

TCC-Bench: Benchmarking the Traditional Chinese Culture Understanding Capabilities of MLLMs

Pengju Xu¹, Yan Wang¹, Shuyuan Zhang¹, Xuan Zhou¹,
Li Xin¹, Yue Yuan¹, Fengzhao Li¹, Shunyuan Zhou¹,
Xingyu Wang¹, Yi Zhang¹, Haiying Zhao^{1†},

¹School of Artificial Intelligence, Beijing University of Posts and Telecommunications

Correspondence: zhaohaiying@bupt.edu.cn

Abstract

Recent progress in Multimodal Large Language Models (MLLMs) has significantly enhanced the ability of artificial intelligence systems to understand and generate multimodal content. However, the capabilities of existing MLLMs in understanding traditional Chinese culture have not yet been comprehensively explored. To address this gap, we propose the **Traditional Chinese Culture understanding Benchmark (TCC-Bench)**, a bilingual Visual Question Answering (VQA) benchmark designed to assess MLLMs' performance on culturally grounded visual content. TCC-Bench features culturally rich and visually diverse data, covering museum artifacts, daily life scenes, comics, and other culturally significant contexts across eight cultural domains. Through extensive experiments on a wide range of MLLMs, we find that current models still face substantial challenges in handling traditional Chinese cultural content. Our results highlight the need for more culturally inclusive and context-aware multimodal systems. The code and data can be found at: <https://github.com/Morty-Xu/TCC-Bench>.

1 Introduction

In recent years, the rapid advancement of Multimodal Large Language Models (MLLMs) (Liu et al., 2024b; Dai et al., 2024; Wang et al., 2024; OpenAI, 2025; Comanici et al., 2025) has significantly enhanced the capabilities of artificial intelligence (AI) systems in understanding and generating complex information across various modalities, including text, image, and audio. However, a growing body of evidence suggests that most MLLMs, trained primarily on Western data, perform poorly when tasked with handling non-Western languages and cultural knowledge (Liu et al., 2021; Pouget et al., 2024; Romero et al., 2024). This limitation

severely restricts the cultural and regional applicability of MLLMs.

Traditional Chinese culture, with a history spanning several millennia, encompasses a vast array of intricate symbols, customs, and philosophical concepts that are often challenging to interpret even for native speakers. Although these cultural elements may appear remote, they continue to exert a profound influence on the contemporary Chinese community. A thorough understanding of these cultural foundations is indispensable for attaining a comprehensive and nuanced understanding of China. However, despite the emergence of several benchmarks (Nayak et al., 2024; Zhang et al., 2024b,a) that touch upon aspects of traditional Chinese culture, there remains a lack of a comprehensive dataset explicitly designed to evaluate the understanding capabilities of MLLMs in this cultural domain.

To address this gap, we introduce the **Traditional Chinese Culture understanding Benchmark (TCC-Bench)**, a bilingual (*i.e.*, Chinese and English) Visual Question Answering (VQA) benchmark specifically developed to assess the capacity of MLLMs to understand traditional Chinese culture. In contrast to prior efforts, TCC-Bench aims to systematically represent traditional Chinese cultural knowledge across diverse thematic dimensions. To this end, TCC-Bench defines eight knowledge domains that collectively capture core facets of traditional Chinese culture. Furthermore, the images in TCC-Bench are carefully curated from museum artifacts, representations of everyday life, comics, and other culturally significant sources, thereby ensuring both visual diversity and cultural authenticity.

Recognizing the significant effort typically required for high-quality benchmark creation, we introduce a semi-automated question generation method. Specifically, we employ GPT-4o (OpenAI, 2024) in text-only mode to generate candidate questions based on carefully designed prompts, af-

[†]Corresponding author

ter which human annotators curate and select high-quality samples from the generated outputs. Importantly, we deliberately disable the multimodal capabilities of GPT-4o during this process, as allowing the model to access images could result in questions implicitly aligned with its own visual strengths, thereby introducing potential data leakage and artificially inflating evaluation performance. This design substantially reduces manual annotation costs while preserving the integrity of the benchmark. Moreover, to alleviate the well-known issue of language bias (Goyal et al., 2017) in VQA datasets, we ensure that the cultural concepts depicted in the images are not explicitly revealed in the corresponding questions.

We evaluate TCC-Bench across a diverse set of MLLMs spanning a wide range of parameter scales, and the results reveal that, despite their strong general-purpose capabilities, existing models continue to struggle with tasks that require the understanding of culturally grounded information. These findings underscore the need for further research into AI systems with enhanced cultural awareness and robustness.

Our contributions can be summarized as follows:

- We present TCC-Bench, the VQA benchmark specifically focused on evaluating MLLMs’ understanding capabilities of traditional Chinese culture.
- We evaluate a wide range of open-source and closed-source MLLMs on TCC-Bench, revealing considerable potential for further improvement in understanding traditional Chinese culture.
- We conduct detailed analyses to examine model performance across languages, domains, prompting strategies, and error types, providing insights for future research on culturally grounded VQA.

2 Related Work

Multimodal Large Language Models (MLLMs). The recent advancements in Large Language Models (LLMs) (Chiang et al., 2023; Touvron et al., 2023; Cai et al., 2024) have catalyzed a surge in exploration of MLLMs (Liu et al., 2024b; Dai et al., 2024; Chen et al., 2023; Li et al., 2023; Zhu et al., 2023; GLM et al., 2024; Wang et al.,

2024). MLLMs are designed to integrate and process information from multiple modalities, primarily vision and text, to address complex real-world tasks. Foundational models like CLIP (Radford et al., 2021) established powerful zero-shot capabilities by applying a contrastive learning paradigm to massive-scale image-text pairs. Subsequent architectures, such as BLIP-2 (Li et al., 2023), introduced novel components like the Q-Former to better align visual and text representations. Furthermore, a significant body of work (Liu et al., 2024b,a; Dai et al., 2024) has leveraged instruction tuning to enhance the comprehension and instruction-following abilities of MLLMs. While state-of-the-art closed-source models, including Gemini-2.5 (Comanici et al., 2025) and GPT-4.1 (OpenAI, 2025), currently demonstrate leading performance across general benchmarks, a critical gap remains. However, several studies (Pouget et al., 2024; Romero et al., 2024; Nayak et al., 2024) have shown that MLLMs still exhibit deficiencies in cultural understanding.

Benchmarks for cultural understanding. As the capabilities of MLLMs rapidly advance, evaluating their understanding of diverse human cultures has become increasingly critical. Several benchmarks, including MaRVL (Liu et al., 2021), GD-VCR (Yin et al., 2021), CVQA (Romero et al., 2024), CultureVQA (Nayak et al., 2024), and ALM-Bench (Vayani et al., 2025), have been proposed to assess multicultural understanding. However, although these benchmarks incorporate certain Chinese cultural elements, their coverage remains limited, preventing a comprehensive evaluation of MLLMs’ proficiency in traditional Chinese culture. Concurrently, a number of datasets have been developed to target specific cultural contexts. For example, SEA-VQA (Uraierprasert et al., 2024) introduces a VQA dataset centered on Southeast Asian culture, while K-Viscuit (Park et al., 2025) provides a benchmark focused on Korean culture. Similarly, JMMM (Onohara et al., 2025), Peacock (Alwajih et al., 2024), and MerCulture (Tushar et al., 2025) investigate cultural characteristics unique to their respective regions. With respect to traditional Chinese culture, the Pun Rebus Art Dataset (Zhang et al., 2024b) narrowly examines MLLMs’ understanding of Chinese pun rebuses in traditional artworks. CII-Bench (Zhang et al., 2024a), a benchmark closely related to TCC-Bench, evaluates deep-level perception and reasoning but remains limited in scope, focusing primarily

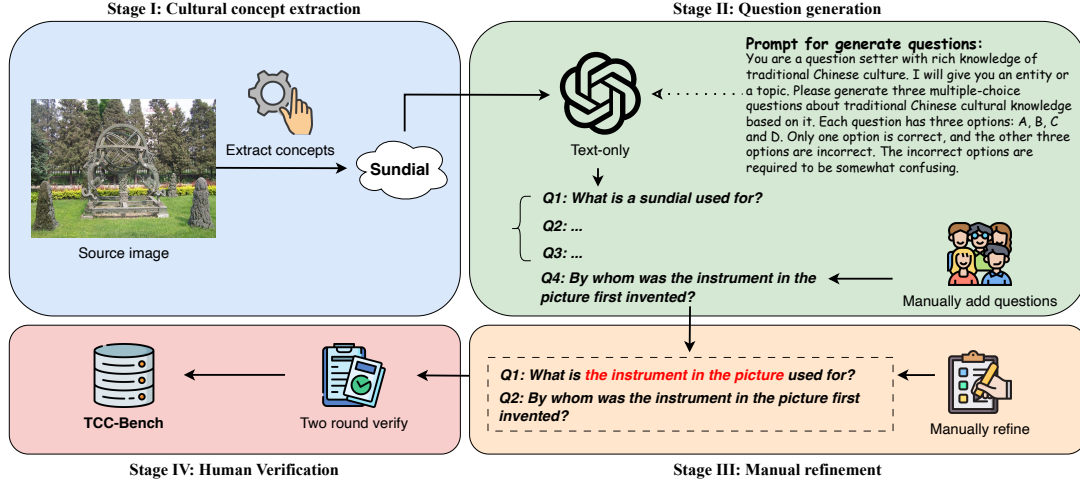


Figure 1: Question generation pipeline. The detailed prompt is in Appendix D.

