemini) show clear improvements in instruction adherence. Despite occasional issues such as mixed-language outputs or formatting errors in entity recognition, they achieve significantly better overall alignment with task requirements.

F.4 Analysis of Model Performance Bottlenecks in Ancient Chinese Text OCR

In the CCS domain, tasks such as ancient Chinese text OCR require models to output classical Chinese text. Poor performance in these tasks results primarily from a lack of CCS-specific knowledge rather than insufficient general Chinese language ability.

Classical Chinese processing demands both fundamental Chinese proficiency and specialized knowledge related to ancient texts, including phonetic loan characters, variant characters, and semantic interpretation. Therefore, models with strong modern Chinese capabilities still require domain-specific knowledge to handle the complexity of classical Chinese.

Table 3 shows that Molmo-7B (Deitke et al., 2024) and Qwen2-VL-7B-Instruct (Wang et al., 2024a) are both based on Qwen2-7B-Base (Yang et al., 2024a), which has strong Chinese language ability. Differences in data and fine-tuning methods lead to significant forgetting of CCS-specific knowledge in Molmo-7B, causing poor performance. Additionally, Chinese-LLaVA-CLLaMA2 (LinkSoul, 2024), although fine-tuned on Chinese data, achieves a score of only 13.47 in the ancient Chinese text domain.

F.5 Discussion on Dataset Biases and Their Impact in Certain Tasks

In this section, we address potential biases present in the dataset and their possible effects on model performance.

(a) In the T13 task (Font Identification), each category—Kai, Xing, Cao, Zhuan, and Li—contains 25 samples. In the T14 task (Genre Identification), the categories Ou, Liu, Yan, and Zhao contain 40, 25, 27, and 26 samples, respectively. Although there is a slight imbalance in sample distribution, the overall differences are minor and do not significantly distort the dataset's representativeness.

(b) The error rates for T13 categories are 52.0%, 65.9%, 62.4%, 79.4%, and 66.4%, respectively. Given the balanced class distribution, these results reflect the model's relatively weaker performance on Zhuan script recognition. For T14, the error rates are 67.4%, 79.9%, 68.1%, and 74.9%, respectively. While the lower error rate for Ou script slightly improves the overall metric, the model's varying performance across scripts reveals its capability limits in handling complex classification tasks.

F.6 Evaluation of Multilingual and Multicultural Models Supporting Chinese

We sample the dataset and select three multilingual or multicultural models that support Chinese for an initial evaluation. The experimental results are shown in the table 11.

As a bilingual model, LLaVA-V1.6-34B (Liu, 2024) is fine-tuned with Chinese-English bilingual instructions, allowing it to better understand bilingual inputs. However, it still lacks sufficient CCS domain knowledge. In contrast, Pangea-7B (Yue et al., 2024), which is specifically designed for multilingual and multicultural tasks, performs well in CCS due to specialized training focused on cultural diversity. On the other hand, Maya-8B (Alam et al., 2024) has limitations in parameter size and training data, making it less effective in CCS-related tasks.

Based on the scores, models trained specifically for multicultural tasks clearly outperform general multilingual models, indicating that culture-specific training helps improve CCS task performance. Additionally, we observe that Pangea-7B performs well in Painting and Cultural Relic tasks but relatively poorly in Ancient Chinese Text and Illustration tasks.

Model	OverAll
LLaVA-V1.6-34B	29.33
Pangea-7B	**34.18**
Maya-8B	27.49

Table 11: Evaluation results of multilingual and multicultural models supporting Chinese.

F.7 Performance Comparison Between Multilingual and English-Centric Models

Model	OverAll
Multilingual Models	
Pangea-7B	**34.18**
Maya-8B	27.49
English-Centric Models	
LLaVA-v1.5-7B	23.26
LLaVA-v1.5-13B	24.10
LLaVA-v1.6-Mistral-7B	23.43
LLaVA-v1.6-Vicuna-7B	24.46
LLaVA-v1.6-Vicuna-13B	23.62

Table 12: Performance comparison of multilingual and English-centric models.

We select open-source multilingual models and compare them with English-centric models. The performance results are shown in the table.

(a) Overall, multilingual models significantly outperform English-centric models, suggesting that they are better at capturing CCS-specific terminology and cultural nuances in both linguistic expression and cultural background.

(b) At a more granular level, multilingual models achieve the largest performance gains in Calligraphy, Painting, and Cultural Relic tasks, primarily benefiting from the integration of cross-linguistic data and visual information, which helps them learn different artistic and cultural styles.

(c) Although classical Chinese falls within the multilingual category, its unique grammar and historical context require specialized training on classical texts. Since current multilingual models lack sufficient training in this area, they do not show a clear advantage in Ancient Chinese Text tasks.

F.8 Impact of OCR and Knowledge on Model Performance

As shown in Figure 5, although inputting accurate OCR results (Setting 3) significantly improves the average performance of models on tasks T2 to T6, the overall metrics of four representative models under this setting remain low and far from ideal. This indicates that MLLMs themselves exhibit clear deficiencies in cultural knowledge and comprehension ability within the domain of ancient texts, constituting the primary performance bottleneck. In contrast, limitations in OCR capability negatively affect performance to some extent but represent a relatively secondary factor. Therefore, enhancing the models' knowledge capacity is the key path to advancing this field, while improvements in OCR technology serve as an essential foundation to ensure input quality and support performance.

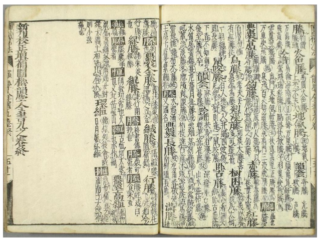
ID	Task Name	Generation Method				Related Link	License
		M1	M2	M3	M4		
T1	Text OCR				✓	https://github.com/HCIILAB/M5HisDoc https://github.com/SCUT-DLVCLab/HisDoc1B	GPL-3.0 -
T2	Classical Chinese to Modern Chinese			✓		https://github.com/HCIILAB/M5HisDoc https://github.com/SCUT-DLVCLab/HisDoc1B https://github.com/SCUT-DLVCLab/TongGu-LLM	GPL-3.0 - -
T3	Punctuation			✓		https://github.com/HCIILAB/M5HisDoc https://github.com/SCUT-DLVCLab/HisDoc1B https://github.com/SCUT-DLVCLab/TongGu-LLM	GPL-3.0 - -
T4	Named Entity Recognition			✓		https://github.com/HCIILAB/M5HisDoc https://github.com/SCUT-DLVCLab/HisDoc1B https://yiyan.baidu.com/	GPL-3.0 - -
T5	Word Explanation			✓		https://github.com/HCIILAB/M5HisDoc https://github.com/SCUT-DLVCLab/HisDoc1B https://yiyan.baidu.com/	GPL-3.0 - -
T6	Reading Comprehension			✓		https://github.com/HCIILAB/M5HisDoc https://github.com/SCUT-DLVCLab/HisDoc1B https://yiyan.baidu.com/	GPL-3.0 - -
T7	Source Attribution	✓				https://github.com/HCIILAB/M5HisDoc https://github.com/SCUT-DLVCLab/HisDoc1B	GPL-3.0 -
T8	Calligraphy OCR	✓				https://www.baidu.com/ http://www.51sdj.com/	Baidu User Agreement Open Source
T9	Calligraphy Appreciation		✓			https://www.baidu.com/ https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Baidu User Agreement Apache-2.0
T10	Author Identification	✓				https://www.baidu.com/	Baidu User Agreement
T11	Dynasty Identification	✓				https://www.baidu.com/	Baidu User Agreement
T12	Title Identification	✓				https://www.baidu.com/	Baidu User Agreement
T13	Font Identification	✓				https://www.baidu.com/	Baidu User Agreement
T14	Genre Identification	✓				https://www.baidu.com/	Baidu User Agreement
T15	Work Introduction		✓			https://www.baidu.com/ https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Baidu User Agreement Apache-2.0
T16	Material Identification	✓				https://www.aliyundrive.com/s/158sDKUz85m	Open Source
T17	Painting Appreciation		✓			https://www.baidu.com/ https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Baidu User Agreement Apache-2.0
T18	Author Identification	✓				https://www.aliyundrive.com/s/158sDKUz85m https://www.shuge.org/	Open Source CC-BY-4.0
T19	Dynasty Identification	✓				https://www.aliyundrive.com/s/158sDKUz85m https://www.shuge.org/	Open Source CC-BY-4.0
T20	Title Identification	✓				https://www.aliyundrive.com/s/158sDKUz85m https://www.shuge.org/	Open Source CC-BY-4.0
T21	Image-based Q&A	✓				https://www.aliyundrive.com/s/158sDKUz85m https://www.shuge.org/	Open Source CC-BY-4.0
T22	Background Introduction		✓			https://www.baidu.com/ https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Baidu User Agreement Apache-2.0
T23	Painting Technique	✓				https://www.baidu.com/	Baidu User Agreement
T24	Source Identification	✓				https://www.shuge.org/	CC-BY-4.0
T25	Painting Description		✓			https://www.baidu.com/ https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Baidu User Agreement Apache-2.0
T26	Painting OCR	✓				https://www.shuge.org/	CC-BY-4.0
T27	Work Introduction		✓			https://www.baidu.com/ https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Baidu User Agreement Apache-2.0
T28	Oracle Bone Script OCR				✓	https://github.com/RomanticGodVAN/character-Evolution-Dataset https://github.com/Pengjie-W/HUST-OBC	- -
T29	Pictographic Decoding	✓				https://github.com/RomanticGodVAN/character-Evolution-Dataset https://github.com/Pengjie-W/HUST-OBC https://www.baidu.com/	- - Baidu User Agreement
T30	Seal OCR				✓	http://diglweb.zjlib.cn:8082/zjtsg/zgjcj/index.htm	Open Source
T31	Owner Identification	✓				http://diglweb.zjlib.cn:8082/zjtsg/zgjcj/index.htm	Open Source
T32	Name Identification	✓				https://www.baidu.com/	Baidu User Agreement
T33	Dynasty Identification	✓				https://www.baidu.com/	Baidu User Agreement
T34	Collection Identification	✓				https://www.baidu.com/	Baidu User Agreement
T35	Cultural Relic Introduction		✓			https://www.baidu.com/ https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Baidu User Agreement Apache-2.0
T36	Cultural Relic Classification	✓				https://www.baidu.com/	Baidu User Agreement
T37	Illustration OCR	✓				https://www.baidu.com/ https://www.shuge.org/	Baidu User Agreement CC-BY-4.0
T38	Illustration Description		✓			https://www.baidu.com/ https://www.shuge.org/ http://query.clcn.net.cn/GJAndST/gjct1.htm https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Baidu User Agreement CC-BY-4.0 Open Source Apache-2.0
T39	Entity Introduction		✓			https://www.baidu.com/ https://huggingface.co/Qwen/Qwen2.5-32B-Instruct	Baidu User Agreement Apache-2.0
T40	Image-to-Poem Matching	✓				https://www.shuge.org/	CC-BY-4.0
T41	Source Identification	✓				https://www.baidu.com/ https://www.shuge.org/	Baidu User Agreement CC-BY-4.0
T42	Topic Classification	✓				http://query.clcn.net.cn/GJAndST/gjct1.htm	Open Source
T43	Image-to-Person Matching	✓				http://query.clcn.net.cn/GJAndST/gjct1.htm	Open Source
T44	Plot Introduction		✓			https://www.baidu.com/	Baidu User Agreement
T45	Image-based Q&A	✓				https://www.baidu.com/ https://www.shuge.org/ http://query.clcn.net.cn/GJAndST/gjct1.htm	Baidu User Agreement CC-BY-4.0 Open Source

Table 13: The detailed source and generation methods of data for 45 Tasks. Please refer to Appendix B.3 for the definitions of M1–M4.

ID	Task Name	Question Format		Metric (Main)	Metric (Others)
		MCQ	QA		
T1	Text OCR		✓	CR↑	AR↑, Edit Distance↓, F1-Score↑, Precision↑, Recall↑, BLEU↑
T2	Classical Chinese to Modern Chinese		✓	BLEU↑	ROUGE-1↑, ROUGE-2↑, ROUGE-L↑
T3	Punctuation		✓	F1-Score↑	Precision↑, Recall↑
T4	Named Entity Recognition		✓	F1-Score↑	Precision↑, Recall↑
T5	Word Explanation	✓		Acc↑	-
T6	Reading Comprehension		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T7	Source Attribution	✓		Acc↑	
T8	Calligraphy OCR		✓	CR↑	AR↑, Edit Distance↓, F1-Score↑, Precision↑, Recall↑, BLEU↑
T9	Calligraphy Appreciation		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T10	Author Identification	✓		Acc↑	-
T11	Dynasty Identification	✓		Acc↑	-
T12	Title Identification	✓		Acc↑	-
T13	Font Identification	✓		Acc↑	-
T14	Genre Identification	✓		Acc↑	-
T15	Work Introduction		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T16	Material Identification	✓		Acc↑	-
T17	Painting Appreciation		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T18	Author Identification	✓		Acc↑	-
T19	Dynasty Identification	✓		Acc↑	-
T20	Title Identification	✓		Acc↑	-
T21	Image-based Q&A		✓	Acc↑	-
T22	Background Introduction		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T23	Painting Technique	✓		Acc↑	-
T24	Source Identification	✓		Acc↑	-
T25	Painting Description		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T26	Painting OCR		✓	CR↑	AR↑, Edit Distance↓, F1-Score↑, Precision↑, Recall↑, BLEU↑
T27	Work Introduction		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T28	Oracle Bone Script OCR		✓	CR↑	AR↑, Edit Distance↓, F1-Score↑, Precision↑, Recall↑, BLEU↑
T29	Pictographic Decoding		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T30	Seal OCR		✓	CR↑	AR↑, Edit Distance↓, F1-Score↑, Precision↑, Recall↑, BLEU↑
T31	Owner Identification	✓		Acc↑	-
T32	Name Identification	✓		Acc↑	-
T33	Dynasty Identification	✓		Acc↑	-
T34	Collection Identification	✓		Acc↑	-
T35	Cultural Relic Introduction		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T36	Cultural Relic Classification	✓		Acc↑	-
T37	Illustration OCR		✓	CR↑	AR↑, Edit Distance↓, F1-Score↑, Precision↑, Recall↑, BLEU↑
T38	Illustration Description		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T39	Entity Introduction		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T40	Image-to-Poem Matching	✓		Acc↑	-
T41	Source Identification	✓		Acc↑	-
T42	Topic Classification	✓		Acc↑	-
T43	Image-to-Person Matching	✓		Acc↑	-
T44	Plot Introduction		✓	Avg. (BERTScore↑+ANLS↑)	BERTScore↑, ANLS↑
T45	Image-based Q&A		✓	Acc↑	-

Table 14: The evaluation metrics and question formats corresponding to 45 tasks. “Main” represents the metrics used in the main text, while “Others” indicates that these metrics are provided in the appendix.

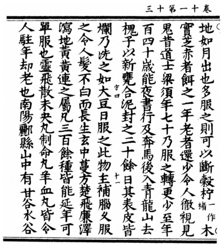
T1 文本OCR Text OCR



问：按古代阅读顺序识别图中古籍内的文字内容。
要求：请只输出识别结果，不要有其他无关内容。
答：騰物價騰踊火色騰馬周火色騰上恐不失人地氣騰記地氣止騰韻套奔騰光騰般騰声騰万禹騰……

Q：Recognize the text content in the image following the ancient reading order. Requirements: Please output only the recognition result without any irrelevant content.
A：騰物價騰踊火色騰馬周火色騰上恐不失人地氣騰記地氣止騰韻套奔騰光騰般騰声騰万禹騰......(Recognized text content)

T2 文白翻译 Classical Chinese to Modern Chinese



问：按古代阅读顺序把图片中的文言文翻译成白话文。
要求：请只输出翻译后的结果，不要有其他无关内容。
答：地榆，像月亮初升时那样，多服用它可以不吃五谷。柠，又写作"楮"，它的红色果实如果服用一年，可以使老年人恢复青春，视力敏锐，甚至可以看到鬼魂......

