

Dunhuang-Bench: How Well Do MLLMs Understand Cultural Heritage?

Junyi Yuan¹, Jian Zhang¹, Tianxiu Yu²,
Yanlin Zhou², Xiaobo Jin¹, Qiufeng Wang¹, Fangyu Wu^{1*},

¹Xi'an Jiaotong-Liverpool University, ²Dunhuang Academy

Correspondence: fangyu.wu02@xjtlu.edu.cn

Abstract

Dunhuang art, a cornerstone of global heritage, demands fine-grained visual perception anchored by specialized cultural knowledge. Given the strong performance of multimodal large language models (MLLMs) on generic multimodal benchmarks, *to what extent can they understand artifacts from Dunhuang art that are grounded in cultural context?* To this end, we construct Dunhuang-Bench, a large-scale benchmark comprising 486 images and 22,970 QA pairs. It incorporates diverse task formats to evaluate MLLMs' cultural understanding: Question Answering with Text Description, Multi-turn Dialogue, and Question Answering with Choices. Guided by Panofsky's theory of iconology, we design two tasks including visual perception and knowledge reasoning for the evaluation of content understanding. In addition, we follow the theory of formal analytic tradition to design another task of artistic appreciation in our Dunhuang-Bench. Extensive evaluations of 20 mainstream MLLMs on Dunhuang-Bench reveal a consistent performance drop from perception and appreciation to reasoning. Moreover, CoT and few-shot prompting show marginal or negative impact, highlighting the limits of prompting-based improvements. Dunhuang-Bench thus provides a challenging benchmark for advancing multimodal cultural understanding. The code and data of this paper are available at <https://github.com/yyyjyy/Dunhuang-Bench>.

1 Introduction

Dunhuang art, a millennium-old cultural treasure, stands as a quintessential representative of global cultural heritage (Wang and Wu, 2023). Recent advances in Multimodal Large Language Models (MLLMs) have led to extensive evaluations on Chinese cultural benchmarks (Liu et al., 2025; Romero et al., 2024), indicating substantial potential for

*Corresponding author.



Figure 1: A comparison between existing culture benchmarks and Dunhuang-Bench, illustrating how theory-grounded multi-task evaluation: visual perception, knowledge reasoning, and artistic appreciation.

digital preservation and cultural inheritance. However, the capabilities of MLLMs in Dunhuang art remain largely unevaluated. Originating from the Mogao Caves and related grotto sites along the Silk Road, Dunhuang art comprises large-scale murals and polychrome sculptures shaped by centuries of historical exchange (Whitfield et al., 2015). Its rich visual diversity, dense symbolic conventions, and deep grounding in art-historical and religious knowledge make it a challenging domain for multimodal understanding. Moreover, the development of benchmarks in this domain has been constrained by data scarcity and high annotation costs. Consequently, it remains unclear how well current MLLMs actually perform on Dunhuang art.

To fill this gap, we introduce Dunhuang-Bench, the first theory-grounded multimodal benchmark consisting of 486 images and 22,970 QA pairs from

authoritative Dunhuang art sources. It is designed to comprehensively evaluate MLLMs’ capabilities of cultural understanding in Dunhuang art. As shown in Fig. 1, Dunhuang-Bench includes three complementary tasks: *Task 1: Visual Perception* evaluates recognition of objective visual attributes through single-turn, image-based questions; *Task 2: Multi-turn Knowledge Reasoning* evaluates the ability to integrate visual evidence and expert cultural knowledge across multiple dialogue turns; and *Task 3: Artistic Appreciation* measures the capability to discriminate fine-grained stylistic differences in multiple-choice settings. In addition, our Dunhuang-Bench is theory-grounded. Specifically, both Task 1 and Task 2 are guided by Panofsky’s theory of iconology (Panofsky, 2018) for the evaluation of perception and reasoning, while task 3 are constructed from the theory of formal analytic tradition (Arnheim, 1954).

Compared to existing cultural benchmarks shown in Table 1, Dunhuang-Bench has three key features. First, Dunhuang-Bench is specifically designed to evaluate the MLLM in Dunhuang Art Culture, which stands as a quintessential representative of Chinese cultural heritage (Wang and Wu, 2023; Zhang et al., 2025b). Second, Dunhuang-Bench is substantially larger in scale. It is about twice the size of the previously largest dataset, CVQA (Romero et al., 2024), and nearly tenfold that of CulturalVQA (Nayak et al., 2024), ensuring broader evaluative coverage. Finally, while most current benchmarks (e.g., CVQA (Romero et al., 2024), SCB (Satar et al., 2025)) are restricted to Question Answering with Choices (QA-C), Dunhuang-Bench provides a broader task format: Question Answering with Text Description (QA-T) for visual perception, Multi-turn Dialogue for knowledge reasoning, and QA-C for artistic appreciation. Notably, it is the only theory-grounded benchmark in this field, enabling a more systematic and rigorous assessment of MLLMs.

We evaluate 20 mainstream MLLMs, including closed-source systems such as Gemini-2.5 (Comanici et al., 2025) and GPT-5.1 (OpenAI, 2025b), as well as open-source model families including the InternVL (Zhu et al., 2025) and Qwen-VL (Bai et al., 2025b) series. To support these evaluations, we propose a unified evaluation protocol with LLM-based scoring for Task 2 and a Normalized Rank Score for Task 3. The experimental results indicate that current MLLMs have significant room for improvement. The top-performing InternVL3.5-

Name	Domain	Scale	Task Formats			Theory-grounded
			QA-T	Dialogue	QA-C	
CVQA (Romero et al., 2024)	Global Culture	10,374	⊗	⊗	⊗	⊗
K-Viscuit (Park et al., 2025)	Korean Culture	420	⊗	⊗	⊗	⊗
CulturalVQA (Nayak et al., 2024)	Global Culture	2,328	⊗	⊗	⊗	⊗
FoodieQA (Li et al., 2024)	Chinese Food Culture	403	⊗	⊗	⊗	⊗
SCB (Satar et al., 2025)	Southeast Asian Culture	3,178	⊗	⊗	⊗	⊗
MCS-Bench (Liu et al., 2025)	Traditional Chinese Culture	6,500	⊗	⊗	⊗	⊗
Pun Rebus Art (Zhang et al., 2025c)	Traditional Chinese Culture	3,033	⊗	⊗	⊗	⊗
Dunhuang-Bench (Ours)	Dunhuang Art Culture	22,970	⊗	⊗	⊗	⊗

Table 1: Comparison of culture-oriented multimodal benchmarks. We provide three task formats including QA-T (Question Answering with Text Description) for the *Visual Perception* task, Multi-turn Dialogue for the *Knowledge Reasoning* task, and QA-C (Question Answering with Multiple Choices: MCQ) for the *Artistic Appreciation* task.

241B-A28B (Wang et al., 2025a) achieves an average score below 50.9% shown in Section 5. Notably, Chain-of-Thought (Wei et al., 2022) prompting unexpectedly degrades performance, particularly in tasks demanding fine-grained visual perception. Moreover, a few-shot prompting exhibits unstable effects and may decrease performance on knowledge reasoning and artistic appreciation. Overall, these results indicate that current MLLMs face fundamental challenges in performing cultural understanding in Dunhuang art.

Our main contributions can be summarized:

- We introduce Dunhuang-Bench, the first large-scale theory-grounded benchmark for Dunhuang art. It evaluates MLLMs’ cultural understanding in visual perception, knowledge reasoning, and artistic appreciation.
- We evaluate cultural understanding across 22,970 QA pairs using three distinct task formats: QA-T, multi-turn dialogue, and single-choice MCQs (i.e., QA-C), ensuring a comprehensive evaluation of MLLM capabilities.
- We design a unified evaluation protocol that introduces LLM-based scoring for multi-turn knowledge reasoning and a Normalized Rank Score for artistic appreciation.
- We evaluate 20 mainstream open- and closed-source MLLMs on Dunhuang-Bench, revealing significant performance gaps and providing valuable insights for future work on cultural heritage.

2 Related Work

Multi-modal Large Language Models. Recent advances in MLLMs have led to strong

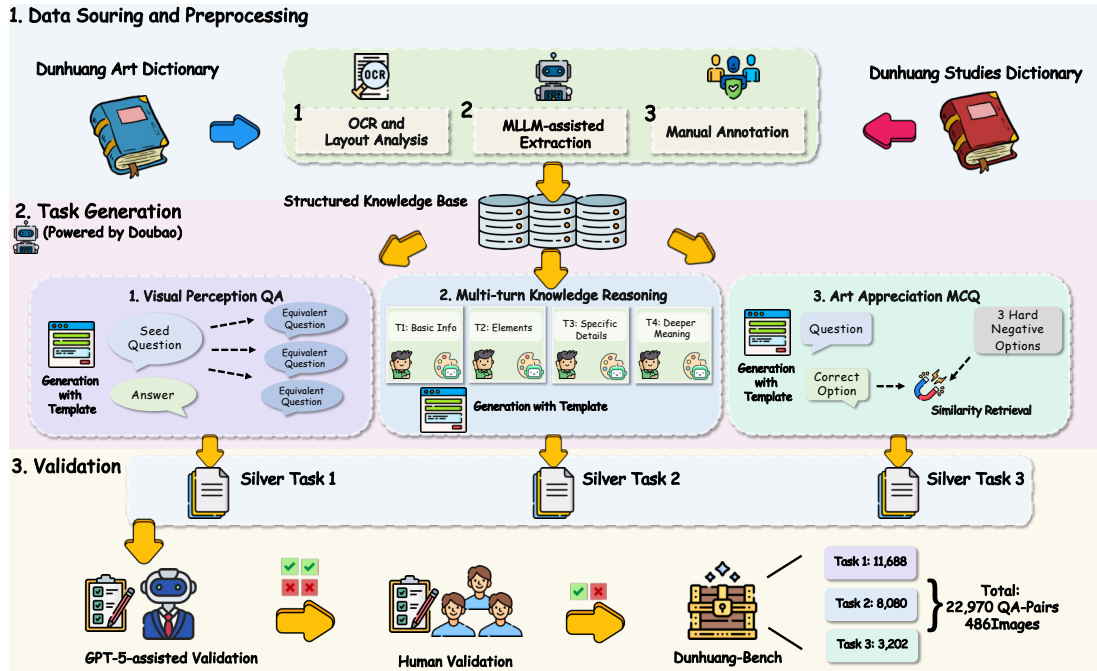


Figure 2: Overview of the Dunhuang-Bench construction pipeline, illustrating data sourcing and preprocessing from Dunhuang art resources, task generation (Task 1 visual perception, Task 2 multi-turn knowledge reasoning, and Task 3 artistic appreciation), and a hierarchical validation process combining GPT-5–assisted and human validation.