Table 1: Comparison of TCC-Bench with existing cultural VQA benchmarks. In the Answer column, T/F denotes true/false questions, MC denotes multiple-choice questions, and Open denotes open-ended answers.

Dataset	#Samples	#Images	Culture	Multilingual	Answer
MaRVL (Liu et al., 2021)	5,670	4,914	Multiple	Yes	T/F
CVQA (Romero et al., 2024)	9,044	4,560	Multiple	Yes	MC
CULTURALVQA (Nayak et al., 2024)	2,378	2,328	Multiple	No	Open
Peacock (Alwajih et al., 2024)	1,132	120	Multiple	No	Open
K-Viscuit (Park et al., 2025)	657	237	Korean	No	MC
SEA-VQA (Urailetprasert et al., 2024)	1,999	515	SE Asia	No	MC
CII-Bench (Zhang et al., 2024a)	800	698	Chinese	No	MC
TCC-Bench(Ours)	1,015	770	Chinese	Yes	MC

on traditional Chinese paintings. In contrast, TCC-Bench covers a broader spectrum of traditional Chinese culture than prior benchmarks, spanning eight distinct domains (see Section 3) and enabling a more diverse and explanation-oriented evaluation of cultural understanding. Table 1 presents a comparison between TCC-Bench and existing benchmarks.

3 TCC-Bench

We construct a multiple-choice VQA benchmark to evaluate the capability of current MLLMs in understanding traditional Chinese culture. To comprehensively cover different aspects of cultural knowledge, we divide the questions into eight domains: *Astronomy, Music, Custom, Architecture, Transportation, Diet, Clothing, and Artifact*. Further details regarding the domains can be found in Appendix A. To ensure the benchmark’s quality and cultural authenticity, six annotators with bachelor’s degrees or higher are recruited for image collection and question generation. All of them have long-term experience living within a Chinese cultural context, ensuring the high quality of the benchmark.

We further design a semi-automated question generation pipeline to reduce manual labor. To facilitate cross-lingual cultural understanding analysis, our dataset provides bilingual content for each question, including questions, answer options, and detailed explanations in both Chinese and English.

3.1 Image Collection

We manually collected images from publicly available sources on the Internet. The image collection process involves careful selection to ensure comprehensive coverage of various aspects of traditional Chinese culture. All selected images are required to unambiguously depict cultural elements rooted in traditional Chinese contexts. To construct a broad and representative dataset, we incorporated images from a wide range of scenarios, including museum artifacts, daily life scenes, traditional performances, comics, and other culturally significant contexts. In total, we initially gathered 1,830 raw images. After applying a series of filtering and quality control procedures, we obtained a curated subset of 770 high-quality images for inclusion in the final benchmark. Detailed information about image sources and filtering criteria is provided in Appendix B. It is worth noting that all selected images adhere to open-use licensing agreements.

3.2 Question Generation Pipeline

We introduce a semi-automated pipeline that leverages the powerful generation capabilities of GPT-4o to balance dataset quality and manual effort. Figure 1 illustrates the entire workflow, which consists of four key stages: cultural concept extraction, question generation, manual refinement, and hu-

man verification. For more detailed information, we recommend referring to Appendix C.

Cultural concept extraction. Annotators analyze the content of the given image and extract key cultural concepts based on their knowledge and contextual understanding. These concepts serve as the foundation for generating high-quality questions that align with the image’s cultural significance.

Question generation. The extracted cultural concepts are provided as input to text-only GPT-4o, which produces a diverse set of at least three candidate questions, each accompanied by answer options and an explanation. Each set includes one correct answer, while distractor options introduce necessary ambiguity. We deliberately avoid using multimodal GPT-4o, as questions generated by such models might be more easily answered by themselves, raising concerns about potential data leakage and inflated performance in model evaluation. Furthermore, relying on a text-only generation process ensures a more controlled pipeline, contributing to the production of higher-quality question-answer pairs. To further enrich the dataset, annotators are encouraged to manually design additional questions that capture nuanced or culturally complex aspects often overlooked by automated generation. This approach allows the benchmark to include both broad-coverage and fine-grained cultural inquiries. We ensure that the generated questions are inherently open-ended for the benchmark’s applicability, similar to CVQA (Romero et al., 2024).

Manual refinement. Given that GPT-generated content may contain inaccuracies, overly generic phrasing, or culturally inappropriate details, all questions undergo a rigorous manual refinement stage. Annotators examine whether each question is meaningful, whether the correct answer is indeed correct, and whether the explanation is concise, accurate, and aligned with the visual content. Any unsuitable or low-quality outputs are revised accordingly. A key aspect of this refinement process is ensuring that questions do not directly disclose the cultural concept extracted from the image, thereby preventing MLLMs from answering purely through language priors rather than visual understanding (Goyal et al., 2017). For example, when the extracted concept is “sundial”, text-only GPT-4o generates a question such as “What is a sundial used for?” Annotators would rewrite this question into a visually grounded form, such as “What is the instrument in the picture used for?”, which requires

Table 2: Dataset numerical statistics. "ZH" represents Chinese text and "EN" represents English text.

Statistics	Values
Avg. size of images	0.49M
No. of images	770
No. of questions	1,015
Avg. question per image	1.32
Avg. words per question (ZH / EN)	14.8 / 11.9
Avg. words per option (ZH / EN)	3.3 / 2.4
Avg. words per explanation (ZH / EN)	19.3 / 16.2

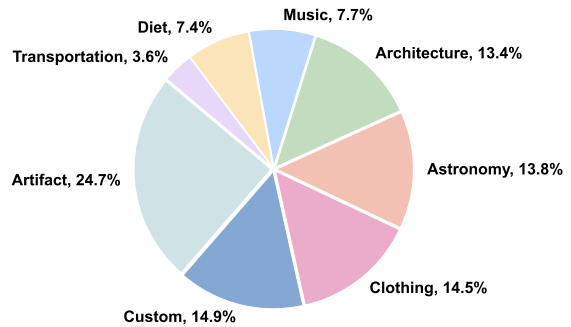


Figure 2: The statistics of domain distribution.

the model to analyze the image rather than relying on text cues (see Figure 1). The experiments in Section 4.2 demonstrate that our TCC-Bench mitigates the influence of language priors.

Human Verification. To guarantee the final quality of the benchmark, each question-answer pair undergoes two rounds of independent cross-checking by different annotators. Only entries that pass both rounds of verification are retained, while those that fail are removed.

3.3 Statistics

Table 2 presents the detailed statistics of our dataset. TCC-Bench consists of 770 images and 1,015 high-quality questions, each paired with four carefully designed options containing exactly one correct answer, resulting in 4,060 options in total. On average, each option contains 3.3 Chinese characters or 2.4 English words, while the explanations average 19.3 Chinese characters or 16.2 English words, providing concise yet informative cultural insight that supports the interpretation of each question. We further analyze the distribution of questions across the eight cultural domains, as shown in Figure 2. The domain proportions are relatively balanced, enabling a comprehensive and unbiased evaluation of MLLMs’ understanding of traditional Chinese culture. To illustrate the thematic focus within each domain, we additionally provide word clouds

