ArtChat: An AI-Driven Conversational Agent for Art Appreciation at the National Palace Museum

Eric Chien, Audrey Hsiao, Kelly Hung, Chu Yun Chu
Group 22

Abstract

Traditional audio guides in museums often provide static, one-way descriptions that overlook the contextual nuance visitors seek. We introduce ***ArtChat****, a single-turn vision–language assistant that answers free-form questions about artworks exhibited at Taiwan's National Palace Museum. Starting from three open-source vision–language models (Qwen2-VL-7B, Qwen2.5-VL-3B, BLIP-2 OPT-2.7B), we apply* parameter-efficient *QLoRA fine-tuning on just 1 000 image–question–answer pairs scraped from the museum website. The entire adaptation fits on a single NVIDIA A100 GPU through 4-bit weight quantisation and low-rank adapters. To evaluate linguistic generalisation without leaving the domain, we propose a 200-question paraphrase benchmark that re-phrases held-out training queries while keeping the original images and answers. Preliminary results show that our fine-tuned models substantially outperform zero-shot baselines on ROUGE-L and BERTScore, and qualitative analysis confirms their ability to recover artist, period, and stylistic details. ArtChat demonstrates that low-cost adaptation of open VLMs can deliver culturally informed, conversational guidance for museum visitors.*

1. Introduction

Interactive and personalized guidance is essential to bring artworks to life. Traditional museum guides and audio tours are often static and fail to convey the deeper context behind art movements and masterpieces. Our project, **ArtChat**, is motivated by our strong connection with Taiwan's cultural heritage and focuses on enhancing the visitor experience at the National Palace Museum. By leveraging state-of-the-art multimodal AI techniques, we aim to create a conversational agent that synthesizes artwork metadata with visual cues to generate rich, contextual answers. In contrast to multi-turn dialogue systems, our design is based on single-turn interaction where each input image and question pair produces a complete, informative response.

2. Related Work

Vision–language pre-training. Contrastive language–image pre-training (CLIP) demonstrated that large-scale image–text pairs yield transferable multimodal features. Generative successors such as BLIP and BLIP-2 freeze the vision encoder and attach a language decoder to enable both understanding and generation [3].

Instruction-tuned multimodal assistants. Adapting large language models to follow multimodal instructions has produced conversational systems including LLaVA [4], and the Qwen-VL family [5]. These models provide strong open-source baselines that can be further specialised with lightweight fine-tuning.

Parameter-efficient adaptation. LoRA introduces low-rank adapters to update only a small subset of weights during fine-tuning [2]. QLoRA combines LoRA with 4-bit weight quantisation, enabling single-GPU adaptation of billion-parameter models with minimal performance loss [1]. Our work adopts QLoRA to adapt three state-of-the-art vision–language models.

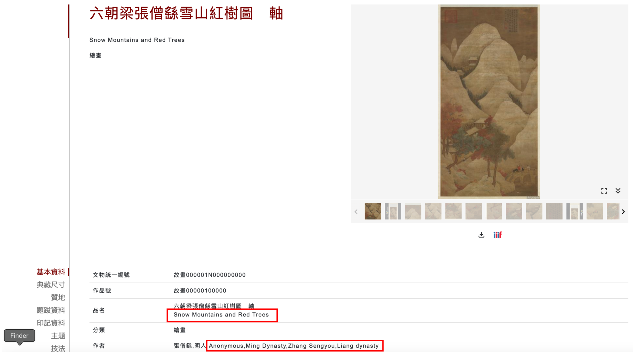
Positioning of our work. We bridge these lines of research by (i) applying *parameter-efficient* QLoRA to adapt open VLMs (Qwen2-VL-7B, Qwen2.5-VL-3B, BLIP-2 OPT-2.7B) with just 1 000 museum-specific QA pairs, and (ii) introducing a paraphrase-based evaluation set that tests linguistic generalisation while remaining within the cultural-heritage domain.

