HanMoVLM: Large Vision-Language Models for Professional Chinese Artistic Painting Evaluation

Anonymous ACL submission

Abstract

While Large Vision-Language Models (VLMs) demonstrate impressive general visual capabilities, they remain unable to offer professional evaluation within specific artistic domains. This evaluation deficit creates a critical bottleneck for improving Chinese painting generation, as text-to-image models lack expert feedback to distinguish genuinely superior outputs. To bridge this gap, we introduce **HanMo-Bench**, a new dataset featuring authentic auction-grade masterpieces and AI-generated works with expert-validated annotations grounded in real-world market valuations, and **HanMoVLM**, which leverages an expert-validated Chain-of-Thought (CoT) that addresses artistic misalignment by injecting domain-specific knowledge and bridges the understanding-evaluation gap through structured reasoning from content identification to three-tier evaluation. Furthermore, we design a reward function structurally aligned with the CoT to provide granular supervision over each reasoning step, ensuring faithful expert-level reasoning. With the evaluation capability established, HanMoVLM naturally extends as a verifier for Test-time Scaling, enabling generative models to select the most artistically superior outputs from multiple candidates. Experimental results and human studies confirm that HanMoVLM achieves high consistency with professional experts and significantly improves Chinese painting generation quality.

1 Introduction

Large Vision-Language Models (VLMs) have demonstrated extraordinary capabilities in general visual understanding and logical reasoning. Although VLMs can identify "mountains" or "rivers" across diverse artistic styles in paintings, including Chinese paintings, they still lack the domain-specific artistic reasoning required to evaluate non-Western visual art (Liu et al., 2025; Zhang et al., 2025a; Yu et al., 2025, 2026a; Tao et al., 2024). In this study, we take Chinese painting as the subject of investigation due to its rich artistic context. Unlike conventional visual aesthetics, its evaluation does not rely on realism or linear perspective, but rather a unique system of judgment standards rooted in cultural depth (Zhang et al., 2025b). Consequently, evaluating such artworks remains inaccessible to non-experts, as it requires expert-level cognition honed through years of training in the Chinese arts. While current VLMs can recognize visual elements in these paintings, they still struggle to evaluate them on par with domain experts.

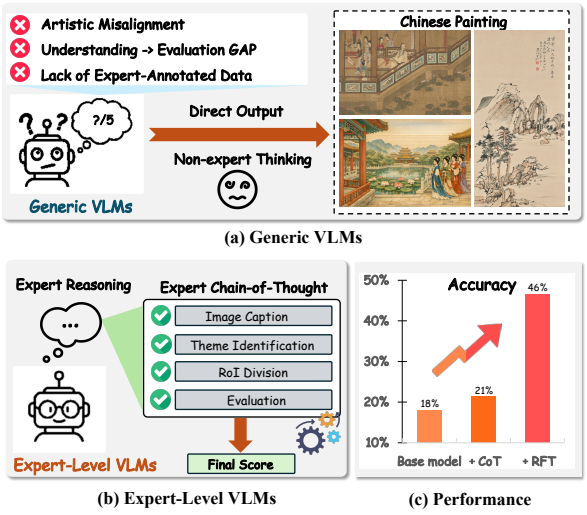

Figure 1: The motivation of our study. **(a)** Generic VLMs suffer from artistic misalignment, understanding-evaluation gap, and lack of expert-annotated data when handling domain-specific artistic content such as Chinese paintings; **(b)** The expert-level VLMs introduce the structured expert CoT and reward function, to produce reliable final scores; **(c)** Experimental results show that our CoT and RFT significantly improve performance.

As shown in Fig. 1, three interconnected challenges underlie this limitation: ① **Artistic Misalignment**, where the generic model lacks domain-specific knowledge of Chinese painting aesthetic systems, making it unable to assess paintings according to professional standards.

② **Understanding-Evaluation GAP**, where content recognition alone is insufficient for professional assessment—identifying a painting's elements does not equate to evaluating it against theme-specific, multi-tier criteria. ③ **Lack of Expert-Annotated Data**, there is a lack of high-quality annotated data with expert-level reasoning under the review of domain experts.

These challenges further create a critical bottleneck for artistic painting generation: without a reliable evaluation signal, T2I models lack the feedback to improve artistic quality, and no automated method exists to distinguish genuinely superior results from superficially pleasing ones. Bridging the evaluation gap is thus a prerequisite for advancing generation quality, not merely an end in itself.

To address these evaluation challenges and their cascading effect on generation, we present HanMoVLM[1], which evaluates Chinese paintings through a professional Chain-of-Thought (Wei et al., 2022; Wang et al.; Zhou et al.). Specifically, an expert-level CoT tackles ① Artistic Misalignment by injecting domain-specific knowledge from global theme identification to local RoI analysis; theme-specific criteria and the three-tier evaluation method bridge ② Understanding-Evaluation GAP by translating recognition into professional judgment; and HanMo-Bench addresses ③ Lack of Expert-Annotated Data with expert-validated annotations. To ensure the model faithfully follows each CoT step rather than hallucinating, we design a reward function structurally aligned with the CoT, each component supervises a corresponding reasoning step, and employ GRPO (Guo et al., 2025) for Reinforcement Fine-Tuning. Upon establishing reliable evaluation through RFT, we further extend HanMoVLM to artistic generation via test-time scaling: HanMoVLM serves as an external verifier for the Best-of-N (Cobbe et al., 2021) strategy, sampling multiple candidates from existing T2I models and selecting the one with the highest evaluation score. Experiments on HanMo-Bench validate the effectiveness of the proposed pipeline.

Our contributions are as follows:

- We construct HanMo-Bench, a dataset with expert-validated annotations for evaluating VLMs on artistic painting assessment (③).
- We propose HanMoVLM with an expert-level CoT that integrates global-local analysis, theme-specific criteria, and three-tier evaluation to address artistic misalignment (①).
- We design a CoT-aligned reward function where each component supervises a corresponding CoT step to bridge understanding-evaluation gap (②).
- We employ HanMoVLM as an external verifier for test-time scaling of T2I generation, with human studies confirming high consistency with art experts.

2 Related Work

The rapid development of vision-language models also raises concerns about artistic values and knowledge(Nayak et al., 2024; Romero et al., 2024; Xu et al., 2025). Most VLMs and T2I models are pre-trained on internet-scale datasets that exhibit a heavy Western-centric bias, leading to the "artistic blindness" when encountering non-Western visual semantics, especially Chinese artistic painting.

2.1 Visual Artistic Understanding

Many efforts have been devoted to evaluating whether VLMs can accurately interpret content from different arts (Yuan et al., 2023; Cetinic and She, 2022; Hayashi et al., 2024; Padó and Thomas, 2025). Representative work on artistic domain like the artistic and aesthetic dataset(SemArt (Garcia and Vogiatzis, 2018), Artemis (Achlioptas et al., 2021), Ava-dataset (Murray et al., 2012), Artimuse (Cao et al., 2025), WikiArt (Saleh and Elgammal, 2015), and OmniArt (Strezoski and Worring, 2018)), benchmarks (Alfarano et al., 2025; Yu et al., 2026b), artistic caption (Bai et al., 2021), artistic interpretation (Bin et al., 2024; Padó and Thomas, 2025), artistic evaluation (Zheng et al., 2025; Strafforello et al., 2025), proves that existing VLMs struggle to understand the specific artists (especially non-Western arts) (Liu et al., 2025; Zhang et al., 2025a; Yu et al., 2025, 2026a; Tao et al., 2024). However, these studies focus on capturing artistic content or translating the artistic symbol, without any consideration of expert-level artistic evaluation within a specific artistic background. Our study is the first to investigate this task and focus on Chinese art painting, whose evaluation requires extensive artistic training to reach the expert level (Wan et al., 2024).

2.2 Visual Artistic Generation

Visual artistic generation is a longstanding research topic in generative models (Wang et al., 2025a; Ko

[1]HanMo(翰墨) refers to brush and ink, symbolizing scholarship and traditional Chinese artistic expression.

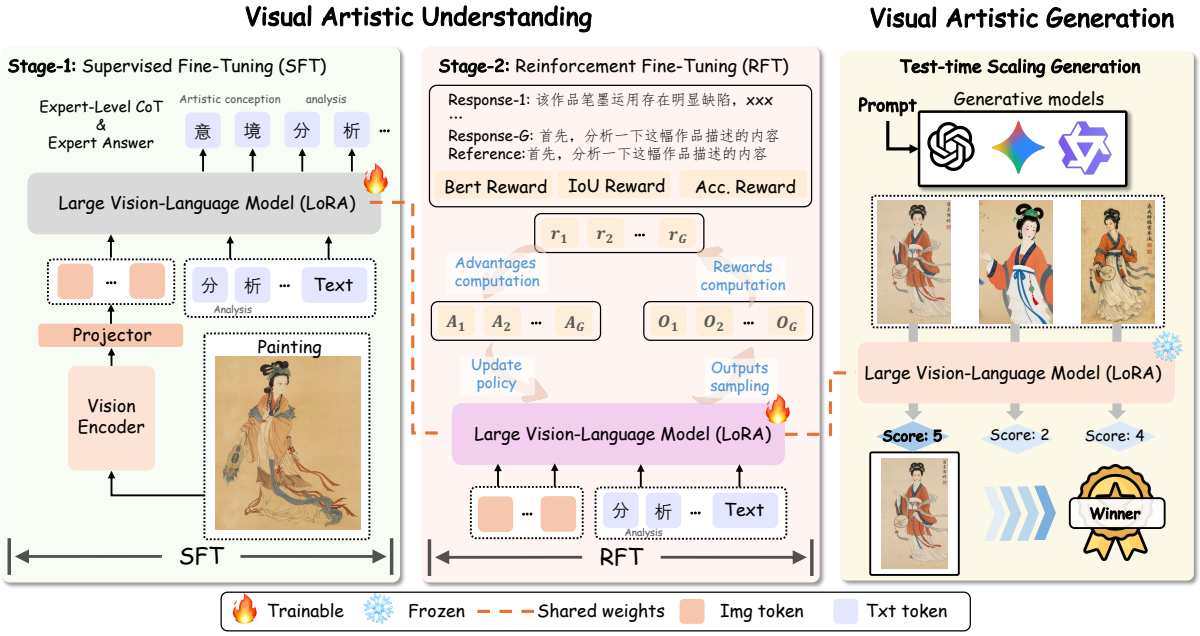

Figure 2: The overall framework of HanMoVLM. For visual artistic understanding, we finetune HanMoVLM with expert-level CoT and expert answer via SFT, and then perform RFT based on the expert reward function. For visual artistic generation, we apply the frozen HanMoVLM as the external verifier to evaluate the images.

et al., 2023; Gong et al., 2025; He et al., 2018). To introduce artistic concepts, some methods(Zhao et al., 2025; Yang et al., 2024; Liao et al., 2023; Tang et al., 2025) apply LoRA(Hu et al., 2022) or ControlNet(Zhang et al., 2023) to adopt a certain style into the content and thus introduce artistic concepts. However, although generative models produce impressive visual results, the outputs do not fully align with aesthetic values shaped by specific artistic contexts. For example, the artistic value of Chinese painting does not fully derive from realism or accurate linear perspective. Besides, the evaluation of artistic generation still relies heavily on labor-intensive validation by human experts, which prevents timely feedback and automated evaluation during training.