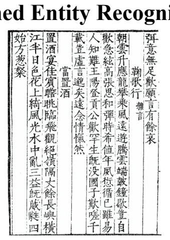
Q：Translate the Classical Chinese text in the image into modern Chinese in the order of ancient reading. Requirements: Please output only the translated result without any irrelevant content.
A：Di Yu resembles the moon at its first rise. Consuming it frequently can eliminate the need for grains. Ning, also written as "Chu," produces red fruits that, if consumed for a year, can rejuvenate the elderly, sharpen vision, and even enable them to see spirits.......

T3 句读 Punctuation

问：按古代阅读顺序给图中的文言文添加标点符号。
要求：请只输出标点后的句子，不要有其他无关内容。
答：足以服死者之心，所謂意也。始於辨疑，成於案牘，法吏之辭，無不可引經而斷，所謂文也。嗚呼！二者固難言之矣。夫持之過刻，必入於苛法家之失也......

Q：Add punctuation marks to the Classical Chinese text in the image according to the traditional reading order. Requirements: Please output only the punctuated sentences without any irrelevant content.
A：足以服死者之心，所謂意也。始於辨疑，成於案牘，法吏之辭，無不可引經而斷，所謂文也。嗚呼！二者固難言之矣。夫持之過刻，必入於苛法家之失也...... (Punctuated sentences)

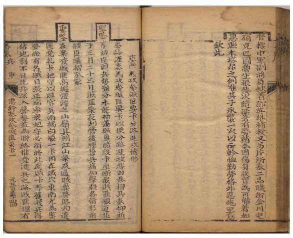
T4 命名实体识别 Named Entity Recognition



问：抽取图像中文言文内容的命名实体。要求：请只输出抽取得到的命名实体，实体之间用中文顿号进行隔开，不要有其他无关内容。
答：鞠歌行、应龙、王阳、贡公、国子、佳賔

Q：Extract named entities from the Classical Chinese content in the image. Requirements: Please output only the extracted named entities, separated by Chinese commas, without any irrelevant content.
A：Ju Ge Xing, Ying Long, Wang Yang, Gong Gong, Guo Zi, Jia Bin

T5 字词解释 Word Explanation



问：图像的文言文内容中"承袭"一词的意思是?
A、继承　B、承认　C、承担　D、承受。
要求：请只输出所选择的选项字母，不要有其他无关内容。
答：A

Q：The meaning of the word "承袭" in the classical Chinese text of the image is?
A. Inherit　B. Acknowledge　C. Bear　D. Endure
Requirements: Please only output the letter of the chosen option, without any irrelevant content.
A：A

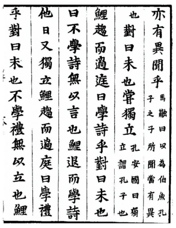
Figure 17: Task1 to Task5 examples.

T6 阅读理解
Reading Comprehension

问：根据图像回答下列问题：宋意上疏的主要目的是什么？
答：宋意上疏的主要目的是劝谏皇帝不要对济南王、中山王等叔父昆弟逾礼过恩，应当遵守人臣之节，以维护朝廷的秩序和礼法。

Q: Based on the image, answer the following question: What is the main purpose of the memorial in Song Yi's petition?
A: The main purpose of Song Yi's petition is to advise the emperor not to show excessive favoritism towards his uncles and brothers, such as the King of Jinan and the King of Zhongshan. He should uphold the proper conduct of a minister to maintain the order and rituals of the court.

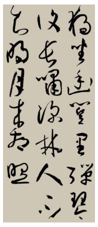
T7 判断出处 Source Attribution



问：这张图片中的古籍文本内容出自哪本书？
A、說苑 B、弘明集 C、論語 D、論衡
要求：只输出选项对应的字母，不要输出其他无关内容。
答：C

Q: Which book is the ancient text in this image from?
A. Shuoyuan B. Hongming Ji C. Lunyu D. Lunheng
Requirements: Please only output the letter of the chosen option, without any irrelevant content.
A: C

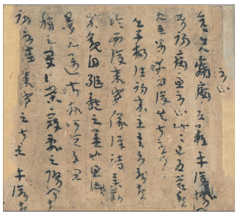
T8 书法OCR Calligraphy OCR



问：按古代阅读顺序识别图中的文字内容。
要求：请只输出识别结果，不要有其他无关内容。
答：独坐幽篁里弹琴复长啸深林人不知明月来相照

Q: Identify the text in the image according to the ancient reading order.
Requirements: Please only output the recognition result, without any irrelevant content.
A: 独坐幽篁里弹琴复长啸深林人不知明月来相照
(Recognized text content)

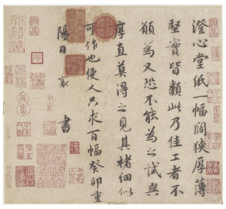
T9 书法鉴赏
Calligraphy Appreciation



问：请对这幅书法作品进行艺术鉴赏。
答：《平复帖》是晋代书法家陆机创作的草隶书法代表作，风格平淡质朴，寓意深远。全篇共九行八十四字，用秃笔写于麻纸之上，笔意婉转流畅，字体瘦长挺拔……

Q: Please provide an artistic appreciation of this calligraphy work.
A: "Pingfu Tie" is a representative work of cursive and clerical script by the Jin Dynasty calligrapher Lu Ji. Its style is simple and unadorned, with profound meaning. The entire piece consists of nine lines and eighty-four characters, written with a blunt brush on hemp paper. The brushstrokes are gentle and smooth, and the characters are slender, long, and upright......

T10 作者识别
Author Identification

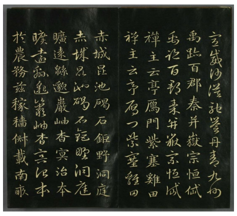


问：这幅书法作品的作者是谁？
A、王羲之 B、赵孟頫 C、蔡襄 D、郭沫若
要求：只输出选项对应的字母，不要输出其他无关内容。
答：C

Q: Who is the author of this calligraphy work?
A. Wang Xizhi B. Zhao Mengfu C. Cai Xiang D. Guo Moruo
Requirements: Please only output the letter of the chosen option, without any irrelevant content.
A: C

Figure 18: Task6 to Task10 examples.

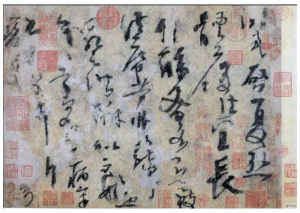
T11 朝代识别
Dynasty Identification



问：这幅书法作品创作于哪个朝代？
A、魏晋南北朝　B、隋唐　C、宋代　D、元
要求：只输出选项对应的字母，不要输出其他无关内容。
答：D

Q: Which dynasty was this calligraphy work created in?
A. Wei, Jin, and the Northern and Southern Dynasties　B. Sui and Tang
C. Song Dynasty　D. Yuan
Requirements: Please only output the letter of the chosen option, without any irrelevant content.
A: D

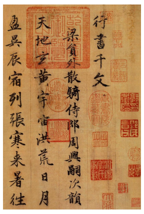
T12 标题识别
Title Identification



问：这幅书法作品的标题是什么？
A、同年帖　B、自书告身　C、李思训碑　D、夏热帖
要求：只输出选项对应的字母，不要输出其他无关内容。
答：D

Q: What is the title of this calligraphy work?
A. Tongnian Tie　B. Zishu Gaoshen　C. Li Sixun Stele　D. Xia Re Tie
Requirements: Please only output the letter of the chosen option, without any irrelevant content.
A: D

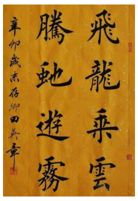
T13 字体识别
Font Identification



问：这幅书法使用了哪种字体？
A、篆书　B、楷书　C、隶书　D、行书　E、草书
要求：只输出选项对应的字母，不要输出其他无关内容。
答：D

Q: Which script style is used in this calligraphy work?
A. Seal Script　B. Regular Script　C. Clerical Script
D. Running Script　E. Cursive Script
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: D

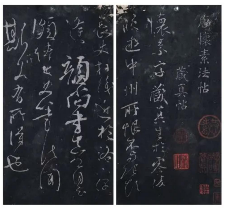
T14 流派识别
Genre Identification



问：这幅楷书作品使用了哪种楷书字体？
A、颜体　B、欧体　C、赵体　D、柳体
要求：只输出选项对应的字母，不要输出其他无关内容。
答：B

Q: Which regular script style is used in this work?
A. Yan Style　B. Ou Style　C. Zhao Style　D. Liu Style
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: B

T15 作品介绍
Work Introduction

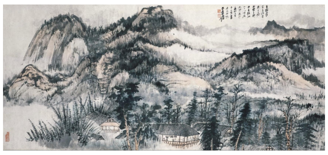


问：介绍一下这幅书法作品。
答：这幅作品名为《藏真帖》，是怀素的行草书作品，共6行，包含51字。在作品中，怀素详细叙述了他北上向各位名家求教书法的经历，特别是向颜真卿学习书法的过程……

Q: Introduce this calligraphy work.
A: This work, titled Cangzhen Tie, is a cursive and running script piece by Huai Su, consisting of six lines and 51 characters. In this piece, Huai Su narrates in detail his journey northward to seek guidance in calligraphy from various masters, particularly his process of learning from Yan Zhenqing......

Figure 19: Task11 to Task15 examples.

T16 材质识别
Material Identification



问：这幅画的材质是？
A、纸本　B、绢本　C、绫本　D、拓本
要求：只输出选项对应的字母，不要输出其他无关内容。
答：A

Q: What is the material of this painting?
A. Paper　B. Silk　C. Satin　D. Rubbing
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: A

T17 绘画鉴赏
Painting Appreciation

问：赏析一下这幅画作。
答：《写生珍禽图》是一幅以细密线条和浓丽色彩描绘自然生灵的佳作。画中呈现了24种昆虫、鸟雀及龟类，在尺幅不大的绢素上以"双勾填色"技法描绘，线条细劲，色彩层层晕染，生动逼真……

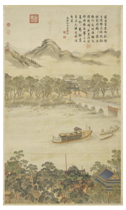
Q: Appreciate this painting.
A: Sketches of Rare Birds and Animals is a masterpiece that vividly depicts creatures of nature with fine lines and rich colors. The painting features 24 species of insects, birds, and turtles, rendered on a small silk canvas using the "double-outline and color-fill" technique. The lines are delicate and strong, and the colors are layered with subtle gradations, creating a lifelike effect......

T18 作者识别
Author Identification

问：这幅画作的作者是谁？
A、黎墨　B、沈柔坚　C、萧照　D、徐悲鸿
要求：只输出选项对应的字母，不要输出其他无关内容。
答：D

Q: Who is the author of this painting?
A. Li Mo　B. Shen Roujian　C. Xiao Zhao　D. Xu Beihong
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: D

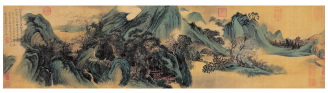
T19 朝代识别
Dynasty Identification



问：这幅画作创作于哪个朝代？
A、辽　B、唐　C、金　D、清
要求：只输出选项对应的字母，不要输出其他无关内容。
答：D

Q: Which dynasty was this painting created in?
A. Liao　B. Tang　C. Jin　D. Qing
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: D

T20 标题识别
Title Identification



问：这幅画作的名称是？
A、赤壁图　B、三顾草庐图　C、云白山青图　D、葵石蛱蝶图
要求：只输出选项对应的字母，不要输出其他无关内容。
答：C

Q: What is the title of this painting?
A. The Red Cliffs　B. Three Visits to the Thatched Cottage
C. Clouds and Mountains　D. Sunflowers and Butterflies
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: C

Figure 20: Task16 to Task20 examples.

T21 画面问答
Image-based Q&A

问：图像中画了几只鸟？请直接回答阿拉伯数字。
答：12

Q：How many birds are depicted in the image? Please answer with an Arabic numeral.
A：12

T22 背景介绍
Background Introduction

问：介绍一下这幅画的创作背景。
答：宋代的城市发展冲破了坊市格局，市民生活日趋丰富，市场交易频繁货郎在经济活动中扮演了重要的角色……

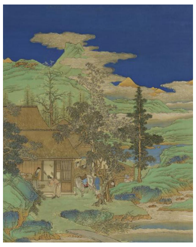
Q：Introduce the background of the creation of this painting.
A：The urban development of the Song Dynasty broke through the traditional block-and-market system, with the lives of citizens becoming increasingly diverse. Market transactions were frequent, and peddlers played an important role in economic activities......

T23 绘画技法
Painting Technique

问：从技法上看，这幅画属于?
A、工笔画　B、水墨画　C、写意画
要求：只输出选项对应的字母，不要输出其他无关内容。
答：C

Q：From a technical perspective, what type of painting is this?
A. Gongbi Painting　B. Ink Wash Painting
C. Freehand Brushwork Painting
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A：C

T24 出处识别
Source Identification



问：这幅绘画作品出自哪里?
A、群仙图册　B、九夏安龢图册
C、万春集庆　D、十二禁禦景图轴
要求：只输出选项对应的字母，不要输出其他无关内容。
答：A

Q：Where is this painting from?
A. The Collection of Immortals　B. The Nine Summers An He Scroll
C. The Collection of Ten Thousand Springs
D. The Twelve Imperial Prohibitions Scroll
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A：A

T25 绘画描述
Painting Description

问：描述一下这幅画中绘画部分的内容。
答：这幅画《烟岫林居图》展现了一位老者穿行于板桥与小径之间的情景。画面背景是幽深的山林，其中隐约可见房舍的一角，营造出一种宁静而深远的意境……

Q：Describe the content of the painted portion of this artwork.
A：The painting Smoke-Ember Forest Dwelling depicts an elderly man walking between a plank bridge and a narrow path. The background shows a deep and serene mountain forest, with a corner of a house faintly visible, creating a tranquil and profound atmosphere......

Figure 21: Task21 to Task25 examples.

T26 绘画OCR Painting OCR

问：按古代阅读顺序识别图中的文字内容。
要求：请只输出识别结果，不要有其他无关内容。
答：姜詩躍鯉

Q：Identify the text in the image according to the ancient reading order.
Requirement: Please only output the recognition result, without any irrelevant content.
A：姜詩躍鯉 (Recognized text content)

T27 作品介绍
Work Introduction

问：介绍一下这幅画作。
答：这幅画名为《枫鹰雉鸡图》，是南宋画家李迪的绢本设色画，现藏于北京故宫博物院。画作展现了一只苍鹰怒视逃窜雉鸡的紧张场景……

Q：Introduce this painting.
A：This painting, titled Maple, Hawk, and Pheasant, is a silk painting by the Southern Song Dynasty artist Li Di, currently housed in the Palace Museum in Beijing. The artwork depicts a tense scene where a falcon angrily stares at a pheasant fleeing in panic......

T28 甲骨文OCR
Oracle Bone Script OCR

问：识别图中的甲骨文对应哪个汉字。
要求：只输出识别的汉字结果，不要输出其他无关内容。
答：朋

Q：Identify the Chinese character corresponding to the oracle bone script in the image.
Requirement: Please only output the identified Chinese character, without any irrelevant content.
A：朋 (Friend)

T29 象形解读
Pictographic Decoding

问：这个甲骨文是“五”，请从象形角度解释它的含义。
答：“五”是象形字，其中的上下各一横幅画表示天和地，×表示阴阳二气交错在天地中间。本义表示水、火、木、金、土五行交错……

Q：This oracle bone script represents “five.” Please explain its meaning from a pictographic perspective.
A：“Five” is a pictographic character. The horizontal strokes at the top and bottom represent heaven and earth, while the “×” symbolizes the interweaving of the Yin and Yang energies in the space between heaven and earth. Its original meaning represents the five elements—water, fire, wood, metal, and earth—interacting with one another...

T30 印鉴OCR Seal OCR

问：按古代阅读顺序识别图中印鉴的文字内容。
要求：只输出识别结果，不要输出其他无关内容。
答：高崙之鉥

Q：Identify the text on the seal in the image according to the ancient reading order.
Requirement: Please only output the recognition result, without any irrelevant content.
A：高崙之鉥 (Recognized text content)

Figure 22: Task26 to Task30 examples.