performance from closed-source models such as GPT (Achiam et al., 2023) and Gemini (Team et al., 2024) on vision–language tasks. While open-source community has developed a diverse family of MLLMs, including InternVL (Wang et al., 2025a), Qwen-VL (Bai et al., 2025b) and DeepSeek-VL (Wu et al., 2024) series. While recent works have begun to evaluate MLLMs in cultural heritage contexts (Liu et al., 2025; Zhang et al., 2025a), and explore improving semantic understanding with large language models (Hu et al., 2026; Zhang et al., 2025b), it remains unclear how well these models perform on Dunhuang art. This gap calls for benchmarks that explicitly evaluate cultural understanding in Dunhuang art.

Cultural Multimodal Benchmarks. Recent benchmarks assess cultural understanding through image description and generation, such as Crossmodal3600 (Thapliyal et al., 2022) and CUBE (Kannen et al., 2024), as well as benchmarks targeting cultural semantics, symbolism, and contextual cues across modalities (Nayak et al., 2024; Romero et al., 2024; Schneider et al., 2025; Park et al., 2025; Satar et al., 2025; Yuan et al., 2025). Within Chinese cultural domain, existing benchmarks focusing on areas such as food culture (Li et al., 2024), everyday culturally grounded scenes (Wang et al., 2025b), or selected aspects

of traditional culture heritage (Liu et al., 2025). However, these benchmarks overlook the cultural understanding central to Dunhuang art.

3 Dunhuang-Bench

We introduce Dunhuang-Bench, a multimodal benchmark for cultural understanding over Dunhuang art. It consists of 486 images and 22,970 QA pairs across three tasks, including 11,688 visual perception questions, 2,020 multi-turn dialogues totaling 8,080 QA pairs for knowledge reasoning, and 3,202 multiple-choice questions for artistic appreciation. Fig. 2 illustrates the benchmark construction pipeline, consisting of Data Sourcing and Preprocessing, Task Generation, and Validation.

3.1 Theory-grounded Framework for Cultural Understanding

Dunhuang-Bench is informed by two complementary art-historical theories. Panofsky’s iconology models progressively deeper understanding of pictorial content, from pre-iconographical description to iconographical and iconological interpretation, focusing on culturally grounded meaning (Panofsky, 1918). Since this theory does not explicitly address artistic analysis, we additionally adopt the formal analytic tradition (Arnheim, 1954), which emphasizes composition, color, and stylistic struc-

ture as core dimensions of visual interpretation.

Accordingly, *Task 1 (Visual Perception)* corresponds to pre-iconographical description, *Task 2 (Knowledge Reasoning)* aligns with iconographical and iconological interpretation, and *Task 3 (Artistic Appreciation)* is grounded in formal analysis, together defining theory-grounded multimodal understanding in Dunhuang art. More details can be found in Appendix C.

3.2 Data Sourcing and Preprocessing

We construct Dunhuang-Bench from two authoritative sources: *Dunhuang Art Dictionary* and *Dunhuang Studies Dictionary*. First, we collect 486 image-text pairs from the *Dunhuang Art Dictionary*, covering representative Dunhuang artworks including murals, painted sculptures, and silk paintings. Second, we curate a set of 5,780 stylistic terms from the *Dunhuang Studies Dictionary*. Text descriptions are processed using Dolphin OCR (Feng et al., 2025) and manually corrected by Dunhuang studies experts. All images are converted to RGB and resized to an average resolution of 310×341 pixels. Finally, we employ Doubao-Seed-1-6-Thinking-250715 (Seed, 2025) (referred to as Doubao-1.6) to parse the verified descriptions into a **structured knowledge base** including 16,540 fields. This shared representation serves as the foundation for all three tasks.

3.3 Task 1: Visual Perception

Aligning with the pre-iconographical level of iconology theory (Panofsky, 2018), Task 1 (Visual Perception) evaluates the fundamental recognition capabilities of MLLMs. We derive the question-answer pairs from the “objective visual facts” within our structured knowledge base, encompassing four subcategories: *Counting*, *Color*, *Direction*, and *Behavior Recognition*. Specifically, answers are directly instantiated from factual fields, while Doubao-1.6 is employed to generate corresponding questions conditioned on these fixed answers. To enhance linguistic diversity, we paraphrase 4,374 initial seed questions into three semantically equivalent variants each, totaling 17,496 silver-standard VQA samples.

3.4 Task 2: Multi-turn Knowledge Reasoning

Beyond visual recognition in Task 1, our Task 2 evaluates multi-turn cultural knowledge reasoning. This task assesses an MLLM’s ability to integrate visual evidence with structured cultural knowledge

through sequential inquiry. Each dialogue comprises four turns that progressively advance from basic identification to culturally grounded interpretation, with each turn conditioned on the preceding context. This design aligns with the second (*iconographical analysis*) and third (*iconological interpretation*) levels of iconology theory. To ensure the conversations are both analytically structured and natural, we simulate a pedagogical scenario: an art history student seeking a guided explanation of artwork from an expert (MLLM).

For each dialogue turn, ground-truth answers are formulated as lists of phrases or terms extracted from our structured knowledge base, covering narrative events, key figures, and culturally grounded interpretation. This design requires models to progressively synthesize visual evidence with domain-specific knowledge. We prompt Doubao-1.6 to generate the “student” queries that logically connect within a four-turn flow:

1. **Turn 1 (Basic Info):** Queries high-level meta-data across four dimensions: *name*, *source*, *dynasty*, and *type*.
2. **Turn 2 (Elements):** A question asking to identify key cultural elements. Identifies key cultural and iconographical elements.
3. **Turn 3 (Specific Details):** A follow-up question focusing on a specific element identified in the previous turn.
4. **Turn 4 (Deeper Meaning):** A final probe into the symbolic meaning or overarching themes.

This pipeline produces 2,430 silver-standard multi-turn dialogues (9,720 QA pairs in total).

3.5 Task 3: Artistic Appreciation

Our Task 3 is grounded in the theory of formal analytic tradition (Arnheim, 1954), evaluating artistic appreciation through composition and layout, color usage, and form and style. For each artwork, Doubao-1.6 is prompted with the corresponding field in the structured knowledge base to generate an MCQ stem and a correct answer phrase. In total, this stage produces 4,010 silver-standard MCQs.

Hard-Negative Retrieval. Distractor options are retrieved from stylistic option fields in the structured knowledge base and indexed for category-constrained vector similarity search. Given the correct answer, we select hard negatives from fixed

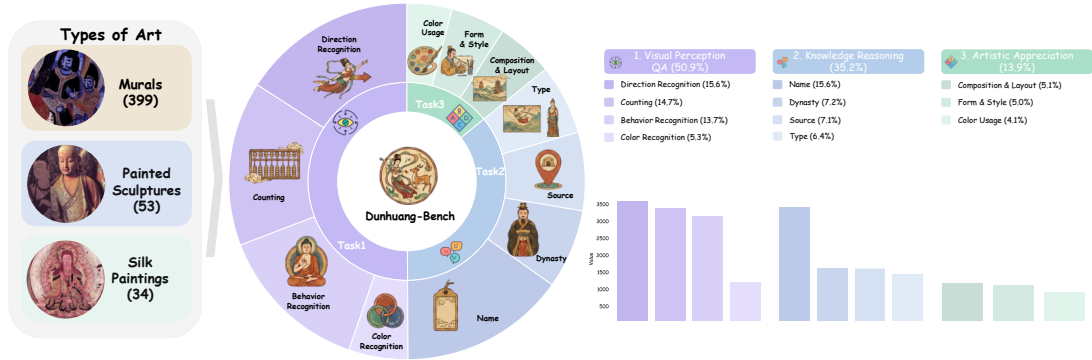


Figure 3: Hierarchical task structure and question distribution of Dunhuang-Bench, illustrating the three-task design with associated capability dimensions. The right panel reports the count and proportion of questions.

intermediate similarity ranks (10th, 20th, and 30th nearest neighbors).

3.6 LLM-assisted Human Validation

We adopt a unified two-stage validation pipeline across all tasks, combining GPT-5.1-assisted screening with expert human validation to ensure visual correctness, cultural accuracy, and stylistic rigor. In the first stage, GPT-5.1 applies task-specific criteria to filter low-quality candidates; in the second stage, all GPT-approved samples are jointly reviewed by five annotators with expertise in Dunhuang studies. Consequently, Task 1 was refined from 17,496 silver samples to 11,804 GPT-approved and 11,688 final human-validated instances. Similarly, Task 2 was pruned from 9,720 to 8,080 dialogue turns, and Task 3 from 4,010 to 3,202 MCQs.

3.7 Dataset Statistics

Dunhuang-Bench consists of 486 images and 22,970 QA pairs. As shown in Fig. 3, it spans 11,688 visual-perception pairs (Task 1), 2,020 multi-turn reasoning dialogues (8,080 pairs; Task 2), and 3,202 artistic appreciation MCQs (Task 3). These statistics ensure a balanced and hierarchical evaluation of MLLM cultural competence.