3 Method

3.1 HanMoVLM Framework

In this section, we present the overall framework of HanMoVLM, which addresses the three challenges identified in Section 1. As illustrated in Fig. 2, the proposed method consists of two training stages, SFT and RFT, to integrate the expert reasoning and responses of paintings in HanMoVLM. The model takes painting as visual input and produces structured textual analysis, enabling both semantic understanding and aesthetic evaluation. Specifically, the expert-level CoT (Sec. 3.2) tackles ① Artistic Misalignment by injecting domain-specific knowledge and ② Understanding-Evaluation GAP by translating recognition into professional judgment; the expert reward function (Sec. 3.3) is structurally aligned with the CoT to ensure each reasoning step is faithfully executed rather than hallucinated; and HanMo-Bench (Sec. 4) addresses ③ Lack of Expert-Annotated Data. With the evaluation capability established, HanMoVLM naturally extends to artistic generation, serving as an external verifier to assess the quality of generated images.

3.2 Artistic Expert-Level Chain-of-Thought

To address ① Artistic Misalignment, our core idea is to design a CoT that mimics the perspective of Chinese art experts, injecting the domain-specific knowledge that generic VLMs lack. To realize an expert-level CoT, we first establish a structured evaluation procedure based on the guidance of experts, which follows a progression from global understanding to detailed assessment as shown in Fig. 3. Specifically, HanMoVLM first identifies the overall content of the given painting and determines its precise theme within the system of Chinese painting in Landscape (山水), Flower&Bird (花鸟), and Figure (人物). Then, the HanMoVLM conducts a detailed analysis of the main visual elements or the local regions of interest, focusing on critical details and primary visual focal points. After completing global and local assessments, the HanMoVLM selects the appropriate evaluation criteria according to the theme of the painting, as the assessment standards vary significantly across different themes. This theme-specific evalu-

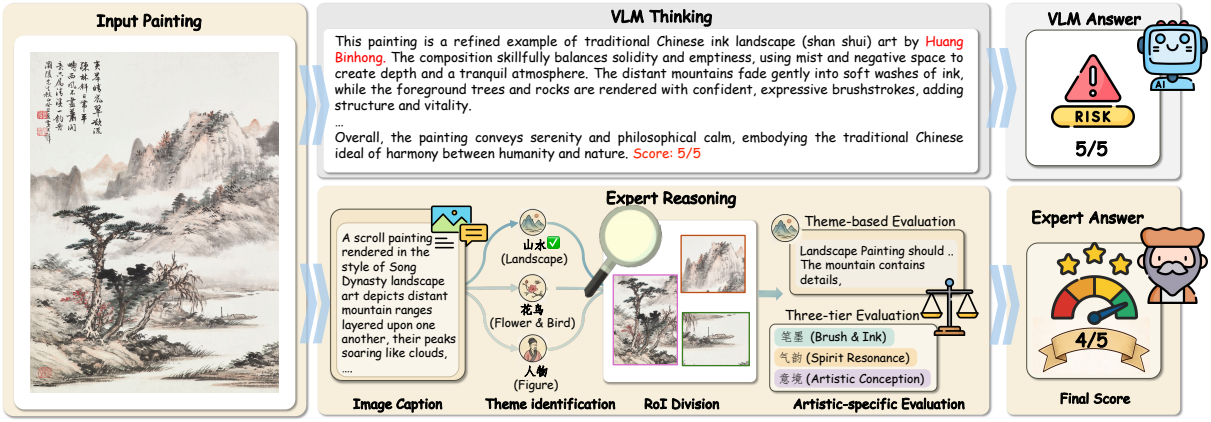

Figure 3: The expert-level Chain-of-Thought in our HanMoVLM. When the generic VLM evaluates Chinese paintings, the model tends to generate outputs based on general knowledge (non-expert) rather than following professional guidelines. As a result, its analysis does not reflect the experts' procedures, leading to a decline in terms of professional level and reliability.

ation mechanism directly bridges ② Understanding-Evaluation GAP, enabling the model to translate content recognition into professional evaluation. Finally, the painting is evaluated based on the typical three-tier criteria of Chinese painting, which are Brush&Ink (笔墨), Spirit Resonance (气韵), and Artistic Conception (意境). After integrating all these professional analyses, the HanMoVLM produces the final score.

3.3 Expert Reward Function

While the expert-level CoT injects domain knowledge and establishes a structured evaluation procedure, SFT alone cannot guarantee that the model faithfully follows each step—the model may produce reasoning that appears structured but deviates from expert judgment at specific steps. To address this, we design a reward function that is structurally aligned with the CoT: each reward component directly supervises a corresponding reasoning step (as described in Sec 3.2), ensuring that the model not only produces the right final answer but also reasons correctly at every stage. Specifically, the CoT identifies theme and RoI as critical intermediate steps, and the reward function enforces their correctness through dedicated supervision signals: R_{mIoU} for RoI localization, R_{BERT} for semantic consistency of each reasoning part, and R_{acc} for the final score. This design transforms the reward function from a generic accuracy signal into a structured enforcement mechanism that keeps the reasoning chain on track.

Accuracy Reward. This reward supervises the *final scoring step* of the CoT, ensuring the output score is consistent with the ground-truth:

$$R_{\text{acc}} = 1 - \frac{|pred_score - gt_score|}{5}, \quad (1)$$

BERT Reward. This reward supervises the *intermediate reasoning steps* of the CoT—including theme identification, theme-specific evaluation, and three-tier evaluation—by measuring the semantic similarity between sampled responses $\hat{x}$ and reference responses x. We encode the parsed K parts using a pre-trained BERT model(Zhang et al.):

$$R_{\text{BERT}} = \frac{1}{K}\sum_{j=1}^{K} \text{BERTScore}(pred_j, gt_j) \quad (2)$$

IoU Reward. This reward supervises the *RoI localization step* of the CoT, ensuring that the model correctly identifies and describes the critical visual regions. We compute the mIoU between the predicted RoIs and the ground-truth bounding boxes in the reference responses. For each predicted box, we match it with the gt boxes that achieve maximum IoU as its corresponding target. Moreover, since this task is an open-vocabulary detection, we also compute the BERTScore for the description in each pair of boxes.

$$j^*(i) = \arg\max_j IoU(pred_box_i, gt_box_j),$$
$$R_{\text{mIoU}} = \frac{1}{N}\sum_{i=1}^{N}\Big[IoU(pred_box_i, gt_box_{j^*(i)}) + \text{BERTScore}(pred_i, gt_{j^*(i)})\Big], \quad (3)$$

Therefore, the overall reward function, R_{final}, is a weighted sum of all the rewards, where each component supervises a specific CoT step:

$$R_{final} = w_{\text{acc}} \cdot R_{\text{acc}} + w_{\text{BERT}} \cdot R_{\text{BERT}} + w_{\text{mIoU}} \cdot R_{\text{mIoU}}. \quad (4)$$

where the w_{acc}, w_{BERT} and w_{mIoU} are the coefficients of the weight in different rewards. This

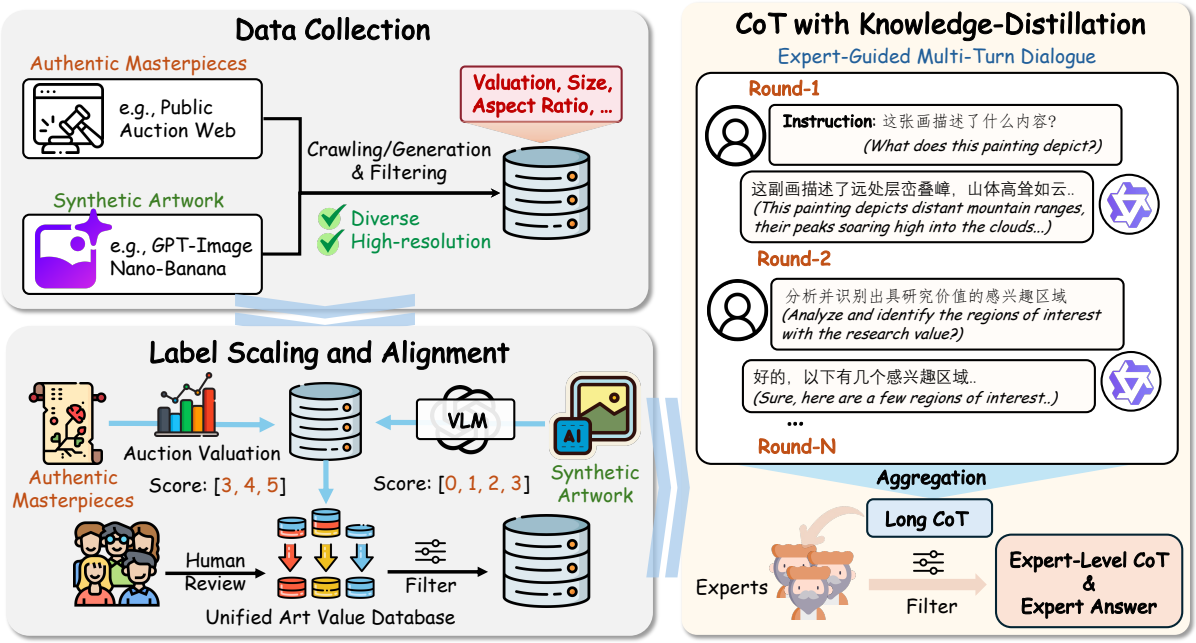

Figure 4: The construction pipeline of HanMo-Bench.

structural alignment ensures that improving one reward directly strengthens the corresponding reasoning step, making the entire evaluation pipeline more reliable.