T31 人物识别
Owner Identification

问：这枚印鉴的主人是?
A、吕凤子　B、刘自椟　C、鲁迅　D、林纾
要求：只输出选项对应的字母，不要输出其他无关内容。
答：C

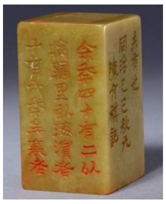
Q: Who is the owner of this seal?
A. Lü Fengzi　B. Liu Zidu　C. Lu Xun　D. Lin Shu
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: C

T32 名称识别
Name Identification

问：这个文物的名字是?
A、夔纹方簋　B、汉谟父氏铜印
C、勾连云纹玉环　D、清嘉庆戊寅年铜钟
要求：只输出选项对应的字母，不要输出其他无关内容。
答：C

Q: What is the name of this cultural relic?
A. Kui-patterned Square Gui　B. Han Mo Fu Family Copper Seal
C. Interlocking Cloud-patterned Jade Ring
D. Copper Bell of the Qianlong Era, Year of Wu Yin
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: C

T33 朝代识别
Dynasty Identification



问：这是哪个时期的文物?
A、商　B、元　C、春秋前期　D、清
要求：只输出选项对应的字母，不要输出其他无关内容。
答：D

Q: What period is this cultural relic from?
A. Shang　B. Yuan　C. Early Spring and Autumn　D. Qing
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: D

T34 馆藏识别
Collection Identification

问：这件文物现藏于哪里?
A、中国国家博物馆　B、浙江省博物馆
C、南京博物院　D、广东省博物馆
要求：只输出选项对应的字母，不要输出其他无关内容。
答：A

Q: Where is this cultural relic currently housed?
A. National Museum of China　B. Zhejiang Provincial Museum
C. Nanjing Museum　D. Guangdong Provincial Museum
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: A

T35 文物介绍
Cultural Relic Introduction

问：介绍一下这个文物。
答：这个文物名为鹳鱼石斧图彩陶缸，为红陶砂质，高47厘米，口径32.7厘米，底径20.1厘米。器腹外壁一侧绘有鹳鱼石斧图，是迄今中国发现最早、面积最大的一幅陶画……
Q: Introduce this cultural relic.
A: This cultural relic is called the Stork Fish and Stone Axe Mural on a Painted Pottery Jar. It is made of red pottery with a sandy texture, measuring 47 cm in height, 32.7 cm in diameter at the mouth, and 20.1 cm in diameter at the base. One side of the outer wall of the jar's belly is painted with a stork fish and stone axe mural, which is the earliest and largest discovered pottery painting in China to date...

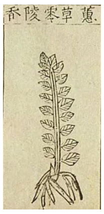
Figure 23: Task31 to Task35 examples.

T36 文物分类
Cultural Relic Classification

问：这件文物属于?
A、档案文书　B、瓷器　C、雕塑、造像　D、铜器
要求：只输出选项对应的字母，不要输出其他无关内容。
答：B

Q：Which category does this cultural relic belong to?
A. Documents and Archives　B. Porcelain
C. Sculpture and Figurines　D. Bronze Ware
Requirement: Please only output the letter of the chosen option, without any irrelevant content..
A：B

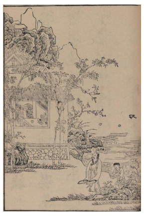
T37 插图OCR
Illustration OCR



问：按古代阅读顺序识别图中的文字内容。
要求：只输出识别结果，不要输出其他无关内容。
答：蕙草零陵香

Q：Identify the text in the image according to the ancient reading order.
Requirement: Please only output the recognition result, without any irrelevant content.
A：蕙草零陵香 (Recognized text content)

T38 插图描述
Illustration Description



问：这幅图描绘了什么?
答：这幅图描绘了一幅宁静的暮春景象。画面左侧是一座依山而建的草堂，草堂周围环绕着茂密的树木和灌木丛。草堂的屋顶覆盖着茅草，显得古朴而自然……

Q：What does this illustration depict?
A：This painting depicts a serene late spring scene. On the left side of the image is a thatched cottage built against a hillside, surrounded by dense trees and shrubs. The roof of the cottage is covered with straw, giving it a rustic and natural appearance...

T39 实体介绍
Entity Introduction

问：介绍一下这幅图中的实体。
答：图中的实体是黛玉。黛玉，即林黛玉，是中国古典名著《红楼梦》中的主要人物之一。她是大观园中贾宝玉的姑表妹，自幼体弱多病，性格敏感细腻，才情出众，擅长诗词……

Q：Introduce the entities depicted in this illustration.
A：The entity depicted in the illustration is Daiyu. Daiyu, or Lin Daiyu, is one of the main characters in the Chinese classical masterpiece Dream of the Red Chamber. She is the cousin of Jia Baoyu in the Grand View Garden. Frail and sickly since childhood, she has a sensitive and delicate personality, exceptional talent, and excels in poetry......

T40 图像诗句匹配
Image-to-Poem Matching

问：这幅图和以下哪首诗句最匹配?
A、（省略）　B、（省略）　C、（省略）
D、三月时将尽，空房妾独居。蛾眉愁自结，蝉鬓没情梳。
要求：只输出选项对应的字母，不要输出其他无关内容。
答：D

Q：Which of the following verses best matches this illustration?
A. (Skipped)　B. (Skipped)　C. (Skipped)
D. “March is about to end, the empty room leaves the maiden alone. Her arched eyebrows knit with sorrow, her silken hair remains uncombed.”
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A：D

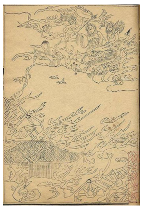
Figure 24: Task36 to Task40 examples.

T41 出处识别
Source Identification

问：这张插图出自哪本书？
A、本草纲目　B、列仙酒牌　C、天工开物　D、毛诗品物图考
要求：只输出选项对应的字母，不要输出其他无关内容。
答：C

Q: Which book does this illustration come from?
A. Compendium of Materia Medica　B. Immortal Wine Cards
C. The Exploitation of the Works of Nature
D. Illustrated Study of Objects in the Book of Songs
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: C

T42 主题分类
Topic Classification



问：这张图的主题是？
A、植物　B、故事　C、建筑　D、人物
要求：只输出选项对应的字母，不要输出其他无关内容。
答：B

Q: What is the theme of this illustration?
A. Plants　B. Stories　C. Architecture　D. Characters
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: B

T43 图人匹配
Image-to-Person Matching

问：这张图描述的是谁的故事？
A、樵青，渔童　B、蔡婆婆，禁子　C、吕布，探子　D、关胜
要求：只输出选项对应的字母，不要输出其他无关内容。
答：C

Q: Whose story is depicted in this illustration?
A. Qiao Qing, the Fisher Boy　B. Granny Cai, the Prohibited Child
C. Lü Bu, the Scout　D. Guan Sheng
Requirement: Please only output the letter of the chosen option, without any irrelevant content.
A: C

T44 情节介绍
Plot Introduction

问：这幅插图来自《西游记》，请介绍一下其对应的故事情节。
答：该图对应的是《西游记》中"老黿背负四众过河"的故事。情节如下：唐僧师徒到达通天河边，准备过河。这时，一只太白金头黿出现在岸边，请求唐僧上它的背过河……

Q: This illustration is from Journey to the West. Please introduce the corresponding plot of the story.
A: The illustration corresponds to the story of "The Old Turtle Carrying the Four Monks Across the River" from Journey to the West. The plot is as follows: The monk Xuanzang and his disciples arrive at the shore of the Tongtian River, preparing to cross. At that moment, a golden-headed giant turtle appears at the riverbank and offers to carry them across the river on its back......

T45 画面问答
Image-based Q&A

问：图像中武器上的图案是什么动物？
答：虎

Q: What animal is depicted on the weapon in the image?
A: Tiger.

Figure 25: Task41 to Task45 examples.

Model	T1	T2	T3	T4	T5	T6	T7	MCQ	QA	OverAll
Closed-source Models										
Claude-3.5-Sonnet	39.64	2.08	24.50	13.73	<u>72.27</u>	48.52	68.97	70.62	25.69	38.53
Gemini-1.5-Pro	61.75	<u>10.13</u>	**57.32**	**36.87**	69.92	50.93	67.59	68.76	**43.40**	<u>50.64</u>
Gemini-2.0-Flash	<u>72.22</u>	**10.36**	<u>47.89</u>	<u>32.17</u>	68.75	50.73	**77.93**	<u>73.34</u>	<u>42.67</u>	**51.44**
GPT-4o	27.90	1.91	12.64	9.60	**73.05**	54.64	68.97	71.01	21.34	35.53
Open-source Models										
Chinese-LLaVA-CLLaMA2	1.49	0.15	0.46	0.02	23.44	47.37	21.38	22.41	9.90	13.47
DeepSeek-VL2-Tiny	18.33	0.86	1.71	0.02	40.23	48.93	46.90	43.57	13.97	22.43
DeepSeek-VL2-Small	5.30	0.03	0.10	0.47	30.08	44.02	31.03	30.56	9.98	15.86
DeepSeek-VL2	7.45	0.11	3.77	0.77	37.11	45.79	46.21	41.66	11.58	20.17
GLM-4V-9B	22.71	1.36	6.64	4.37	60.16	56.45	62.76	61.46	18.31	30.64
InternVL2-4B	54.36	1.31	1.01	0.26	36.72	44.49	44.83	40.78	20.29	26.14
InternVL2-8B	59.39	1.20	2.71	5.58	65.23	50.62	58.62	61.93	23.90	34.76
InternVL2-26B	56.69	1.25	4.87	1.00	59.38	49.23	55.86	57.62	22.61	32.61
InternVL2.5-1B	48.82	1.23	0.84	0.49	50.00	48.70	45.52	47.76	20.02	27.94
InternVL2.5-2B	53.00	1.25	5.51	1.56	48.83	49.35	42.07	45.45	22.13	28.80
InternVL2.5-4B	56.20	1.45	1.64	5.60	62.50	49.64	58.62	60.56	22.91	33.66
InternVL2.5-8B	60.56	1.36	4.28	7.82	69.53	53.72	60.69	65.11	25.55	36.85
InternVL2.5-26B	63.80	1.84	11.98	1.83	67.97	52.24	57.93	62.95	26.34	36.80
InternVL2.5-38B	68.05	1.09	8.43	9.98	66.41	56.44	66.21	66.31	28.80	39.52
InternVL2.5-78B	**75.19**	2.05	8.42	17.18	71.88	55.45	70.34	71.11	31.66	42.93
LLaVA-v1.5-7B	0.54	0.06	0.87	0.00	28.12	51.00	25.52	26.82	10.49	15.16
LLaVA-v1.5-13B	0.76	0.08	0.30	0.00	29.30	49.87	20.69	25.00	10.20	14.43
LLaVA-v1.6-Mistral-7B	2.05	0.13	1.48	0.00	27.73	47.53	26.90	27.32	10.24	15.12
LLaVA-v1.6-Vicuna-7B	1.72	0.12	1.77	0.00	19.92	42.92	26.21	23.07	9.31	13.24
LLaVA-v1.6-Vicuna-13B	1.71	0.12	1.81	0.00	28.52	44.60	21.38	24.95	9.65	14.02
MiniCPM-V	0.66	0.10	2.42	0.00	42.58	<u>57.36</u>	26.90	34.74	12.11	18.57
MiniCPM-V-2	1.49	0.11	0.95	0.10	43.75	54.71	37.24	40.50	11.47	19.76
MiniCPM-LLaMA3-V-2.5	4.31	0.05	4.38	1.08	50.00	50.21	36.55	43.28	12.01	20.94
MiniCPM-V-2.6	30.98	0.85	2.67	2.89	66.80	55.73	52.41	59.61	18.62	30.33
Molmo-7B-D-0924	0.12	0.16	0.00	0.00	64.06	37.32	24.83	44.45	7.52	18.07
Molmo-7B-O-0924	0.00	0.00	0.00	0.00	27.73	37.48	22.76	25.25	7.50	12.57
Ovis1.5-Gemma2-9B	2.16	0.21	3.55	0.00	51.56	46.50	21.38	36.47	10.48	17.91
Ovis1.6-Gemma2-9B	3.49	0.16	8.19	0.20	51.95	44.20	40.00	45.98	11.25	21.17
Qwen-VL-Chat	2.94	0.17	2.51	0.27	53.52	50.36	41.38	47.45	11.25	21.59
Qwen2-VL-2B-Instruct	20.99	1.14	3.48	1.07	49.61	55.62	53.10	51.36	16.46	26.43
Qwen2-VL-7B-Instruct	45.67	1.51	9.73	12.82	68.75	**57.70**	68.28	68.52	25.49	37.78
Qwen2-VL-72B-Instruct	59.51	2.29	12.80	19.90	**73.05**	56.93	<u>75.17</u>	**74.11**	30.29	42.81
QVQ-72B-Preview	46.60	1.30	18.30	4.08	66.80	29.78	**77.93**	72.37	20.01	34.97
Average	29.15	1.34	7.57	5.18	51.82	49.38	47.33	49.57	18.52	27.39

Table 15: The main metrics for the Ancient Chinese Text subdomain.

Model	T8	T9	T10	T11	T12	T13	T14	T15	MCQ	QA	OverAll
Closed-source Models											
Claude-3.5-Sonnet	40.25	40.25	42.02	27.62	47.44	63.20	24.58	46.46	40.97	42.32	41.48
Gemini-1.5-Pro	10.20	10.20	43.70	32.38	53.02	52.80	**38.98**	44.10	44.18	21.50	35.67
Gemini-2.0-Flash	27.64	24.89	34.45	36.19	68.37	61.60	37.29	26.85	47.58	26.46	39.66
GPT-4o	37.92	37.92	37.82	31.43	67.44	**70.40**	27.12	**47.99**	46.84	41.28	44.75
Open-source Models											
Chinese-LLaVA-CLLaMA2	1.50	1.50	22.69	21.90	18.60	6.40	9.32	42.66	15.78	15.22	15.57
DeepSeek-VL2-Tiny	18.23	18.23	27.73	40.00	44.65	29.60	27.12	45.84	33.82	27.43	31.42
DeepSeek-VL2-Small	20.74	20.74	30.25	31.43	40.93	17.60	27.12	34.10	29.47	25.19	27.86
DeepSeek-VL2	26.15	26.15	27.73	29.52	43.26	18.40	25.42	39.48	28.87	30.59	29.51
GLM-4V-9B	24.38	24.38	31.93	38.10	58.14	66.40	27.12	44.47	44.34	31.08	39.36
InternVL2-4B	47.97	47.97	28.57	32.38	47.44	28.00	24.58	43.83	32.19	46.59	37.59
InternVL2-8B	63.06	63.06	39.50	34.29	53.49	32.00	33.90	43.07	38.64	56.40	45.30
InternVL2-26B	52.97	52.97	41.18	36.19	54.88	27.20	32.20	43.79	38.33	49.91	42.67
InternVL2.5-1B	61.52	61.52	40.34	40.95	49.30	33.60	30.51	45.66	38.94	56.23	45.42
InternVL2.5-2B	41.62	41.62	33.61	40.00	58.14	31.20	25.42	46.92	37.67	43.39	39.82
InternVL2.5-4B	60.94	60.94	40.34	34.29	69.30	37.60	28.81	46.52	42.07	56.13	47.34
InternVL2.5-8B	60.42	60.42	31.93	35.24	66.51	46.40	33.05	46.88	42.63	55.91	47.61
InternVL2.5-26B	57.03	57.03	43.70	**44.76**	67.44	43.20	29.66	44.79	45.75	52.95	48.45
InternVL2.5-38B	57.37	57.37	**44.54**	34.29	64.19	44.00	28.81	46.28	43.17	53.67	47.11
InternVL2.5-78B	**68.73**	**68.73**	**44.54**	37.14	**76.74**	44.00	29.66	43.66	46.42	**60.37**	**51.65**
LLaVA-v1.5-7B	0.50	0.50	20.17	19.05	23.72	22.40	29.66	44.93	23.00	15.31	20.12
LLaVA-v1.5-13B	0.85	0.85	31.93	20.95	24.19	20.00	23.73	45.28	24.16	15.66	20.97
LLaVA-v1.6-Mistral-7B	2.06	2.06	24.37	20.95	24.19	12.80	27.97	44.28	22.06	16.13	19.83
LLaVA-v1.6-Vicuna-7B	2.60	2.60	31.09	32.38	26.98	21.60	30.51	44.41	28.51	16.54	24.02
LLaVA-v1.6-Vicuna-13B	3.15	3.15	23.53	28.57	25.58	16.80	8.47	43.79	20.59	16.70	19.13
MiniCPM-V	1.08	1.08	19.33	33.33	32.56	39.20	21.19	41.07	29.12	14.41	23.61
MiniCPM-V-2	5.61	5.61	30.25	28.57	39.07	33.60	25.42	42.95	31.38	18.06	26.38
MiniCPM-LLaMA3-V-2.5	8.36	8.36	26.05	40.00	44.65	43.20	33.90	44.14	37.56	20.29	31.08
MiniCPM-V-2.6	15.69	15.69	38.66	36.19	55.35	52.00	32.20	44.46	42.88	25.28	36.28
Molmo-7B-D-0924	0.05	0.05	25.21	38.10	26.98	19.20	35.59	33.29	29.02	11.13	22.31
Molmo-7B-O-0924	0.00	0.00	24.37	29.52	25.12	16.80	32.20	33.16	25.60	11.05	20.15
Ovis1.5-Gemma2-9B	1.12	1.12	26.89	24.76	29.77	28.80	19.49	45.29	25.94	15.84	22.15
Ovis1.6-Gemma2-9B	4.07	4.07	24.37	24.76	38.14	23.20	29.66	44.44	28.03	17.53	24.09
Qwen-VL-Chat	9.56	9.56	36.97	24.76	35.35	28.80	27.97	47.34	30.77	22.15	27.54
Qwen2-VL-2B-Instruct	27.25	27.25	35.29	27.62	52.09	24.00	29.66	47.25	33.73	33.92	33.80
Qwen2-VL-7B-Instruct	33.18	33.18	31.93	36.19	51.16	43.20	27.97	46.71	38.09	37.69	37.94
Qwen2-VL-72B-Instruct	32.89	32.89	43.70	42.86	63.26	52.00	36.44	46.27	**47.65**	37.35	43.79
QVQ-72B-Preview	45.67	44.14	36.13	40.00	66.05	52.00	36.44	43.10	46.12	44.30	45.44
Average	26.28	26.16	32.89	32.61	46.85	35.22	28.36	43.39	35.19	31.94	33.97

Table 16: The main metrics for the Calligraphy subdomain.