4 Experimental Setup

4.1 Evaluated Models

We extensively evaluate 20 MLLMs, including 4 commercial closed-source models (Gemini-2.5-Pro/Flash (Comanici et al., 2025), GPT-5.1 (OpenAI, 2025b), and Doubao-Seed-1.6-Thinking (Seed, 2025)) as well as 16 open-source MLLMs, including Qwen2.5-VL (Bai et al., 2025b) and Qwen3-VL (Bai et al., 2025a), InternVL3 (Zhu

et al., 2025) and InternVL3.5 (Wang et al., 2025a), GLM-4V (GLM et al., 2024), and DeepSeek-VL-2-Tiny (Wu et al., 2024). We set the temperature to 0.0 for stable evaluation. Open-source models up to 9B parameters are deployed locally on two NVIDIA RTX A6000 GPUs (48GB each), while larger models are accessed via APIs.

4.2 Evaluation Protocol and Metrics

For the visual perception task, counting questions are evaluated using numeric accuracy, while behavior recognition, direction recognition, and color recognition are assessed using the mean of BERTScore (Zhang et al., 2020) and ANLS (Biten et al., 2019). For the knowledge reasoning task, evaluation focuses on attribute-level reasoning rather than error propagation across turns. To this end, each turn is provided with the **ground-truth answers from all previous turns**. Since these gold answers are stored as structured string lists, we convert them into fluent natural-language sentences using Doubao-1.6 during preprocessing. Model outputs are then scored by a GPT-5-mini (OpenAI, 2025a) evaluator following a structured checklist that assesses factual grounding and reasoning coherence. Artistic appreciation is evaluated via Accuracy and Normalized Rank Score (NRS), which measures the proximity of predicted options to the ground truth (see Appendix A for details).

4.3 CoT and Few-shot Prompting Setup

We evaluate Qwen2.5-VL-7B-Instruct, InternVL3-8B, GLM-4V-9B, Doubao-Seed-1.6-Thinking, and Qwen3VL-235B-A22B-Instruct under Chain-of-Thought (CoT) prompting. Specifically, models are instructed to generate intermediate reasoning steps prior to the final answer, following a unified prompt template. Few-shot evaluation is further

conducted on Qwen2.5-VL-7B and InternVL3-8B using two-shot exemplars from the *Dunhuang Art Dictionary*. These external examples serve as in-context guidance and are strictly disjoint from our benchmark to prevent data leakage.

5 Experiment and Analysis

5.1 Main Results

Overall performance remains limited on Dunhuang art VQA. As shown in Table 2, the strongest model, InternVL3.5-241B-A28B, achieves an overall score of 50.9% averaged across 3 tasks, while the mean performance across all models is 43.92%. Among the three tasks, Task 2 exhibits the lowest average performance, with a mean score of only 19.2%. Additionally, the weakest model, DeepSeek-VL-2-Tiny, attains an overall score of 30.57% and achieves merely 14.07% on Task 2, indicating particularly limited performance in multi-turn cultural knowledge reasoning.

Open-source models show emerging advantages over closed-source systems. InternVL3.5-241B-A28B achieves the highest overall score of 50.9%, surpassing all evaluated closed-source models, and is only slightly behind Gemini-2.5-Pro on Task 2, where it obtains 25.5% compared to 26.7%. On Task 3, Qwen3-VL-235B-A22B-Thinking and InternVL3-78B reach 83.5% and 83.8%, respectively, exceeding the performance of closed-source models such as Gemini-2.5-Pro at 78.7%. These results indicate that recent open-source MLLMs have become competitive in Dunhuang art domain.

Scaling improves overall performance but not consistently across tasks. Parameter scaling yields clear overall gains within families: Qwen2.5-VL rises from 37.28% (3B) to 45.15% (72B), and InternVL3 increases from 41.55% (2B) to 47.93% (78B). However, larger models do not uniformly outperform their smaller counterparts on all subtasks. These inconsistencies indicate that scaling alone is insufficient for robust cultural understanding in Dunhuang art, and can even exacerbate errors without reliable visual or cultural grounding.

Reasoning may amplify misgrounded cultural interpretations. Table 2 shows that increased model capacity or explicit reasoning may not consistently benefit culturally sensitive tasks. For instance, while Qwen3-VL-235B-A22B-Thinking achieves a higher overall average than its Instruct

variant (49.31% vs. 48.78%), its performance on Task 2 drops significantly (22.52% vs. 25.05%). These results show that misgrounded visual or cultural cues can be amplified by deeper reasoning, revealing the limits of reasoning-centric approaches for understanding Dunhuang art.

Framework-level changes in MLLMs produce non-uniform cultural improvements. InternVL3.5 adopts a revised framework that incorporates mechanisms such as Cascade RL and enhanced text-centric training data, reflecting a design shift toward improved general reasoning (Wang et al., 2025a). However, these framework-level changes do not consistently enhance performance on culturally grounded visual subtasks. At smaller scales, InternVL3-2B outperforms InternVL3.5-2B on Task 2 (16.92% vs. 14.18%). This suggests that design choices favoring reasoning and text capabilities may incur a trade-off in fine-grained visual perception, particularly at lower parameter counts. In contrast, at sufficiently large scales, increased parameter capacity can compensate for these trade-offs: InternVL3.5-241B-A28B achieves the highest overall average score (50.90%), outperforming smaller InternVL3 variants across all tasks. This suggests that framework redesign mainly redistributes modeling capacity, whereas performance gains on Dunhuang art likely stem from increased model scale.

5.2 Subtask Analysis

Task 1. Task 1 evaluates visual perception, covering four subcategories: Counting, Behavior Recognition, Direction Recognition, and Color Recognition. As shown in Table 2, models perform relatively better on color recognition and counting, which rely on global visual cues (e.g., CR averages 48.75%), but exhibit noticeably lower accuracy on subtasks such as behavior recognition and direction recognition (BR and DR average 33.89% and 36.75%, respectively). This gap holds across models and reflects the challenging visual grounding demands of Dunhuang art.

Task 2. Task 2 is evaluated under a gold-history setting to isolate intrinsic stage difficulty. As shown in Table 3, models benefit from structured context in early turns, with accuracy increasing from Turn 1 to Turn 2 (e.g., Gemini-2.5-Pro: from 16.08% to 29.99%). However, performance drops noticeably at Turn 3 across models (e.g., from 29.99% to 19.95%), revealing a bottleneck in fine-grained de-

Model	Task 1 (Visual Perception)					Task 2 (Knowledge Reasoning)					Task 3 (Artistic Appreciation)			Average	
	Total	CT	BR	DR	CR	Total	SC	NM	DY	TY	Total	CU	LY		FS
Closed-source Models															
Gemini-2.5-Pro	40.01	39.63	35.54	36.27	48.58	26.70	25.19	27.12	27.42	27.07	78.68	71.77	81.33	82.93	48.46
Gemini-2.5-Flash	39.14	34.80	36.04	36.28	49.42	22.27	21.70	21.83	23.64	21.92	77.95	70.68	81.41	81.77	46.45
GPT-5.1	39.02	36.97	33.04	36.86	49.22	23.37	22.84	23.31	23.95	23.40	82.78	73.83	86.92	87.58	48.39
Doubao-Seed-1.6-Thinking	42.08	38.03	36.48	38.37	51.43	23.52	24.80	23.26	24.19	21.81	80.96	75.14	83.39	84.36	48.85
Open-source Models															
Qwen2.5VL-3B-Instruct	34.77	24.61	31.03	34.07	49.37	13.20	11.84	12.16	16.64	12.15	63.86	58.09	65.66	67.83	37.28
Qwen2.5VL-7B-Instruct	36.15	32.23	31.91	35.12	45.33	12.55	11.87	11.21	14.78	12.33	66.67	60.22	69.19	70.60	38.46
Qwen2.5VL-72B-Instruct	37.99	37.00	32.46	36.22	46.29	17.91	17.41	16.54	19.44	18.23	79.54	73.94	82.10	82.57	45.15
Qwen3VL-8B-Instruct	38.80	35.96	35.18	38.17	45.89	19.40	18.32	18.07	20.73	20.48	73.85	64.73	76.68	80.14	44.02
Qwen3VL-30B-A3B-Instruct	41.18	39.72	34.94	39.12	50.92	22.35	22.21	20.92	24.21	22.05	77.56	71.77	79.78	81.14	47.03
Qwen3VL-32B-Instruct	39.02	33.15	34.52	38.75	49.66	24.30	24.22	23.53	24.97	24.47	79.14	72.51	81.98	82.94	47.49
Qwen3VL-235B-A22B-Instruct	40.16	35.10	35.37	39.73	50.42	25.05	24.46	24.19	25.56	26.00	81.12	75.14	83.38	84.84	48.78
Qwen3VL-235B-A22B-Thinking	41.89	37.90	37.71	38.78	53.17	22.52	22.97	22.41	23.69	20.99	83.53	77.85	86.32	86.42	49.31
InternVL3-2B	36.25	31.01	31.66	35.49	46.83	16.92	15.56	15.57	19.49	17.06	71.48	65.15	70.91	78.37	41.55
InternVL3-8B	38.91	33.26	33.00	37.40	51.98	16.11	16.63	14.76	17.86	15.18	74.11	63.95	78.83	79.54	43.04
InternVL3-78B	40.94	42.89	33.88	39.31	47.69	19.02	19.74	16.96	21.62	17.76	83.84	77.20	86.49	87.85	47.93
InternVL3.5-1B	35.18	26.57	31.75	34.88	47.51	12.39	10.31	11.82	14.10	13.34	57.89	49.51	57.75	66.40	35.15
InternVL3.5-2B	37.31	31.72	33.01	34.71	49.80	14.18	13.34	13.01	16.86	13.50	64.91	55.27	66.09	73.37	38.80
InternVL3.5-241B-A28B	42.21	41.02	36.16	37.88	53.79	25.50	25.71	24.43	26.36	25.49	84.98	77.63	88.04	89.28	50.90
Deepseek-VL-2-Tiny	33.59	25.09	32.01	32.12	45.13	14.07	12.68	12.18	16.16	15.28	44.04	39.74	42.60	49.78	30.57
GLM-4V-9B	35.94	33.71	32.03	35.39	42.65	12.59	13.46	10.94	13.24	12.70	73.95	64.93	76.94	79.98	40.83
Average	38.53	34.52	33.89	36.75	48.75	19.20	18.76	18.21	20.75	19.06	74.04	66.95	76.29	78.88	43.92

Table 2: Unified category-wise evaluation results. Task 1 includes Direction Recognition (DR), Counting (CT), Behavior Recognition (BR), and Color Recognition (CR); Task 2 includes Name (NM), Dynasty (DY), Source (SC), and Type (TY); Task 3 includes Composition & Layout (CU), Form & Style (FS), and Color Usage (LY). Best results are highlighted in **bold**.