3.4 Test-time Scaling for Artistic Generation

As discussed in Section 1, the lack of a reliable evaluation signal is a critical bottleneck for improving the artistic quality of Chinese painting generation. Once the VLM is trained to provide expert-level reasoning and answers, it naturally serves as the missing verifier for T2I models. To this end, we introduce a test-time scaling strategy, which requires no additional training or fine-tuning of the generative model and achieves consistent performance gains by increasing computation only at inference time. Given a textual input x, the T2I model G_θ generates N sampled paintings y under the fixed parameter θ, which can be formatted as:

$$\{y_1, y_2, \cdots, y_N\} = G_\theta(x), \tag{5}$$

The VLM then takes the generated image y and outputs a scalar compatibility score, which can be formatted as follows:

$$s_i = \text{VLM}(y_i), \tag{6}$$

where s_i is the expert response rating from HanMoVLM. The final generated image y^* is selected by maximizing s_i as follows:

$$y^* = \arg \max_{y_i \in \{y_1, y_2, \cdots, y_N\}} \text{VLM}(y_i). \tag{7}$$

This select process effectively performs distributional search guided by a multi-modal semantic objective, significantly enhancing the quality and artistic correctness of the T2I outputs.

4 HanMo-Bench

To address the lack of expert-annotated data (③), we create a novel dataset, named HanMo-Bench, which provides high-quality annotations with expert-validated reasoning for professional evaluation of Chinese Painting. This dataset contains 13k Chinese paintings across different themes (Landscape, Flowers&Birds, and Figure). The trainset and testset contain 13,162 and 600 paintings. The whole construction pipeline is shown in Fig. 4.

4.1 Data Collection

To ensure the representation of both historical artistic value and the generative capabilities of modern models, we collect data from two primary sources.

Authentic Masterpieces We collect high-resolution images of authentic Chinese paintings from publicly available auction house records, along with their professional valuations. These data are the representation of high artistic value Chinese paintings. While valuations are inevitably modulated by provenance, artist fame or market dynamics, these values serve as a high-fidelity proxy for human preference at the market scale, reflecting aesthetic consensus on Chinese art.

Synthetic Artwork. We employ the LLMs (GPT-5.1(OpenAI, 2026a) and Gemini-3-Flash(Google, 2026a)) to generate a diverse set of prompts for T2I models. Specifically, we ask LLMs to emphasize traditional Chinese artistic styles, detailed visual elements, painting techniques, and thematic content, thereby improving the quality of generated Chinese paintings. Then, we generate images with various aspect ratios using GPT-Image-1.5(OpenAI,

2026b) and Nano-Banana(Google, 2026b).

4.2 Label Scaling and Alignment

To bridge the gap between valuation and generative quality, we scale the value of these paintings, ranging from 0 to 5. For **Authentic Masterpieces**, we normalize their auction valuations into three tiers, assigning scores of 3, 4, and 5 to the top $60 \sim 100\%$, top $10 \sim 60\%$, and top 10%, respectively, representing different levels of artistic and historical significance. This is because the auctioneer's valuation process is actually a human assessment and reflects the value of the paintings. For **Synthetic Artwork**, we apply a VLM (Gemini-3-Pro(Google, 2026a)) to roughly categorize the outputs into four quality levels (0-3) under the expert CoT. This process has been reviewed by human experts to ensure that no artistic misalignment is introduced. The overlapping at score 3 is designed to simulate the current reality, where high-quality AI-generated images have attained a level approaching that of some early masterpieces, thereby increasing the difficulty for models to distinguish them.

4.3 Chain-of-Thought Construction

To construct professional reasoning, we introduce a structured CoT generation using Qwen-Max(Bai et al., 2025) via multi-turn dialogues under the suggestion by Chinese artist. Since large-scale expert annotations is unfeasible, we utilize expert-defined criteria to guide CoT generation. Three experts specializing in Chinese art are tasked with evaluating the generated CoT to ensure its quality and consistency. The VLM is pre-conditioned on the ground-truth score and authenticity to ensure the generated reasoning is consistent with the final evaluation. The final CoT is concatenated with image caption, ROI, theme identification, theme-specific evaluation, typical three-tier evaluation on Chinese painting, and the final rating.

5 Experiment

5.1 Experimental Setup

Training. Experiments are conducted on the open source base model Qwen3-VL-Instruct(Bai et al., 2025) in 4B and 8B versions with 8 NVIDIA H20 96GB GPUs. We employ LoRA training with a rank of 64 and an alpha of 64, respectively. For SFT, we train the model for 3 epochs with a batch size of 16 without weight decay. The initial learning rate is set to 1e-5. For RFT, we train our model with an initial learning rate of 1e-6 and a batch size of 32 for 2,000 steps. The number of samples for each prompt is set to 8. Please refer to supplementary for details and prompt formats.

Artistic generation. For generations, we adopt the test-time scaling strategy for training-free T2I generation. We apply the Best-of-N for each T2I model to sample multiple outputs and then employ HanMoVLM to select the best. The number of N is set to 8. We choose the leading T2I models (i.e., Nano-Banana(Google, 2026b), GPT-Image-1.5(OpenAI, 2026b), Seedream-4.0(Seedream et al., 2025), Qwen-Image(Wu et al., 2025)) for sampling. Please refer to Appendix for evaluation metrics.

5.2 Main Results

Visual Artistic Understanding. To validate the effectiveness of expert-level CoT and reward functions in HanMoVLM, we compare our methods with other existing models(Google, 2026a; OpenAI, 2026a; Wang et al., 2025b; Huang et al., 2026; Hong et al., 2025) with their own thinking mode and our expert-level CoT. As shown in Tab. 1, the expert-level CoT leads to different performance gains across various models. For example, it exhibits a relative improvement of 18.70% on Gemini-3-Flash. This suggests that the expert-level CoT effectively guides the VLMs to understand and analyze Chinese paintings from an artistic perspective and eventually outputs the corresponding score. In addition, the MAE and RMSE values for GPT-Family and several models are relatively high, which indicates a significant discrepancy between the predicted ratings and the ground-truth ratings. Also, SFT and RFT with the designed reward function significantly improve the performance (3% improvement in terms of accuracy). It should be noted the top-1 accuracy is a hard metric with any slight deviation (e.g., predicting 4 for a ground truth of 5) is a complete failure. We also report the MAE, which remains consistently around 0.6, demonstrates that the model's predictions are highly concentrated around the ground truth.

Moreover, the BERTScore improved by 49%, indicating that the model's reasoning on the painting is increasingly aligning with expert interpretations. We also report the improvement on RoI division and description, as shown in Tab. 2. The fine-tuned model demonstrates greater alignment with experts in segmenting and describing RoIs.

Visual Artistic Generation. We present the visual generation results in Fig. 5 and Fig. 8, respectively. It can be seen that images generated under the same

Table 1: The comparison of HanMoVLM with other models. and indicate the best and second best results.

Model	Size	MAE ↓	RMSE ↓	Top-1 Acc.(%) ↑	BERTScore ↑
Proprietary Models					
Gemini-3-Flash(Google, 2026a)	-	0.9883	1.3472	31.17	–
+*Thinking*	-	0.9633	1.4355	34.83	0.3052
+*Expert CoT*	-	0.9466	1.3351	**37.00**	0.3035
Gemini-3-Pro(Google, 2026a)	-	0.9866	1.3589	33.83	–
+ *Thinking*	-	0.9783	1.3311	35.17	0.2855
+ *Expert CoT*	-	**0.9416**	**1.3297**	36.83	**0.3101**
GPT-4(Achiam et al., 2023)	-	1.8023	2.2387	17.06	–
+*Expert CoT*	-	1.7491	2.1683	17.41	0.3049
GPT-5.1(OpenAI, 2026a)	-	1.7742	2.2123	17.73	–
+*Thinking*	-	1.8204	2.5200	17.95	0.1648
+*Expert CoT*	-	1.7671	2.1471	18.41	0.3073
Open-Source Models					
InternVL-3.5(Wang et al., 2025b)	4B	1.9591	2.4382	13.83	0.3441
Qwen3-VL-Instruct(Bai et al., 2025)	4B	1.9734	2.4055	7.90	–
+ *Thinking*	4B	1.9493	2.4465	17.89	0.2836
+*Expert CoT*	4B	1.9341	2.4495	19.44	0.2919
HanMoVLM (SFT)	4B	0.6850	0.9210	39.17	0.5506
HanMoVLM (RFT)	4B	**0.6632**	**0.9002**	**42.50**	**0.5568**
STEP3-VL(Huang et al., 2026)	10B	2.2950	2.7823	15.67	0.2335
GLM4.6V-Flash(Hong et al., 2025)	9B	2.0495	2.5466	15.50	0.3517
InternVL-3.5(Wang et al., 2025b)	8B	1.8866	2.3579	17.83	0.3395
Qwen3-VL-Instruct(Bai et al., 2025)	8B	1.5582	1.9534	18.53	–
+*Thinking*	8B	1.5050	1.9370	19.00	0.3402
+*Expert CoT*	8B	1.5124	1.9476	21.50	0.3757
+*Finetune only label*	8B	1.3255	1.7091	22.83	-
HanMoVLM (SFT)	8B	0.6450	0.9027	43.67	0.5531
HanMoVLM (RFT)	8B	**0.6310**	**0.8912**	**46.67**	**0.5598**

Table 2: The comparison in mIoU and Theme identification of our VLM with other models.

Model	Size	mIoU	BERTScore	Theme
Qwen3-VL-Instruct	4B	0.0821	0.2852	0.7346
HanMoVLM (SFT)	4B	0.2708	**0.4963**	0.8305
HanMoVLM (RFT)	4B	**0.2917**	0.4911	**0.8411**
Qwen3-VL-Instruct	8B	0.1401	0.3214	0.7218
HanMoVLM (SFT)	8B	0.2835	0.4872	0.8355
HanMoVLM (RFT)	8B	**0.3193**	**0.5126**	**0.8515**

prompt ranging from 0 to 5 points exhibit a significant gap in the quality of painting. Additionally, comparing the images across different groups reveals that artworks with lower scores exhibit distinct AI artifacts or areas that clearly deviate from Chinese painting aesthetics. More results are available in supplementary.

5.3 Ablation Study

For the ablation study, we conduct experiments to investigate how the reward function improves the HanMoVLM-4B in different metrics. As shown in Tab. 3, the model achieves the best performance on all metrics when all three rewards are combined in RFT. The results consistently show that the absence of a specific reward leads to a performance drop in its corresponding metric. For example, omitting the R_{acc} reduces the accuracy from 42.5% to 39.82%. Similarly, omitting R_{mIoU} causes the mIoU to decrease from 0.2917 to 0.2728.