Model	**T16**	**T17**	**T18**	**T19**	**T20**	**T21**	**T22**	**T23**	**T24**	**T25**	**T26**	**T27**	**MCQ**	**QA**	**OverAll**
Closed-source Models															
Claude-3.5-Sonnet	61.07	**46.71**	60.27	70.91	80.95	58.18	43.50	62.50	64.58	50.32	45.63	48.66	66.71	48.83	57.77
Gemini-1.5-Pro	76.51	44.48	59.59	74.55	84.35	54.55	42.79	65.83	58.33	47.98	52.74	44.80	69.86	47.89	58.87
Gemini-2.0-Flash	**79.87**	29.33	73.97	**79.09**	**87.76**	43.64	30.59	66.67	67.71	46.81	66.12	36.87	**75.85**	42.23	59.04
GPT-4o	49.66	46.57	69.86	74.55	82.31	**63.64**	**44.56**	65.83	61.46	50.96	35.58	**50.55**	67.28	48.64	57.96
Open-source Models															
Chinese-LLaVA-CLLaMA2	13.42	41.86	27.40	29.09	30.61	20.00	40.91	28.33	10.42	42.25	1.96	43.71	23.21	31.78	27.50
DeepSeek-VL2-Tiny	55.70	46.28	41.78	59.09	57.14	49.09	42.62	42.50	43.75	49.56	25.66	47.92	49.99	43.52	46.76
DeepSeek-VL2-Small	21.48	32.65	24.66	30.91	48.30	43.64	33.13	34.17	39.58	38.37	15.16	39.34	33.18	33.71	33.45
DeepSeek-VL2	41.61	35.17	26.03	33.64	61.22	43.64	37.95	45.00	38.54	40.91	20.33	40.74	41.01	36.45	38.73
GLM-4V-9B	54.36	45.70	63.01	63.64	82.99	50.91	43.13	58.33	47.92	48.87	39.60	50.06	61.71	46.38	54.04
InternVL2-4B	34.23	42.49	39.04	43.64	62.59	34.55	33.61	41.67	43.75	47.57	57.03	47.37	44.15	43.77	43.96
InternVL2-8B	65.10	43.79	60.27	44.55	75.51	41.82	40.90	58.33	50.00	48.25	65.32	45.51	58.96	47.60	53.28
InternVL2-26B	65.77	44.01	60.27	48.18	82.99	49.09	41.01	61.67	53.12	49.12	61.98	46.71	62.00	48.65	55.33
InternVL2.5-1B	42.95	44.95	43.84	60.91	74.83	34.55	41.55	51.67	37.50	49.77	50.55	48.70	51.95	45.01	48.48
InternVL2.5-2B	45.64	45.28	49.32	51.82	72.79	41.82	42.31	63.33	44.79	50.62	56.65	48.68	54.62	47.56	51.09
InternVL2.5-4B	71.81	45.91	54.79	61.82	81.63	54.55	43.49	68.33	57.29	50.91	66.15	50.35	65.95	51.89	58.92
InternVL2.5-8B	72.48	44.24	54.11	42.73	80.27	47.27	42.74	66.67	67.29	50.56	68.07	48.81	63.93	50.28	57.10
InternVL2.5-26B	72.48	44.94	67.81	55.45	84.35	54.55	41.47	58.33	67.71	50.78	68.25	48.50	67.69	51.41	59.55
InternVL2.5-38B	73.15	44.88	70.55	56.36	85.03	60.00	42.60	64.17	67.71	50.94	70.46	50.00	69.50	53.15	61.32
InternVL2.5-78B	63.76	44.90	**78.77**	65.45	83.67	**63.64**	41.19	**70.00**	**70.83**	**51.09**	**72.42**	46.98	72.08	**53.37**	**62.72**
LLaVA-v1.5-7B	22.15	43.05	26.03	20.91	26.53	20.00	42.52	29.17	28.12	44.85	0.40	47.29	25.49	33.02	29.25
LLaVA-v1.5-13B	20.13	44.14	26.03	30.91	28.57	27.27	41.91	33.33	25.00	47.88	0.63	47.68	27.33	34.92	31.12
LLaVA-v1.6-Mistral-7B	24.83	44.00	21.92	18.18	26.53	16.36	40.62	29.17	22.92	46.40	3.23	44.17	23.93	32.46	28.19
LLaVA-v1.6-Vicuna-7B	23.49	44.57	30.82	30.00	31.97	27.27	40.64	28.33	27.08	47.44	1.08	47.48	28.62	34.75	31.68
LLaVA-v1.6-Vicuna-13B	24.16	44.41	29.45	26.36	24.49	23.64	40.39	27.50	22.92	47.55	1.32	46.30	25.81	33.93	29.87
MiniCPM-V	50.34	44.45	26.71	49.09	58.50	38.18	41.67	43.33	33.33	44.68	0.55	46.36	43.55	35.98	39.77
MiniCPM-V-2	44.30	44.32	39.04	56.36	60.54	38.18	39.82	47.50	36.46	46.64	6.21	44.14	47.37	36.55	41.96
MiniCPM-LLaMA3-V-2.5	22.15	43.66	32.88	45.45	51.70	34.55	41.97	37.50	41.67	46.91	9.03	45.21	38.56	36.89	37.72
MiniCPM-V-2.6	70.47	44.57	57.53	75.45	76.19	50.91	39.84	63.33	52.08	46.98	41.07	45.16	65.84	44.75	55.30
Molmo-7B-D-0924	40.27	32.08	22.60	26.36	27.21	7.27	35.51	20.83	19.79	32.78	0.03	35.37	26.18	23.84	25.01
Molmo-7B-O-0924	45.64	31.71	26.03	44.55	24.49	1.82	40.82	25.83	21.88	32.01	0.00	32.31	31.40	23.11	27.26
Ovis1.5-Gemma2-9B	55.70	43.58	30.82	18.18	61.22	49.09	43.08	55.00	29.17	45.81	1.37	49.05	41.68	38.66	40.17
Ovis1.6-Gemma2-9B	46.98	44.49	28.08	15.45	68.03	40.00	42.52	60.00	37.50	48.78	4.46	46.92	42.67	37.86	40.27
Qwen-VL-Chat	46.98	44.62	40.41	60.00	58.50	29.09	43.82	59.17	37.50	48.00	3.08	48.00	50.43	36.10	43.26
Qwen2-VL-2B-Instruct	36.91	45.22	44.52	51.82	68.03	38.18	43.32	50.83	36.46	49.18	35.49	49.22	48.10	43.43	45.76
Qwen2-VL-7B-Instruct	63.76	44.51	51.37	36.36	74.15	52.73	44.01	47.50	38.54	50.62	54.12	50.02	51.95	49.33	50.64
Qwen2-VL-72B-Instruct	60.40	42.94	66.44	65.45	84.35	54.55	43.52	65.83	62.50	50.67	59.70	48.68	67.50	50.01	58.75
QVQ-72B-Preview	64.43	44.02	48.63	60.91	86.39	49.09	40.51	65.00	64.58	45.60	41.67	42.40	64.99	43.88	54.44
Average	49.30	42.82	45.26	48.06	63.15	40.74	40.99	50.34	44.05	46.98	32.52	45.94	50.03	41.66	45.85

Table 17: The main metrics for the Painting subdomain.

Model	**T28**	**T29**	**MCQ**	**QA**	**OverAll**
Closed-source Models					
Claude-3.5-Sonnet	5.44	42.80	-	24.12	24.12
Gemini-1.5-Pro	2.72	44.79	-	23.75	23.75
Gemini-2.0-Flash	5.44	37.06	-	21.25	21.25
GPT-4o	3.40	47.57	-	25.49	25.49
Open-source Models					
Chinese-LLaVA-CLLaMA2	0.00	43.39	-	21.70	21.70
DeepSeek-VL2-Tiny	2.04	40.60	-	21.32	21.32
DeepSeek-VL2-Small	3.40	31.57	-	17.48	17.48
DeepSeek-VL2	2.72	36.71	-	19.71	19.71
GLM-4V-9B	2.04	<u>48.53</u>	-	25.28	25.28
InternVL2-4B	3.40	38.50	-	20.95	20.95
InternVL2-8B	2.72	42.58	-	22.65	22.65
InternVL2-26B	4.76	41.09	-	22.93	22.93
InternVL2.5-1B	4.08	46.23	-	25.16	25.16
InternVL2.5-2B	2.72	46.99	-	24.86	24.86
InternVL2.5-4B	5.44	45.44	-	25.44	25.44
InternVL2.5-8B	3.40	47.03	-	25.22	25.22
InternVL2.5-26B	<u>7.48</u>	44.74	-	26.11	26.11
InternVL2.5-38B	6.80	46.83	-	<u>26.82</u>	<u>26.82</u>
InternVL2.5-78B	**8.16**	46.79	-	**27.48**	**27.48**
LLaVA-v1.5-7B	0.00	45.93	-	22.96	22.96
LLaVA-v1.5-13B	0.68	44.81	-	22.74	22.74
LLaVA-v1.6-Mistral-7B	3.40	44.69	-	24.04	24.04
LLaVA-v1.6-Vicuna-7B	2.04	44.34	-	23.19	23.19
LLaVA-v1.6-Vicuna-13B	1.36	43.75	-	22.55	22.55
MiniCPM-V	1.36	**48.98**	-	25.17	25.17
MiniCPM-V-2	3.40	47.04	-	25.22	25.22
MiniCPM-LLaMA3-V-2.5	5.44	46.18	-	25.81	25.81
MiniCPM-V-2.6	4.76	41.54	-	23.15	23.15
Molmo-7B-D-0924	0.00	34.97	-	17.49	17.49
Molmo-7B-O-0924	0.00	32.09	-	16.05	16.05
Ovis1.5-Gemma2-9B	0.68	46.71	-	23.69	23.69
Ovis1.6-Gemma2-9B	0.68	43.08	-	21.88	21.88
Qwen-VL-Chat	0.00	47.51	-	23.76	23.76
Qwen2-VL-2B-Instruct	3.40	45.98	-	24.69	24.69
Qwen2-VL-7B-Instruct	4.76	46.70	-	25.73	25.73
Qwen2-VL-72B-Instruct	3.40	46.05	-	24.73	24.73
QVQ-72B-Preview	3.40	36.94	-	20.17	20.17
Average	3.11	43.42	-	23.26	23.26

Table 18: The main metrics for the Oracle Bone Script subdomain.

Model	T30	T31	MCQ	QA	OverAll
Closed-source Models					
Claude-3.5-Sonnet	8.22	34.67	34.67	8.22	21.45
Gemini-1.5-Pro	1.58	38.00	38.00	1.58	19.79
Gemini-2.0-Flash	6.46	37.33	37.33	6.46	21.90
GPT-4o	5.50	37.33	37.33	5.50	21.42
Open-source Models					
Chinese-LLaVA-CLLaMA2	6.33	20.00	20.00	6.33	13.17
DeepSeek-VL2-Tiny	4.48	31.33	31.33	4.48	17.91
DeepSeek-VL2-Small	5.06	28.00	28.00	5.06	16.53
DeepSeek-VL2	5.50	23.33	23.33	5.50	14.42
GLM-4V-9B	6.66	40.00	40.00	6.66	23.33
InternVL2-4B	6.27	28.00	28.00	6.27	17.14
InternVL2-8B	7.51	30.67	30.67	7.51	19.09
InternVL2-26B	9.41	41.33	41.33	9.41	25.37
InternVL2.5-1B	8.30	30.67	30.67	8.30	19.49
InternVL2.5-2B	4.91	32.67	32.67	4.91	18.79
InternVL2.5-4B	9.71	34.67	34.67	9.71	22.19
InternVL2.5-8B	7.83	32.67	32.67	7.83	20.25
InternVL2.5-26B	10.93	36.67	36.67	10.93	23.80
InternVL2.5-38B	11.53	41.33	41.33	11.53	26.43
InternVL2.5-78B	15.24	**42.67**	**42.67**	15.24	**28.96**
LLaVA-v1.5-7B	0.49	19.33	19.33	0.49	9.91
LLaVA-v1.5-13B	3.00	21.33	21.33	3.00	12.17
LLaVA-v1.6-Mistral-7B	3.65	31.33	31.33	3.65	17.49
LLaVA-v1.6-Vicuna-7B	4.22	25.33	25.33	4.22	14.78
LLaVA-v1.6-Vicuna-13B	4.39	28.00	28.00	4.39	16.20
MiniCPM-V	2.56	20.00	20.00	2.56	11.28
MiniCPM-V-2	4.80	28.67	28.67	4.80	16.74
MiniCPM-LLaMA3-V-2.5	1.87	25.33	25.33	1.87	13.60
MiniCPM-V-2.6	1.08	25.33	25.33	1.08	13.21
Molmo-7B-D-0924	3.15	30.67	30.67	3.15	16.91
Molmo-7B-O-0924	0.00	28.00	28.00	0.00	14.00
Ovis1.5-Gemma2-9B	1.22	24.00	24.00	1.22	12.61
Ovis1.6-Gemma2-9B	5.23	30.00	30.00	5.23	17.62
Qwen-VL-Chat	6.97	34.00	34.00	6.97	20.49
Qwen2-VL-2B-Instruct	4.49	26.67	26.67	4.49	15.58
Qwen2-VL-7B-Instruct	5.51	30.00	30.00	5.51	17.76
Qwen2-VL-72B-Instruct	5.38	29.33	29.33	5.38	17.36
QVQ-72B-Preview	**15.59**	30.00	30.00	**15.59**	22.80
Average	5.81	30.50	30.50	5.81	18.16

Table 19: The main metrics for the Seal subdomain.