Model	Turn 1	Turn 2	Turn 3	Turn 4
Closed-source Models				
Gemini-2.5-Pro	16.08	29.99	19.95	41.13
Gemini-2.5-Flash	13.94	21.82	15.03	37.95
GPT-5.1	13.03	23.81	17.82	38.77
Doubao-Seed-1.6-Thinking	13.71	24.92	14.89	40.48
Open-source Models				
Qwen2.5VL-3B-Instruct	7.85	14.32	6.66	23.15
Qwen2.5VL-7B-Instruct	9.43	10.33	7.13	22.16
Qwen2.5VL-72B-Instruct	12.43	14.73	11.39	31.85
Qwen3VL-8B-Instruct	12.46	18.31	11.89	33.69
Qwen3VL-30B-A3B-Instruct	12.12	22.67	13.83	39.55
Qwen3VL-32B-Instruct	13.98	24.46	15.57	42.48
Qwen3VL-235B-A22B-Instruct	15.39	25.60	16.05	42.34
Qwen3VL-235B-A22B-Thinking	13.56	24.53	14.55	37.48
InternVL3-2B	9.43	19.85	8.93	28.29
InternVL3-8B	11.18	16.21	9.07	26.86
InternVL3-78B	12.82	16.91	12.85	31.80
InternVL3.5-1B	6.84	12.72	7.02	22.41
InternVL3.5-2B	9.02	14.23	7.50	25.01
InternVL3.5-241B-A28B	14.98	23.22	19.44	43.40
Deepseek-VL-2-Tiny	12.86	12.81	8.56	20.30
GLM-4V-9B	10.46	9.46	8.77	20.18

Table 3: Turn-wise score (%) on Task 2. Best results are highlighted in **bold**.

tail reasoning. Accuracy rebounds around Turn 4, mainly due to earlier semantic scaffolding rather than independent reasoning.

Task 3. In Task 3, evaluations based on hard accuracy indicate that overall model performance remains limited. However, when adopting the Normalized Rank Score (NRS), all models exhibit a substantial performance increase, resulting in a consistent gap between the two metrics. As shown in Table 4, strong models typically achieve NRS scores in the high 80% or above (e.g., GPT-5.1: 89.30%). This indicates that models can capture coarse stylistic cues but struggle to distinguish closely related Dunhuang art terms.

As a complementary evaluation, we reformulate Task 3 by hiding candidate options and requiring models to answer in a blind manner. Under this set-

Model	Total	CU	LY	FS
Closed-source Models				
Gemini-2.5-Pro	86.91	83.06	88.35	89.31
Gemini-2.5-Flash	86.41	82.63	88.53	88.08
GPT-5.1	89.30	83.68	91.62	92.61
Doubao-Seed-1.6-Thinking	88.80	85.74	89.99	90.68
Open-source Models				
Qwen2.5VL-3B-Instruct	77.15	73.83	78.43	79.18
Qwen2.5VL-7B-Instruct	78.93	74.86	80.67	81.26
Qwen2.5VL-72B-Instruct	88.04	84.47	89.56	90.08
Qwen3VL-8B-Instruct	84.13	78.93	85.47	88.00
Qwen3VL-30B-A3B-Instruct	86.59	83.35	87.75	88.68
Qwen3VL-32B-Instruct	87.42	83.61	88.89	89.77
Qwen3VL-235B-A22B-Instruct	88.62	85.18	89.85	90.82
Qwen3VL-235B-A22B-Thinking	89.99	86.61	91.51	91.84
InternVL3-2B	82.35	79.04	81.50	86.51
InternVL3-8B	84.05	77.45	87.21	87.49
InternVL3-78B	90.23	86.79	91.57	92.34
InternVL3.5-1B	73.56	68.08	74.10	78.49
InternVL3.5-2B	77.92	71.15	79.06	83.56
InternVL3.5-241B-A28B	91.06	87.08	92.63	93.48
Deepseek-VL-2-Tiny	63.39	60.22	62.68	67.26
GLM-4V-9B	84.32	79.04	85.83	88.08

Table 4: Task 3 results (%) of Normalized Rank Score. Best results are highlighted in **bold**.

Model	Total	CU	CL	FS
Gemini-2.5-Flash	60.99	62.11	61.62	59.25
Gemini-2.5-Pro	62.14	64.71	60.15	61.57
GPT-5.1	61.08	61.35	60.33	61.57
InternVL3.5-241B-A28B	54.57	57.33	54.99	51.39

Table 5: Task 3 results (%) under the blind-answer evaluation. Best results are highlighted in **bold**.

ting, performance drops to around 55%–62% (Table 5). This gap indicates that the high performance observed in the multiple-choice setting is partially facilitated by option visibility, which constrains the answer space and provides implicit semantic cues. The substantial degradation under blind answering further highlights the intrinsic difficulty of art appreciation, which demands precise aesthetic discrimination beyond option-level matching.

Model	Task1 (Visual Perception)					Task2 (Knowledge Reasoning)					Task3 (Art Appreciation)				Average
	Total	CT	BR	DR	CR	Total	SC	NM	DY	TY	Total	CU	LY	FS	
Qwen2.5-VL-7B-Instruct (w/o CoT)	36.15	32.23	31.91	35.12	45.33	12.55	11.87	11.21	14.78	12.33	66.67	60.22	69.19	70.60	38.46
Qwen2.5-VL-7B-Instruct (w/ CoT)	35.47	31.76	31.35	34.84	43.92	9.85	8.86	8.58	11.74	10.23	64.11	56.38	68.28	67.68	36.48
	-0.68	-0.47	-0.56	-0.28	-1.41	-2.70	-3.01	-2.63	-3.04	-2.10	-2.56	-3.84	-0.91	-2.92	-1.98
InternVL3-8B (w/o CoT)	38.91	33.26	33.00	37.40	51.98	16.11	16.63	14.76	17.86	15.18	74.11	63.95	78.83	79.54	43.04
InternVL3-8B (w/ CoT)	37.26	30.18	33.54	35.77	49.53	13.67	13.62	12.52	15.41	13.12	70.18	59.28	75.47	75.78	40.37
	-1.65	-3.08	+0.54	-1.63	-2.45	-2.44	-3.01	-2.24	-2.45	-2.06	-3.93	-4.67	-3.36	-3.76	-2.67
GLM-4V-9B (w/o CoT)	35.94	33.71	32.03	35.39	42.65	12.59	13.46	10.94	13.24	12.70	73.95	64.93	76.94	79.98	40.83
GLM-4V-9B (w/ CoT)	34.61	29.38	31.57	34.73	42.74	12.64	13.17	10.53	13.98	12.89	72.98	63.95	76.25	78.73	40.08
	-1.33	-4.33	-0.46	-0.66	+0.09	+0.05	-0.29	-0.41	+0.74	+0.19	-0.97	-0.98	-0.69	-1.25	-0.75
Doubao-Seed-1.6-Thinking (w/o CoT)	41.08	38.03	36.48	38.37	51.43	23.52	24.80	23.26	24.19	21.81	80.96	75.14	83.39	84.36	48.52
Doubao-Seed-1.6-Thinking (w/ CoT)	39.73	33.54	36.29	38.07	51.04	24.58	25.72	24.40	25.04	23.17	81.27	74.81	83.65	85.34	48.53
	-1.35	-4.49	-0.19	-0.30	-0.39	+1.06	+0.92	+1.14	+0.85	+1.36	+0.31	-0.33	+0.26	+0.98	+0.01
Qwen3VL-235B-A22B-Instruct (w/o CoT)	40.16	35.10	35.37	39.73	50.42	25.05	24.46	24.19	25.56	26.00	81.12	75.14	83.38	84.84	48.78
Qwen3VL-235B-A22B-Instruct (w/ CoT)	39.25	34.77	35.21	38.78	48.26	24.47	24.12	23.37	25.81	24.56	80.76	74.70	82.87	84.72	44.79
	-0.91	-0.33	-0.16	-0.95	-2.16	-0.58	-0.34	-0.82	+0.25	-1.44	-0.36	-0.44	-0.51	-0.12	-3.99

Table 6: Unified category-wise results (%) for models with CoT variants (w/ CoT). Cells annotated with + indicate performance gains relative to the non-CoT baseline (w/o CoT), whereas - denotes performance degradation.

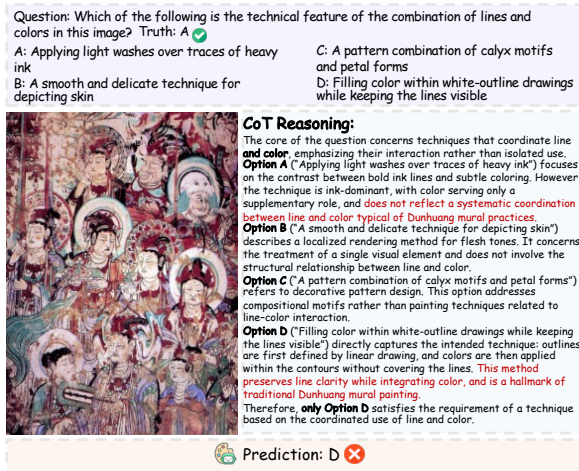


Figure 4: CoT reasoning of InternVL3-8B on a specific case of Task 3; the red highlight denotes the key failure step. Chinese version is provided in the Appendix H.