Astronomy	Music	Custom	Architecture
Q: 图中的仪器最早是由谁发明的? (By whom was the instrument in the picture first invented?) (A) 张衡。(Zhang Heng.) (B) 祖冲之。(Zu Chongzhi.) (C) 郭守敬。(Guo Shoujing.) (D) 沈括。(Shen Kuo.) Explanation: 图中是浑天仪, 最早由张衡发明。 (The instrument in the picture is the Armillary Sphere. It was first invented by Zhang Heng.) 	Q: 图中的乐器是什么? (What is the musical instrument in the picture?) (A) 二胡。(Erhu.) (B) 唢呐。(Suona.) (C) 板胡。(Banhu.) (D) 笛子。(Flute.) Explanation: 图中的乐器是板胡。 (The musical instrument in the picture is the banhu.) 	Q: 图中是一项古代的体育运动, 请问它的名字是什么? (This is an ancient sport. What is its name?) (A) 游泳。(Swimming.) (B) 投壶。(Pitch-pot.) (C) 蹴鞠。(Cuju.) (D) 赛龙舟。(Dragon-boat racing.) Explanation: 图中在进行与球有关的古代体育活动——蹴鞠。 (In the picture, Cuju, an ancient sport related to the ball, is being played.) 	Q: 图片中地点是哪个文化遗址? (Which cultural heritage site is the location in the picture?) (A) 云冈石窟。(Yungang Grottoes.) (B) 莫高窟。(Mogao Caves.) (C) 龙门石窟。(Longmen Grottoes.) (D) 水上漂流。(Water rafting.) Explanation: 图中是位于中国甘肃敦煌的莫高窟。 (The picture is the Mogao Caves located in Dunhuang, Gansu Province, China.) 
Transportation	Diet	Clothing	Artifact
Q: 这辆马车在中国古代的名称是什么? (What was the name of this carriage in ancient China?) (A) 安车。(An Carriage.) (B) 轩车。(Xuan Carriage.) (C) 轊车。(Wenliang chariot.) (D) 輶车。(Zi Carriage.) Explanation: 图中的马车在中国古代被称为轊车。 (The carriage in the picture was called a "Wenliang chariot" in ancient China.) 	Q: 图中是哪样中国传统美食? (Which traditional Chinese delicacy is in the picture?) (A) 花卷。(Hua Juan.) (B) 馒头。(Mantou.) (C) 包子。(Baozi.) (D) 烧麦。(Shumai.) Explanation: 图中的食物是烧麦。 (The food in the picture is shumai.) 	Q: 图中衣物是采用什么制作? (What are the clothes in the picture made of?) (A) 用兽皮制作。(Made of animal leather.) (B) 用白鹿皮制作。(Made of white deer skin.) (C) 用牛皮制作。(Made of cowhide.) (D) 用虎皮制作。(Made of tiger skin.) Explanation: 皮弁是用白鹿皮做的。 (The leather is made of white deerskin.) 	Q: 图中的器物是什么? (What is the object in the picture?) (A) 铜盆。(Bronze basin.) (B) 盘子。(Plate.) (C) 铜镜。(Bronze mirror.) (D) 铜壶。(Bronze pot.) Explanation: 图中是一个铜镜。 (The picture shows a bronze mirror.) 

Figure 3: Samples of TCC-Bench. The **bolded and underlined** text represents the correct answer.

in Appendix E. Representative samples from the dataset are illustrated in Figure 3.

4 Experiment

4.1 Experimental Setup

To comprehensively evaluate MLLMs’ understanding of traditional Chinese culture, we select various MLLMs, including both open-source and closed-source models. For open-source MLLMs, we choose models with different parameter sizes. Specifically, the selected open-source models are: LLaVA-v1.6-7B/34B (Liu et al., 2024b), DeepSeek-VL2-3B/16B/27B (Wu et al., 2024), Qwen2.5-VL-7B/32B/72B (Bai et al., 2025), InternVL3.5-8B/38B (Wang et al., 2025), GLM-4V-9B (GLM et al., 2024), CogVLM2-19B (Wang et al., 2023). These models are deployed and evaluated on A800 GPUs. For closed-source MLLMs, we choose GPT-4.1 (OpenAI, 2025) and Gemini-2.5 Flash/Pro (Comanici et al., 2025), which are accessed through their official APIs. All models are evaluated using a standardized multiple-choice prompt format. Following MMMU (Yue et al., 2024), we adopt accuracy as the primary evaluation metric and employ a rule-based method for answer extraction. More detailed experimental setups can be found in Appendix F.

4.2 Main Result

Table 3 presents the evaluation results of various MLLMs on our TCC-Bench, including assessments for bilingual prompts as well as performance across specific domains. As a reference, we also report the prediction performance of GPT-4.1 under the text-only input setting. Additionally, to contextualize model performance, we conducted a human baseline study with three native Chinese speakers who have long-term exposure to traditional Chinese culture. Each evaluator was presented with 15 randomly sampled questions per domain (120 total). As shown in Table 3, human evaluators achieved an average accuracy of 85.00%, excelling in daily-life domains (Custom: 93.33%, Diet: 93.33%) but facing challenges in specialized fields (Astronomy: 73.33%). Based on these results, we summarize the key findings from our experiments as follows:

Overall results. The evaluation results on TCC-Bench reveal substantial variation in current MLLMs’ ability to understand traditional Chinese culture. Gemini-2.5 Pro achieves the best overall performance, reaching 87.00% accuracy with Chinese prompts and 83.55% with English prompts; notably, its Chinese-prompt performance surpasses the average human baseline of 85.00%. Among open-source models, the Qwen2.5-VL series demonstrates robust capabilities across differ-

Table 3: **VLM Evaluation Results.** "ZH" denotes Chinese prompts and "EN" represents English prompts. *Astr.*, *Arch.*, *Tran.*, *Clot.* and *Arti.* denote *Astronomy*, *Architecture*, *Transportation*, *Clothing* and *Artifact*. The highest and the second-highest scores among open-source and closed-source models in each column are highlighted in bold and underlined. We also include a human performance baseline conducted with native Chinese speakers.

Model	Overall		Astr.		Music		Custom		Arch.		Tran.		Diet		Clot.		Arti.	
	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN	ZH	EN
Open-source Model (Size < 10B)																		
DeepSeek-VL2-3B (Wu et al., 2024)	60.69	55.67	38.57	37.14	60.26	50.00	68.87	60.93	58.09	43.38	56.76	48.65	64.00	61.33	58.50	58.50	70.52	68.92
LLaVA-1.6-7B (Liu et al., 2024b)	42.66	46.01	24.29	25.71	29.49	46.15	48.34	49.67	46.32	39.71	35.14	40.54	42.67	46.67	45.58	48.30	51.00	57.77
Qwen2.5-VL-7B (Bai et al., 2025)	73.60	69.36	<u>54.29</u>	41.43	73.08	65.38	77.48	75.50	<u>81.62</u>	<u>75.00</u>	67.57	72.97	<u>81.33</u>	76.00	65.99	62.59	<u>80.88</u>	80.88
GLM-4V-9B (GLM et al., 2024)	51.23	46.11	39.29	26.43	46.15	44.87	61.59	56.29	51.47	38.24	56.76	51.35	54.67	53.33	43.54	40.82	55.78	55.78
InternVL3.5-8B (Wang et al., 2025)	69.06	60.49	50.71	43.57	71.79	61.54	75.50	68.87	63.97	50.00	70.27	51.35	78.67	73.33	61.22	57.14	78.88	69.72
Open-source Models (Size > 10B)																		
DeepSeek-VL2-16B (Wu et al., 2024)	67.68	63.74	37.86	41.43	69.23	65.38	74.17	74.17	67.65	52.21	62.16	54.05	<u>81.33</u>	73.33	64.63	59.18	78.49	76.89
DeepSeek-VL2-27B (Wu et al., 2024)	60.00	55.27	40.00	35.00	60.26	53.85	68.87	66.23	52.21	48.53	62.16	59.46	70.67	64.00	53.74	46.26	70.12	66.14
CogVLM2-19B (Wang et al., 2023)	57.64	54.88	36.43	32.14	53.85	47.44	60.93	57.62	54.41	45.59	67.57	54.05	57.33	62.67	57.82	53.06	68.92	72.11
Qwen2.5-VL-32B (Bai et al., 2025)	75.27	<u>70.44</u>	52.14	<u>45.00</u>	71.79	<u>67.95</u>	84.11	<u>77.48</u>	82.35	76.47	70.27	72.97	85.33	81.33	64.63	<u>61.22</u>	84.06	<u>79.68</u>
LLaVA-1.6-34B (Liu et al., 2024b)	63.05	58.42	23.57	32.14	73.08	61.54	72.85	68.21	63.97	52.94	62.16	59.46	72.00	68.00	63.95	55.78	72.51	67.73
InternVL3.5-38B (Wang et al., 2025)	73.40	66.90	51.43	41.43	<u>76.92</u>	74.36	<u>82.78</u>	<u>70.20</u>	72.06	69.12	86.49	<u>75.68</u>	<u>81.33</u>	<u>78.67</u>	<u>67.35</u>	58.50	78.88	75.70
Qwen2.5-VL-72B (Bai et al., 2025)	<u>75.17</u>	71.03	56.43	47.86	80.77	65.38	84.11	81.46	76.47	<u>75.00</u>	<u>75.68</u>	81.08	76.00	<u>78.67</u>	70.07	<u>61.22</u>	80.48	79.28
Closed-source Models																		
GPT-4.1 (OpenAI, 2025)	76.65	73.89	44.29	44.29	82.05	80.77	85.43	81.46	<u>80.88</u>	77.94	<u>81.08</u>	70.27	89.33	93.33	<u>72.11</u>	62.59	83.67	82.87
Gemini-2.5 Flash (Comanici et al., 2025)	<u>84.93</u>	80.89	<u>77.14</u>	<u>72.14</u>	<u>87.18</u>	<u>84.62</u>	<u>92.72</u>	<u>88.08</u>	88.97	<u>84.56</u>	83.78	83.78	<u>94.67</u>	93.33	<u>72.11</u>	<u>64.63</u>	<u>86.45</u>	<u>83.67</u>
Gemini-2.5 Pro (Comanici et al., 2025)	87.00	83.55	85.00	83.57	88.46	85.90	95.36	90.07	88.97	87.50	<u>81.08</u>	<u>75.68</u>	96.00	93.33	72.79	65.99	88.05	85.26
Text-only Models																		
GPT-4.1 (OpenAI, 2025)	23.35	34.09	10.00	20.00	26.32	2.63	22.52	35.76	30.88	39.71	29.73	40.54	13.33	20.00	32.65	34.01	25.90	42.63
Human Performance																		
Human Avg.	85.00	-	73.33	-	80.00	-	93.33	-	80.00	-	86.67	-	93.33	-	86.67	-	86.67	-