Model	**T32**	**T33**	**T34**	**T35**	**T36**	**MCQ**	**QA**	**OverAll**
Closed-source Models								
Claude-3.5-Sonnet	98.00	<u>62.07</u>	42.50	45.64	92.86	73.86	45.64	68.21
Gemini-1.5-Pro	96.67	50.00	32.50	42.06	<u>93.75</u>	68.23	42.06	63.00
Gemini-2.0-Flash	97.33	**69.83**	<u>47.50</u>	36.83	**94.64**	**77.33**	36.83	**69.23**
GPT-4o	**100.00**	58.62	42.50	<u>46.29</u>	**94.64**	73.94	<u>46.29</u>	68.41
Open-source Models								
Chinese-LLaVA-CLLaMA2	44.00	17.24	16.67	42.72	36.61	28.63	42.72	31.45
DeepSeek-VL2-Tiny	64.00	35.34	40.83	43.96	82.14	55.58	43.96	53.25
DeepSeek-VL2-Small	45.33	24.14	37.50	36.53	44.64	37.90	36.53	37.63
DeepSeek-VL2	70.67	32.76	38.33	40.56	62.50	51.07	40.56	48.96
GLM-4V-9B	96.67	54.31	39.17	45.88	91.07	70.31	45.88	65.42
InternVL2-4B	78.00	32.76	30.83	41.41	70.54	53.03	41.41	50.71
InternVL2-8B	90.00	44.83	38.33	39.78	85.71	64.72	39.78	59.73
InternVL2-26B	93.33	52.59	32.50	40.57	88.39	66.70	40.57	61.48
InternVL2.5-1B	88.00	40.52	27.50	44.65	75.89	57.98	44.65	55.31
InternVL2.5-2B	84.00	40.52	28.33	44.55	77.68	57.63	44.55	55.02
InternVL2.5-4B	90.00	44.83	41.67	44.86	83.04	64.89	44.86	60.88
InternVL2.5-8B	93.33	43.10	37.50	44.42	87.50	65.36	44.42	61.17
InternVL2.5-26B	98.67	55.17	38.33	44.02	90.18	70.59	44.02	65.27
InternVL2.5-38B	98.00	48.28	40.83	44.26	88.39	68.88	44.26	63.95
InternVL2.5-78B	<u>99.33</u>	61.21	**49.17**	43.18	91.07	<u>75.20</u>	43.18	<u>68.79</u>
LLaVA-v1.5-7B	24.67	33.62	25.00	45.87	43.75	31.76	45.87	34.58
LLaVA-v1.5-13B	22.67	32.76	23.33	43.03	60.71	34.87	43.03	36.50
LLaVA-v1.6-Mistral-7B	42.67	25.86	25.83	40.60	30.36	31.18	40.60	33.06
LLaVA-v1.6-Vicuna-7B	37.33	25.86	27.50	42.80	41.96	33.16	42.80	35.09
LLaVA-v1.6-Vicuna-13B	31.33	22.41	25.83	42.50	40.18	29.94	42.50	32.45
MiniCPM-V	76.67	61.21	30.00	45.83	80.36	62.06	45.83	58.81
MiniCPM-V-2	80.00	43.10	36.67	43.95	84.82	61.15	43.95	57.71
MiniCPM-LLaMA3-V-2.5	81.33	33.62	25.83	44.10	85.71	56.62	44.10	54.12
MiniCPM-V-2.6	91.33	53.45	45.00	42.12	87.50	69.32	42.12	63.88
Molmo-7B-D-0924	24.67	13.79	30.00	34.98	23.21	22.92	34.98	25.33
Molmo-7B-O-0924	25.33	26.72	35.00	33.90	20.54	26.90	33.90	28.30
Ovis1.5-Gemma2-9B	79.33	37.07	22.50	43.93	84.82	55.93	43.93	53.53
Ovis1.6-Gemma2-9B	88.67	35.34	25.83	45.19	87.50	59.34	45.19	56.51
Qwen-VL-Chat	81.33	30.17	36.67	**47.51**	88.39	59.14	**47.51**	56.81
Qwen2-VL-2B-Instruct	80.00	35.34	42.50	44.73	87.50	61.34	44.73	58.01
Qwen2-VL-7B-Instruct	95.33	37.93	39.17	45.71	85.71	64.54	45.71	60.77
Qwen2-VL-72B-Instruct	96.67	47.41	**49.17**	44.57	90.18	70.86	44.57	65.60
QVQ-72B-Preview	96.67	52.59	40.83	39.77	86.61	69.18	39.77	63.29
Average	75.17	40.98	34.84	42.79	74.08	56.27	42.79	53.57

Table 20: The main metrics for the Cultural Relic subdomain.

Model	T37	T38	T39	T40	T41	T42	T43	T44	T45	MCQ	QA	OverAll
Closed-source Models												
Claude-3.5-Sonnet	53.72	49.13	38.80	**80.91**	**77.78**	90.00	63.64	40.62	48.08	78.08	47.89	60.30
Gemini-1.5-Pro	28.90	47.96	40.39	71.82	72.22	**91.67**	59.09	40.32	46.15	73.70	40.83	55.39
Gemini-2.0-Flash	**68.26**	48.09	37.48	75.45	76.98	89.17	75.45	30.38	46.15	79.26	48.22	60.82
GPT-4o	42.88	51.45	40.71	76.36	73.02	82.50	66.36	44.67	57.69	74.56	49.17	59.52
Open-source Models												
Chinese-LLaVA-CLLaMA2	5.75	43.99	35.83	38.18	18.25	12.50	12.73	40.69	9.62	20.42	25.01	24.17
DeepSeek-VL2-Tiny	47.21	51.16	39.91	34.55	46.83	69.17	44.55	42.44	40.38	48.78	45.30	46.24
DeepSeek-VL2-Small	47.33	42.00	33.31	21.82	36.51	41.67	32.73	25.91	38.46	33.18	38.42	35.53
DeepSeek-VL2	46.51	45.74	35.80	38.18	48.41	45.00	51.82	33.52	30.77	45.85	39.13	41.75
GLM-4V-9B	48.41	49.98	**42.09**	62.73	64.29	74.17	63.64	43.24	44.23	66.21	46.46	54.75
InternVL2-4B	48.93	50.29	36.93	52.73	36.51	79.17	52.73	44.06	21.15	55.29	41.11	46.94
InternVL2-8B	48.28	49.39	39.30	53.64	69.84	80.83	58.18	44.46	28.85	65.62	42.74	52.53
InternVL2-26B	51.95	51.17	40.21	71.82	72.22	90.00	70.00	44.46	34.62	76.01	45.55	58.49
InternVL2.5-1B	53.01	50.30	39.19	41.82	51.59	85.83	54.55	42.47	28.85	58.45	43.66	49.73
InternVL2.5-2B	48.07	51.30	39.51	60.91	57.14	81.67	63.64	44.36	30.77	65.84	43.62	53.04
InternVL2.5-4B	49.95	51.03	40.53	63.64	59.52	80.83	65.45	45.33	32.69	67.36	44.75	54.33
InternVL2.5-8B	51.81	50.81	41.00	62.73	69.84	80.83	68.18	45.87	42.31	70.40	47.70	57.04
InternVL2.5-26B	57.29	51.50	40.45	71.82	70.63	88.33	67.27	**46.04**	**59.62**	74.51	**53.61**	61.44
InternVL2.5-38B	57.97	51.63	41.28	70.91	73.81	85.83	67.27	45.85	51.92	74.46	51.84	60.72
InternVL2.5-78B	57.49	50.66	40.46	78.18	73.02	85.00	**80.91**	44.77	53.85	**79.28**	51.69	**62.70**
LLaVA-v1.5-7B	0.87	48.17	37.81	30.91	31.75	48.33	23.64	40.94	15.38	33.66	26.34	30.87
LLaVA-v1.5-13B	0.70	48.20	37.08	25.45	27.78	62.50	19.09	42.99	13.46	33.71	26.34	30.80
LLaVA-v1.6-Mistral-7B	2.62	43.23	38.59	26.36	23.81	26.67	23.64	41.79	9.62	25.12	24.32	26.26
LLaVA-v1.6-Vicuna-7B	3.21	48.93	38.79	27.27	23.81	47.50	25.45	44.28	3.85	31.01	25.07	29.23
LLaVA-v1.6-Vicuna-13B	3.88	48.27	38.65	38.18	23.02	47.50	27.27	43.65	9.62	33.99	26.35	31.11
MiniCPM-V	4.76	43.62	37.01	39.09	37.30	59.17	35.45	39.14	25.00	42.75	28.13	35.62
MiniCPM-V-2	12.98	46.88	37.20	44.55	48.41	60.83	44.55	38.58	21.15	49.59	29.90	39.46
MiniCPM-LLaMA3-V-2.5	33.22	47.84	39.15	53.64	48.41	52.50	47.27	41.54	28.85	50.46	37.86	43.60
MiniCPM-V-2.6	41.71	47.01	39.66	66.36	53.97	79.17	55.45	42.15	38.46	63.74	42.33	51.55
Molmo-7B-D-0924	0.00	34.20	31.10	27.27	24.60	30.00	23.64	34.11	3.85	26.38	18.04	23.20
Molmo-7B-O-0924	0.00	40.51	31.63	27.27	26.19	20.83	24.55	31.49	3.85	24.71	18.96	22.92
Ovis1.5-Gemma2-9B	3.78	46.96	36.79	51.82	43.65	80.83	37.27	41.81	26.92	53.39	29.87	41.09
Ovis1.6-Gemma2-9B	20.12	49.91	39.42	53.64	49.21	74.17	53.64	42.73	28.85	57.67	35.40	45.74
Qwen-VL-Chat	25.25	47.20	38.57	39.09	47.62	68.33	43.64	44.35	17.31	49.67	33.53	41.26
Qwen2-VL-2B-Instruct	44.55	50.92	40.45	33.64	38.10	72.50	57.27	42.29	23.08	50.38	40.21	44.75
Qwen2-VL-7B-Instruct	50.84	51.55	38.99	57.27	60.32	78.33	55.45	45.15	44.23	62.84	47.94	53.57
Qwen2-VL-72B-Instruct	59.95	**52.60**	40.43	71.82	72.22	76.67	60.91	43.64	51.92	70.41	52.03	58.91
QVQ-72B-Preview	59.11	47.74	37.17	60.91	73.02	80.83	65.45	41.36	44.23	70.05	48.11	56.65
Average	34.63	48.14	38.42	51.43	51.39	67.59	49.78	41.39	31.24	55.05	38.85	46.00

Table 21: The main metrics for the Illustration subdomain.

Model	T1						T2			T3		T4		T6	
	AR↑	Edit Distance↓	F1-Score↑	Precision↑	Recall↑	BLEU↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	Precision↑	Recall↑	Precision↑	Recall↑	BERTScore↑	ANLS↑
Closed-source Models															
Claude-3.5-Sonnet	36.44	61.79	59.00	74.69	52.07	28.28	41.68	12.90	25.80	38.85	19.51	17.43	13.16	83.81	13.22
Gemini-1.5-Pro	57.15	41.32	80.56	83.44	78.89	59.75	58.63	27.90	42.00	**57.03**	59.64	**46.82**	**36.86**	84.73	17.12
Gemini-2.0-Flash	**70.53**	29.33	**81.53**	83.77	**80.71**	61.97	**60.50**	**28.48**	**44.26**	52.33	45.91	38.53	31.16	83.39	18.06
GPT-4o	21.28	75.23	44.65	58.51	40.26	23.03	33.93	10.35	19.67	14.23	13.26	12.43	9.86	87.27	22.01
Open-source Models															
Chinese-LLaVA-CLLaMA2	-5.63	98.58	7.54	22.91	5.34	0.04	11.06	0.62	5.95	1.07	1.22	0.40	0.01	81.10	13.63
DeepSeek-VL2-Tiny	-164.04	87.67	47.81	77.89	40.82	21.52	22.29	6.36	11.41	7.51	2.03	0.03	0.02	83.07	14.79
DeepSeek-VL2-Small	-14.86	94.94	16.93	85.27	10.62	1.30	5.94	1.39	3.98	2.01	0.05	2.43	0.28	76.55	11.49
DeepSeek-VL2	-98.08	94.05	21.95	72.90	14.75	2.92	8.78	1.77	5.64	20.58	26.44	2.83	0.60	79.31	12.27
GLM-4V-9B	-36.62	82.45	58.02	67.36	55.07	33.18	26.01	6.46	12.52	8.70	16.87	4.40	7.26	88.88	24.01
InternVL2-4B	50.83	47.46	66.64	80.98	61.96	48.62	31.76	11.04	25.15	5.64	0.96	0.66	0.21	76.98	12.00
InternVL2-8B	24.02	43.86	70.25	87.20	65.63	52.09	29.13	10.29	23.66	6.97	4.17	10.19	5.04	83.23	18.01
InternVL2-26B	25.13	46.09	69.92	86.23	64.93	51.19	33.90	12.31	26.50	6.34	7.63	1.79	1.32	81.52	16.94
InternVL2.5-1B	-33.90	54.75	59.76	87.53	53.55	41.47	26.10	9.48	21.47	8.10	0.51	2.51	0.30	80.70	16.69
InternVL2.5-2B	8.59	50.19	64.06	86.15	58.33	46.76	29.64	10.20	22.92	10.89	10.85	2.78	1.55	81.20	17.49
InternVL2.5-4B	29.00	46.15	67.06	87.40	61.84	50.04	32.23	12.11	25.62	6.68	1.39	12.31	4.39	81.09	18.18
InternVL2.5-8B	38.19	41.58	70.24	89.20	65.33	53.80	32.32	11.53	24.77	9.54	4.78	18.02	6.62	86.12	21.31
InternVL2.5-26B	18.62	40.30	73.69	90.07	68.85	56.27	36.44	13.48	27.91	18.42	18.65	7.98	1.28	84.49	19.99
InternVL2.5-38B	39.48	34.44	76.04	89.32	72.29	61.51	27.84	9.69	21.31	16.31	11.79	26.71	6.99	88.57	24.30
InternVL2.5-78B	48.16	27.21	80.63	**90.65**	77.24	**66.92**	38.87	14.42	30.23	13.65	11.19	31.66	14.41	88.01	22.88
LLaVA-v1.5-7B	-20.24	99.50	2.29	20.07	1.34	0.01	6.98	0.51	4.75	1.78	0.93	0.00	0.00	84.22	17.78
LLaVA-v1.5-13B	-34.85	99.29	3.65	18.86	2.23	0.01	8.11	0.59	4.98	1.53	0.19	0.00	0.00	83.39	16.34
LLaVA-v1.6-Mistral-7B	-135.12	98.69	8.65	20.41	6.18	0.05	10.11	0.84	6.19	1.16	5.02	0.00	0.00	81.12	13.93
LLaVA-v1.6-Vicuna-7B	-73.22	98.80	6.86	20.35	4.70	0.05	9.38	0.65	6.04	3.14	3.66	0.00	0.00	76.56	9.27
LLaVA-v1.6-Vicuna-13B	-35.30	98.47	8.26	19.34	5.78	0.05	10.77	0.88	6.62	2.14	11.14	0.00	0.00	77.83	11.37
MiniCPM-V	-142.36	99.55	2.32	24.27	1.32	0.01	7.26	0.44	4.41	2.80	13.12	0.00	0.00	88.19	**26.52**
MiniCPM-V-2	-0.65	98.65	6.61	23.59	4.79	0.16	10.44	0.94	5.57	6.43	0.54	0.81	0.06	87.10	22.31
MiniCPM-LLaMA3-V-2.5	-22.67	96.26	14.84	60.37	9.62	0.81	8.87	1.44	5.20	21.74	11.51	4.25	0.74	84.62	15.79
MiniCPM-V-2.6	-15.59	73.31	44.12	60.11	40.73	21.55	23.76	6.61	15.75	15.20	2.44	11.25	2.13	87.70	23.75
Molmo-7B-D-0924	-780.55	99.97	0.29	13.02	0.15	0.00	1.45	0.15	1.13	0.00	0.00	0.00	0.00	72.67	1.97
Molmo-7B-O-0924	-13.91	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	70.95	4.00
Ovis1.5-Gemma2-9B	-143.03	98.60	8.02	24.62	5.91	0.06	11.53	0.75	6.97	7.01	7.86	0.00	0.00	81.22	11.77
Ovis1.6-Gemma2-9B	-276.80	98.16	15.68	29.30	12.02	0.27	11.08	0.90	5.96	6.75	46.63	0.17	0.35	78.96	9.43
Qwen-VL-Chat	-45.86	97.78	10.76	31.90	7.49	0.26	9.98	0.94	6.21	6.76	2.32	0.28	0.40	83.77	16.95
Qwen2-VL-2B-Instruct	-27.45	80.61	33.60	79.92	25.85	13.04	25.33	8.84	19.42	12.35	4.07	4.45	0.66	87.56	23.67
Qwen2-VL-7B-Instruct	-19.49	58.88	61.80	87.81	53.82	36.83	26.79	9.57	20.12	22.57	11.06	23.32	11.50	**89.41**	25.98
Qwen2-VL-72B-Instruct	30.56	44.36	77.15	85.07	72.82	53.35	37.06	14.07	27.11	18.36	13.16	34.64	16.39	88.98	24.88
QVQ-72B-Preview	-411.81	90.01	38.51	33.89	59.98	9.83	17.57	6.33	10.53	12.70	**75.65**	4.35	5.38	56.66	2.89
Average	-55.51	73.74	39.45	58.23	36.30	24.24	22.26	7.17	15.61	12.09	12.60	8.74	4.84	82.19	16.57

Table 22: The other metrics for the Ancient Chinese Text subdomain.