5.3 Impact of CoT Reasoning

Introducing CoT can not yield consistent improvements across tasks, and may lead to degraded performance on fine-grained visual tasks. As shown in Table 6, enabling CoT may reduce MLLMs' performance on Task 1, such as Qwen2.5-VL-7B (36.15% to 35.47%) and InternVL3-8B (38.91% to 37.26%), indicating that explicit reasoning chains cannot compensate for weak visual grounding. A similar trend is observed in Task 2, where Qwen2.5-VL-7B drops from 12.55% to 9.85%, suggesting that additional reasoning steps may amplify misaligned cultural or perceptual cues. Also, Task 3 shows no systematic benefit from CoT either, with models such as Qwen2.5-VL-7B decreasing from 66.67% to 64.11%. As illustrated in Fig. 4, CoT encourages the model to rely on canonical stylistic definitions instead of image-specific visual evidence, leading to incorrect judgments. Overall, in the Dunhuang context, CoT effectiveness is

constrained by visual and cultural grounding rather than reasoning form.

5.4 Effect of Few-shot Demonstrations

As shown in Fig. 5, few-shot prompting yields modest and task-dependent effects rather than uniform improvements. The overall gains are mainly driven by Task 1, which focuses on visual perception. For example, Qwen2.5VL-7B-Instruct improves its Task1 total from 36.15% (0-shot) to 39.40% (4-shot), and InternVL3-8B increases from 38.91% (0-shot) to 39.01% (4-shot), suggesting that few-shot demonstrations help models better align with task formats and surface-level visual patterns. In contrast, performance on Task 2 slightly degrades under few-shot prompting (e.g., InternVL3-8B drops from 16.11% to 15.66%), while Task 3 remains negative. Overall, few-shot prompting is effective for perception grounded in direct visual evidence, but provides limited and unstable benefits for knowledge reasoning and artistic appreciation.

6 Conclusion

We present Dunhuang-Bench, a theory-grounded multimodal benchmark for evaluating cultural understanding in MLLMs through Dunhuang art. It is a large-scale benchmark comprising 486 images and 22,970 QA pairs, integrating three task formats: QA with Text Description, Multi-turn Dialogue, and QA with Choices. Following Panofsky's iconological theory, we design Task 1 and Task 2 to evaluate visual perception and knowledge reasoning, while Task 3 is grounded in the formal analytic tradition to assess artistic appreciation. Experiments on 20 mainstream MLLMs highlighting the challenges of culturally grounded multimodal understanding. Additionally, we provide an in-depth analysis of CoT and few-shot prompting across

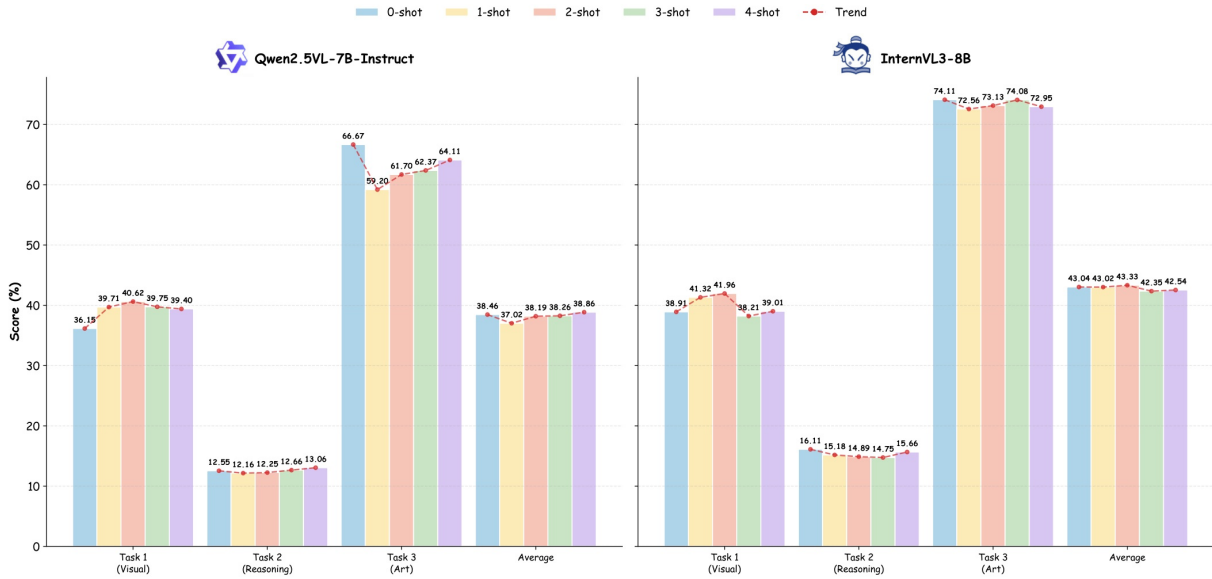


Figure 5: Few-shot evaluation results of Qwen2.5VL-7B-Instruct and InternVL3-8B.

tasks, offering insights into future research directions. We believe Dunhuang-Bench can serve as a foundational benchmark for advancing cultural understanding in Dunhuang art.

Limitations

Although we analyze the effects of CoT and few-shot prompting, our study is limited to inference-time prompting strategies. We do not investigate domain-specific fine-tuning of MLLMs on Dunhuang art, such as continued pretraining, which may further enhance culturally grounded visual understanding and therefore remain important directions for future work.

Ethical Consideration

Dunhuang art embodies profound religious, historical, and artistic significance. When using the Dunhuang-Bench dataset, special care must therefore be taken to avoid historical bias and cultural misinterpretation in modeling and evaluation. This dataset is intended solely for academic research to assess multimodal models on culturally grounded visual understanding and reasoning. All data sources comply with copyright requirements, and the dataset must not be used for commercial or unethical purposes.

Acknowledgments

This research was funded by the Joint Fund of the National Natural Science Foundation of

China under Grant No. U2574209, the National Natural Science Foundation of China under Grant No. 62276258, the Zhejiang Province Leading Earth Goose Program under Grant No. 2026LDC01028(XC), and the Suzhou Science and Technology Development Planning Programme under Grant No. ZX2023176.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Rudolf Arnheim. 1954. *Art and visual perception: A psychology of the creative eye*. Univ of California Press.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. [Qwen2.5-vl technical report](#). *arXiv preprint arXiv:2502.13923*.
- Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís 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.

- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Hao Feng, Shu Wei, Xiang Fei, Wei Shi, Yingdong Han, Lei Liao, Jinghui Lu, Binghong Wu, Qi Liu, Chunhui Lin, and 1 others. 2025. Dolphin: Document image parsing via heterogeneous anchor prompting. *arXiv preprint arXiv:2505.14059*.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.
- Peiyu Hu, Wayne Lu, and Jia Wang. 2026. From ids to semantics: A generative framework for cross-domain recommendation with adaptive semantic tokenization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14874–14882.
- Nithish Kannan, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: cultural competence in text-to-image models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 13716–13747.
- Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, and 1 others. 2024. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095.
- Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025. Mcs-bench: A comprehensive benchmark for evaluating multimodal large language models in chinese classical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10435–10492.
- 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.
- OpenAI. 2025a. Gpt-5 system card. <https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf>. System-level documentation for the GPT-5 model.
- OpenAI. 2025b. Gpt-5.1 instant and gpt-5.1 thinking system card addendum. https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf. System-level documentation for GPT-5.1 Instant and GPT-5.1 Thinking models.
- Erwin Panofsky. 2018. *Studies in iconology: Humanistic themes in the art of the renaissance*. Routledge.
- 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.
- 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 *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 11479–11505.
- Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22238–22254.
- Florian Schneider, Carolin Holtermann, Chris Biemann, and Anne Lauscher. 2025. **GIMMICK: Globally inclusive multimodal multitask cultural knowledge benchmarking**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9605–9668, Vienna, Austria. Association for Computational Linguistics.
- ByteDance Seed. 2025. Seed1. 6 tech introduction. Accessed on September, 28:2025.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
- Ashish V Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

- Yihui Wang and Xiaodong Wu. 2023. Current progress on murals: distribution, conservation and utilization. *Heritage Science*, 11(1):61.
- Yuxuan Wang, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, Wanxiang Che, and Hongyang Chen. 2025b. Cvlue: A new benchmark dataset for chinese vision-language understanding evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8196–8204.
- 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.
- Roderick Whitfield, Susan Whitfield, and Neville Agnew. 2015. *Cave Temples of Mogao at Dunhuang: Art History on the Silk Road*. Getty Publications.
- Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-v12: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.
- Junyi Yuan, Jian Zhang, Fangyu Wu, Huanda Lu, Dongming Lu, and Qiufeng Wang. 2025. Towards cross-modal retrieval in chinese cultural heritage documents: Dataset and solution. In *International Conference on Document Analysis and Recognition*, pages 570–586. Springer.
- Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, and 1 others. 2025a. Can mllms understand the deep implication behind chinese images? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14402.
- Jian Zhang, Junyi Guo, Junyi Yuan, Huanda Lu, Yanlin Zhou, Fangyu Wu, Qiufeng Wang, and Dongming Lu. 2025b. Llm-driven completeness and consistency evaluation for cultural heritage data augmentation in cross-modal retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19418–19428.
- 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*.
- Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Kiana Avestimehr, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth Narayanan, and 1 others. 2025c. Creating a lens of chinese culture: A multimodal dataset for chinese pun rebus art understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22473–22487.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internv13: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Evaluation Metrics

A.1 Exact-match Accuracy for Numerical Questions

For numerical questions (e.g., counting), we parse a number from both the ground truth and prediction (Arabic digits or common Chinese numerals). The score is:

$$\text{Acc}_{\text{num}} = \mathbb{I}[\hat{n} = n]. \quad (1)$$

Where n and \hat{n} are the ground-truth and predicted numbers, respectively, and $\mathbb{I}[\cdot]$ denotes the indicator function.

