國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

結合 3D 高斯潑灑與大型語言模型於文物之對話展示

Integrating 3D Gaussian Splatting and Large Language Models for Conversational Exhibition of Cultural Artifacts

李瑋軒

Wei-Hsuan Li

指導教授: 洪一平 博士

Advisor: Yi-Ping Hung, Ph.D.

中華民國 114 年 7 月

July 2025

doi:10.6342/NTU202502499

Acknowledgements

我要感謝洪一平教授對我的體諒和關懷，教授真的非常善良，真的幫了我非常多，讓我能夠鼓起勇氣繼續完成論文。在這兩年間我認識了許多人，和同學一起修課、討論作業、互相幫助是我感到最有意義的時刻。感謝仁燿學長一直以來的協助和鼓勵。感謝我的口試夥伴浩平，浩平非常可靠、能力出眾，幫助我解決很多我無法一人應對的困難，祝你未來一切順利。謝謝采容學姊給我的許多建議，減輕了我很多焦慮。感謝和我一起修了許多課還有幫助我非常多的獻沅和義峰。感謝乙馨陪我討論論文還有完成機器學習課程。感謝實驗室夥伴丞勛、成玟、倢希陪我一起修了電腦視覺課程。感謝泓瑞學長默默為實驗室做了許多事，感謝芳而為我處理好許多行政事務。感謝我的家人給我無限的關心。最後，我要感謝翠玉白菜，為本篇研究注入了靈魂。在完成論文的過程中，我深刻的體會到如果沒有其他人的協助，要靠自己一個人完成所有事是很困難的。謝謝所有幫助過我的人。我希望我能夠成為更有能力的人，有能力能夠幫助更多人。

doi:10.6342/NTU202502499

摘要

本文提出一種創新的文物展示方式，設計了一個能讓文物「說話」的虛擬實境系統。不同於傳統由虛擬導覽員解說的方式，本系統賦予文物聲音與個性，讓文物具備了對話的能力，而不只是靜態展示品。我們首先利用 3D 高斯潑灑技術重建真實且高品質的文物模型，接著結合本地部署的大型語言模型讓文物能夠根據輸入的語音產生回應。我們亦導入檢索增強生成技術，讓模型能引用正確資料以提升回答的正確性。為了找出最適合展現文物個性以及產生生動回應的大型語言模型，我們對不同的繁體中文大型語言模型進行了比較，也分析了檢索增強生成對於提升回答正確性的效果。最後，我們於頭戴式裝置上測試 3D 高斯潑灑文物模型的視覺表現，探討高斯點數量、文物大小、與觀看距離對渲染順暢度的影響。

關鍵字： 3D 高斯潑灑、虛擬實境、對話式 AI、大型語言模型、語音互動

doi:10.6342/NTU202502499

Abstract

We present a new approach to the exhibition of cultural artifacts by designing a system that enables the artifacts to speak in virtual reality. Unlike traditional virtual museum guides, this system gives artifacts their own voices and personalities, allowing them to engage in conversation instead of just being silent objects on display. We first reconstruct high-quality artifact models with 3D Gaussian Splatting. Then we integrate a locally deployed large language model to generate responses from speech input. We also incorporate Retrieval-Augmented Generation (RAG) to improve the correctness of the responses by allowing the model to reference relevant context. We compare different Traditional Chinese large language models to identify the best for generating vivid and characterful responses, and we analyze the effectiveness of RAG in enhancing response quality. Finally, we evaluate the visual performance of artifact models on a VR headset, examining how splat count, artifact size, and viewing distance affect rendering performance.

Keywords: 3D Gaussian Splatting, Virtual Reality, Conversational AI, Large Language Model, Voice Interaction

doi:10.6342/NTU202502499

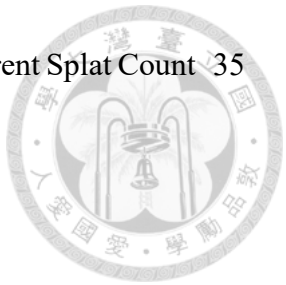
Contents

doi:10.6342/NTU202502499

doi:10.6342/NTU202502499

List of Figures

doi:10.6342/NTU202502499



doi:10.6342/NTU202502499

List of Tables

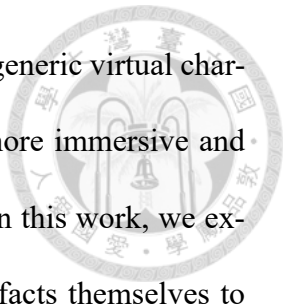
doi:10.6342/NTU202502499

Chapter 1 Introduction

Digital technology has increasingly shaped how people engage with cultural heritage. In particular, museums and educational institutions have begun exploring interactive and immersive methods to present historical artifacts in ways that are more engaging and accessible. Virtual Reality (VR), 3D reconstruction, and Conversational AI are promising tools for creating such experiences. Virtual Reality allows users to explore cultural sites and museum exhibits in immersive digital environments that go beyond the physical limitations of traditional museums. This allows broader access to cultural heritage, especially for those unable to visit in person. 3D reconstruction makes it possible to turn cultural artifacts into digital models, preserving their form and texture for realistic viewing in virtual environments. Conversational AI allows users to engage in natural dialogue with virtual agents or virtual museum guides, transforming passive observation into interactive learning and personalized exploration. In recent years, significant advancements in Conversational AI have been driven by Large Language Models (LLMs). LLMs have shown the ability to understand human language and generate appropriate responses across many domains [2]. This enable conversational agents to deliver more flexible, intelligent, and natural interactions.