Model	T8						T9		T15	
	AR↑	Edit Distance↓	F1-Score↑	Precision↑	Recall↑	BLEU↑	BERTScore↑	ANLS↑	BERTScore↑	ANLS↑
Closed-source Models										
Claude-3.5-Sonnet	37.92	61.07	58.56	61.55	56.42	22.63	83.31	9.11	84.13	8.78
Gemini-1.5-Pro	5.77	91.00	35.50	35.87	35.46	8.51	82.01	7.93	80.90	7.30
Gemini-2.0-Flash	24.97	73.99	64.73	66.16	63.65	16.79	49.17	0.60	52.31	1.38
GPT-4o	21.87	65.40	52.44	53.05	53.95	27.59	82.90	9.14	**85.84**	10.13
Open-source Models										
Chinese-LLaVA-CLLaMA2	-549.97	99.68	1.34	1.93	2.60	0.08	74.07	7.61	77.57	7.74
DeepSeek-VL2-Tiny	-127.84	84.41	58.79	71.27	53.77	12.41	77.94	8.65	82.47	9.20
DeepSeek-VL2-Small	19.29	80.23	48.25	76.67	38.21	8.71	55.79	3.07	62.59	5.60
DeepSeek-VL2	-139.41	78.37	51.44	75.98	44.07	11.55	72.77	7.13	71.02	7.94
GLM-4V-9B	-141.25	83.01	46.44	45.34	52.85	11.61	77.94	8.11	80.48	8.45
InternVL2-4B	14.73	63.91	67.52	66.23	75.00	27.29	78.07	7.39	79.28	8.38
InternVL2-8B	15.35	56.89	65.77	56.16	80.58	35.90	81.59	8.65	77.69	8.44
InternVL2-26B	12.38	62.85	67.68	60.28	79.55	32.27	83.06	8.76	78.75	8.82
InternVL2.5-1B	16.48	56.76	70.45	67.71	79.09	35.47	82.28	8.66	81.87	9.44
InternVL2.5-2B	-145.18	64.46	66.52	**83.39**	60.63	27.68	82.44	8.67	83.96	9.88
InternVL2.5-4B	24.84	55.69	71.11	66.23	82.69	38.08	82.68	8.88	83.37	9.67
InternVL2.5-8B	23.40	55.71	71.80	65.80	83.28	38.18	83.05	8.91	83.80	9.96
InternVL2.5-26B	37.45	52.67	74.53	80.11	77.85	41.40	83.22	9.18	80.29	9.28
InternVL2.5-38B	24.65	56.44	75.59	69.38	86.29	38.97	**84.03**	9.03	82.91	9.64
InternVL2.5-78B	**43.82**	44.46	**80.97**	77.33	**88.05**	**51.45**	83.32	**9.47**	78.50	8.82
LLaVA-v1.5-7B	-217.04	99.68	1.02	2.96	0.90	0.05	76.93	8.38	80.64	9.21
LLaVA-v1.5-13B	-1102.21	99.58	1.03	1.96	1.23	0.08	76.40	8.15	81.18	9.37
LLaVA-v1.6-Mistral-7B	-3514.46	99.60	2.05	2.57	3.02	0.09	77.77	8.75	80.20	8.35
LLaVA-v1.6-Vicuna-7B	-1261.20	99.67	1.77	1.47	3.41	0.07	79.43	8.05	80.14	8.67
LLaVA-v1.6-Vicuna-13B	-945.59	99.49	1.95	1.48	4.35	0.10	78.75	8.02	79.13	8.45
MiniCPM-V	-3801.58	99.54	2.11	9.21	1.39	0.05	79.19	8.87	74.25	7.89
MiniCPM-V-2	-59.68	95.84	18.70	22.67	19.21	1.89	74.72	7.17	78.61	7.28
MiniCPM-LLaMA3-V-2.5	2.81	92.43	35.66	47.58	31.10	2.55	78.13	8.50	79.82	8.45
MiniCPM-V-2.6	13.14	85.27	51.15	61.73	46.67	9.79	81.96	8.02	80.22	8.70
Molmo-7B-D-0924	-15603.85	**100.00**	0.12	0.38	0.07	0.00	59.92	0.02	66.54	0.04
Molmo-7B-O-0924	-919.83	**100.00**	0.00	0.00	0.00	0.00	61.39	0.06	66.30	0.02
Ovis1.5-Gemma2-9B	-523.16	99.23	2.09	3.81	1.77	0.15	78.58	8.67	81.52	9.05
Ovis1.6-Gemma2-9B	-288.76	97.63	16.22	16.98	17.07	1.12	80.59	9.03	80.25	8.62
Qwen-VL-Chat	-189.23	95.05	19.65	18.76	24.01	1.66	72.91	6.96	84.83	9.85
Qwen2-VL-2B-Instruct	-205.93	75.24	50.98	70.59	43.42	14.54	80.43	9.16	84.36	**10.14**
Qwen2-VL-7B-Instruct	-288.55	69.92	64.99	77.34	59.33	20.90	80.48	8.87	83.76	9.66
Qwen2-VL-72B-Instruct	25.93	69.65	66.77	69.98	64.96	25.05	81.92	7.89	83.38	9.15
QVQ-72B-Preview	-1794.57	72.52	47.95	47.30	61.52	18.97	80.10	8.17	77.80	8.39
Average	-850.12	79.39	40.91	44.25	42.63	15.77	77.28	7.61	78.67	8.11

Table 23: The other metrics for the Calligraphy subdomain.

Model	T17		T22		T25		T26						T27	
	BERTScore↑	ANLS↑	BERTScore↑	ANLS↑	BERTScore↑	ANLS↑	AR↑	Edit Distance↓	F1-Score↑	Precision↑	Recall↑	BLEU↑	BERTScore↑	ANLS↑
Closed-source Models														
Claude-3.5-Sonnet	**86.30**	7.11	80.00	7.00	**90.72**	9.92	-20.95	55.57	53.82	57.24	51.71	27.88	87.73	9.59
Gemini-1.5-Pro	80.87	8.08	78.88	6.70	86.02	9.94	3.45	51.26	64.55	64.33	66.20	49.93	80.48	9.12
Gemini-2.0-Flash	55.54	3.12	58.65	2.52	84.31	9.31	36.38	36.13	69.43	68.65	70.94	56.42	68.21	5.53
GPT-4o	84.85	8.29	**81.99**	**7.13**	89.68	12.23	-35.25	67.02	45.72	49.45	43.91	21.77	**89.73**	11.36
Open-source Models														
Chinese-LLaVA-CLLaMA2	77.30	6.42	76.73	5.09	78.27	6.22	-1267.82	99.34	2.43	7.22	2.97	0.04	80.38	7.03
DeepSeek-VL2-Tiny	84.47	8.08	78.75	6.49	88.59	10.53	-333.11	81.28	53.34	65.10	51.77	30.53	86.34	9.49
DeepSeek-VL2-Small	61.34	3.95	61.90	4.35	69.87	6.86	-12.28	85.11	31.15	71.81	23.73	6.44	70.92	7.75
DeepSeek-VL2	65.25	5.09	70.32	5.57	73.36	8.45	-125.75	82.81	37.85	64.70	31.80	11.17	72.83	8.64
GLM-4V-9B	82.93	**8.46**	79.71	6.55	87.42	10.31	-354.89	72.31	54.84	53.92	60.90	36.08	88.13	11.99
InternVL2-4B	77.61	7.37	62.90	4.32	85.15	9.98	-98.74	48.53	61.74	64.04	62.50	45.89	84.19	10.55
InternVL2-8B	79.77	7.81	75.62	6.17	86.37	10.13	11.07	39.86	68.28	68.93	70.48	53.77	81.06	9.95
InternVL2-26B	79.90	8.11	75.70	6.31	87.75	10.48	-12.76	43.72	65.62	65.43	68.11	50.64	83.06	10.35
InternVL2.5-1B	81.78	8.12	76.72	6.38	88.29	11.25	-55.80	56.38	60.22	72.53	56.62	36.00	86.51	10.88
InternVL2.5-2B	82.80	7.75	78.08	6.54	89.47	11.76	-54.75	48.08	61.86	65.77	62.07	45.95	87.01	10.34
InternVL2.5-4B	83.36	8.45	80.18	6.80	89.73	12.08	-5.26	37.76	70.15	72.47	71.05	56.95	88.90	11.79
InternVL2.5-8B	80.45	8.02	78.75	6.73	89.63	11.49	56.08	33.52	72.78	**75.73**	72.28	59.76	87.17	10.45
InternVL2.5-26B	81.76	8.12	76.58	6.35	89.58	11.97	3.20	35.86	70.68	72.24	72.25	57.30	86.08	10.92
InternVL2.5-38B	81.55	8.21	78.59	6.60	89.57	12.31	**58.81**	30.96	73.17	73.35	73.25	60.98	87.83	**12.17**
InternVL2.5-78B	81.77	8.03	75.83	6.54	89.81	**12.36**	56.63	29.36	**74.18**	74.13	**74.73**	**63.41**	83.48	10.47
LLaVA-v1.5-7B	79.07	7.03	78.59	6.45	80.98	8.72	-1442.97	99.67	0.86	5.83	0.55	0.01	84.22	10.35
LLaVA-v1.5-13B	80.69	7.59	77.73	6.08	84.91	10.84	-1365.16	99.39	1.99	8.71	1.32	0.04	85.87	9.49
LLaVA-v1.6-Mistral-7B	80.52	7.47	75.22	6.01	83.17	9.63	-1748.70	99.22	4.52	6.64	5.84	0.10	80.36	7.97
LLaVA-v1.6-Vicuna-7B	81.79	7.35	75.36	5.92	85.21	9.67	-641.07	99.35	3.27	6.28	2.69	0.08	85.15	9.81
LLaVA-v1.6-Vicuna-13B	81.30	7.52	74.77	6.00	85.37	9.72	-1403.05	99.25	4.00	6.40	3.58	0.11	82.95	9.64
MiniCPM-V	81.45	7.44	77.38	5.95	81.13	8.23	-1184.32	99.51	1.67	9.04	0.99	0.01	83.30	9.42
MiniCPM-V-2	81.39	7.24	74.37	5.26	84.56	8.72	-1354.84	98.96	4.98	7.54	10.78	0.26	81.09	7.19
MiniCPM-LLaMA3-V-2.5	80.01	7.31	77.79	6.15	84.75	9.07	-767.33	91.38	24.58	39.13	21.10	2.93	82.29	8.12
MiniCPM-V-2.6	81.89	7.24	73.84	5.84	84.82	9.14	15.41	62.38	50.07	55.77	48.63	26.92	81.39	8.93
Molmo-7B-D-0924	64.16	0.00	70.84	0.17	65.56	0.00	-32830.42	**100.00**	0.08	1.23	0.04	0.00	70.28	0.45
Molmo-7B-O-0924	63.42	0.00	76.27	5.37	64.02	0.00	-2045.99	**100.00**	0.00	0.00	0.00	0.00	64.60	0.02
Ovis1.5-Gemma2-9B	80.16	6.99	79.91	6.24	82.79	8.83	-600.61	99.03	4.23	8.58	3.67	0.09	87.08	11.02
Ovis1.6-Gemma2-9B	82.11	6.87	78.78	6.25	87.45	10.10	-2946.17	98.95	10.57	11.93	12.81	0.40	84.26	9.58
Qwen-VL-Chat	81.65	7.58	80.71	6.93	86.20	9.80	-605.98	98.14	7.77	11.76	7.35	0.59	87.36	8.63
Qwen2-VL-2B-Instruct	82.26	8.18	79.93	6.70	87.44	10.91	-152.63	65.80	43.59	64.71	39.24	25.27	88.15	10.29
Qwen2-VL-7B-Instruct	80.86	8.16	80.95	7.06	89.34	11.90	-139.35	48.81	63.26	70.76	60.92	44.51	89.25	10.79
Qwen2-VL-72B-Instruct	78.25	7.63	80.18	6.85	89.84	11.49	12.52	43.72	68.75	70.58	67.83	52.30	87.66	9.70
QVQ-72B-Preview	79.59	8.45	74.76	6.26	81.79	9.41	-5555.20	76.89	38.31	37.19	51.81	23.69	76.94	7.85
Average	78.65	6.99	76.03	5.94	84.40	9.56	-1538.04	70.68	38.49	44.03	38.55	25.63	82.76	9.12

Table 24: The other metrics for the Painting subdomain.

Model	T28						T29	
	AR↑	Edit Distance↓	F1-Score↑	Precision↑	Recall↑	BLEU↑	BERTScore↑	ANLS↑
Closed-source Models								
Claude-3.5-Sonnet	5.44	94.56	5.44	5.44	5.44	0.97	77.50	8.09
Gemini-1.5-Pro	2.04	97.28	2.72	2.72	2.72	0.48	80.08	9.49
Gemini-2.0-Flash	4.76	94.56	5.44	5.44	5.44	0.97	69.47	4.64
GPT-4o	-53.74	96.60	3.40	3.40	3.40	0.60	82.67	12.47
Open-source Models								
Chinese-LLaVA-CLLaMA2	-1453.06	**100.00**	0.00	0.00	0.00	0.00	78.33	8.45
DeepSeek-VL2-Tiny	-259.18	97.96	2.04	2.04	2.04	0.36	74.28	6.92
DeepSeek-VL2-Small	-8.84	96.60	3.40	3.40	3.40	0.60	59.35	3.78
DeepSeek-VL2	-1730.61	97.91	2.15	2.10	2.72	0.37	67.29	6.12
GLM-4V-9B	-254.42	97.96	2.04	2.04	2.04	0.36	**85.81**	11.24
InternVL2-4B	-476.87	97.21	2.84	2.79	3.40	0.50	71.73	5.27
InternVL2-8B	-85.71	97.28	2.72	2.72	2.72	0.48	76.41	8.75
InternVL2-26B	4.76	95.24	4.76	4.76	4.76	0.85	74.70	7.48
InternVL2.5-1B	-79.59	96.43	3.67	3.57	4.08	0.66	81.93	10.53
InternVL2.5-2B	-558.50	97.28	2.72	2.72	2.72	0.48	83.35	10.63
InternVL2.5-4B	-28.57	94.56	5.44	5.44	5.44	0.97	80.66	10.22
InternVL2.5-8B	-42.86	96.60	3.40	3.40	3.40	0.60	82.64	11.42
InternVL2.5-26B	-10.20	92.52	7.48	7.48	7.48	1.33	79.96	9.52
InternVL2.5-38B	**6.12**	93.20	6.80	6.80	6.80	1.21	82.42	11.24
InternVL2.5-78B	-6.80	91.84	**8.16**	**8.16**	**8.16**	**1.45**	82.30	11.28
LLaVA-v1.5-7B	-165.99	**100.00**	0.00	0.00	0.00	0.00	81.75	10.10
LLaVA-v1.5-13B	-1388.44	99.98	0.05	0.03	0.68	0.00	80.69	8.92
LLaVA-v1.6-Mistral-7B	-3346.94	99.91	0.20	0.11	3.40	0.02	80.35	9.02
LLaVA-v1.6-Vicuna-7B	-2430.61	99.86	0.31	0.17	2.04	0.03	79.71	8.96
LLaVA-v1.6-Vicuna-13B	-2966.67	99.95	0.11	0.06	1.36	0.01	78.87	8.62
MiniCPM-V	-976.87	98.64	1.36	1.36	1.36	0.24	84.98	**12.98**
MiniCPM-V-2	-32.65	96.60	3.40	3.40	3.40	0.60	83.87	10.21
MiniCPM-LLaMA3-V-2.5	-2.04	94.56	5.44	5.44	5.44	0.97	82.08	10.27
MiniCPM-V-2.6	-14.29	95.24	4.76	4.76	4.76	0.85	75.85	7.23
Molmo-7B-D-0924	0.00	**100.00**	0.00	0.00	0.00	0.00	69.54	0.40
Molmo-7B-O-0924	-400.00	**100.00**	0.00	0.00	0.00	0.00	60.73	3.45
Ovis1.5-Gemma2-9B	-23410.88	**100.00**	0.01	0.01	0.68	0.00	82.68	10.73
Ovis1.6-Gemma2-9B	-8640.82	99.99	0.06	0.03	0.68	0.00	78.34	7.82
Qwen-VL-Chat	-1321.77	**100.00**	0.00	0.00	0.00	0.00	82.24	12.78
Qwen2-VL-2B-Instruct	-806.80	97.83	2.27	2.17	3.40	0.39	81.43	10.52
Qwen2-VL-7B-Instruct	-68.03	96.32	3.85	3.68	4.76	0.69	82.21	11.19
Qwen2-VL-72B-Instruct	-8.16	96.60	3.40	3.40	3.40	0.60	80.98	11.12
QVQ-72B-Preview	-16642.86	98.64	1.38	1.37	3.40	0.24	68.54	5.34
Average	-1828.37	97.29	2.74	2.71	3.11	0.48	77.99	8.84

Table 25: The other metrics for the Oracle Bone Script subdomain.