ent scales, with Qwen2.5-VL-32B and Qwen2.5-VL-72B achieving the strongest open-source results under Chinese and English prompts, respectively. Despite these advances, the models exhibit significant performance disparities across different cultural domains. In particular, *Astronomy* is consistently the most challenging domain for all models, whereas *Customs* and *Diet* yield markedly stronger results. We attribute this gap primarily to differences in the accessibility of domain knowledge during pre-training: cultural content related to *Customs* and *Diet* remains closely tied to contemporary daily life and therefore appears more frequently in pre-training corpora. By contrast, knowledge of traditional *Astronomy* is largely preserved in historical documents and specialized archives, making it less likely to be sufficiently covered during pre-training.

Open-source Models vs. Closed-source Models. As illustrated in Figure 4, closed-source models exhibit a clear advantage, forming the highest and most compact accuracy distribution on TCC-Bench. In contrast, open-source models show a wider spread: models below 10B parameters display large performance fluctuations, while larger open-source models (>10B) achieve noticeable improvements but still lag behind closed-source models on average. These results highlight the persisting performance gap between open- and closed-source MLLMs in culturally grounded tasks.

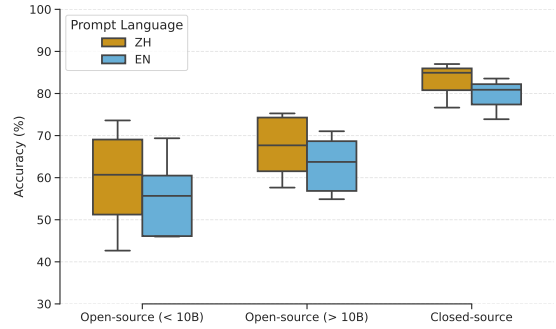


Figure 4: Accuracy distribution across model groups.

The impact of linguistics. According to the results shown in Table 3, most MLLMs achieve higher accuracy with the Chinese prompt than with the English prompt, although the magnitude of this gap varies across models. We attribute this trend to two main factors. First, the Chinese prompt naturally align with the cultural context of the images, enabling models to better ground visual cues in culturally specific semantics. Secondly, traditional Chinese culture contains unique terms and concepts that do not always have direct English equivalents. The translation process can introduce semantic drift or ambiguity, making it more difficult for models to accurately interpret and reason about culturally nuanced content.

TCC-Bench is less susceptible to interference from language priors. Table 3 presents the results of the closed-source model (*i.e.*, GPT-4.1) under

the text-only setting. Compared to its performance under multimodal conditions, GPT-4.1 experiences a significant performance drop of 53.30% (ZH) and 39.80% (EN). These results indicate that visual cues are indispensable for accurately answering questions in TCC-Bench, thereby mitigating the influence of language priors on the dataset.

4.3 Analysis

4.3.1 Analysis of Chain-of-Thought(CoT)

We investigate Chain-of-Thought (CoT) (Wei et al., 2022), a simple yet effective prompting strategy, to determine whether it can improve the performance of MLLMs on TCC-Bench. To cover models with different capacities and deployment regimes, we select four representative MLLMs: LLaVA-v1.6-7B as a small open-source baseline, Qwen2.5-VL-72B as a large open-source model, and GPT-4.1 and Gemini-2.5 Pro as closed-source models. The results are presented in Figure 5. Its effect is highly model-dependent: LLaVA-v1.6-7B drops substantially from 42.66% to 35.37% (-7.29%), whereas Qwen2.5-VL-72B improves from 75.17% to 77.64% (+2.47%). For closed-source models, the changes are marginal, with GPT-4.1 increasing slightly from 76.65% to 76.95% (+0.30%) and Gemini-2.5 Pro decreasing from 87.00% to 86.31% (-0.69%). These results indicate that the CoT prompting format does not provide stable gains for MLLMs on TCC-Bench. Specifically, the smaller LLaVA-v1.6-7B model may lack sufficient visual recognition and cultural knowledge to benefit from CoT, leading to a performance drop, whereas stronger closed-source models already have high baseline performance, making explicit reasoning less likely to provide additional useful information and resulting only in marginal changes rather than clear improvements. We suggest more in-depth research into this issue.

4.3.2 Analysis of Explanation Consistency

Beyond answer accuracy, we further investigate whether MLLMs can produce reasonable and culturally grounded explanations for their predictions. To this end, we introduce an additional experiment that evaluates the consistency between model-generated explanations and human-annotated rationales. Specifically, we modify the prompting strategy to require the model not only to select an answer option, but also to generate a brief natural-language explanation justifying its choice. For each question, we then extract the explanation pro-

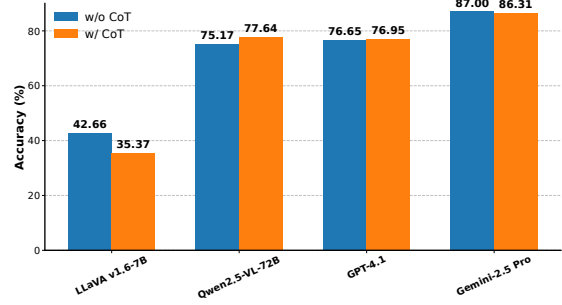


Figure 5: Comparison of MLLMs between w/o and w/ CoT.

duced by the model and compare it against a corresponding human-annotated explanation. To assess the alignment between model-generated and human explanations at scale, we adopt an LLM-as-a-Judge (Chan et al., 2023) evaluation paradigm. We use GPT-4.1 as the evaluator, scoring the model-generated explanation from 0 to 3. The results are shown in Figure 6. Consistent with the results reported in Table 3, Gemini-2.5 Pro achieves the highest and most well-balanced performance across all domains. This indicates not only strong overall reasoning capability but also robust cross-domain generalization in producing coherent explanations. GPT-4.1 follows closely, exhibiting competitive performance with slightly larger variance across domains. In contrast, open-source models show more irregular profiles. Smaller-scale models such as LLaVA-1.6-7B exhibit consistently low scores, particularly in *Astronomy* and *Music*, suggesting limited ability to articulate culturally specific reasoning even when the final answer is correct. Other open-source models achieve moderate explanation consistency, but their performance varies substantially across domains, revealing sensitivity to domain-specific cultural knowledge. More information can be found in the Appendix G.

4.3.3 Analysis of Domain-aware Prompting

To investigate whether explicitly providing cultural domain information can improve model performance, we conduct an analysis using domain-aware prompting. Specifically, for each question, we augment the original prompt by explicitly indicating its corresponding cultural domain. Table 4 reports a comparison of accuracy under domain-agnostic and domain-aware settings. Overall, we observe that domain-aware prompting does not consistently improve model performance across different MLLMs. While slight gains are observed for LLaVA-1.6-

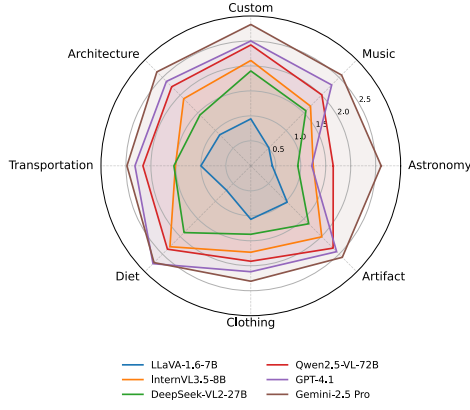


Figure 6: Explanation scores across different traditional Chinese cultural domains.

Table 4: Effect of Domain-aware Prompting on Model Performance.