A.2 ANLS

For non-numerical answers, we compute ANLS based on the normalized Levenshtein distance between the ground-truth answer g and the model prediction p :

$$d_{\text{norm}} = \frac{\text{LD}(g, p)}{\max(|g|, |p|)}, \quad (2)$$

where $\text{LD}(\cdot, \cdot)$ denotes the Levenshtein edit distance, and $|\cdot|$ is the string length. With a threshold $\tau = 0.5$, the ANLS score is defined as:

$$\text{ANLS} = \begin{cases} 1 - d_{\text{norm}}, & d_{\text{norm}} < \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

A.3 BERTScore

We measure semantic similarity using BERTScore and report the F1 score between the predicted answer p and the ground-truth answer g :

$$\text{BERTScore} = \text{F1}_{\text{BERT}}(p, g), \quad (4)$$

where $\text{F1}_{\text{BERT}}(\cdot, \cdot)$ denotes the token-level F1 score computed from contextualized embeddings.

A.4 Category Score

Each category is reduced to a single scalar score:

$$\text{Score}_c = \begin{cases} \text{Acc}_{\text{num}}, & c \in \mathcal{C}_{\text{num}}, \\ \frac{\text{ANLS} + \text{BERTScore}}{2}, & \text{otherwise.} \end{cases} \quad (5)$$

A.5 Aggregated Scores

Average Counting Accuracy. We average Acc_{num} over numerical categories.

Total Score. We average category scores over all categories:

$$\text{TotalScore} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{Score}_c. \quad (6)$$

A.6 Keyword-based Multi-turn Dialogue Scoring

For Task2 multi-turn knowledge reasoning, we adopt a keyword-based automatic scoring scheme that evaluates both factual coverage and response quality, following the evaluation prompt in Appendix E.

For a given dialogue turn, let $\mathcal{W} = \{w_1, \dots, w_M\}$ denote the set of required keywords or key phrases specified for that turn. Each keyword w_i is assigned a discrete coverage score

$$s(w_i) \in \{1.0, 0.5, 0.0\}, \quad (7)$$

corresponding to *Yes*, *Partial*, and *No* coverage, respectively. The keyword coverage score is computed as

$$\text{Cov} = \frac{1}{M} \sum_{i=1}^M s(w_i), \quad (8)$$

where M is the number of required keywords for the turn.

In addition, each turn receives an overall dialogue quality score $Q \in \{1, \dots, 10\}$, predicted by the evaluator model according to a fixed rubric that considers linguistic fluency and contextual coherence, with priority given to keyword recall.

The final turn-level score is defined as the product of coverage and quality:

$$\text{Score}_{\text{turn}} = \text{Cov} \cdot \frac{Q}{10}, \quad (9)$$

which ensures that fluent but factually incomplete responses receive lower scores.

For a dialogue with T turns, the dialogue-level score is obtained by averaging turn-level scores:

$$\text{Score}_{\text{dialogue}} = \frac{1}{T} \sum_{t=1}^T \text{Score}_{\text{turn}}^{(t)}. \quad (10)$$

A.7 Normalized Rank Score (NRS)

For Task3 multiple-choice questions, we report a *Normalized Rank Score* (NRS) to measure how well a model’s predicted option aligns with the ground-truth option under a semantic similarity ranking.

For each question, let K denote the number of options (fixed to $K = 4$ in our benchmark), g the ground-truth option key, and \hat{o} the model-predicted option key. We embed the ground-truth option text t_g into a normalized vector $\mathbf{e}(t_g)$. For each option

k with text t_k , its similarity to the ground truth is defined as

$$s(k) = \begin{cases} 1, & k = g, \\ \cos(\mathbf{e}(t_k), \mathbf{e}(t_g)), & k \neq g. \end{cases} \quad (11)$$

All options are sorted by $s(k)$ in descending order. Let $\text{rank}(\hat{o}) \in \{1, \dots, K\}$ denote the rank of the predicted option. The NRS is computed as

$$\text{NRS} = \begin{cases} \frac{K - \text{rank}(\hat{o})}{K - 1} \cdot 100, & K > 1, \\ 100, & K = 1. \end{cases} \quad (12)$$

B Supplementary Experimental Results

Model	BR		DR		CR	
	BERTScore	ANLS	BERTScore	ANLS	BERTScore	ANLS
Closed-source Models						
Gemini-2.5-Pro	65.53	5.54	66.62	5.92	78.29	18.87
Gemini-2.5-Flash	66.22	5.86	66.03	6.52	80.28	18.56
GPT-5.1	64.53	1.56	67.53	6.20	80.38	18.06
GPT-5.1 (Step-by-step)	64.32	1.95	66.66	5.29	79.82	17.58
Doubao-Seed-1.6	66.54	6.43	68.37	8.37	81.20	21.67
Open-source Models						
Qwen2.5VL-3B-Instruct	61.58	0.49	64.78	3.37	82.40	16.34
Qwen2.5VL-7B-Instruct	62.40	1.41	65.90	4.34	78.68	11.99
Qwen2.5VL-72B-Instruct	63.66	1.26	66.86	5.59	79.45	13.13
Qwen3VL-8B-Instruct	65.83	4.53	68.25	8.10	77.90	13.88
Qwen3VL-30B-A3B-Instruct	66.04	3.84	69.15	9.10	81.45	20.39
Qwen3VL-32B-Instruct	65.44	3.59	68.78	8.73	80.37	18.94
Qwen3VL-235B-A22B-Instruct	66.29	4.44	69.73	9.72	81.29	19.56
Qwen3VL-235B-A22B-Thinking	66.82	8.59	68.30	9.26	83.15	23.19
InternVL3-2B	62.50	0.81	66.09	4.89	80.38	13.29
InternVL3-8B	64.25	1.74	67.55	7.24	83.04	20.92
InternVL3-78B	64.85	2.91	69.22	9.39	79.92	15.46
InternVL3.5-1B	62.54	0.96	65.10	4.66	80.76	14.26
InternVL3.5-2B	63.56	2.46	64.93	4.50	81.93	17.67
InternVL3.5-8B	62.59	0.70	65.39	3.59	80.60	16.60
InternVL3.5-241B-A28B	66.16	6.15	68.15	7.62	82.70	24.87
Deepseek-VL-2-Tiny	62.71	1.31	62.52	1.72	77.27	12.99
GLM-4V-9B	62.82	1.25	65.53	5.25	75.29	10.00

Table 7: Task 1 results (excluding Counting): BERTScore and ANLS (%).

Table 7 reports closed-source models achieve stronger and more stable performance, while top open-source models show competitive results across categories.

Shot	Task1 (Visual Perception)					Task2 (Knowledge Reasoning)					Task3 (Art Appreciation)					Average
	Total	CT	BR	DR	CR	Total	SC	NM	DY	TV	Total	CU	LY	FS		
Qwen2.5VL-7B-Instruct																
0-shot	36.15	32.23	31.91	35.12	45.33	12.55	11.87	11.21	14.78	12.33	66.67	60.22	69.19	70.60	38.46	
1-shot	39.71	36.88	32.79	37.31	51.87	12.16	11.03	11.17	14.05	12.39	59.20	50.49	62.05	65.06	37.02	
2-shot	40.62	39.45	32.76	37.48	52.79	12.25	11.02	11.41	14.40	12.18	61.70	53.20	64.97	66.93	38.19	
	+4.47	+7.22	+0.85	+2.36	+7.46	-0.30	-0.85	+0.20	-0.73	-0.15	+4.97	-7.02	-4.22	-3.67	-0.27	
InternVL3-8B																
0-shot	38.91	33.26	33.00	37.40	51.98	16.11	16.63	14.76	17.86	15.18	74.11	63.95	78.83	79.54	43.04	
1-shot	41.32	37.03	34.64	39.56	54.03	15.18	13.89	13.25	16.25	17.36	72.56	59.93	79.00	78.73	43.02	
2-shot	41.96	38.27	35.15	39.44	54.99	14.89	14.25	13.66	15.23	16.42	73.13	59.50	78.83	81.05	43.33	
	+3.00	+5.81	+2.15	+2.04	+3.01	-1.32	-2.38	-1.10	-2.63	+1.28	-0.98	-4.44	+0.00	+1.51	+0.29	

Table 8: Few-shot unified evaluation results across tasks.

Table 8 reports few-shot unified evaluation results across Task1-3 for representative models, showing scores under different shot settings.

C Preliminary: Theoretical Foundations

This work draws on two complementary art-historical traditions to ground cultural understanding and visual interpretation.

Iconology. Following Panofsky, iconology (Panofsky, 2018) conceptualizes image understanding as a hierarchical process with three levels. *Pre-iconographical description* concerns the recognition of directly observable visual elements such as figures, actions, and spatial relations. *Iconographical analysis* interprets these elements in terms of conventional motifs, themes, or subjects grounded in shared cultural knowledge, while *iconological interpretation* further situates them within broader historical, symbolic, and cultural frameworks. This theory provides a systematic account of how visual representations convey meaning beyond surface appearance. It motivates the design of **Task 1 (Visual Perception)**, aligned with pre-iconographical description, and **Task 2 (Knowledge Reasoning)**, aligned with iconographical and iconological interpretation.