Table 3: The ablation study of using different rewards.

R_{acc}	R_{BERT}	R_{mIoU}	Acc.(%)↑	BERTScore↑	mIoU↑
✓		✓	42.18	0.5529	0.2914
✓	✓		42.36	0.5492	0.2728
	✓	✓	39.82	0.5562	0.2902
✓	✓	✓	**42.50**	**0.5568**	**0.2917**

5.4 Human Study

To validate the consistency between our HanMoVLM's rating and human preferences, we conduct a human study where 43 researchers with backgrounds in Chinese painting are asked to rank the images sampled from the T2I models. We com-

Figure 5: Generated results in different ratings groups (zoom in for details).

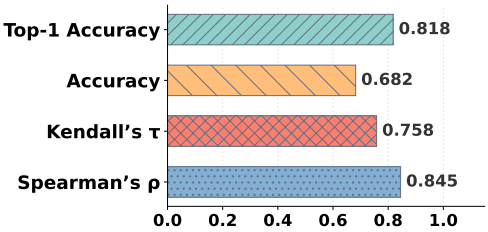

Figure 6: Consistency between HanMoVLM ratings and human rankings.

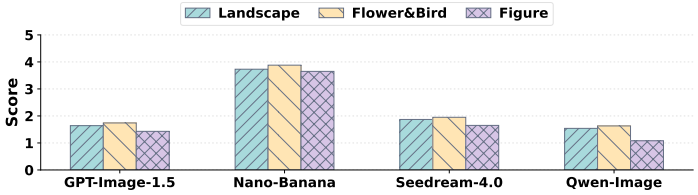

Figure 7: Comparison of T2I models across different themes.

pare the ranking results with the scores produced by HanMoVLM. As shown in Fig. 6, the ranking order obtained by HanMoVLM for the painting rating approaches that of experts. This indicates that the proposed HanMoVLM is capable of modeling expert judgments. Additionally, Kendall's τ and Spearman's ρ correlation coefficients reached 0.758 and 0.845, respectively, suggesting that the HanMoVLM aligns with human preference.

5.5 Analysis

We conduct a study on the performance of different leading T2I models across different themes. The results, as shown in Fig. 7, demonstrate that Nano-Banana performs best in generating Chinese painting, significantly outperforming GPT-Image-1.5(OpenAI, 2026b), Seedream-4.0(Seedream et al., 2025), and Qwen-Image(Wu et al., 2025), with a lead of approximately 2 points. Moreover, the T2I models achieve better results in Flowers&Birds painting than in other themes. For more analysis, please refer to the Appendix.

6 Conclusion

In this paper, we address the problem that current generic VLMs cannot accurately evaluate paintings in the Chinese artistic domain due to artistic misalignment, the gap between understanding and evaluation, and the lack of expert-annotated data. To tackle these challenges, we propose HanMoVLM, which leverages a structured expert-level CoT to inject domain-specific knowledge and translate recognition into professional evaluation. Furthermore, we introduce a new reward mechanism to ensure reasoning reliability in the RFT. Besides, we introduce HanMo-Bench, a novel and rigorous benchmark featuring authentic masterpieces and AI-generated works with expert-validated annotations. We further demonstrate that the established evaluation capability naturally extends to empower generation: HanMoVLM serves as an external verifier for Test-time Scaling, effectively improving the artistic quality of generated paintings. This evaluation-to-generation paradigm highlights that bridging the evaluation gap is not an end in itself, but a prerequisite for advancing artistic generation.

7 Limitations

The proposed HanMoVLM currently participates in test-time scaling as a frozen external verifier at inference time. A promising direction is to incorporate the evaluation feedback into the training loop of T2I models (e.g, VLM-based RL), enabling end-to-end optimization of generation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579.

Andrea Alfarano, Lorenzo Venturoli, and Darío Negueruela del Castillo. 2025. Vqart-bench: A semantically rich vqa benchmark for art and cultural heritage. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 406–416. IEEE.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5422–5432.

Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. 2024. Gallerygpt: Analyzing paintings with large multimodal models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7734–7743.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, and 1 others. 2025. Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*.

Eva Cetinic and James She. 2022. Understanding and creating art with ai: Review and outlook. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 18(2):1–22.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Noa Garcia and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multimodal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

Yue Gong, Shanyuan Liu, Liuzhuozheng Li, Jian Zhu, Bo Cheng, Liebucha Wu, Xiaoyu Wu, Yuhang Ma, Dawei Leng, and Yuhui Yin. 2025. Cta-flux: Integrating chinese cultural semantics into high-quality english text-to-image communities. *arXiv preprint arXiv:2508.14405*.

Google. 2026a. Gemini-3.1. https://gemini.google.com. Accessed: 15 March. 2026.

Google. 2026b. Nano-banana. https://deepmind.google/models/gemini-image/. Accessed: 15 March. 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024. Towards artwork explanation in large-scale vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 705–729.

Bin He, Feng Gao, Daiqian Ma, Boxin Shi, and Ling-Yu Duan. 2018. Chipgan: A generative adversarial network for chinese ink wash painting style transfer. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1172–1180.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Ailin Huang, Chengyuan Yao, Chunrui Han, Fanqi Wan, Hangyu Guo, Haoran Lv, Hongyu Zhou, Jia Wang, Jian Zhou, Jianjian Sun, and 1 others. 2026. Step3-vl-10b technical report. *arXiv preprint arXiv:2601.09668*.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663.

Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, pages 919–933.

Qisheng Liao, Gus Xia, and Zhinuo Wang. 2023. Calliffusion: Chinese calligraphy generation and style transfer with diffusion modeling. *arXiv preprint arXiv:2305.19124*.

Shudong Liu, Yiqiao Jin, CHENG LI, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. CultureVLM: Characterizing and improving cultural understanding of vision-language models for over 100 countries. In *Workshop of Computer Vision and Pattern Recognition Conference*.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Empirical Methods in Natural Language Processing*, pages 5769–5790.

OpenAI. 2026a. Gpt-5.1. https://openai.com/index/gpt-5-1/. Accessed: 15 March. 2026.

OpenAI. 2026b. Gpt-image-1.5. https://openai.com/index/new-chatgpt-images-is-here/. Accessed: 15 March. 2026.

Sebastian Padó and Kerstin Thomas. 2025. Artwork interpretation with vision language models: A case study on emotions and emotion symbols. In *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, pages 1–12.

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 *Neural Information Processing Systems*, pages 11479–11505.

Babak Saleh and Ahmed Elgammal. 2015. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, and 1 others. 2025. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*.

Ombretta Strafforello, Derya Soydaner, Michiel Willems, Anne-Sofie Maerten, and Stefanie De Winter. 2025. Have large vision-language models mastered art history? In *International Conference on Image Analysis and Processing*, pages 524–544. Springer.

Gjorgji Strezoski and Marcel Worring. 2018. Omniart: a large-scale artistic benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–21.

Shan Tang, Wenhua Qian, Peng Liu, and Jinde Cao. 2025. Art creator: Steering styles in diffusion model. *Neurocomputing*, 626:129511.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Jing Wan, Hao Zhang, Jun Zou, Ao Zou, Yubin Chen, Qingyang Zeng, Xinrong Li, and Qiya Wang. 2024. Wumkg: a chinese painting and calligraphy multimodal knowledge graph. *Heritage Science*, 12(1):1–18.

Bingyuan Wang, Qifeng Chen, and Zeyu Wang. 2025a. Diffusion-based visual art creation: A survey and new perspectives. *ACM Computing Surveys*, 57(10):1–37.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, and 1 others. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*.

Pengju Xu, Yan Wang, Shuyuan Zhang, Xuan Zhou, Xin Li, Yue Yuan, Fengzhao Li, Shunyuan Zhou, Xingyu Wang, Yi Zhang, and 1 others. 2025. Tcc-bench: Benchmarking the traditional chinese culture understanding capabilities of mllms. *arXiv preprint arXiv:2505.11275*.

Hyemin Yang, Heekyung Yang, and Kyungha Min. 2024. Artfusion: A diffusion model-based style synthesis framework for portraits. *Electronics*, 13(3):509.

Haorui Yu, Ramon Ruiz-Dolz, and Qiufeng Yi. 2025. A structured framework for evaluating and enhancing interpretive capabilities of multimodal llms in culturally situated tasks. *Findings of the Association for Computational Linguistics: EMNLP*, 2025:1945–1971.

Haorui Yu, Xuehang Wen, Fengrui Zhang, and Qiufeng Yi. 2026a. Cross-cultural expert-level art critique evaluation with vision-language models. *arXiv preprint arXiv:2601.07984*.

Haorui Yu, Diji Yang, Hang He, Fengrui Zhang, and Qiufeng Yi. 2026b. Vulca-bench: A multicultural vision-language benchmark for evaluating cultural understanding. *arXiv preprint arXiv:2601.07986*.

Zhengqing Yuan, Yunhong He, Kun Wang, Yanfang Ye, and Lichao Sun. 2023. Artgpt-4: Towards artistic-understanding large vision-language models with enhanced adapter. *arXiv preprint arXiv:2305.07490*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 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. 2025a. 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.

Wei Zhang, Wong Kam-Kwai, Biying Xu, Yiwen Ren, Yuhuai Li, Yingchaojie Feng, Minfeng Zhu, and Wei Chen. 2025b. Cultiverse: Towards cross-cultural understanding for paintings with large language model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6710–6719.

Yifei Zhao, Ziqi Liang, Yingrui Qiu, and Xiaona Wang. 2025. A novel flexible identity-net with diffusion models for painting-style generation. *Scientific Reports*, 15(1):27896.

Chanjin Zheng, Zengyi Yu, Yilin Jiang, Mingzi Zhang, Xunuo Lu, Jing Jin, and Liteng Gao. 2025. Artmentor: Ai-assisted evaluation of artworks to explore multimodal large language models capabilities. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and 1 others. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Evaluation Metrics

To further investigate the accuracy and analysis capability of the HanMoVLM, we apply multiple metrics in this task. For the final score prediction, we use the accuracy to assess whether the VLM can correctly assign each input painting to its grade. Besides, since this task is actually a regression problem rather than a classification task, we employ Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for evaluation. When computing accuracy for the cases of (0 vs. 5) and (4 vs. 5), both predictions are treated as incorrect. However, the prediction of 4 is clearly closer to the ground-truth score of 5, so the MAE and RMSE are applied to reflect the gap between the true values and the estimated scores. For models using expert-level CoT, we also report the BERTScore between the response and the reference answer to verify whether there are similar reasoning processes. Finally, we evaluate the mIoU for RoI localization and the accuracy of theme classification, as well as the BERTScore in each RoI.