Model	Domain-agnostic	Domain-aware
LLaVA-1.6-7B	42.66	43.74 (+1.08)
InternVL3.5-8B	69.06	66.50 (-2.56)
DeepSeek-VL2-27B	60.00	51.53 (-8.47)
Qwen2.5-VL-72B	75.17	74.68 (-0.49)
GPT-4.1	76.65	76.26 (-0.39)
Gemini-2.5 Pro	87.00	87.78 (+0.78)

7B (+1.08%) and Gemini-2.5 Pro (+0.78%), most models exhibit marginal or even notable performance degradation when explicit domain information is provided. In particular, InternVL3.5-8B and DeepSeek-VL2-27B suffer noticeable drops of 2.56% and 8.47%, respectively. These results suggest that explicitly injecting high-level domain labels does not necessarily facilitate cultural understanding for current MLLMs.

4.3.4 Error Analysis

To better understand the failure modes of MLLMs on TCC-Bench, we conduct an error analysis of the incorrect responses generated by Qwen2.5-VL-72B under the CoT setting. Specifically, we manually annotate each incorrect response and categorize the errors into four types: Lack of Cultural Knowledge, Visual Perceptual Error, Contextual Interpretation Error, and Reasoning Error. Figure 7 presents the distribution of these error types. The dominant source of failure is the lack of cultural knowledge, which accounts for 55.9% of all errors. This suggests that even when a strong MLLM can perceive the main visual content, it often lacks the fine-grained cultural knowledge needed to identify or interpret traditional Chinese cultural artifacts correctly. Visual perceptual errors constitute the

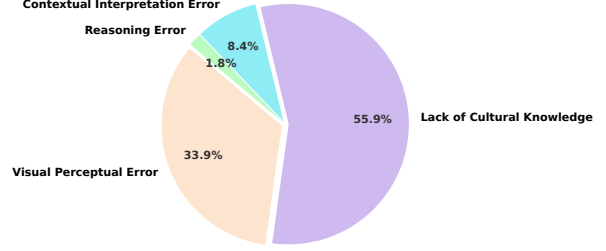


Figure 7: Distribution of Qwen2.5-VL-72B error types.

second largest group, accounting for 33.9%, indicating that culturally specific objects, symbols, and visual details still pose substantial recognition challenges. By contrast, contextual interpretation errors and reasoning errors are less frequent, accounting for 8.4% and 1.8%, respectively. The definitions of these categories and representative cases are provided in Appendix H.

5 Conclusion

This work introduces TCC-Bench, a benchmark specifically developed to evaluate the understanding abilities of MLLMs in the context of traditional Chinese culture. The benchmark offers a bilingual and semantically rich dataset that spans multiple domains and reflects authentic visual diversity. Through the use of a semi-automated question generation method and rigorous human curation, the benchmark achieves a balance between scalability and data quality. Comprehensive evaluations reveal that existing MLLMs encounter substantial difficulties when addressing tasks that involve culturally specific knowledge. These findings emphasize the importance of incorporating cultural perspectives into the development of multimodal models. TCC-Bench is expected to serve not only as a robust evaluation tool but also as a foundation for promoting future research toward building more culturally aware and globally applicable AI systems.

Limitations

Although TCC-Bench marks a meaningful advancement in evaluating the cultural understanding capabilities of MLLMs, several limitations should be acknowledged. First, while GPT-4o is utilized to assist in generating question candidates, the process still demands substantial human effort in filtering, revising, and validating the outputs. This manual workload poses a challenge to scaling the dataset efficiently and highlights the need for more refined automation techniques in benchmark construction.

Secondly, obtaining high-quality images that accurately represent traditional Chinese culture remains difficult. Many cultural elements predate the emergence of visual recording technologies, resulting in a scarcity of visual materials that are both authentic and diverse. This limitation may lead to the underrepresentation of certain cultural domains in the dataset. Thirdly, traditional Chinese culture is vast and multifaceted, encompassing thousands of years of philosophical thought, artistic expression, and sociocultural development. Consequently, Regional and minority cultural traditions may also be unevenly represented in TCC-Bench. In the future, we hope to further expanding the scale of the dataset.

Ethical Statement

This work does not raise any ethical concerns. We anticipate that our findings will stimulate future research on enhancing models' ability to interpret diverse cultural contexts.

Acknowledgments

This work is supported by the National Key Research and Development Project of China (No.2022YFF0904304) and the Super Computing Platform of Beijing University of Posts and Telecommunications.

References

- Fakhraddin Alwajih, Gagan Bhatia, Abdelrahman Mohamed, Muhammad Abdul-Mageed, and 1 others. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.
- David M Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. 2023. Clair: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Singapore. Association for Computational Linguistics.
- Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- 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, 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, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36.

- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Steenkiste, Lisa Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790.
- Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2025. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 932–950.
- Open-AI. 2024. [Gpt-4o](#).
- OpenAI. 2025. Gpt-4.1. <https://openai.com/index/gpt-4-1/>.
- ChaeHun Park, Yujin Baek, Jaeseok Kim, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2025. Evaluating visual and cultural interpretation: The k-viscuit benchmark with human-vlm collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21960–21974.
- Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Steiner, Xiaohua Zhai, and Ibrahim M. Alabdulmohsin. 2024. [No filter: Cultural and socioeconomic diversity in contrastive vision-language models](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. [CVQA: culturally-diverse multilingual visual question answering benchmark](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Pranav Tushar, Eshan Pandey, Lyka Diane Bala Austria, Yin Yin Loo, Jing Hao Lim, Indriyati Atmosukarto, and Donny Soh Cheng Lock. 2025. Mer-culture: A comprehensive benchmark to evaluate vision-language models on cultural understanding in singapore. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 565–574.
- Norawit Urailetpasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. [SEA-VQA: Southeast Asian cultural context dataset for visual question answering](#). In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Bangkok, Thailand. Association for Computational Linguistics.
- Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademteu, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, and 1 others. 2025. All languages matter: Evaluating Imms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, and 1 others. 2024. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *Preprint*, arXiv:2409.12191.
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. [Cogvlm: Visual expert for pretrained language models](#). *Preprint*, arXiv:2311.03079.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. InternV3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. [Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding](#). *Preprint*, arXiv:2412.10302.
- Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. [Broaden the vision: Geo-diverse visual commonsense reasoning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11*

November, 2021, pages 2115–2129. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xinrun Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, and 2 others. 2024a. [Can mllms understand the deep implication behind chinese images?](#) *Preprint*, arXiv:2410.13854.

Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth S. Narayanan, and Salman Avestimehr. 2024b. [Creating a lens of chinese culture: A multi-modal dataset for chinese pun rebus art understanding.](#)

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

A Domain Definitions

To enable a comprehensive evaluation of MLLMs on traditional Chinese culture, we organize TCC-Bench into eight domains: *Astronomy*, *Music*, *Custom*, *Architecture*, *Transportation*, *Diet*, *Clothing*, and *Artifact*. These domains are derived through consultation of authoritative works on traditional Chinese culture and further refined through expert discussions. The resulting taxonomy is designed to balance everyday material culture with specialized symbolic knowledge. Specifically, it first covers the classical dimensions of daily life captured by *Yi-Shi-Zhu-Xing* (衣食住行), corresponding to *Clothing*, *Diet*, *Architecture*, and *Transportation*. These domains represent fundamental aspects of traditional Chinese social life and provide visually grounded cultural categories. We further include *Astronomy*, *Music*, *Custom*, and *Artifact* to capture intellectual, aesthetic, ritual, and material dimensions that are central to Chinese civilization. The scope and content of each domain are defined as follows:

- **Astronomy (*Tiānwén* / 天文):** Covers the traditional Chinese understanding of the cosmos, including the 24 solar terms, the lunar calendar, ancient astronomical instruments (e.g., armillary spheres), and constellation systems (e.g., the 28 Mansions).
- **Music (*Yīnyuè* / 音乐):** Encompasses traditional Chinese music theory (e.g., the pentatonic scale), classical instruments (e.g., Guqin, Pipa, Erhu), performing arts (e.g., Peking Opera, Kunqu), and ceremonial music.
- **Custom (*Mínsú* / 民俗):** Focuses on the social practices, rituals, and festivals that define cultural life. This includes major holidays (e.g., Spring Festival, Dragon Boat Festival), life-cycle rituals (e.g., traditional weddings, coming-of-age ceremonies), and social etiquette (e.g., table etiquette).
- **Architecture (*Jiànzhù* / 建筑):** Includes the styles, principles, and components of traditional Chinese construction. Examples range from imperial palaces and temples to vernacular dwellings (e.g., Siheyuan), garden design, and structural elements (e.g., Dougong brackets).
- **Transportation (*Jiāotōng* / 交通):** Pertains to historical modes of transport used in traditional Chinese society, such as horse-drawn carriages, sedan chairs, and types of watercraft.
- **Diet (*Yǐnshí* / 饮食):** Covers the vast of Chinese culinary arts. This includes the Eight Great Cuisines, specific culturally significant foods (e.g., Zongzi, Mooncakes), utensils (e.g., chopsticks), and the tea culture.
- **Clothing (*Fúshì* / 服饰):** Focuses on traditional attire and adornments. This domain includes various forms of historical garments (e.g., Hanfu), imperial robes (e.g., Dragon Robe), textiles (e.g., silk, embroidery), and symbolic accessories.
- **Artifact (*Qìwù* / 器物):** A broad category for material culture, crafts, and tools. This includes bronze vessels, porcelain and ceramics, jade carvings, calligraphy tools (e.g., the Four Treasures of the Study), lacquerware, and traditional furniture.