Formal Analytic Tradition. The formal analytic tradition adopted in this work follows Arnheim’s perceptual understanding of art (Arnheim, 1954), which understands artworks primarily through the organization of visual form. This perspective emphasizes composition, color relations, balance, and structural coherence as central carriers of artistic meaning, rather than explicit narrative or iconographic content. By focusing on how visual elements are arranged and perceived as a whole, it provides a theoretical foundation for evaluating sensitivity to style and aesthetic structure. In our benchmark, this tradition directly grounds **Task 3 (Artistic Appreciation)**.

Together, these two theories offer complementary foundations that align with the three tasks, supporting a structured evaluation of cultural understanding in Dunhuang art.

D Validation Results

Task	Silver	GPT-5.1 Approved	GPT Pass (%)	Human-approved	Human Pass (%)
Task 1	17,496	11,804	67.47	11,688	99.02
Task 2	9,720	8,892	91.48	8,080	90.87
Task 3	4,010	3,571	89.05	3,202	89.67

Table 9: Overview of LLM-assisted and human validation across the three tasks. Pass rates are computed relative to the previous stage.

Table 9 summarizes the LLM-assisted and human validation process across tasks. The reported pass rates indicate the proportion of samples retained at each stage relative to the previous filtering step, reflecting both the effectiveness of automatic screening and the reliability of human validation.

E Task 2 Evaluation Prompt

```
Prompt: Keyword-based Multi-turn Dialogue Evaluation (Chinese)

角色说明:
你是一位敦煌学专家兼语言学家,负责评估多模态大模型在轮文化对话中的回答质量。
你的目标是进行高效、稳健、可复现的自动化评测。

输入信息:
• 当前对话轮次
• 历史对话上下文
• 当前用户提问
• 该轮对话中必须覆盖的关键信息点(关键词或关键词列表)
• 待评估的模型回答

评测任务一:关键词覆盖核对
请逐一检查模型回答是否覆盖了给定的每一个关键信息点,并为每个关键词生成一个检查项:
• Yes (1.0): 明确提及该关键词,或使用语义等价的准确表达。
• Partial (0.5): 涉及相关概念,但表述不完整或存在轻微偏差。
• No (0.0): 未提及该点,或出现相互矛盾的错误信息。

评测任务二:整体质量与连贯性评分
请给出一个1-10的整数评分,重点考虑以下维度:
1. 语言流畅性:表达是否自然、通顺;
2. 上下文连贯性:是否正确理解并承接历史对话内容。

输出格式要求:
请仅输出一个合法的JSON对象,不包含任何解释性文字或标记:

{
  "checklist_results": [
    {"keyword": "关键词1", "result": "Yes"},
    {"keyword": "关键词2", "result": "Partial"}
  ],
  "quality_score":
```

Figure 6: Prompt for LLM-based scoring of multi-turn cultural dialogue responses, used to evaluate keyword coverage and overall response quality in Task 2 (Chinese Version).

```
Prompt: Keyword-based Multi-turn Dialogue Evaluation (English)

Role Description:
You are an expert in Dunhuang studies and linguistics, responsible for evaluating the response quality of multimodal large language models in multi-turn cultural dialogues.

Your goal is to conduct efficient, robust, and reproducible automated evaluation.

Input Information:
• Current dialogue turn
• Dialogue history context
• Current user question

Required key information points for the current turn (a list of keywords or key phrases)
• Model-generated response to be evaluated

Evaluation Task 1: Keyword Coverage Verification
Please examine whether the model response covers each required key information point and generate a checklist item for every keyword:
• Yes (1.0): The keyword is explicitly mentioned, or an accurate semantically equivalent expression is used.
• Partial (0.5): The related concept is mentioned, but the expression is incomplete or contains minor deviations.
• No (0.0): The keyword is not mentioned, or the response contains contradictory or incorrect information.

Evaluation Task 2: Overall Quality and Coherence Scoring
Please assign an integer score from 1 to 10, focusing on the following dimensions:
1. Linguistic fluency: whether the response is natural and well-formed;
2. Contextual coherence: whether the response correctly understands and follows the dialogue history.

Output Format Requirements:
Please output only a valid JSON object, without any explanatory text or markers:

{
  "checklist_results": [
    {"keyword": "keyword1", "result": "Yes"},
    {"keyword": "keyword2", "result": "Partial"}
  ],
  "quality_score":
```

Figure 7: Prompt for LLM-based scoring of multi-turn cultural dialogue responses, used to evaluate keyword coverage and overall response quality in Task 2 (English Version).

F Task-specific Data Generation Prompts

```
Task 1
角色说明:
你是一名视觉问答出题助手。输入是一条来自结构化知识库的客观视觉事实,对应一件敦煌艺术作品。
任务目标:
生成单轮、图像可验证的问答对,用于评估模型对基础视觉属性的感知能力。
生成规则:
1. 问题仅围绕数量、姿态、空间位置、构图和颜色之一。
2. 禁止使用具体名称或专有名词。
3. 答案必须与输入的视觉事实逐字一致,不得改写。
4. 禁止生成任何涉及尺寸或比例的问题。
输出格式: {"question": "...", "answer": "..."}

Task 2
角色说明:
你是一名艺术史学习场景中的对话生成助手。
正在帮助一位艺术史专业的学生为一件特定的敦煌艺术品生成多轮视觉问答对话数据。
输入是一条来自结构化知识库的文物知识条目。
任务目标:
生成一个四轮对话,用于评估模型在多轮交互中整合视觉线索与专业文化知识的能力。
对话结构(必须遵循):
1. Turn 1: 基本信息(名称、来源、朝代或类型之一)。
2. Turn 2: 关键人物或文化要素识别。
3. Turn 3: 针对某一要素的细节追问。
4. Turn 4: 主题、象征或整体意义。
生成规则:
1. 所有答案必须严格来自输入的结构化知识。
2. 答案需要是列表形式,内容为结构化知识条目中的短语或词组。
3. 不得引入任何外部背景或推测性解释。
4. 对话语义递进、逻辑连贯。
输出格式: [{"question": "...", "answer": "..."}], [{"question": "...", "answer": "..."}], [{"question": "...", "answer": "..."}], [{"question": "...", "answer": "..."}]

Task 3
角色说明:
你是一名艺术史测评题目生成助手。输入是一条包含形式与风格分析信息的结构化知识条目。
任务目标:
生成一条选择题的题目和一个正确选项,评估模型对构图、色彩或风格差异的区分能力。
生成规则:
1. 题目必须明确对应构图、色彩或风格三者之一。
2. 正确选项必须直接来源于输入描述。
3. 避免主观或情感化表述,强调形式证据与术语使用的规范性。
输出格式: {"question": "...", "option": "..."}]
```

Figure 8: Task definitions and data construction prompts for the three tasks in Dunhuang-Bench, including Task 1 with single-turn visual QA, Task 2 with multi-turn dialogue, and Task 3 with single-choice questions, illustrating the distinct evaluation objectives and constraints of each task (Chinese Version).

```
Task 1
Role Description:
You are a visual question answering assistant. The input is an objective visual fact from a structured knowledge base, corresponding to a Dunhuang art work.
Mission Objectives:
Single-round, image-verifiable question-answer pairs are generated, which are used to evaluate the model's ability to perceive underlying visual attributes.
Generate rules:
1. The question revolves only around one of quantity, pose, spatial location, composition, and color.
2. The use of specific names or proper nouns is prohibited.
3. The answer must be consistent with the input word by word, and must not be rewritten.
4. Refrain from making any questions about size or scale.
Output format: {"question": "...", "answer": "..."}

Task 2
Role Description:
You're a dialogue generation assistant in an art history learning scenario. Is helping an art history student generate multiple rounds of visual question answering dialogue data for a specific Dunhuang artwork.
The input is a cultural relic knowledge entry from the structured knowledge base.
Mission Objectives:
A four-wheeled dialogue is generated and used to evaluate the model's ability to integrate visual cues with specialized cultural knowledge across multiple rounds of interaction.
Dialogue structure (must be followed):
Turn 1: Basic information (name, origin, dynasty, or one of the types).
Turn 2: Identification of key figures or cultural elements.
Turn 3: Ask questions about an element in detail.
Turn 4: Theme, symbol, or overall meaning.
Generate rules:
1. All answers must come strictly from the input structured knowledge.
2. The answer needs to be in the form of a list of phrases or phrases from a structured knowledge item.
3. No external background or inferential interpretation should be introduced.
4. The dialogue should be progressive and logically coherent.
Output format: [{"question": "...", "answer": "..."}], [{"question": "...", "answer": "..."}], [{"question": "...", "answer": "..."}], [{"question": "...", "answer": "..."}]

Task 3
Role Description:
You are an art history quiz generation assistant. The input is a structure that contains information about form and style analysis.
Knowledge items.
Mission Objectives:
Generate a multiple-choice question and a correct choice, and evaluate the model's discrimination of differences in composition, color, or style.
Generate rules:
1. The title must clearly correspond to one of three: composition, color, or style.
2. The correct option must be directly derived from the input description.
3. Avoid subjective or emotional statements and emphasize formal evidence and the normative use of terminology.
Output format: {"question": "...", "option": "..."}]
```

Figure 9: Task definitions and data construction prompts for the three tasks in Dunhuang-Bench, including Task 1 with single-turn visual QA, Task 2 with multi-turn dialogue, and Task 3 with single-choice questions, illustrating the distinct evaluation objectives and constraints of each task (English Version).