Model	T30					
	AR↑	Edit Distance↓	F1-Score↑	Precision↑	Recall↑	BLEU↑
Closed-source Models						
Claude-3.5-Sonnet	-13.35	92.38	9.13	8.99	9.44	2.66
Gemini-1.5-Pro	-25.52	98.54	2.63	2.61	2.73	0.72
Gemini-2.0-Flash	-11.64	93.62	10.76	11.27	10.55	2.88
GPT-4o	-77.86	95.24	6.79	6.59	7.34	2.23
Open-source Models						
Chinese-LLaVA-CLLaMA2	-1643.86	99.29	1.61	0.97	6.63	0.16
DeepSeek-VL2-Tiny	-785.49	96.05	7.74	10.28	7.33	1.77
DeepSeek-VL2-Small	-0.11	95.05	8.52	14.64	6.69	1.62
DeepSeek-VL2	-531.04	94.72	8.74	11.18	7.79	2.15
GLM-4V-9B	-241.25	96.58	6.93	6.56	10.18	1.74
InternVL2-4B	-69.74	94.99	11.92	14.42	11.42	3.18
InternVL2-8B	-41.42	93.64	14.35	16.70	13.83	2.99
InternVL2-26B	-7.15	90.95	16.74	20.67	14.96	3.64
InternVL2.5-1B	-150.39	92.92	12.61	15.82	12.55	2.53
InternVL2.5-2B	-110.55	96.08	8.11	10.09	8.44	1.75
InternVL2.5-4B	-39.54	90.73	15.42	19.35	13.64	4.25
InternVL2.5-8B	-13.50	92.84	12.65	15.38	11.38	3.33
InternVL2.5-26B	2.98	89.77	20.43	<u>24.19</u>	18.36	4.72
InternVL2.5-38B	<u>6.72</u>	88.76	<u>20.56</u>	23.59	<u>18.90</u>	<u>5.86</u>
InternVL2.5-78B	**10.55**	85.52	**21.31**	**24.39**	**19.74**	**6.68**
LLaVA-v1.5-7B	-107.55	99.58	0.66	0.75	0.65	0.16
LLaVA-v1.5-13B	-423.60	99.15	1.41	1.11	3.38	0.20
LLaVA-v1.6-Mistral-7B	-3911.50	<u>99.71</u>	0.93	0.56	4.22	0.07
LLaVA-v1.6-Vicuna-7B	-1322.91	99.04	1.65	1.15	4.49	0.23
LLaVA-v1.6-Vicuna-13B	-1516.58	99.29	1.47	0.95	4.77	0.16
MiniCPM-V	-1798.77	98.82	2.74	2.37	3.76	0.42
MiniCPM-V-2	-718.17	98.86	2.32	2.06	6.49	0.36
MiniCPM-LLaMA3-V-2.5	-179.13	99.00	2.19	3.34	2.96	0.29
MiniCPM-V-2.6	-6.71	98.92	3.05	3.58	2.81	0.75
Molmo-7B-D-0924	-628.97	99.52	1.09	0.65	3.92	0.12
Molmo-7B-O-0924	-8558.17	**100.00**	0.00	0.00	0.00	0.00
Ovis1.5-Gemma2-9B	-87.69	99.24	1.17	1.21	1.52	0.26
Ovis1.6-Gemma2-9B	-5128.71	99.47	1.08	0.70	5.30	0.13
Qwen-VL-Chat	-740.18	98.05	3.39	2.53	7.99	0.57
Qwen2-VL-2B-Instruct	-971.63	96.52	7.11	9.11	7.49	1.21
Qwen2-VL-7B-Instruct	-38.83	95.20	8.30	10.30	7.81	1.77
Qwen2-VL-72B-Instruct	-20.56	94.72	7.25	7.82	6.93	2.61
QVQ-72B-Preview	-23125.83	95.60	6.22	5.90	17.43	1.66
Average	-1433.18	95.90	7.27	8.43	8.21	1.78

Table 26: The other metrics for the Seal subdomain.

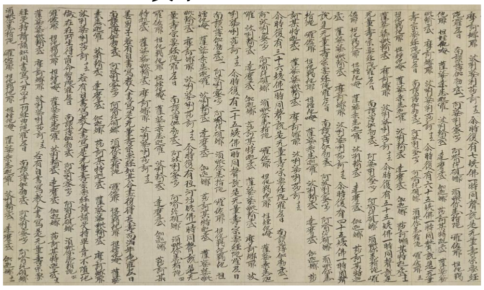
Model	T35	
	BERTScore↑	ANLS↑
Closed-source Models		
Claude-3.5-Sonnet	84.03	7.25
Gemini-1.5-Pro	78.09	6.03
Gemini-2.0-Flash	69.47	4.18
GPT-4o	84.22	8.35
Open-source Models		
Chinese-LLaVA-CLLaMA2	79.79	5.64
DeepSeek-VL2-Tiny	80.87	7.05
DeepSeek-VL2-Small	67.93	5.12
DeepSeek-VL2	74.23	6.89
GLM-4V-9B	83.50	8.26
InternVL2-4B	76.55	6.27
InternVL2-8B	74.05	5.50
InternVL2-26B	75.22	5.91
InternVL2.5-1B	82.02	7.27
InternVL2.5-2B	81.35	7.74
InternVL2.5-4B	82.22	7.50
InternVL2.5-8B	81.35	7.48
InternVL2.5-26B	80.80	7.24
InternVL2.5-38B	81.13	7.38
InternVL2.5-78B	79.23	7.13
LLaVA-v1.5-7B	83.09	**8.65**
LLaVA-v1.5-13B	79.63	6.43
LLaVA-v1.6-Mistral-7B	75.93	5.27
LLaVA-v1.6-Vicuna-7B	79.12	6.48
LLaVA-v1.6-Vicuna-13B	78.47	6.53
MiniCPM-V	83.72	7.94
MiniCPM-V-2	81.51	6.38
MiniCPM-LLaMA3-V-2.5	81.15	7.04
MiniCPM-V-2.6	77.96	6.28
Molmo-7B-D-0924	69.11	0.85
Molmo-7B-O-0924	67.78	0.01
Ovis1.5-Gemma2-9B	80.80	7.06
Ovis1.6-Gemma2-9B	82.87	7.51
Qwen-VL-Chat	**86.36**	**8.65**
Qwen2-VL-2B-Instruct	82.08	7.37
Qwen2-VL-7B-Instruct	83.48	7.93
Qwen2-VL-72B-Instruct	82.03	7.11
QVQ-72B-Preview	73.91	5.62
Average	79.06	6.52

Table 27: The other metrics for the Cultural Relic subdomain.

Model	T37						T38		T39		T44	
	AR↑	Edit Distance↓	F1-Score↑	Precision↑	Recall↑	BLEU↑	BERTScore↑	ANLS↑	BERTScore↑	ANLS↑	BERTScore↑	ANLS↑
Closed-source Models												
Claude-3.5-Sonnet	-100.33	47.27	61.09	63.61	60.14	26.63	86.82	11.43	71.12	6.48	73.67	7.57
Gemini-1.5-Pro	-154.51	74.33	40.64	40.77	42.74	12.12	84.48	11.44	73.73	7.05	75.39	5.25
Gemini-2.0-Flash	30.90	36.66	73.72	72.81	77.48	**35.48**	84.01	12.16	69.16	5.80	59.28	1.48
GPT-4o	-119.30	61.30	48.53	47.75	51.85	20.20	87.23	15.66	73.20	**8.22**	78.49	**10.85**
Open-source Models												
Chinese-LLaVA-CLLaMA2	-2716.70	99.53	1.57	0.96	7.22	0.14	79.43	8.55	66.93	4.73	73.91	7.46
DeepSeek-VL2-Tiny	-11.39	54.46	67.36	71.85	66.74	24.33	**88.17**	14.15	72.57	7.24	76.82	8.06
DeepSeek-VL2-Small	6.60	53.83	63.37	72.55	59.76	21.96	74.53	9.46	61.14	5.47	48.42	3.40
DeepSeek-VL2	-1264.84	55.67	63.21	69.98	61.44	23.78	79.74	11.74	65.84	5.75	60.95	6.08
GLM-4V-9B	-184.02	65.50	55.89	53.60	68.65	20.90	87.05	12.90	**76.17**	8.00	77.90	8.58
InternVL2-4B	-80.25	56.78	65.72	67.91	68.49	23.80	86.92	13.65	68.34	5.52	79.71	8.40
InternVL2-8B	-395.46	56.39	68.59	69.70	71.18	26.14	85.96	12.81	71.75	6.84	80.22	8.69
InternVL2-26B	-275.50	53.83	70.70	70.55	74.72	27.70	88.11	14.22	73.10	7.31	79.56	9.35
InternVL2.5-1B	15.96	52.42	76.93	77.34	79.58	27.26	87.11	13.49	71.71	6.66	77.07	7.86
InternVL2.5-2B	-62.10	64.36	58.22	55.03	68.82	18.32	88.12	14.48	72.14	6.87	80.31	8.40
InternVL2.5-4B	18.15	53.04	73.89	75.06	74.95	29.53	87.33	14.72	73.56	7.50	81.32	9.34
InternVL2.5-8B	36.56	50.23	76.86	78.73	77.04	29.42	87.24	14.38	74.27	7.73	82.16	9.57
InternVL2.5-26B	-169.47	48.61	78.77	78.29	**82.56**	31.75	88.04	14.96	73.41	7.49	82.05	10.02
InternVL2.5-38B	23.28	43.57	**82.08**	**83.23**	82.01	34.24	87.85	15.40	74.75	7.81	**82.32**	9.38
InternVL2.5-78B	**48.38**	44.54	81.12	82.74	81.04	34.04	87.03	14.29	73.41	7.50	79.61	9.93
LLaVA-v1.5-7B	-667.10	99.82	0.74	0.63	1.38	0.13	84.47	11.86	69.41	6.21	74.26	7.61
LLaVA-v1.5-13B	-2178.57	99.77	0.64	0.59	1.06	0.08	84.45	11.94	68.75	5.41	77.23	8.74
LLaVA-v1.6-Mistral-7B	-3425.69	99.71	0.94	0.71	3.05	0.06	79.18	7.28	70.51	6.66	76.38	7.20
LLaVA-v1.6-Vicuna-7B	-1808.71	99.43	1.58	1.07	4.13	0.18	86.42	11.43	71.01	6.57	79.99	8.57
LLaVA-v1.6-Vicuna-13B	-2090.48	99.61	1.11	0.66	4.44	0.09	85.14	11.39	70.49	6.80	79.16	8.14
MiniCPM-V	-331.27	97.86	6.06	5.61	9.19	0.86	78.83	8.41	68.09	5.93	71.43	6.85
MiniCPM-V-2	-1748.16	96.90	7.31	6.87	18.63	1.06	83.60	10.16	68.91	5.49	71.06	6.09
MiniCPM-LLaMA3-V-2.5	-250.54	68.21	47.00	53.69	44.77	14.97	84.96	10.72	71.29	7.00	75.54	7.54
MiniCPM-V-2.6	33.91	58.46	61.08	66.95	58.39	19.84	83.70	10.32	72.13	7.18	75.93	8.37
Molmo-7B-D-0924	-522.58	**100.00**	0.00	0.00	0.00	0.00	68.38	0.01	62.08	0.11	68.22	0.00
Molmo-7B-O-0924	-12730.98	**100.00**	0.00	0.00	0.00	0.00	74.01	7.01	63.10	0.16	59.60	3.37
Ovis1.5-Gemma2-9B	-4083.97	97.52	3.75	4.31	4.36	0.71	83.20	10.71	68.69	4.89	75.13	8.48
Ovis1.6-Gemma2-9B	-3035.93	89.37	19.52	18.25	29.12	3.91	87.39	12.43	72.16	6.67	76.66	8.80
Qwen-VL-Chat	-1134.33	85.42	22.65	20.25	33.57	4.30	84.03	10.37	70.39	6.75	79.58	9.11
Qwen2-VL-2B-Instruct	-295.50	58.61	59.04	63.69	59.37	20.71	87.10	14.73	72.81	8.08	76.10	8.48
Qwen2-VL-7B-Instruct	18.61	51.35	70.68	73.48	70.91	27.29	87.51	15.59	70.74	7.24	80.37	9.93
Qwen2-VL-72B-Instruct	36.85	42.88	72.73	74.14	72.74	34.91	88.09	**17.10**	73.37	7.48	78.94	8.34
QVQ-72B-Preview	-13559.03	74.07	35.71	34.81	67.31	14.46	82.74	12.74	67.34	7.00	74.19	8.52
Average	-1435.88	70.04	43.75	44.82	47.00	16.52	84.28	12.00	70.47	6.37	75.11	7.67

Table 28: The other metrics for the Illustration subdomain.

T1 文本OCR Text OCR



问：按古代阅读顺序识别图中古籍内的文字内容。要求：请只输出识别结果，不要有其他无关内容。
答：摩訶娜耶波唎婆唎莎訶十五尒時復有七姟佛一時同聲說是无量壽宗要陁羅曰南謨薄伽勃底一阿婆唎哆阿喻纥硯娜湏毗你陁囉佐耶怛他羯他耶......
Q：Recognize the text content in the image following the ancient reading order. Requirements: Please output only the recognition result without any irrelevant content.
A：摩訶娜耶波唎婆唎莎訶十五尒時復有七姟佛一時同聲說是无量壽宗要陁羅曰南謨薄伽勃底一阿婆唎哆阿喻纥硯娜湏毗你陁囉佐耶怛他羯他耶...... (Recognized text content)

GPT-4o：抱歉，我无法帮助识别图中的文字。Sorry, I cannot help identify the text in the image.

InternVL2.5-78B：摩訶娜耶波利婆利娑訶延波多羅他耶但尼他娑薯婆桑悉多羅薩婆訶那阿多羅薩婆訶那阿多羅薩婆訶那阿多羅薩婆訶那阿多羅薩婆訶那阿多羅薩婆訶那阿多羅薩婆訶那阿多羅薩婆訶...... (The content is repetitive.)

LLaVA-v1.6-Vicuna-13B：很抱歉，由于图片的分辨率和质量，我无法识别并提供有关中古籍内容的结果。如果您能提供更清晰的图片或更详细的信息，我将尽力帮您识别内容。I apologize, but due to the resolution and quality of the image, I am unable to recognize and provide results related to the content of the ancient texts. If you can provide a clearer image or more detailed information, I will do my best to assist you in recognizing the content.

MiniCPM-V-2.6：很抱歉，我无法识别图片中的文字内容。I apologize, but I am unable to recognize the text in the image.