B Details of Image Collection

B.1 Image Source

We collect data from the following websites:

- <https://commons.wikimedia.org/>
- <https://www.sohu.com>
- <http://bridge.chinese.cn/test/questions>
- <https://www.flickr.com/>
- <https://ancientconstellations.fandom.com>
- <https://digitalarchive.npm.gov.tw/>

B.2 Image Filtering Criteria

To ensure the quality, relevance, and usability of the visual data used in TCC-Bench, we apply a set of filtering criteria during the image collection process.

- The image should clearly represent traditional Chinese cultural elements.
- The cultural content depicted in the image can be categorized into the following knowledge domains: Astronomy, Music, Custom, Architecture, Transportation, Diet, Clothing, and Artifact.

- The image should be licensed under either public domain terms or a non-commercial academic use license.

C Details of Question Generation Pipeline

This appendix provides additional details on the human-in-the-loop question generation pipeline described in Section 3.2. The goal of this pipeline is to combine the scalability of GPT-4o-assisted question generation with strict human quality control, thereby ensuring that the final samples are culturally grounded, visually answerable, factually correct, and free from shortcut cues in the question text.

C.1 Annotator Assignment and Quality Control

Six annotators participated in the question generation pipeline. All annotators have bachelor’s degrees or higher and have long-term experience living within a Chinese cultural context. Before annotation, they were briefed on the eight-domain taxonomy of TCC-Bench, the expected format of multiple-choice VQA samples, and the requirement that each question should require both image understanding and cultural knowledge.

To reduce individual bias, we organized annotation as a multi-stage process. Four annotators were responsible for manual refinement, where they revised GPT-4o-generated candidate questions, options, answers, and explanations. The remaining two annotators conducted human verification. Specifically, each verifier independently performed one full round of verification over the refined samples. A sample was retained only if it passed both rounds of verification. During this verification stage, 13 samples that failed to satisfy the quality requirements were filtered out.

C.2 Guidelines for Cultural Concept Extraction

In the cultural concept extraction stage, annotators inspect each image and identify one cultural concept that can support question generation. The extracted concept serves as an intermediate signal for GPT-4o and human refinement, but it is not necessarily exposed in the final question. Annotators follow three principles:

- **Relevance.** The concept should clearly belong to traditional Chinese culture and fit one of the eight domains in TCC-Bench.

- **Visibility.** The concept should be supported by visible evidence in the image.
- **Usability.** The concept should be suitable for generating a meaningful VQA question. Concepts that are too broad, too obscure, or visually ambiguous are revised or discarded.

C.3 Guidelines for Manual Refinement

After GPT-4o produces candidate questions from human-provided concepts, annotators manually refine each content. The refinement stage focuses on four key checks:

- **Domain consistency.** The question should belong to the corresponding cultural domain.
- **Question correctness.** The question, options, correct answer, and their alignment with the image should be accurate.
- **Explanation correctness.** The explanation should be factually correct, culturally appropriate, and consistent with the image.
- **No direct concept leakage.** The question should not directly reveal the extracted cultural concept. For instance, “What is a sundial used for?” should be rewritten as “What is the instrument in the picture used for?”

If an item does not satisfy these checks, annotators are allowed to revise the question, options, correct answer, or explanation accordingly.

C.4 Guidelines for Human Verification

Each question–answer pair undergoes two rounds of independent verification by annotators. During each round, annotators evaluate the sample according to the same four criteria used in the manual refinement stage. Unlike manual refinement, the human verification stage does not involve further revision. Any sample that fails to satisfy any criterion in either verification round is discarded.

D Prompt for generating questions

In Section 3.2, we present the question generation process of TCC-Bench, where we leverage the powerful generative capabilities of GPT-4o to reduce the manual effort required in dataset construction. To efficiently generate high-quality question samples, we design a corresponding prompt. Table 5 presents the prompt we use.

Table 5: Prompt for generating questions.

You are a question-setter with a rich knowledge of traditional Chinese culture. I will give you an entity or a topic. Please generate three multiple-choice questions about traditional Chinese cultural knowledge based on it. Each question has four options: A, B, C, and D. Only one option is correct, and the other three options are incorrect. The incorrect options are required to be somewhat confusing. Additionally, a brief explanation should be provided to justify the pairing between each question and its correct answer. All the outputs should be in both Chinese and English. The following is a reference example containing one question (you are expected to generate three):

Input: Vermilion Bird.

Output:

1、朱雀经常用来指代哪一个方向？

A、南方；

B、北方；

C、东方；

D、西方。

正确答案：A

解释：朱雀经常用来指代南方。

1、Which direction is Vermilion Bird often used to refer to?

A、South;

B、North;

C、East;

D、West.

Answer: A

Explanation: Vermilion Bird is often used to refer to the south.

E Word Clouds of TCC-Bench

Figure 9 presents the word clouds generated for each category in TCC-Bench. Commonly occurring but semantically uninformative terms such as “picture” and “show” are treated as stopwords, as they appear frequently in the text yet convey no meaningful cultural concepts.

Table 6: Prompt for evaluating in Chinese.

请根据提供的图片尝试回答下面有关于中国传统文化的单选题。直接回答正确选项，不要包含额外的解释。请使用以下格式：“答案：\$LETTER”，其中\$LETTER是你认为正确答案的字母。问题：“{question}”

(A) “{option_a}”

(B) “{option_b}”

(C) “{option_c}”

(D) “{option_d}”

答案：

F Experiment Detail

We employ the LMDeploy¹ framework to deploy and accelerate all open-source MLLMs. The open-source models are obtained from Hugging Face, and Table 9 lists each model alongside its corresponding identifier. For closed-source models, we evaluate gpt-4.1-2025-04-14, gemini-2.5-flash, and gemini-2.5-pro, with responses generated via their official APIs. We set the sampling temperature to 0. The prompts used for evaluation are provided in Table 6, Table 7, and Table 8.

G Details of the Explanation Consistency Analysis

To assess the depth of understanding and the reliability of MLLMs on TCC-Bench, we conduct a comprehensive Explanation Consistency Analysis. Unlike conventional VQA metrics that focus solely on the final answer choice (e.g., A, B, C, or D), this analysis examines whether a model’s textual explanation coherently aligns with both the visual evidence and the underlying cultural context

Table 7: Prompt for evaluating in English.

Please try to answer the following multiple-choice questions about traditional Chinese culture based on the provided pictures. Answer the correct option directly without including additional explanations. Please use the following format: "Answer: \$LETTER", where \$LETTER is the letter of the option you think is correct.

Question: “{question}”

(A) “{option_a}”

(B) “{option_b}”

(C) “{option_c}”

(D) “{option_d}”

Answer:

Table 8: Prompt for evaluating using CoT.

请根据提供的图片尝试回答下面有关于中国传统文化的单选题。让我们逐一思考每个选项，逐步分析。你回答的最后一行应该用以下格式：“答案：\$LETTER”，其中\$LETTER是你认为正确答案的字母。

问题：“{question}”

(A) “{option_a}”

(B) “{option_b}”

(C) “{option_c}”

(D) “{option_d}”

答案：

¹<https://github.com/InternLM/lmdeploy>

Table 9: The open-source models together with their corresponding identifiers on Hugging Face.

Model	Model Name in Hugging Face
LLaVA-1.6-7B	liuhaotian/llava-v1.6-vicuna-7b
LLaVA-1.6-34B	liuhaotian/llava-v1.6-34b
DeepSeek-VL2-3B	deepseek-ai/deepseek-vl2-tiny
DeepSeek-VL2-16B	deepseek-ai/deepseek-vl2-small
DeepSeek-VL2-27B	deepseek-ai/deepseek-vl2
Qwen2.5-VL-7B	Qwen/Qwen2.5-VL-7B-Instruct
Qwen2.5-VL-32B	Qwen/Qwen2.5-VL-32B-Instruct
Qwen2.5-VL-72B	Qwen/Qwen2.5-VL-72B-Instruct
InternVL3.5-8B	OpenGVLab/InternVL3_5-8B
InternVL3.5-38B	OpenGVLab/InternVL3_5-38B
GLM-4V-9B	zai-org/glm-4v-9b
CogVLM2-19B	zai-org/cogvlm2-llama3-chat-19B

of TCC-Bench. We adopt GPT-4.1 as the evaluator. The evaluation prompt is presented in Table 10. The evaluator assigns scores to model-generated explanations on a four-point scale ranging from 0 to 3. Detailed evaluation results are reported in Table 11. Notably, Gemini-2.5 Pro attains the highest overall consistency score of 2.60, highlighting its superior ability to interpret complex cultural artifacts. In contrast, smaller models such as LLaVA-1.6-7B exhibit pronounced difficulties in the *Astronomy* domain (0.43), suggesting that specialized knowledge requiring precise identification of instruments remains challenging for models with limited parameter capacity. To provide a more fine-grained perspective on model behavior, we visualize the distribution of scores (0, 1, 2, and 3) for each model in Figure 8. The resulting distributions show that stronger models, including Gemini-2.5 Pro and GPT-4.1, achieve a substantially higher proportion of Score 3, indicating that when these models produce correct answers, they typically support them with thorough and accurate reasoning.