G CoT and Few-shot Prompts

Task 1

Zero-shot CoT
以下是一道基于图像的单一问答问题，请一步一步地思考，并在最后给出答案。
问题: {question}
答案: {answer}

Few-shot
以下是若干示例，请参考示例的格式回答最后一道问题。
示例 1
问题: {demo_q1}
答案: {demo_a1}

问题: {question}
答案: {answer}

Task 2

Zero-shot CoT
以下是一段多轮对话中的当前轮问题，请结合给定对话历史一步一步地思考，并在最后给出本轮答案。
对话历史: {dialogue_history}
当前问题: {question}
答案: {answer}

Few-shot
以下是若干示例对话片段，请参考示例在给定历史上下文下回答最后一道问题。
示例 1
对话历史: {demo_history1}
当前问题: {demo_q1}
答案: {demo_a1}

对话历史: {dialogue_history}
当前问题: {question}
答案: {answer}

Task 3

Zero-shot CoT
以下是一道单项选择题，请一步一步地思考，并在最后给出答案。
问题: {question}
选项: {options_str}
答案: {answer}

Few-shot
以下是若干示例，请参考示例的格式回答最后一道单项选择题。
示例 1
问题: {demo_q1}
选项: {demo_opt1}
答案: {demo_a1}

问题: {question}
选项: {options_str}
答案: {answer}

Figure 10: Chain-of-Thought (CoT) and few-shot prompting templates used for Task 1–3 in Dunhuang-Bench (Chinese Version).

Task 1

Zero-shot CoT
The following is a single-round question based on an image. Please think through it step by step and give your answer at the end.
Question: {question}
Answer: {answer}

Few-shot
Here are some examples. Please refer to the format in the examples when answering the last question.
Example 1
Question: {demo_q1}
Answer: {demo_a1}

Question: {question}
Answer: {answer}

Task 2

Zero-shot CoT
The following is the current question in a multi-turn dialogue. Please consider the given dialogue history step by step and provide your answer for this turn at the end.
Dialogue history: {dialogue_history}
Current question: {question}
Answer: {answer}

Few-shot
Here are some example dialogue snippets. Please refer to the examples and answer the last question within the given historical context.
Example 1
Dialogue history: {demo_history1}
Current question: {demo_q1}
Answer: {demo_a1}

Dialogue history: {dialogue_history}
Current question: {question}
Answer: {answer}

Task 3

Zero-shot CoT
The following is a multiple-choice question. Please think through it step by step and give your answer at the end.
Question: {question}
Options: {options_str}
Answer: {answer}

Few-shot
Here are some examples. Please refer to the format in the examples to answer the last multiple-choice question.
Example 1
Question: {demo_q1}
Options: {demo_opt1}
Answer: {demo_a1}

Question: {question}
Options: {options_str}
Answer: {answer}

Figure 11: Chain-of-Thought (CoT) and few-shot prompting templates used for Task 1–3 in Dunhuang-Bench (English Version).

H Case Study

The case shown in Fig. 12 highlights a clear task-level differentiation. The model performs reliably

Turn	Question	Ground Truth	Golden Utterance	Model Answer
1	这件作品是什么类型的?	故事画	这属于故事画类型。	这件作品是古代壁画，描绘了人物群像的壮观场景。
2	这幅画里主要有哪些元素?	华夏皇帝像; 特征	这幅画主要包含华夏皇帝像与特征。	画面中心: 身穿华丽红袍, 头戴冠冕居于中央, 周围簇拥着侍从马从, 背景点缀绿色华盖与繁复纹饰, 人物神态各异, 交相辉映, 凸显威严恢弘的氛围。
3	图里的华夏皇帝像有什么特点?	头戴冠冕, 身穿华服; 昂首阔步; 不可一世; 锦衣朱紫	图中的华夏皇帝头戴冠冕, 身穿青紫朱紫的华服, 昂首阔步的姿态传递着不可一世的威严。	画面中皇帝身穿华丽的红袍, 头戴冠冕, 姿态威严, 双手叉腰于膝前, 周围簇拥着侍从马从, 背景点缀绿色华盖与繁复纹饰, 人物神态各异, 交相辉映, 凸显威严恢弘的氛围。
4	图里华夏皇帝的形象有什么象征意义?	皇帝权威; 世俗权力干预宗教	《帝王问疾图》里的华夏皇帝形象呈现出微妙的姿态, 这种刻画象征着世俗权力对宗教的干预。	图中华夏皇帝形象传递着皇权权威, 昂首阔步的姿态暗喻帝王对法律的尊崇与扶持, 象征世俗权力对宗教精神的压制与干预。

Figure 12: Qualitative case study of InternVL3.5-241B-A28B on Dunhuang-Bench, demonstrating model behaviors across tasks. Green, yellow, and red highlights denote correct, partially correct, and incorrect answers, respectively (Chinese Version).

Turn	Question	Ground Truth	Golden Utterance	Model Answer
1	What type of work is this?	Story painting	This belongs to the story illustration type.	This work is a traditional Chinese painting of a group of figures and a crowd scene.
2	What are the main elements in this picture?	Portrait of a Chinese emperor; courtier	The picture mainly features portraits of Chinese emperors and their attendants.	In the painting, the emperor is dressed in a magnificent robe, surrounded by attendants in various poses. The background is filled with intricate patterns and colors, creating a rich and detailed scene.
3	What are the characteristics of the portrait of a Chinese emperor in the picture?	Wearing a crown and formal robes; dressed in high-ranking and distinctive; dressed in blue robe and red robe	The Chinese emperor in the picture wears a crown and a blue robe with red trim, the elegant demeanor exuding an air of nobility.	In the painting, the emperor is dressed in a magnificent robe, surrounded by attendants in various poses. The background is filled with intricate patterns and colors, creating a rich and detailed scene.
4	What is the symbolic meaning of the image of the Chinese emperor in the picture?	The emperor's serene face; leader; power; authority; with regalia	"The Emperor's Visit to the Slaves" presents an emperor figure, which emphasizes the importance of social power or religion.	The image depicts the Chinese emperor in his crown and robes, displaying his authority and power over the entire nation. The emperor's presence is a central focus, symbolizing the power and stability of the empire.

Figure 13: Qualitative case study of InternVL3.5-241B-A28B on Dunhuang-Bench, demonstrating model behaviors across tasks. Green, yellow, and red highlights denote correct, partially correct, and incorrect answers, respectively (English Version)

on Task 1 by accurately grounding visual perception, and achieves correct decisions on Task 3. In contrast, Task 2 reveals a critical limitation: across dialogue turns, the model increasingly drifts from the gold history, introduces unsupported details, and ultimately reverses the intended symbolic interpretation. This failure stems not from linguistic fluency, but from an inability to maintain consistent grounding between visual evidence, structured cultural knowledge, and turn-wise reasoning.

Question: 关于这幅图像，其线条与色彩配合的技法特点是以下哪项？

Truth: A

A: 焦墨痕中略施微染

C: 柿蒂纹外套叶瓣纹的纹样组合

B: 莹润细腻的肌肤表现手法

D: 在白画线纹内填色敷彩、色不掩线的技法



CoT Reasoning:

首先，题目核心是“线条与色彩配合的技法”，需聚焦于两者的互动关系。逐一分析选项：

1. 选项A：“焦墨痕中略施微染”侧重浓墨线条与淡彩渲染的结合，但“焦墨痕”以墨线为主，“微染”是局部淡彩，未体现“线条与色彩的系统性配合”，且不符合敦煌壁画常见的填色逻辑。

2. 选项B：“莹润细腻的肌肤表现手法”是针对肌肤的局部细节处理，属于单一元素的表现技巧，不涉及“线条与色彩”的配合。

3. 选项C：“柿蒂纹外套叶瓣纹的纹样组合”是图案的结构设计，属于纹样本身的组合方式，与“线条和色彩的技法配合”无关。

4. 选项D：“在白画线纹内填色敷彩、色不掩线的技法”直接对应“线条（白描勾勒的轮廓）”与“色彩（填充于线条内的颜料）”的核心配合——先以白描线条确定轮廓，再在轮廓内填充色彩，且色彩不覆盖线条，保持线条的清晰性。这是敦煌壁画等传统绘画中典型的“线描填色”技法，完全契合“线条与色彩配合”的要求。

综上，只有D选项符合“线条与色彩配合的技法”这一题目核心。


 Prediction: D

Figure 14: CoT reasoning of InternVL3-8B on a specific case; the red highlight marks the key failure step (Chinese Version).

I Screening and Validation Criteria for Dataset Construction

Table 10 summarizes the first-stage GPT-5.1 screening criteria and the second-stage human expert validation criteria for the three tasks.

Table 10: Two-stage quality control criteria for dataset construction across the three tasks.

Task	GPT-5.1 first-stage screening criteria	Second-stage human expert validation
Task 1: Visual Perception (QA-T)	(1) Observability : the question must rely only on directly observable visual facts, not external knowledge or inference. (2) Ambiguity filtering : samples with unclear targets, occlusion, or indistinguishable attributes are removed. (3) Linguistic clarity : questions must be unambiguous and semantically stable across paraphrases.	(1) Visual correctness : verified against the image. (2) Annotation rigor : confirm the question matches the intended visual attribute. (3) Category validity : ensure the sample fits the defined sub-task.
Task 2: Knowledge Reasoning (Multi-turn Dialogue)	(1) Turn coherence : later turns must correctly depend on entities or elements introduced earlier. (2) Evidence grounding : all answers must be supported by entries in the structured knowledge base. (3) Information completeness : key cultural elements are neither missing nor contradictory. (4) Answerability : questions must be answerable using available professional descriptions only.	(1) Cultural accuracy : checked against authoritative sources. (2) Pedagogical plausibility : dialogue progression matches a student–expert inquiry scenario. (3) Content boundary control : no turn introduces unverifiable or external content.
Task 3: Artistic Appreciation (QA-C / MCQ)	(1) Single-correctness : exactly one option must be correct. (2) Category consistency : correct and distractor options must belong to the same stylistic dimension. (3) Distractor quality : remove near-synonymous or trivially incorrect options. (4) Visual support : the question must be answerable based on visual form rather than external art history.	(1) Terminological rigor : confirm stylistic terms are used correctly in the Dunhuang context. (2) Evaluation fairness : ensure the question targets the intended analytic dimension. (3) Discriminability : maintain difficulty without introducing ambiguity.