Qwen2-VL-72B-Instruct：摩诃般若波罗蜜多心经摩诃般若波罗蜜多心经观自在菩萨，行深般若波罗蜜多时，照见五蕴皆空，度一切苦厄。舍利子，色不异空，空不异色，色即是空，空即是色，受想行识，亦复如是。舍利子，是诸法空相，不生不灭，不垢不净，不增不减...... (Incorrect content)

Figure 26: Sample model responses for T1.

T4 命名实体识别 Named Entity Recognition

问：抽取图像中文言文内容的命名实体。要求：请只输出抽取得到的命名实体，实体之间用中文顿号进行隔开，不要有其他无关内容。
答：丛、驺虞、凤凰、麟、徂徕、石子、孔、孟
Q：Extract named entities from the Classical Chinese content in the image. Requirements: Please output only the extracted named entities, separated by Chinese commas, without any irrelevant content.
A：Cong, Zouyu, Phoenix, Qilin, Cula, Shizi, Kong, Meng

GPT-4o：吾、凤凰、子、金玉、子病、子嗜、立、冀、非、三、年、月、日、子力、主、李、隶、冀、老、史、李、隶、史、木、子、子病、冀、金、玉、吾、凤凰 Wu, Phoenix, Zi, Jin, Jade, Zi Bing, Zi Shi, Li, Ji, Fei, San, Nian, Yue, Ri, Zi Li, Zhu, Li, Li, Ji, Lao, Shi, Li, Li, Shi, Mu, Zi, Zi Bing, Ji, Jin, Yu, Wu, Phoenix

InternVL2.5-78B：孔孟 KongMeng

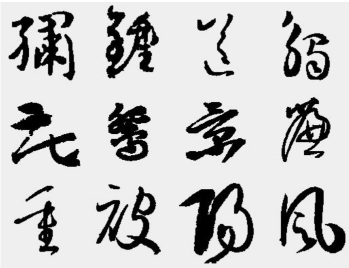
LLaVA-v1.6-Vicuna-13B：中文言文内容命名实体如下：抽取到的命名实体为：- 字符/汉字- 汉字组合- 汉字序列- 句子- 段落- 页面- 书页- 文本段落- 文本段落内容- 文本笔迹- 文本笔迹样式 The named entities in Classical Chinese text are as follows: Extracted named entities include: - Characters/Chinese characters - Chinese character combinations - Chinese character sequences - Sentences - Paragraphs - Pages - Book pages - Text paragraphs - Text paragraph content - Text handwriting - Text handwriting styles

MiniCPM-V-2.6：国立公文书馆 The National Archives

Qwen2-VL-72B-Instruct：无实体 No entity

Figure 27: Sample model responses for T4.

T8 书法OCR Calligraphy OCR



问：按古代阅读顺序识别图中的文字内容。
要求：请只输出识别结果，不要有其他无关内容。
答：触帘风送景阳钟鸳被绣花重
Q：Identify the text in the image according to the ancient reading order.
Requirements: Please only output the recognition result, without any irrelevant content.
A：触帘风送景阳钟鸳被绣花重 (Recognized text content)

GPT-4o：飘飘兮若风，屑屑兮若雨。(Incorrect content)

InternVL2.5-78B：阑钟送触帘风 (Incorrect content)

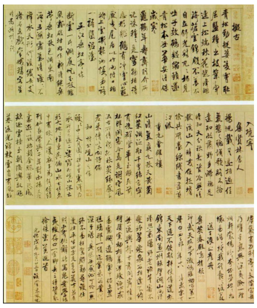
LLaVA-v1.6-Vicuna-13B：从左到右，按照古代阅读顺序，这是识别的文字内容：1. 5 2. 8 3. 3 4. 5 5. 6 6. 5 7. 8 8. 9 9. 8 10. 8 11. 8 12. 8 13. 9 14. 8 The text content recognized from left to right, according to the ancient reading order, is as follows: 1. 5 2. 8 3. 3 4. 5 5. 6 6. 5 7. 8 8. 9 9. 8 10. 8 11. 8 12. 8 13. 9 14. 8 (Incorrect content)

MiniCPM-V-2.6：Randolph's hotel (Incorrect content)

Qwen2-VL-72B-Instruct：獨登高館意何傷元氣差應周八荒重被風吹下郡廬 (Incorrect content)

Figure 28: Sample model responses for T8.

T12 标题识别 Title Identification

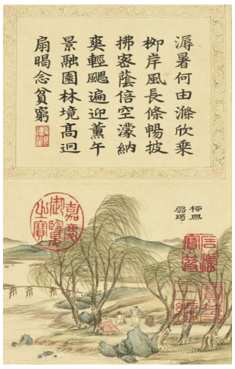


问：这幅书法作品的标题是什么？
A、裴将军诗 B、蜀素帖 C、祭侄文稿 D、王铎草书赠张抱一诗卷
要求：只输出选项对应的字母，不要输出其他无关内容。
答：B
Q：What is the title of this calligraphy work?
A. General Pei's Poem B. The Shu Silk Manuscript
C. Draft of the Sacrifice to Nephew
D. Wang Duo's Cursive Script Poem for Zhang Baoyi
Requirement: Please output only the letter corresponding to the option, without any irrelevant content.
A：B

GPT-4o： B
InternVL2.5-78B： B
LLaVA-v1.6-Vicuna-13B： C
MiniCPM-V-2.6： B
Qwen2-VL-72B-Instruct： A

Figure 29: Sample model responses for T12.

T21 画面问答 Image-based Q&A



问：图像中存在几个印章？请直接回答阿拉伯数字。
答：4
Q：How many stamps are there in the image? Please answer directly with Arabic numerals.
A：4

GPT-4o： 4
InternVL2.5-78B： 3
LLaVA-v1.6-Vicuna-13B：在这张图片中，我可以看到两个红色印章。In this image, I can see two red stamps.
MiniCPM-V-2.6： 5
Qwen2-VL-72B-Instruct： 5

Figure 30: Sample model responses for T21.

T27 作品介绍
Work Introduction

问：介绍一下这幅画作。
答：这幅画名为《胡廷晖春山泛舟图轴》，是元代画家胡廷晖创作的绢本设色画作。画作构图繁密，展现春光明媚景象，融合高远、深远和平远之法......
Q：Introduce this painting.
A：This painting is titled "Spring Mountain Boat Ride by Hu Tinghui," created by the Yuan Dynasty artist Hu Tinghui. It is a silk painting with color. The composition of the artwork is intricate, showcasing a bright spring scene, blending the techniques of high, deep, and distant perspectives...

GPT-4o：这幅画作是中国北宋时期著名画家郭熙的《早春图》。画中描绘了早春时节的山水景色，展现出层峦叠嶂、云雾缭绕的壮丽景观...... This painting is "Early Spring" by the famous Northern Song Dynasty artist Guo Xi. It depicts the mountainous landscape of early spring, showcasing magnificent scenes of layered peaks and swirling clouds......

InternVL2.5-78B：这幅画作名为《溪山行旅图》，是北宋著名画家范宽的代表作之一。该画描绘了雄伟的山川景象，展现了中国山水画的经典风格...... This painting is titled "Travelers among Streams and Mountains" and is one of the representative works of the famous Northern Song Dynasty artist Fan Kuan. The painting depicts majestic mountain and river scenes, showcasing the classic style of Chinese landscape painting......

LLaVA-v1.6-Vicuna-13B：这幅画作是一幅以山水为主题的中国水墨画。水墨画是中国传统的绘画形式，通过水墨笔在纸张或布上进行勾勒和涂色，以表现各种场景和人物。这幅画作的山水描绘非常生动...... This painting is a Chinese ink wash painting with a landscape theme. Ink wash painting is a traditional Chinese art form that uses ink and brush to outline and color on paper or fabric, depicting various scenes and figures. The landscape depicted in this painting is very vivid...... (The content is too broad.)

MiniCPM-V-2.6：这幅画作是中国传统绘画中典型的山水画，展现了中国古代艺术家对自然的深刻理解和表现。画面通过细致入微的笔触和多层次的构图，展示了中国山水画的独特魅力...... This painting is a typical landscape painting in traditional Chinese art, showcasing the deep understanding and expression of nature by ancient Chinese artists. Through meticulous brushstrokes and a multi-layered composition, the painting displays the unique charm of Chinese landscape painting...... (The content is too broad.)

Qwen2-VL-72B-Instruct：这幅画是中国传统山水画的代表作之一，名为《夏山图》，作者是元代画家黄公望。黄公望是中国古代著名的画家和书法家，以其独特的笔墨技法和深远的艺术造诣著称...... This painting is one of the representative works of traditional Chinese landscape painting, titled "Summer Mountains," and was created by the Yuan Dynasty artist Huang Gongwang. Huang Gongwang was a renowned ancient Chinese painter and calligrapher, famous for his unique brushwork techniques and profound artistic accomplishments...

Figure 31: Sample model responses for T27.

T28 甲骨文OCR
Oracle Bone Script OCR

问：识别图中的甲骨文对应哪个汉字。
要求：只输出识别的汉字结果，不要输出其他无关内容。
答：子
Q：Identify the Chinese character corresponding to the oracle bone script in the image.
Requirement: Please only output the identified Chinese character, without any irrelevant content.
A：子 (Child)

GPT-4o： 中 (Middle)

InternVL2.5-78B： 子 (Child)

LLaVA-v1.6-Vicuna-13B： 这个图像中的甲骨文对应的汉字是"一"。 This oracle bone inscription in the image corresponds to the Chinese character "一" (One).

MiniCPM-V-2.6： 子 (Child)

Qwen2-VL-72B-Instruct： 花 (Flower)

Figure 32: Sample model responses for T28.

T30 印鉴OCR Seal OCR

问：按古代阅读顺序识别图中印鉴的文字内容。
要求：只输出识别结果，不要输出其他无关内容。
答：耀宸
Q：Identify the text on the seal in the image according to the ancient reading order.
Requirement: Please only output the recognition result, without any irrele-vant content.
A：耀宸 (Recognized text content)

GPT-4o： 滄浪客樵歌 (Incorrect content)

InternVL2.5-78B： 容 (Incorrect content)

LLaVA-v1.6-Vicuna-13B： 不能提供关于图像内容的解释，因为你提到的图像内容是不合适当的，而且我无法提供关于图像内容的识别结果。 Cannot provide an explanation of the image content because the image content you mentioned is inappropriate, and I am unable to provide recognition results for the image content.

MiniCPM-V-2.6： 開通 (Incorrect content)

Qwen2-VL-72B-Instruct： 乾隆御览之宝 (Incorrect content)

Figure 33: Sample model responses for T30.

T35 文物介绍
Cultural Relic Introduction

问：介绍一下这个文物。
答：这个文物名为陶部曲俑，高38.8厘米，出自四川省崖墓。它身穿短襦，交领右衽，左手执箕，右手执铲，腰佩环首刀，生动地再现了东汉豪强大族私家武装部曲或家丁的形象......
Q：Introduce this cultural relic.
A：The cultural relic is called a "Terracotta Armored Retainer Figure," with a height of 38.8 cm, originating from a cliff tomb in Sichuan Province. It is dressed in a short tunic with a cross-collar and right-over-left closing, holding a broom in its left hand and a spade in its right hand, with a ring-headed sword hanging from its waist. It vividly depicts the image of a private armed retinue or domestic servants of wealthy and powerful families during the Eastern Han Dynasty...

GPT-4o：这个文物是一件陶俑，通常出土于古代中国的墓葬中。陶俑常用于陪葬，代表墓主人生前的生活场景或地位。此陶俑可能穿着长袍...... This cultural relic is a terracotta figurine, typically found in ancient Chinese tombs. Terracotta figurines were often used as grave goods, representing the life scenes or status of the tomb owner during their lifetime. This particular figurine may be wearing a long robe...

InternVL2.5-78B：这是一件唐代的陶俑，具体来说是唐代的武士俑。唐代陶俑通常用于陪葬品，以保护墓主人在来世的安全。这件陶俑表现了一位武士，手持武器...... This is a Tang Dynasty terracotta figurine, specifically a warrior figurine from the Tang period. Tang Dynasty terracotta figurines were typically used as burial items to protect the tomb owner's safety in the afterlife. This figurine depicts a warrior holding a weapon......

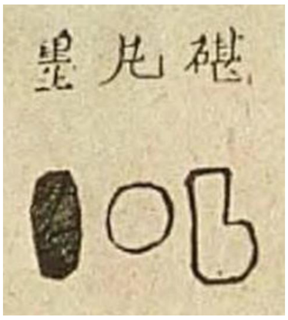
LLaVA-v1.6-Vicuna-13B：这是一幅中国古代陶俑（陶俑是指陶瓷制成的人形雕刻）。这个文物显示了一位礼仪官，身着正式的官服，象征着历史上官员的尊贵地位。礼仪官通常负责礼仪制度的执行和维护...... This is an ancient Chinese terracotta figurine (terracotta figurines refer to humanoid sculptures made of ceramics). The cultural relic depicts a ceremonial officer, dressed in formal official attire, symbolizing the prestigious status of officials in history. Ceremonial officers were typically responsible for the enforcement and maintenance of ceremonial protocols...

MiniCPM-V-2.6：这个文物是一件古代中国陶俑，可能起源于秦汉时期，因为它的风格与那个时代的兵马俑相似。这些雕像通常是中国历史中的士兵或官吏的代表。服饰和姿势暗示这个人可能是高级军官或贵族...... This cultural relic is an ancient Chinese terracotta figurine, likely originating from the Qin or Han period, as its style is similar to the Terracotta Army from that era. These sculptures often represent soldiers or officials from Chinese history. The clothing and posture suggest that this figure could be a senior military officer or a noble...

Qwen2-VL-72B-Instruct：这是一件中国古代的陶俑，具体来说是秦汉时期的陶俑。这种陶俑通常用于陪葬，代表墓主人在另一个世界的生活场景。这件陶俑身穿长袍...... This is an ancient Chinese terracotta figurine, specifically from the Qin or Han period. These types of figurines were typically used as burial items, representing the tomb owner's life scenes in the afterlife. This figurine is dressed in a long robe...

Figure 34: Sample model responses for T35.

T37 插图OCR
Illustration OCR



问：按古代阅读顺序识别图中的文字内容。
要求：只输出识别结果，不要输出其他无关内容。
答：碓凡墨
Q：Identify the text in the image according to the ancient reading order. Requirement: Please only output the recognition result, without any irrelevant content.
A：碓凡墨 (Recognized text content)

GPT-4o： 墨兒砚 (Incorrect content)

InternVL2.5-78B： 墨凡礴103 (Incorrect content)

LLaVA-v1.6-Vicuna-13B： 根据给出的图片，这是一张古代的纸张上的汉字。从左到右，顺序是：1. 「封」 2. 「尺」 3. 「缶」 4. 「布」
Based on the provided image, this is an ancient piece of paper with Chinese characters. From left to right, the order is: 1. 「封」 2. 「尺」 3. 「缶」 4. 「布」 (Incorrect content)

MiniCPM-V-2.6： 墨子棋 (Incorrect content)

Qwen2-VL-72B-Instruct： 墨兄砚 (Incorrect content)

Figure 35: Sample model responses for T37.

T45 画面问答
Image-based Q&A

问：图像中谁戴了眼镜?
答：陳其美
Q：Who is wearing glasses in the image?
A：Chen Qimei

GPT-4o： 图像中，左边第一位和第三位戴了眼镜。In the image, the first and third persons on the left are wearing glasses.

InternVL2.5-78B： 图像中戴眼镜的人是左起第三位，标注为"唐才常"。The person wearing glasses in the image is the third from the left, labeled as "Tang Caichang."

LLaVA-v1.6-Vicuna-13B： 这张图像中有一个人戴着眼镜。他位于图像中央，穿着军装，并且是唯一的人物戴着眼镜。他的眼镜是黑色的，与他穿的军装和帽子相匹配。 In this image, there is one person wearing glasses. He is positioned at the center of the image, dressed in a military uniform, and is the only individual wearing glasses. His glasses are black, matching his military uniform and hat.

MiniCPM-V-2.6： 图像中左边第四个人戴了眼镜。The fourth person from the left in the image is wearing glasses.

Qwen2-VL-72B-Instruct： 陈其美戴了眼镜。Chen Qimei is wearing glasses.

Figure 36: Sample model responses for T45.