3. Methodology

Relevant scripts are available via this link

3.1. Data Collection and Preprocessing

To construct a high-quality dataset for multimodal question answering, we collected data from the official website of the National Palace Museum using a custom crawler based on Selenium. This crawler systematically navigated through



六朝梁張僧繇雪山紅樹圖　軸

Snow Mountains and Red Trees

繪畫

文物統一編號	故畫000001N000000000
作品號	故畫00000100000
品名	六朝梁張僧繇雪山紅樹圖　軸 Snow Mountains and Red Trees
分類	繪畫
作者	張僧繇,明人 Anonymous,Ming Dynasty,Zhang Sengyou,Liang dynasty

Figure 1. National Palace Museum Web Page

collection pages and extracted metadata, including: **artwork titles**, **artist names**, and **associated image URLs**.

To enrich the dataset, we further visited individual artwork pages to extract additional background information about the artwork.

The final dataset was serialized in line-delimited JSONL format, with each entry comprising an image reference and 3–5 question-answer (QA) pairs. These QA pairs were generated using the DeepSeek language model, with prompts crafted to elicit questions emphasizing cultural and historical context over visual characteristics.

Each example was structured into a conversational format compatible with instruction-tuned vision-language models, consisting of:

- a system-level prompt
- a user message containing both the image and a question
- an assistant response with the answer

To streamline image handling, URLs were hashed and mapped to locally stored files.

3.2. Model Architecture and Fine-Tuning

We fine-tuned three distinct vision-language models: **Qwen2-VL-7B-Instruct**, **Qwen2.5-VL-3B-Instruct**, and **BLIP2-OPT-2.7B**.

Each model comprises a visual encoder and a causal language decoder designed for instruction-following tasks grounded in visual and textual inputs.

To enable fine-tuning on constrained hardware, we employed 4-bit quantization (e.g., NF4 format). We adopted QLoRA, a parameter-efficient fine-tuning strategy, wherein:

- LoRA adapters were injected into selected projection layers (e.g., `q_proj`, `v_proj`).
- Only adapter weights were updated during training, while the backbone parameters remained frozen.

3.3. Training Pipeline

The training pipeline was implemented using the Hugging Face `transformers`, `trl`, and `peft` libraries. Core components include:

- **Data Collator:** Each sample is formatted into model-specific instruction templates. Image inputs are preprocessed accordingly, and text inputs are tokenized with padding and masking for loss computation (excluding special tokens).
- **Optimization:** We employed AdamW optimizer with constant or linear learning rate scheduling. Mixed-precision (bfloat16) and gradient accumulation were used to improve efficiency.
- **Configuration:** Training settings (e.g., effective batch size, number of epochs, gradient checkpointing) were adjusted per model and hardware capacity. Checkpoints were saved at regular intervals for reproducibility.

This training pipeline supports scalable and reproducible fine-tuning of vision-language models for educational and cultural QA tasks.

4. Experimental Settings

4.1. Training Data

To fit our single-GPU budget we fine-tune on **1 000** image–question–answer (IQA) examples drawn from the National Palace Museum corpus (see §3). All 1 000 samples are used for optimisation; no further subdivision is made.

4.2. Evaluation Protocol and Rationale

The system is intended to answer questions about *known* artworks. Evaluating on unseen pieces would test a different generalisation axis, so we instead probe **linguistic generalisation**:

1. Select 100 random IQA examples from the training set.
2. Use deepseek-r1 to generate **two** semantically diverse paraphrases of each question.
3. Retain the original image and answer as ground truth.

This procedure yields a **200-question paraphrase test set** whose images the model has already observed, but whose wordings are novel.

4.3. Decoding & Metrics

- **Generation**: temperature 0.8, top-p 1.0, max new tokens 256.

- **Automatic metrics**: ROUGE-L and BERTScore-F_1 averaged over the 200 paraphrase queries.