B Human Evaluation details

In Sec. 5.4, we report the HanMoVLM's rating and human ranking. The 43 participants are students specializing in Traditional Chinese Painting at various high education institutions, and possessed four or five years of formal artistic training. The human study required them to rank 30 representative set of generated paintings, each set consist of 4 images. We calculate the Kendall's $\mathcal{W}$ among multiple human evaluators, resulting in a value of 0.7375. This indicates a high degree of inter-rater agreement, suggesting that the evaluators share consistent perspectives and no significant divergence.

Dim	Num
$N_{evaluators}$	43
N_{sets}	30
Imgs/Set	4
Kendall's $\mathcal{W}$	0.7375

Table 4: Details of human studies.

C Implementation details

The weights of the reward in RFT are set to $w_{\text{acc}} = 10.0, w_{\text{BERT}} = 2.0, w_{\text{mIoU}} = 2.0, w_{\text{format}} = 1.0$. For HanMoVLM training, we apply the cosine decay learning rate scheduler in both SFT and RFT. For the thinking mode, we select models built in thinking capabilities, or explicitly employ a CoT prompting with "*Let's think step by step*".

D Results

We compare the ratings of the proposed HanMoVLM-8B with other image rating models. Specifically, we evaluate each model's ability to rank paintings either by sorting their predicted scores or by directly producing a ranking. The resulting rankings are then compared with the rankings obtained from human studies, which were treated as the ground truth. The result is presented in the Tab. 5. Both Kendall's τ and Spearman's ρ range from -1 to 1, where values closer to 1 indicate a higher similarity between the predicted ranking and the human-provided ranking. The aesthetic score and PickScore have a low τ and ρ value (≤ 0.2), indicating that these two rating models fail to represent the aesthetic standards in the Chinese painting field.

E Analysis

Image quality evaluators. Comparing other image quality evaluator, PickScore (Kirstain et al., 2023) and Aesthetic scores cannot accurately reflect human ranking of Chinese paintings. For example, the painting with the highest aesthetic score (6.4644) and PickScore (20.56) is not the first ranking from humans.

The effects of different type scrolls. We present the results in different types of Chinese painting (立轴, 方幅, 横卷) in Fig. 9. The 立轴 (hanging scroll) refers to the painting designed to be hung on the wall, typically with a greater height than width. The 方幅 (square format) is a square composition that emphasizes balance and symmetry. And the 横卷 (handscroll) is a long and horizontal scroll, highlighting its extended format. The landscape painting themes exhibit relatively stable performance across different types, while the figure painting performance shows variation.

The number of RoI divisions. As shown in Fig. 10, the number of RoIs predicted by VLM is mostly consistent with those in the ground truth. Meanwhile, paintings focusing on specific small-scale subjects like flowers and birds require fewer RoIs compared to other types of paintings. For example, the average number of RoI divisions in the Landscape paintings is 3.25, while that in the Flowers&Birds paintings averages around 2.8.

Table 5: Comparison of ranking accuracy in different models.

Rating Type	Top-1 Accuracy↑	Accuracy↑	Kendall's τ ↑	Spearman's ρ ↑
Aesthetic Score	0.045	0.125	-0.258	-0.282
PickScore (Kirstain et al., 2023)	0.409	0.341	0.106	0.073
HPSv2.1 (Wu et al., 2023)	0.364	0.193	-0.151	-0.173
GPT-4	0.227	0.341	0.197	0.218
GPT-5.1	0.273	0.364	0.121	0.209
HanMoVLM-8B	**0.818**	**0.682**	**0.758**	**0.845**

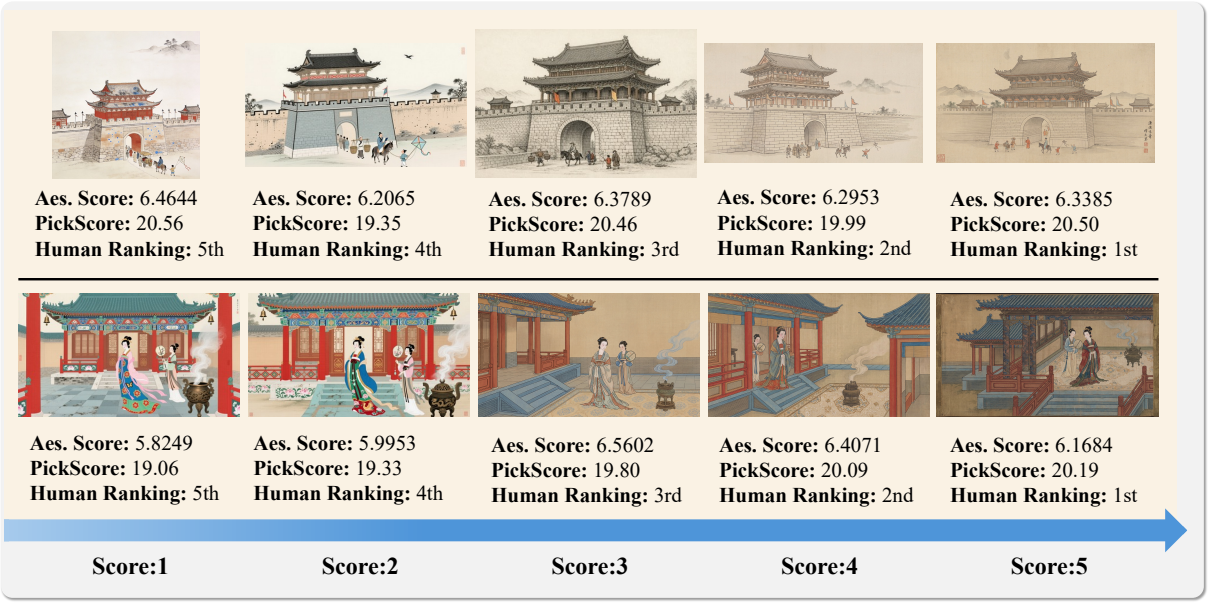

Figure 8: Generated results with the same prompt across different ratings.

The case of RoI. In expert-level CoT, we guide the VLM to segment specific IoU regions. Although the mIoU value is only 0.31, as shown in the Tab. 2, its ability to localize specific semantic regions remains robust. We illustrate a case with an mIoU value of only 0.31 in Fig. 11. Although the ground truth boxes and predicted boxes exhibit differences, the content they describe is similar. For instance, HanMoVLM accurately identified the objects in the painting that should be analyzed, especially since this is an extremely tall hanging scroll, and the bounding box of some targets (e.g., pedestrians) is extremely small.

F HanMo-Bench

HanMo-Bench, as shown in Fig. 12, consisting of different theme based on the advice of the artist experts. Three main themes in Chinese art domain (Landscape, Flowers&Birds, and Figure).

G Visual Results

Fig. 13 shows more visual results grouped by different scores in artistic generation. Although generative models can produce images with high-quality and remarkable details, they still fail to produce authentic Chinese paintings. Generated results with rating 0 exhibit noticeable AI artifacts, as the aesthetics of Chinese painting remain absent in generative models. Besides, the generative model incorrectly blends elements of modern painting and Western realism with the style of Chinese painting (e.g., the dog in the lower-right corner).

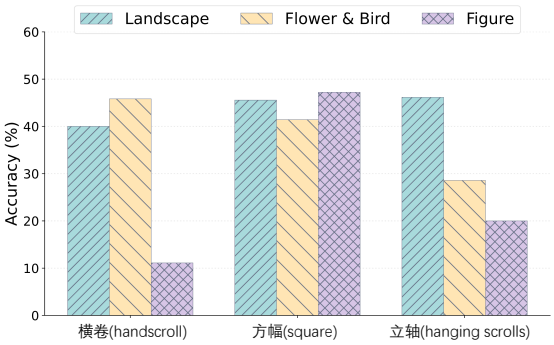

Figure 9: The effects of different types of painting.

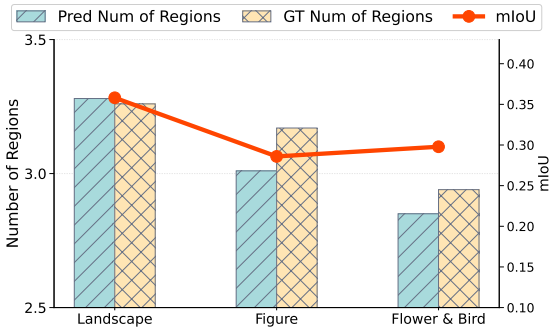

Figure 10: Number of RoI divisions and mIoU across different themes

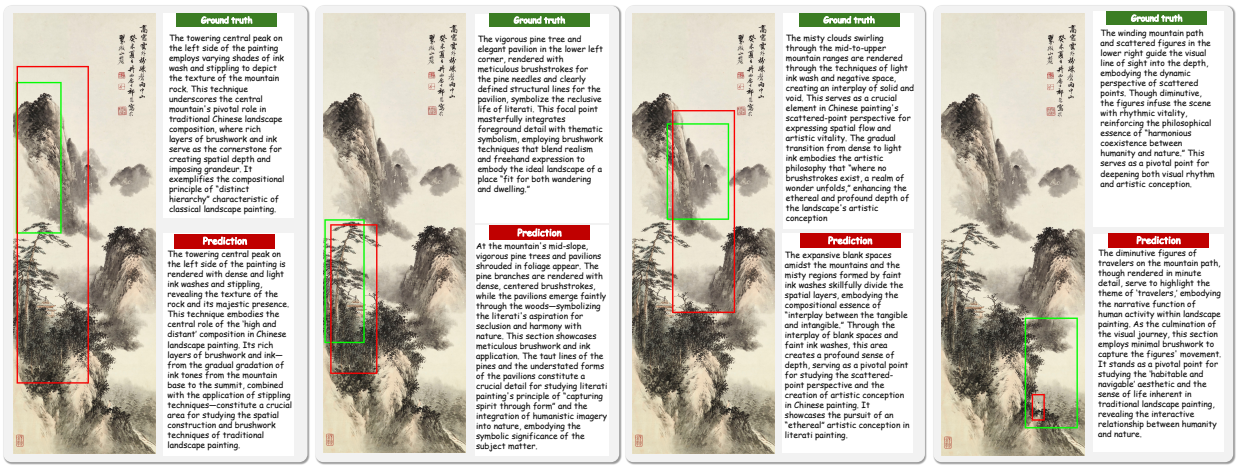

Figure 11: The case of RoI in HanMoVLM and the corresponding predicted description and gt description in HanMoVLM. The IoU of this case is 0.31. Zoom in for more details.