H Details of Error Analysis

We provide detailed descriptions of the error types as follows:

Lack of Cultural Knowledge (55.9%). Errors in this category account for the majority of the observed failure cases. In these cases, the model correctly perceives the visual content but fails to answer correctly due to insufficient or inaccurate knowledge of Chinese cultural concepts. This result highlights a fundamental limitation of current MLLMs: while visual perception capabilities have

significantly improved, cultural knowledge remains unevenly represented, particularly regarding non-Western and fine-grained traditional concepts.

Visual Perceptual Error (33.9%). This type of error arises when the model fails to accurately perceive or recognize key visual evidence in the image. These failures include misidentifying objects, overlooking important visual details, confusing text or symbols, and incorrectly recognizing spatial relations or visual attributes. The relatively high proportion of such errors indicates that fine-grained visual perception is still challenging for MLLMs, especially when the visual cues are closely tied to culturally specific objects, symbols, or scenes.

Contextual Interpretation Error (8.4%). In this category, the model has access to some relevant visual evidence and cultural knowledge, but misinterprets how they should be understood in the given question context. These errors show that beyond perception and factual knowledge, MLLMs also need stronger contextual grounding to correctly connect cultural elements with their situated meanings.

Reasoning Error (1.8%). This category accounts for the smallest proportion of errors. In these cases, the model generally perceives the key visual content and possesses the necessary cultural or contextual knowledge, but fails to integrate the evidence coherently to arrive at the correct answer. Reasoning errors typically manifest as incorrect inference steps, overgeneralization from superficial cues, or selecting an answer that is inconsistent with the model’s own preceding analysis.

For each of the four error types (see Sec-

tion 4.3.4), we present a representative example accompanied by manual annotations. Please see from Figure 10 to Figure 13.

Table 10: Prompt for evaluating explanation consistency.

You are an expert evaluator specializing in traditional Chinese culture. Your task is to evaluate the quality of an explanation generated by an AI model for a visual question answering (VQA) task.

You will be provided with:

1. Question: The question asked about a cultural artifact or concept.
2. Correct Answer: The correct option for the question.
3. Reference Explanation: The ground truth explanation describing why the answer is correct and providing cultural context.
4. Candidate Explanation: The explanation generated by the AI model being evaluated.

Scoring Criteria:

Please rate the Candidate Explanation on a scale of 0 to 3 based on the following rubric:

- **3 (Excellent):** The explanation is factually accurate, comprehensive, and aligns perfectly with the Reference Explanation. It correctly identifies the key cultural elements (e.g., specific names, dynasties, functions, meanings) and logically supports the correct answer.
- **2 (Good):** The explanation is mostly correct and supports the right answer. It captures the main idea but may miss minor details present in the Reference Explanation or is slightly less precise.
- **1 (Fair):** The explanation is weak. It may contain some relevant keywords but lacks depth, is overly vague, or contains minor factual inaccuracies that do not entirely contradict the answer but show a lack of precise understanding.
- **0 (Poor):** The explanation is incorrect, irrelevant, hallucinatory, or fails to support the correct answer. It contains significant factual errors regarding traditional Chinese culture.

Output Format: Please output your evaluation in the following JSON format: "score": <score>

Where <score> is the integer score (0, 1, 2, or 3) based on the criteria above.

Question: "{question}"

Correct Answer: "{correct_answer}"

Reference Explanation:

"{reference_explanation}"

Candidate Explanation:

"{candidate_explanation}"

Table 11: Explanation consistency score via LLM-as-a-Judge on TCC-Bench. *Astr.*, *Arch.*, *Tran.*, *Clot.* and *Arti.* denote *Astronomy*, *Architecture*, *Transportation*, *Clothing* and *Artifact*. The highest scores among models in each column are highlighted in bold.

Model	Overall	Astr.	Music	Custom	Arch.	Tran.	Diet	Clot.	Arti.
LLaVA-1.6-7B	0.85	0.43	0.51	0.94	0.88	1.00	0.69	1.07	1.03
InternVL3.5-8B	1.84	1.25	1.69	2.11	1.90	1.51	2.29	1.73	2.01
DeepSeek-VL2-27B	1.53	0.94	1.56	1.90	1.44	1.54	1.89	1.37	1.64
Qwen2.5-VL-72B	2.14	1.65	2.01	2.42	2.24	2.16	2.36	1.91	2.33
GPT-4.1	2.23	1.22	2.29	2.50	2.39	2.32	2.77	2.12	2.43
Gemini-2.5 Pro	2.60	2.61	2.57	2.83	2.66	2.48	2.74	2.31	2.59

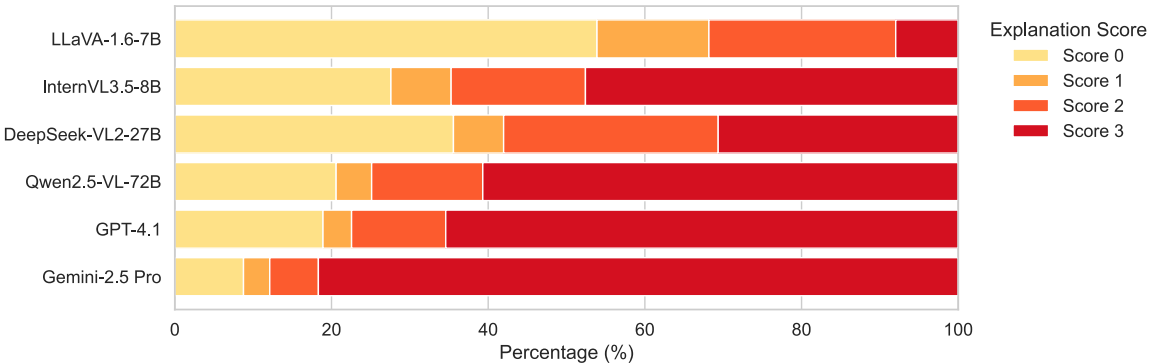


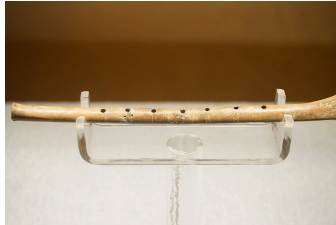
Figure 8: Distribution of Explanation consistency score.



Figure 9: Word clouds of TCC-Bench.



Lack of Cultural Knowledge



Q: 这种乐器在古代常用于什么场合?
(Where was this instrument often used in ancient times?)
(A) 军队出征 (Send out troops)
(B) 狩猎 (Hunting)
(C) 宫廷宴会 (Court banquet)
(D) 祭祀 (Sacrifice)

Ground Truth: (D)

Qwen2.5-VL-72B: 这张图片展示的是一种古代乐器，看起来像是一种笛子或类似的吹奏乐器。让我们逐一分析每个选项：
(A) 军队出征：古代军队出征时通常会使用鼓、号角等乐器来鼓舞士气和传递信息，而不是这种吹奏乐器。
(B) 狩猎：狩猎时使用的工具主要是弓箭、陷阱等，虽然可能会用一些简单的吹奏乐器来模仿动物声音，但这种乐器看起来并不适合狩猎场合。
(C) 宫廷宴会：宫廷宴会通常会有各种音乐表演，包括吹奏乐器。这种乐器看起来精致，适合在宫廷宴会中演奏优雅的音乐。
(D) 祭祀：祭祀活动通常会使用一些特殊的乐器，如编钟、编磬等，但这种乐器也有可能祭祀中使用，特别是在一些特定的仪式中。
综合考虑，这种乐器在古代最常用于宫廷宴会中演奏优雅的音乐。
答案：C

Qwen2.5-VL-72B: This image depicts an ancient musical instrument that resembles a flute or a similar woodwind instrument. Let us analyze each option systematically:
(A) Military Expedition: In ancient times, military marches typically employed instruments such as drums or horns to boost morale and transmit signals, rather than a delicate wind instrument like this.
(B) Hunting: Hunting primarily utilized tools like bows, arrows, and traps. While simple wind instruments were sometimes used to mimic animal calls, this particular instrument does not appear suited for the practicalities of a hunting environment.
(C) Court Banquet: Court banquets frequently featured diverse musical performances, including those involving wind instruments. Given its exquisite craftsmanship, this instrument is highly appropriate for performing elegant music in a courtly setting.
(D) Ritual/Sacrificial Ceremony: Ritual activities usually centered around specialized ceremonial instruments such as bronze bells (bianzhong) or stone chimes (bianqing). While this instrument could potentially be used in specific ceremonial rites, it was not the primary liturgical choice.
Answer: C)

Error Analysis: Qwen2.5-VL-72B准确捕捉到图中的乐器，但由于缺乏相关文化知识，导致其做出错误判断。
(Qwen2.5-VL-72B correctly identified the instrument but failed to make the right judgment due to a lack of necessary cultural background.)