5. Results and Discussion

As noted in Section 4, we use BERTScore and ROUGE to quantitatively evaluate text quality. We also manually reviewed a subset of test outputs to assess aspects beyond automated metrics, such as factual accuracy, coherence, and relevance. For fairness, we used the best-performing checkpoint (highest BERTScore F1) for each model—Qwen2.5-VL-3B and Qwen2-VL-7B—to generate and compare answers against ground truth, providing qualitative insights into each model's optimal performance.

5.1. Quantitative analysis

First, we examine the BERTScore and ROUGE-L for the **Qwen2-VL-7B**'s performance. As shown in Figure 2, both metrics demonstrate a generally upward trend as training progresses, with the BERTScore peaking at checkpoint 250. This indicates that the model's ability to produce semantically meaningful and lexically aligned answers improved steadily during training. We selected the checkpoint with the highest BERTScore for further qualitative human evaluation, as it reflects the model's best semantic generation capacity.

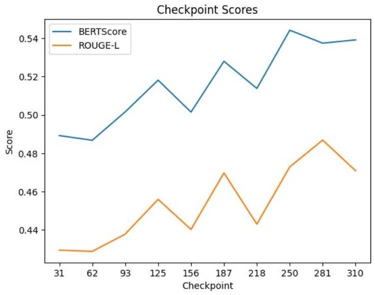


Figure 2. BERTScore and ROUGE-L scores across checkpoints for Qwen2-VL-7B.

To further contextualize our model performance, we also included **Qwen2.5-VL-3B** and **BLIP2-OPT-2.7B** in our evaluation. As shown in Figure 3. Across all stages, Qwen2-VL-7B significantly outperforms the smaller 3B model in both BERTScore and ROUGE-L. BLIP2 achieves moderately better scores than Qwen2.5-VL-3B across both BERTScore and ROUGE-L, but still falls short of Qwen2-VL-7B.

The BLIP2 model shows relatively stable performance across checkpoints, suggesting general consistency in its vision-language reasoning capabilities. However, its lower semantic and lexical scores compared to the fine-tuned Qwen2-VL-7B highlight the benefits of task-specific adaptation and larger model capacity.

Overall, the results reaffirm that model size and fine-tuning have a significant impact on performance in vision-language generation tasks, especially under open-ended conditions where semantic understanding is critical.

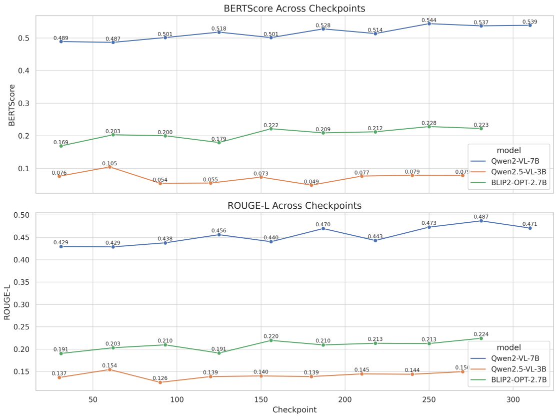


Figure 3. BERTScore and ROUGE-L scores across checkpoints for 3 different VLMs

Qwen2.5-VL-3B maintains low scores (BERTScore ¡ 0.08, ROUGE-L 0.14), showing weak alignment with reference answers. In contrast, Qwen2-VL-7B steadily improves, peaking at a BERTScore of 0.544 (checkpoint 250) and ROUGE-L of 0.487 (checkpoint 281), underscoring the advantage of larger models in handling multimodal semantics and long-range dependencies.

We also argue that BERTScore is more reliable than ROUGE for evaluating open-ended generation, as it better captures semantic similarity despite varied phrasing—making it well-suited for vision-language tasks.

5.2. human qualitative evaluation: Accuracy, Relevance, and Fluency

In addition to automatic metrics, we conducted a qualitative human evaluation of generated responses by manually reviewing a random subset of the test set outputs. We assessed each model output according to three criteria:

1. **Accuracy** – factual correctness of the response (e.g., correct names, numbers).
2. **Relevance** – whether the response directly addresses the visual or textual prompt.
3. **Fluency and Style** – grammaticality, coherence, and appropriateness of language.