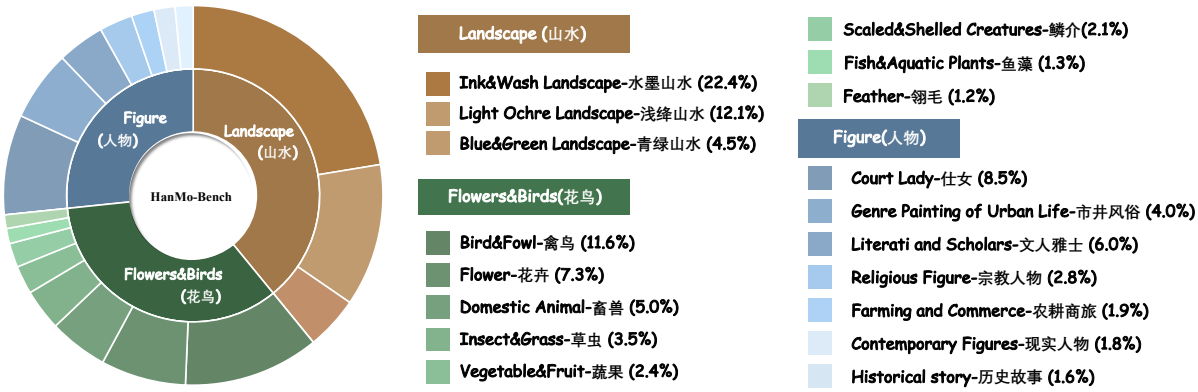

Figure 12: Composition of HanMo-Bench.

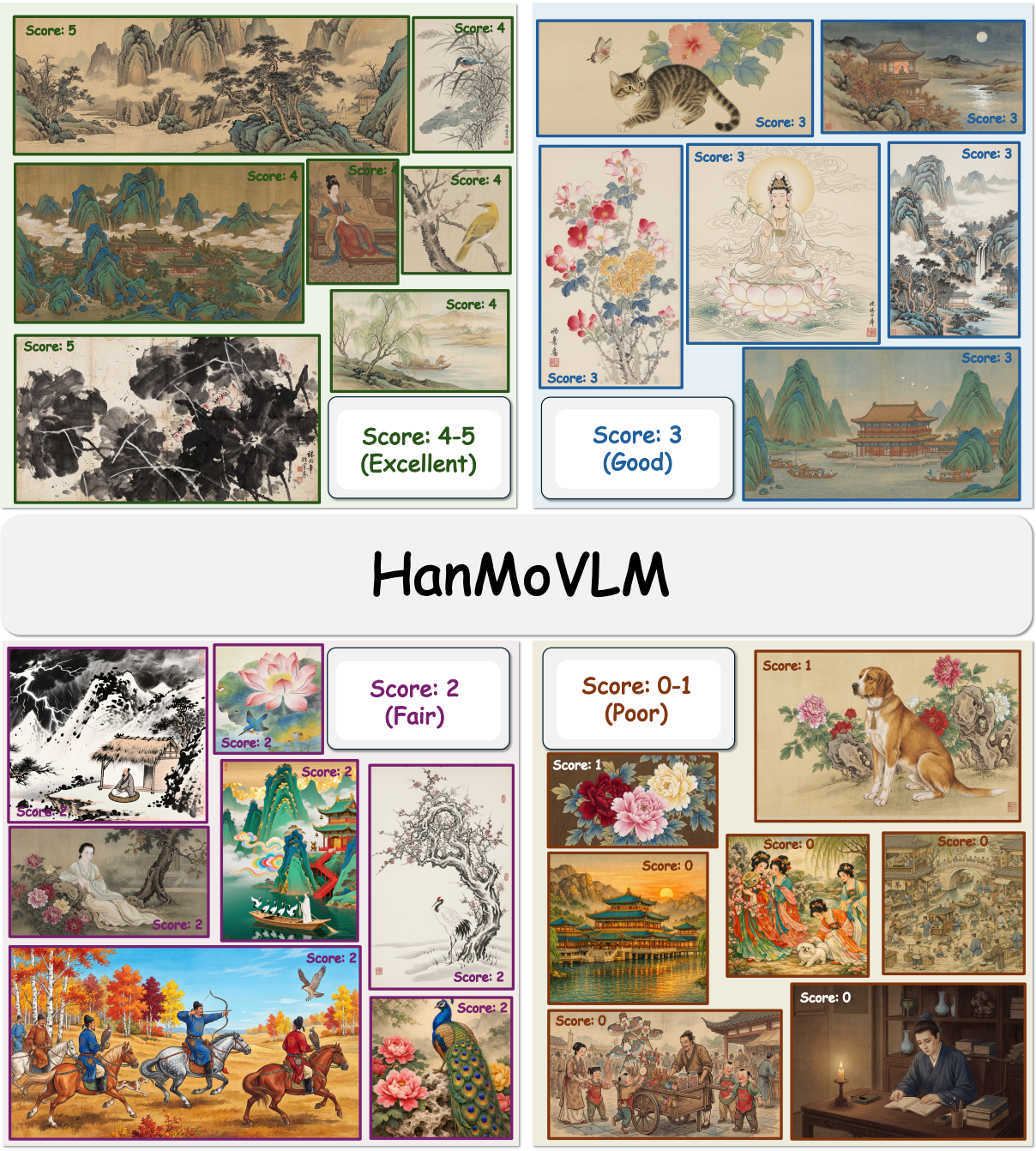

Figure 13: Visual results grouped by the score from our HanMoVLM. Zoom in for details.

H Prompts

The prompts we use in this study are shown as follows. We employ a Chinese prompt in VLM to prevent semantic gaps arising from cross-linguistic differences. We also provide an English translation for readability.

H.1 Expert-Level CoT Prompt

Expert-Level CoT Prompting (Chinese)

你是一位中国传统绘画鉴赏专家，熟悉笔墨技法、中国画美学、艺术史与文人画理论。请对输入的国风绘画图像进行深入、专业、客观的艺术评估。让我们一步一步思考并回复

1. 观察这幅画作描述了什么内容，包括画面主体、构图特点以及可能的艺术风格，使用中文输出，确保描述清晰、详细

2. 请明确这幅作品的具体题材分类（如：山水画、花鸟画、人物画）

人物画（包括历史故事、宗教人物、文人雅士、仕女、市井风俗、农耕商旅、现实人物
山水画（包括青绿山水、水墨山水、浅绛山水）
花鸟画（包括花卉、禽鸟、翎毛、蔬果、草虫、畜兽、鳞介、鱼藻

3. 观察画作的局部内容，确定值得分析的区域，并放大查看

4. 根据画作的题材，以及该题材的评判标准，给出对这幅画的专业评价

人物画题材：首先检验人物造型是否符合结构与比例，动态是否自然协调；其次重点判断人物是否具备清晰的精神气质、情绪表达与性格特征，尤其关注眼神、姿态与整体神态是否生动；同时分析线条是否具有书法性的节奏、骨力与提按变化，用笔是否支撑人物结构；再综合评估画面整体是否呈现连贯流动的生命感与精神统一性，即“气韵生动”；最终以“形神兼备、以神为主、气韵为最高标准”作为优劣判定依据，而非单纯追求形似或细节精度。

山水画题材：首先分析山石、树木、水体等物象的结构是否合理，皴法与笔法是否符合自然形态与山体结构；其次评估构图章法是否完整，散点透视是否自然，虚实、疏密、动静关系是否协调，留白是否有效参与空间营造；再次重点判断笔墨变化是否丰富，干湿浓淡是否层次分明，线条是否具有节奏与骨力；最后综合评估画面是否形成连贯统一的意境空间，是否呈现“可行、可望、可游、可居”的整体境界，并以“气韵生动、境由心造、笔墨当随时代”作为最高优劣判断标准，而非单纯写实或细节复杂。

花鸟画题材：首先分析花卉、禽鸟、草虫、鱼兽等物象的形态结构是否准确自然，比例与动态是否符合生物特征；其次重点判断物象是否具有鲜明的生命感与生动气息，是否体现自然生长节律；再次评估用笔是否灵动有力、富有节奏，墨色与设色是否协调雅致、层次分明；同时分析构图是否疏密有致、主次清晰、虚实得当；最后综合判断作品是否通过物象表达情感与人格象征，即“托物言志、以小见大”，并以“生意盎然、气韵生动、意象统一”作为最高优劣评判依据，而非单纯工细程度或色彩浓艳度。

5. 根据笔墨**-**气韵**-**意境三个层次进行评估。 评判时需兼顾艺术真实感、生命感与文化深度，避免仅以画面精致度、工整度或装饰性效果作为高分依据。对于自然笔触的不完美所产生的生命感、节奏感与艺术真实，应给予正向评价；对于过度工整、机械化、装饰化的画面效果，应保持审慎态度。
原则:
笔墨：关注线条质量、运笔控制、墨色层次与结构塑造，判断其是否自然、稳定、生动，而非仅仅是否干净、精细
气韵：关注画面是否具有内在生命流动感、气场贯通性、节奏变化与动势走向，判断作品是否“活”，而非只是构图完整
意境：关注作品是否营造出诗意空间、情绪表达与文化审美高度，是否引发联想与回味，而非堆砌符号或套用意象模板
在综合评价中，意境权重高于气韵，气韵权重高于笔墨。若意境明显不足，即使笔墨精致，整体评分也不应过高。

6. 根据以上所有的评价，给出这张画作的综合艺术评估分数（**0-5**分），分数必须为整数。

5分：局部与整体在笔墨结构、气韵流动与精神指向上高度同构，意境深远，气韵化生，笔墨随心而不逾法，形成高度统一的生命结构与文化境界，具有不可替代的艺术原创性。
4分：关键局部笔墨精到、气脉贯通，整体节奏自然，意境清远深长，文化气息浓厚，艺术语言成熟稳定，具有鲜明而稳定的艺术品格。
3分：局部具备基本结构与气韵支撑，画面开始“有呼吸”，意境初步成立，但思想深度、文化厚度与局部—整体协同仍有限。
2分：局部描绘精细但重技巧轻生发，整体结构严谨却气韵不足，精神指向薄弱，艺术表达主要停留在技法与形式层面。
1分：构图完整，形象清晰，具备基本绘画表达，局部结构松散，笔墨僵滞，节奏单一，整体气息闭塞，尚未形成完整艺术语言。
0分：画面虽然好看、精致，但缺乏笔墨逻辑、气韵流动与意境生成，整体停留在表层视觉美感，更接近插画或装饰图。

按以下格式输出分数

最终分数**: [**整数分数**]**

Expert-Level CoT Prompting (English)

You are an expert in traditional Chinese painting appreciation, well-versed in brushwork techniques, Chinese painting aesthetics, art history, and literati painting theory. Please provide an in-depth, professional, and objective artistic evaluation of the submitted traditional Chinese-style painting image. Let us consider and respond step by step:

1. Observe and describe the content of this painting, including the main subject, compositional characteristics, and potential artistic style. Output in Chinese, ensuring the description is clear and detailed.