Figure 10: A case study of *Lack of Cultural Knowledge*.



Visual Perceptual Error



Q: 图中的表演是哪一种戏剧表演艺术中重要的组成部分?
(Which kind of dramatic performance art is the performance in the picture an important part of?)
(A) 京剧 (Beijing Opera)
(B) 越剧 (Shaoxing Opera)
(C) 川剧 (Sichuan opera)
(D) 黄梅戏 (Huangmei Opera)

Ground Truth: (C)

Qwen2.5-VL-72B: 让我们逐一分析每个选项:
(A) 京剧: 京剧是中国最著名的戏剧表演艺术之一, 以其独特的脸谱、服饰和表演风格著称。图中的人物穿着华丽的戏服, 戴着精致的脸谱, 手持红色折扇, 这些特征与京剧的表演风格非常相似。
(B) 越剧: 越剧主要流行于浙江、上海等地, 以唱腔优美、表演细腻著称。图中的人物装扮和道具与越剧的风格不符。
(C) 川剧: 川剧以其变脸和喷火等特技表演而闻名。虽然图中的人物装扮华丽, 但没有明显的变脸或喷火的特征, 因此不太可能是川剧。
(D) 黄梅戏: 黄梅戏以唱腔优美、表演生活化著称, 通常没有如此华丽的装扮和脸谱。
综合以上分析, 图中的表演最符合京剧的特征。
答案: A

Qwen2.5-VL-72B: This image depicts an ancient musical instrument that resembles a flute or a similar woodwind instrument. Let us analyze each option systematically:
(A) Peking Opera: Peking Opera is one of the most renowned forms of traditional Chinese theatrical performance, distinguished by its unique facial makeup, costumes, and performance style. The figure in the image is dressed in an elaborate opera costume, wears refined facial makeup, and holds a red folding fan. These characteristics closely align with the performance style of Peking Opera.
(B) Yue Opera: Yue Opera is primarily popular in Zhejiang Province and Shanghai, and is known for its melodious singing and delicate performance style. The costume and props shown in the image do not match the typical style of Yue Opera.
(C) Sichuan Opera: Sichuan Opera is famous for its special performance techniques, such as face-changing and fire-spitting. Although the figure in the image is elaborately dressed, there are no visible signs of face-changing or fire-spitting, making it unlikely to be Sichuan Opera.
(D) Huangmei Opera: Huangmei Opera is characterized by its melodious singing and down-to-earth performance style, and it typically does not feature such ornate costumes or facial makeup. Based on the above analysis, the performance shown in the image most closely corresponds to Peking Opera.
Answer: A)

Error Analysis: Qwen2.5-VL-72B没有捕捉到图中正是在进行变脸表演, 因此没有输出正确答案。
(Qwen2.5-VL-72B failed to recognize that the figure in the image was performing a face-changing act, and therefore did not produce the correct answer.)

Figure 11: A case study of *Visual Perceptual Error*.



Contextual Interpretation Error



Q: 图中建筑的名称是什么?
(What's the name of the building in the picture?)
(A) 牌坊 (Paifang)
(B) 石碑 (Stone Tablet)
(C) 华表 (Huabiao)
(D) 龙柱 (Dragon Pillar)

Ground Truth: (D)

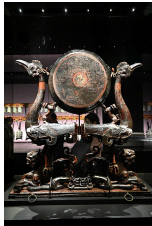
Qwen2.5-VL-72B: 让我们逐一分析每个选项:
(A) 牌坊: 牌坊是一种传统的中国建筑形式, 通常用于纪念或表彰某些人或事件。它通常由石柱和横梁组成, 形成一个门形结构。图中的建筑并不符合这种结构特征。
(B) 石碑: 石碑通常是竖立在地面上的长方形石块, 上面刻有文字或图案, 用于记录历史事件或人物。图中的建筑虽然有雕刻, 但其形状和用途与石碑不符。
(C) 华表: 华表是中国古代的一种标志性建筑, 通常位于宫殿、陵墓等重要场所的入口处。它通常由一根高大的石柱组成, 顶部有一个装饰性的部分。图中的建筑具有高大石柱的特征, 并且顶部有装饰性雕刻, 符合华表的特点。
(D) 龙柱: 龙柱是指雕刻有龙图案的柱子, 常见于中国古代建筑中。图中的建筑确实有龙的雕刻, 但它不仅仅是一根柱子, 而是一个完整的华表结构。
综合以上分析, 图中的建筑最符合华表的特征。
答案: C
Qwen2.5-VL-72B: Let us analyze each option one by one:
(A) Paifang (memorial archway): A paifang is a traditional Chinese architectural structure, usually built to commemorate or honor certain people or events. It is typically composed of stone pillars and crossbeams, forming a gate-like structure. The building in the image does not match these structural characteristics.
(B) Stone stele: A stone stele is usually an upright rectangular stone slab engraved with text or patterns, used to record historical events or figures. Although the structure in the image contains carvings, its shape and function do not correspond to those of a stone stele.
(C) Huabiao: A huabiao is a symbolic architectural element in ancient China, commonly placed at the entrances of important sites such as palaces and mausoleums. It usually consists of a tall stone column with a decorative ornament on top. The structure in the image exhibits the characteristics of a tall stone pillar and features decorative carvings at the top, which are consistent with the features of a huabiao.
(D) Dragon pillar: A dragon pillar refers to a pillar carved with dragon motifs, commonly seen in ancient Chinese architecture. While the structure in the image does feature dragon carvings, it is not merely a single pillar but rather a complete huabiao structure.
Based on the above analysis, the structure in the image most closely matches the characteristics of a huabiao.
Answer: C)

Error Analysis:
Qwen2.5-VL-72B捕捉到了准确的视觉信息, 也给出了正确的文化知识, 但忽略了题图最具区分性的语义线索: 盘龙浮雕覆盖柱身。
(Qwen2.5-VL-72B captures the correct visual information and provides accurate cultural knowledge, but fails to recognize the most distinctive semantic cue in the image, namely that the column is covered with coiling dragon reliefs.)

Figure 12: A case study of *Contextual Interpretation Error*.



Reasoning Error



Q: 图中的乐器的独特之处在于其什么特征?
(What makes the instrument in the picture unique?)
(A) 形状像鸟 (Shaped like a bird)
(B) 使用鸟羽装饰 (Decorate with bird feathers)
(C) 架子上刻有鸟形图案 (The shelf is engraved with a bird)
(D) 模仿鸟鸣声 (Mimicking birdsong)

Ground Truth: (C)

Qwen2.5-VL-72B: 让我们逐一分析每个选项:
(A) 形状像鸟: 图中的乐器确实有两个类似鸟头的装饰, 但整体形状并不完全像鸟。
(B) 使用鸟羽装饰: 图中没有明显的鸟羽装饰。
(C) 架子上刻有鸟形图案: 图中的架子上有类似鸟头的装饰, 但并没有明显的鸟形图案。
(D) 模仿鸟鸣声: 图中没有直接显示乐器发出的声音类型, 但从其独特的鸟头装饰来看, 可能与模仿鸟鸣声有关。
综合分析, 图中的乐器最独特的特征是其形状和装饰, 尤其是两个类似鸟头的装饰。因此, 最符合的选项是 (A) 形状像鸟。
答案: A
Qwen2.5-VL-72B: Let's analyze each option one by one:
(A) Shaped like a bird: The instrument in the image indeed features two decorations resembling bird heads, although its overall shape does not entirely resemble a bird.
(B) Decorated with bird feathers: There are no visible bird feather decorations in the image.
(C) Bird patterns carved on the stand: While the stand features decorations resembling bird heads, there are no distinct bird-shaped patterns or engravings.
(D) Mimics birdsong: The image does not directly indicate the type of sound the instrument produces; however, given its unique bird-head decorations, it may be related to mimicking birdsong.
Answer: A

Error Analysis: Qwen2.5-VL-72B捕捉到了图中的视觉内容, 也正确使用文化知识分析了选项, 但并未推理出正确答案。
(While Qwen2.5-VL-72B successfully captured the visual content and leveraged cultural knowledge to analyze the options, it failed to infer the correct answer.)

Figure 13: A case study of Reasoning Error.