During this human review, we observed a consistent pattern: although the Qwen2.5-VL-3B model often generated longer responses that included more descriptive or embellished language, the content was frequently inaccurate or only loosely related to the input question and image. In contrast, Qwen2-VL-7B responses, while generally shorter and more concise, demonstrated significantly higher factual accuracy and stronger alignment with the ground-truth answers.

This observation aligns with the quantitative results shown in Figure 3, where Qwen2.5-VL-3B scores substantially lower in both BERTScore and ROUGE-L. The human evaluation thus reinforces the conclusion that larger model capacity leads to more reliable and semantically grounded responses in multimodal settings.

To illustrate this difference, we provide a representative example from the test set, where both models are prompted with the same visual input and question. For this data point, we gave the model is a traditional Chinese landscape painting shown in Figures 4, accompanied by a question asking for the artist's name.

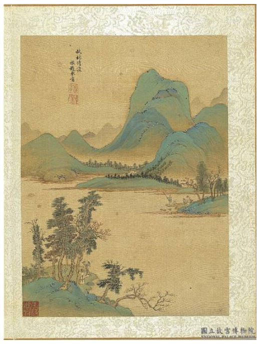


Figure 4. An example data point in test set

Figure 6 shows the comparison between the two models' responses to the same input. From this example, we observe that the Qwen2.5-VL-3B model tends to generate longer responses that incorporate descriptive elaboration, but sometimes at the cost of factual precision. In this case, the model adds biographical details and interprets the painting title, yet risks hallucinating information or introducing less relevant content. On the other hand, the Qwen2-VL-7B model provides a concise and accurate answer that more closely matches the ground truth.

6. Conclusion and Future Work

In this work, we introduced **ArtChat**, a generative, single-turn vision-language model tailored for art appreciation at the National Palace Museum. Our approach leverages pretrained vision and language models with an adapter that is fine-tuned on museum-specific data. Initial experiments show promising improvements in generating accurate and contextually rich descriptions. We also performed a qualitative and quantitative evaluation for our result, which provided a deeper understanding of the differences between the behavior of the two models and the strengths in multimodal understanding in the real world.

Field	Content
Question	Who painted the artwork titled 清王翬畫山水冊仿趙孟頫秋林待渡?
Answer from Qwen2-VL-7B	The artist is Wang Hui (王翬), a prominent Qing-dynasty painter.
Answer from Qwen2.5-VL-3B	The painting is by Wang Hui (1682–1755) in the Qing dynasty, known for his landscapes. The artist’s name is “王翬”, and it is part of an album featuring “仿趙孟頫秋林待渡”(Changhui Zhao Mengfu’s *Autumn River and Waiting Ferry*).
Ground Truth	The artist is Wang Hui (王翬), a prominent Qing-dynasty painter.

Figure 5. Output Comparison

Future work includes:

- Expanding the dataset and further data augmentation.
- Investigating selective unfreezing of the backbone for enhanced domain adaptation.
- Improving inference speed and multi-language TTS integration.
- Conducting user studies with museum visitors for qualitative evaluation.

References

[1] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient fine-tuning of quantized Language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.14314.

[2] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2021. arXiv:2106.09685.

[3] J. Li, D. Li, S. Savarese, and S. C. Hoi. BLIP-2: Bootstrapping language–image pre-training with frozen image encoders and large language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023. arXiv:2301.12597.

[4] H. Liu, P. Li, X. Chen, J. Li, Z. Zhang, Z. Hu, Y. Wang, C. Tao, S. Tang, Y. Chuang, and M. Zhou. Visual instruction tuning. In *arXiv preprint arXiv:2304.08485*, 2023.

[5] Q. Team. Qwen-VL: A versatile vision–language model. arXiv preprint arXiv:2403.05616, 2024.