2. Please specify the category of the theme (e.g., Landscape, Flowers&Birds, Figure):

Figure (including court lady, genre painting of urban life, literati and scholars, religious figure, farming and commerce, contemporary figures, historical story)
Landscape (including Ink&Wash, light ocher, blue&green)
Flowers&Birds: (Including bird&fowl, flower, domestic animal, insect&grass, vegetables&fruits, scaled&shelled creatures, fish&aquatic plants, feather).

3. Observe specific details of the painting, identify areas worthy of analysis, and zoom in for inspection.

4. Based on the genre and its specific criteria, provide a professional critique:

Figure Painting: First, examine if the modeling follows proper structure and proportion and if the movement is natural. Focus on whether the figure possesses a clear spiritual temperament, emotional expression, and character, paying close attention to the eyes, posture, and overall vitality. Analyze whether the linework possesses calligraphic rhythm, "bone-strength", and variations in pressure. Evaluate the "Spirit resonance" using "the integration of form and spirit, prioritizing spirit" as the standard rather than mere likeness or detail.

Landscape Painting: Analyze the structure of rocks, trees, and water. Assess whether the texture strokes and brushwork align with natural forms and geological structures. Evaluate the composition, the use of "shifting perspective," and the balance between void and solid, density and sparseness. Determine if the brush and ink changes are rich (dry/wet/dark/light) and if the white space effectively creates space. The highest standard is the creation of the "Artistic Conception", a space that is "walkable, viewable, travelable, and livable"—prioritizing spirit and the idea that "brush and ink should follow the times."

Flowers&Birds Painting: Analyze the morphological structure of the subjects. Focus on whether they possess a vivid sense of life and natural growth rhythms. Evaluate if the brushwork is agile and rhythmic, and if the coloring is elegant and layered. Assess the composition for clarity of primary and secondary subjects. The ultimate judgment rests on "Expressing Ambition through Objects" and "Vividness of Life", rather than mere technical fineness or vivid colors.

5. Brushwork&Ink, Spirit resonance, and Artistic conception:

Principle:
Brush and Ink: Focus on line quality, brush control, ink layering, and structural modeling; judge whether it is natural, stable, and vivid, rather than merely whether it is clean or refined.

Spirit Resonance: Focus on whether the image possesses an internal sense of life-flow, a coherent "field", rhythmic variations, and momentum; judge whether the work is "alive" rather than just compositionally complete.
Artistic Conception: Focus on whether the work creates a poetic space, emotional expression, and cultural-aesthetic depth; judge whether it evokes association and reflection rather than merely piling up symbols or applying conventional image templates.

In the comprehensive evaluation, Artistic Conception carries more weight than Spirit Resonance, and Spirit Resonance carries more weight than Brush and Ink. If the Artistic Conception is clearly insufficient, the overall score should not be high even if the brush and ink are refined.

6. Based on all the above evaluations, provide a comprehensive artistic assessment score (0-5), which must be an integer.

5 points: Local details and the whole are highly isomorphic in brushwork structure, flow of spirit, and spiritual orientation. The Artistic Conception is profound, Spirit Resonance is transformed into life, and brush and ink follow the heart without violating principles, forming a highly unified life structure and cultural realm with irreplaceable artistic originality.
4 points: Key local details are refined in brush and ink, the spiritual pulse is continuous, the overall rhythm is natural, the Artistic Conception is clear and far-reaching, the cultural atmosphere is thick, and the artistic language is mature and stable, possessing a distinct and stable artistic character.
3 points: Local parts possess basic structure and spiritual support; the image begins to "breathe," and the Artistic Conception is initially established, but depth of thought, cultural thickness, and the synergy between local and whole are still limited.
2 points: Local descriptions are fine but emphasize technique over vitality; the overall structure is rigorous but lacks Spirit Resonance, the spiritual orientation is weak, and artistic expression mainly remains at the level of technique and form.
1 point: The composition is complete and the images are clear, possessing basic pictorial expression; however, local structures are loose, brush and ink are stagnant, the rhythm is singular, and the overall atmosphere is blocked, failing to form a complete artistic language.
0 point: Although the image is attractive or exquisite, it lacks brushwork logic, the flow of spirit, and the generation of the Artistic Conception; the whole remains at the level of surface visual beauty, closer to an illustration or decorative graphic.

Output the score in the following format:

Final rating: [Integer Score]

H.2 Prompt for Generation

Prompt for Text-to-Image models (Chinese)

生成一个聚焦于中国画作的prompt，按格式输出20个详细的、可直接用于文本生成图像模型的prompt
1. 需要充分发挥想象，并且对出现的元素进行详细描述，但不要写出画作的名字
2. 可以指定题材，包括但不限于山水/花鸟/人物
3. 可以指定绘画手法以及色彩
4. 需要符合中国画中的笔墨-气韵-意境的特点
5. 每条prompt应该不少于150词，以中文输出

[Prompt1]: <prompt1>
[Prompt2]: <prompt2>
[Prompt3]: <prompt3>
...
(up to 20 prompts)

Prompt for Text-to-Image models (English)

Generate 20 detailed prompts focused on Chinese painting that can be directly used for text-to-image models. Please output them according to the following format:

1. You need to fully utilize your imagination and provide detailed descriptions of the elements involved, but do not include the titles of specific famous paintings.
2. You may specify the subject matter, including but not limited to landscapes, birds and flowers, or figures.
3. You may specify painting techniques (brushwork) and color schemes.
4. The prompts must align with the characteristics of "Brush and Ink", "Spirit Resonance", and "Artistic Conception" inherent in traditional Chinese painting.
5. Each prompt should be no less than 150 words and must be output in Chinese.

[Prompt1]: <prompt1>
[Prompt2]: <prompt2>
[Prompt3]: <prompt3>
...
(up to 20 prompts)

H.3 CoT Construction

Multi-turn Dialogue for CoT Construction (Round-1)

请尽可能详细描述这幅图像的内容，包括画面主体、构图特点以及可能的艺术风格，使用中文输出，确保描述清晰、详细，不少于100个词。

基于前面的描述，请明确这幅作品的具体题材分类（如：山水画、花鸟画、人物画等）。

人物画（包括历史故事、宗教人物、文人雅士、仕女、市井风俗、农耕商旅、现实人物)
山水画（包括青绿山水、水墨山水、浅绛山水）
花鸟画（包括花卉、禽鸟、翎毛、蔬果、草虫、畜兽、鳞介、鱼藻

(English version)
Please describe the content of this image in as much detail as possible, including the main subject, compositional features, and possible artistic style. Provide the description in Chinese, ensuring it is clear and detailed, with a minimum of 100 words.
Based on the description, specify the category of the theme (e.g., Landscape, Flowers&Birds, Figure):

Figure (including court lady, genre painting of urban life, literati and scholars, religious figure, farming and commerce, contemporary figures, historical story)
Landscape (including Ink&Wash, light ocher, blue&green)
Flowers&Birds: (Including bird&fowl, flower, domestic animal, insect&grass, vegetables&fruits, scaled&shelled creatures, fish&aquatic plants, feather).

Multi-turn Dialogue for CoT Construction (Round-2)

基于上述描述，请你从艺术与视觉结构角度分析画面内容，识别出最具研究价值的感兴趣区域（Region of Interest, ROI），并为每个区域提供精确的bounding box。
你需要仔细看一下这张图像的内容，然后根据图像的内容分析出具体需要有几个感兴趣区域。

分析要求：1. 综合考虑中国画的构图方式（如：主次关系、散点透视）、笔墨技法、题材象征意义。
2. 感兴趣区域可以包括：主要描绘对象、视觉中心、具有显著笔墨特征的局部，视觉中心或视觉动线的关键节点。
3. 每个感兴趣区域需给出明确的语义说明。
4. 不要分析题跋章印或者文字部分，聚焦在画作视觉元素本身

输出格式要求：
请以JSON 格式输出结果，使用归一化的像素坐标（范围0-1），注意精度尽可能准确到小数点后5位，坐标原点为图像左上角。
确保输出为合法JSON，输出一个JSON 对象。
使用中文。

```
{
 "height": {height},
 "width": {width},
 "num_regions": N,
 "regions_of_interest": [
  {
   "label": "区域名称",
   "description": "该区域在中国画中的艺术或研究意义",
   "bounding_box": {
    "x_min": x1,
    "y_min": y1,
    "x_max": x2,
    "y_max": y2
   }
  }
 ]
}
```

(English version)
Based on the description above, please analyze the content of the image from an artistic and visual structure perspective, identify the Regions of Interest (ROI) with the greatest research value, and provide precise bounding boxes for each region.
Please examine the content of this image carefully, and then analyze how many specific Regions of Interest (ROI) are required based on the visual information.

Requirements:
1. Comprehensively consider the compositional methods of Chinese painting (e.g., primary-secondary relationships, shifting perspective), brush and ink techniques, and the symbolic significance of the subject matter.
2. Regions of Interest may include: the primary subjects, the visual center, localized areas with significant brushwork characteristics, and key nodes of the visual center or visual flow.
3. Provide a clear semantic description for each Region of Interest.
4. Do not analyze inscriptions, colophons, or seals; focus exclusively on the visual elements of the artwork itself.

Output Format Requirements:
Please output the results in JSON format using normalized pixel coordinates (range 0-1). Ensure the precision is as accurate as possible to 5 decimal places, with the coordinate origin at the top-left corner of the image.

Ensure the output is a valid JSON object.
Use Chinese for the descriptions.

```
{
 "height": {height},
 "width": {width},
 "num_regions": N,
 "regions_of_interest": [
  {
   "label": "Label of the regions",
   "description": "The artistic significance of this region in Chinese painting.",
   "bounding_box": {
    "x_min": x1,
    "y_min": y1,
    "x_max": x2,
    "y_max": y2
   }
  }
 ]
}
```

Multi-turn Dialogue for CoT Construction (Round-3)

基于前面的描述，根据该题材所属的特定审美标准，对作品进行简要评价。

人物画题材:
首先检验人物造型是否符合结构与比例，动态是否自然协调；其次重点判断人物是否具备清晰的精神气质、情绪表达与性格特征，尤其关注眼神、姿态与整体神态是否生动；同时分析线条是否具有书法性的节奏、骨力与提按变化，用笔是否支撑人物结构；再综合评估画面整体是否呈现连贯流动的生命感与精神统一性，即“气韵生动”；最终以“形神兼备、以神为主、气韵为最高标准”作为优劣判定依据，而非单纯追求形似或细节精度。

山水画题材:
首先分析山石、树木、水体等物象的结构是否合理，皴法与笔法是否符合自然形态与山体结构；其次评估构图章法是否完整，散点透视是否自然，虚实、疏密、动静关系是否协调，留白是否有效参与空间营造；再次重点判断笔墨变化是否丰富，干湿浓淡是否层次分明，线条是否具有节奏与骨力；最后综合评估画面是否形成连贯统一的意境空间，是否呈现“可行、可望、可游、可居”的整体境界，并以“气韵生动、境由心造、笔墨当随时代”作为最高优劣判断标准，而非单纯写实程度或细节复杂度。

花鸟画题材:
首先分析花卉、禽鸟、草虫、鱼兽等物象的形态结构是否准确自然，比例与动态是否符合生物特征；其次重点判断物象是否具有鲜明的生命感与生动气息，是否体现自然生长节律；再次评估用笔是否灵动有力、富有节奏，墨色与设色是否协调雅致、层次分明；同时分析构图是否疏密有致、主次清晰、虚实得当；最后综合判断作品是否通过物象表达情感与人格象征，即“托物言志、以小见大”，并以“生意盎然、气韵生动、意象统一”作为最高优劣评判依据，而非单纯工细程度或色彩浓艳度。

(English version)
Based on the preceding description, provide a brief evaluation of the work according to the specific aesthetic standards applicable to its genre.

Figure Painting:
First, examine if the modeling follows proper structure and proportion and if the movement is natural. Focus on whether the figure possesses a clear spiritual temperament, emotional expression, and character, paying close attention to the eyes, posture, and overall vitality. Analyze whether the linework possesses calligraphic rhythm, "bone-strength", and variations in pressure. Evaluate the "Spirit resonance" using "the integration of form and spirit, prioritizing spirit" as the standard rather than mere likeness or detail.

Landscape Painting:
Analyze the structure of rocks, trees, and water. Assess whether the texture strokes and brushwork align with natural forms and geological structures. Evaluate the composition, the use of "shifting perspective," and the balance between void and solid, density and sparseness. Determine if the brush and ink changes are rich (dry/wet/dark/light) and if the white space effectively creates space. The highest standard is the creation of the "Artistic Conception", a space that is "walkable, viewable, travelable, and livable"—prioritizing spirit and the idea that "brush and ink should follow the times."

Flowers&Birds Painting:
Analyze the morphological structure of the subjects. Focus on whether they possess a vivid sense of life and natural growth rhythms. Evaluate if the brushwork is agile and rhythmic, and if the coloring is elegant and layered. Assess the composition for clarity of primary and secondary subjects. The ultimate judgment rests on "Expressing Ambition through Objects" and "Vividness of Life", rather than mere technical fineness or vivid colors.

Multi-turn Dialogue for CoT Construction (Round-4)

请从笔墨、气韵、意境三个层次进行逐级分析。三者之间存在递进关系:
评判时需兼顾艺术真实感、生命感与文化深度，避免仅以画面精致度、工整度或装饰性效果作为高分依据。对于自然笔触的不完美所产生的生命感、节奏感与艺术真实，应给予正向评价；对于过度工整、机械化、装饰化的画面效果，应保持审慎态度。

原则:
笔墨：关注线条质量、运笔控制、墨色层次与结构塑造，判断其是否自然、稳定、生动，而非仅仅是否干净、精细
气韵：关注画面是否具有内在生命流动感、气场贯通性、节奏变化与动势走向，判断作品是否“活”，而非只是构图完整
意境：关注作品是否营造出诗意空间、情绪表达与文化审美高度，是否引发联想与回味，而非堆砌符号或套用意象模板

在综合评价中，意境权重高于气韵，气韵权重高于笔墨。若意境明显不足，即使笔墨精致，整体评分也不应过高。

请按以下顺序进行分析:
1. 笔墨分析：用一段自然语言，评估线条、运笔、墨色变化与造型结构，指出优点与不足。
2. 气韵分析：评估画面整体生命感、动势、节奏与气场流动，判断其是否生动、有呼吸感。
3. 意境分析：评估作品是否营造出明确审美境界，是否具有情绪感染力与文化韵味，是否在有限视觉信息中，构建出超出画面本身的空间感、情绪感与联想空间

按以下格式输出：
笔墨分析: [笔墨分析结果]
气韵分析: [气韵分析结果]
意境分析: [意境分析结果]

(English version)
Please perform a step-by-step analysis across three levels: Brush and Ink, Spirit Resonance, and Artistic Conception. A progressive relationship exists between these three:
The evaluation must balance artistic authenticity, the sense of vitality, and cultural depth. Avoid using visual refinement, neatness, or decorative effects as the sole basis for high scores. Positive recognition should be given to the sense of life, rhythm, and artistic truth arising from the imperfections of natural brushstrokes; conversely, a cautious stance should be maintained toward visual effects that are overly meticulous, mechanized, or purely decorative.

Principle:
Brush and Ink: Focus on line quality, brush control, ink layering, and structural modeling; judge whether it is natural, stable, and vivid, rather than merely whether it is clean or refined.
Spirit Resonance: Focus on whether the image possesses an internal sense of life-flow, a coherent "field", rhythmic variations, and momentum; judge whether the work is "alive" rather than just compositionally complete.
Artistic Conception: Focus on whether the work creates a poetic space, emotional expression, and cultural-aesthetic depth; judge whether it evokes association and reflection rather than merely piling up symbols or applying conventional image templates. In the comprehensive evaluation, Artistic Conception carries more weight than Spirit Resonance, and Spirit Resonance carries more weight than Brush and Ink. If the Artistic Conception is clearly insufficient, the overall score should not be high even if the brush and ink are refined.
In a comprehensive evaluation, the weight of Artistic Conception is higher than Spirit Resonance, and the weight of Spirit Resonance is higher than Brush and Ink. If the Artistic Conception is clearly insufficient, the overall score should not be high, even if the brush and ink are refined.

Please perform the analysis in the following order:
Brush and Ink Analysis: Use natural language to evaluate line quality, brushwork, ink variations, and structural modeling, noting both strengths and deficiencies.
Spirit Resonance Analysis: Evaluate the overall sense of life, momentum, rhythm, and the flow of the "field" within the frame, judging whether it is vivid and possesses a sense of "breathing."
Artistic Conception Analysis: Evaluate whether the work creates a clear aesthetic conception, possesses emotional resonance and cultural charm, and whether it constructs a sense of space, emotion, and association that transcends the visual information of the canvas itself.

Output in the following format:
Brush and Ink Analysis: [Results of Brush and Ink Analysis]
Spirit Resonance Analysis: [Results of Spirit Resonance Analysis]
Artistic Conception Analysis: [Results of Artistic Conception Analysis]

Multi-turn Dialogue for CoT Construction (Round-5)

在综合图像描述、基于题材的艺术分析和评价、RoI感兴趣区域的描述和评价以及三层次分析后，给出这张画作最终的综合艺术评估分数（0-5分），分数必须为整数。

5分：局部与整体在笔墨结构、气韵流动与精神指向上高度同构，意境深远，气韵化生，笔墨随心而不逾法，形成高度统一的生命结构与文化境界，具有不可替代的艺术原创性。
4分：关键局部笔墨精到、气脉贯通，整体节奏自然，意境清远深长，文化气息浓厚，艺术语言成熟稳定，具有鲜明而稳定的艺术品格。
3分：局部具备基本结构与气韵支撑，画面开始“有呼吸”，意境初步成立，但思想深度、文化厚度与局部—整体协同仍有限。
2分：局部描绘精细但重技巧轻生发，整体结构严谨却气韵不足，精神指向薄弱，艺术表达主要停留在技法与形式层面。
1分：构图完整，形象清晰，具备基本绘画表达，局部结构松散，笔墨僵滞，节奏单一，整体气息闭塞，尚未形成完整艺术语言。
0分：画面虽然好看、精致，缺乏笔墨逻辑、气韵流动与意境生成，整体停留在表层视觉美感，更接近插画或装饰图。

按以下格式输出，不要输出额外内容。
最终分数: [整数分数]

(English version)
5 points: Local details and the whole are highly isomorphic in brushwork structure, flow of spirit, and spiritual orientation. The Artistic Conception is profound, Spirit Resonance is transformed into life, and brush and ink follow the heart without violating principles, forming a highly unified life structure and cultural realm with irreplaceable artistic originality.
4 points: Key local details are refined in brush and ink, the spiritual pulse is continuous, the overall rhythm is natural, the Artistic Conception is clear and far-reaching, the cultural atmosphere is thick, and the artistic language is mature and stable, possessing a distinct and stable artistic character.
3 points: Local parts possess basic structure and spiritual support; the image begins to "breathe," and the Artistic Conception is initially established, but depth of thought, cultural thickness, and the synergy between local and whole are still limited.
2 points: Local descriptions are fine but emphasize technique over vitality; the overall structure is rigorous but lacks Spirit Resonance, the spiritual orientation is weak, and artistic expression mainly remains at the level of technique and form.
1 point: The composition is complete and the images are clear, possessing basic pictorial expression; however, local structures are loose, brush and ink are stagnant, the rhythm is singular, and the overall atmosphere is blocked, failing to form a complete artistic language.
0 point: Although the image is attractive or exquisite, it lacks brushwork logic, the flow of spirit, and the generation of the Artistic Conception; the whole remains at the level of surface visual beauty, closer to an illustration or decorative graphic.
Ouput the score in the following format, do not output irreverent content.
Final rating: [Integer Score]