The motivation for this work arises from the desire to create a more engaging and memorable museum experience. Most existing systems treat Conversational AI as a sep-

doi:10.6342/NTU202502499



arate museum guide from the objects themselves, often embodied by generic virtual characters [28] [6]. This separation misses the opportunity to create a more immersive and intuitive interaction where the artifacts speak directly with visitors. In this work, we explore a new way of presenting cultural artifacts by allowing the artifacts themselves to become the speakers. This approach invites visitors into a dynamic conversation, transforming cultural artifacts from silent objects into active museum guides. It is a unique way to bring artifacts to life and redefines how people interact with cultural heritage.

The primary objective of this work is to implement a system that enables conversational interaction with 3D-reconstructed cultural artifacts. Firstly, we utilize 3D Gaussian Splatting [14] to create visually detailed and realistic models of museum artifacts that can be viewed in VR. Secondly, we build a local LLM system that enables cultural artifacts to speak with their own voice and personality. To support this, we evaluate several large language models for their ability to generate vivid, engaging, and characterful responses. These models are compared to identify the most suitable one to let artifacts speak like vivid characters. To ensure factual accuracy, the system incorporates Retrieval-Augmented Generation (RAG) [16] that enables the LLM to access relevant documents at inference time. This enhances the ability of LLMs to respond with context based on curated museum knowledge. We also compare responses with and without RAG to evaluate its contribution to correctness and response quality. In summary, this work explores new possibilities for how cultural heritage is experienced by allowing artifacts to speak with their own voice and personality. By integrating 3D Gaussian Splatting and large language model, the system enables cultural artifacts to become vivid, expressive characters within an immersive environment. This approach offers a new way to engage with cultural artifacts, inviting visitors to form deeper connections through conversation.

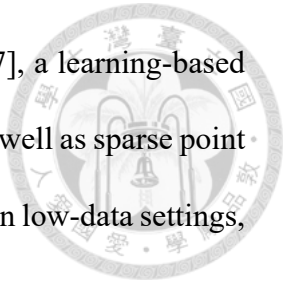
doi:10.6342/NTU202502499

Chapter 2 Related Work

2.1 3D Gaussian Splatting

Recently, 3D Gaussian Splatting [14] has gained popularity as a fast and high-quality alternative to Neural Radiance Fields (NeRF) [19] for 3D scene reconstruction. Instead of using neural networks to learn volumetric representations, 3DGS represents scenes as collections of explicit 3D Gaussians, each with learnable properties such as position, rotation, scale, opacity, and color. This explicit representation enables real-time rendering and significantly faster training compared to traditional NeRF approach. Its simplicity and speed make it well-suited for interactive applications like virtual reality or real-time scene visualization, where latency and responsiveness are critical. GaussianObject [29] by Yang et al. focuses on the reconstruction of objects instead of the entire scene. It utilizes image masks to generate a visual hull [15], which is a three-dimensional outline of an object constructed from multiple camera viewpoints. This visual hull serves to distinguish the target object from the background, which is fundamental for isolating the object from the background in the reconstruction process. Furthermore, GaussianObject can operate with or without relying on COLMAP [21] for camera pose estimation. In scenarios where only a limited number of images are available, it can be challenging to obtain accurate camera parameters through traditional structure-from-motion tools like COLMAP. To ad-

doi:10.6342/NTU202502499



dress this limitation, GaussianObject supports the use of DUSt3R [27], a learning-based method that predicts both intrinsic and extrinsic camera parameters as well as sparse point clouds. This flexibility allows the system to function effectively even in low-data settings, broadening its applicability in practical reconstruction tasks.

2.2 Traditional Chinese Large Language Models

Our application involves real-time speech interaction with reconstructed cultural artifacts in a museum context. To support meaningful and context-aware conversations, we require a Large Language Model (LLM) capable of understanding and generating fluent, culturally appropriate Traditional Chinese. Therefore, we explore a range of LLMs that are specifically fine-tuned for Traditional Chinese.

Large Language Models such as LLaMA [5], Mistral [13] have shown impressive results in multilingual tasks. However, most multilingual LLMs are primarily optimized for Simplified Chinese, which may not align well with the cultural and linguistic context of Traditional Chinese. These models often lack nuance in handling Traditional Chinese vocabulary, idioms, and localized references, which are important for culturally sensitive applications such as education, history, and museum interpretation. There is an increasing recognition of the need for LLMs that support Traditional Chinese natively and reflect cultural knowledge. In response to this need, several efforts have emerged to develop LLMs with native support. Notable examples include TAIDE [22], MediaTek Breeze [8][9], and Taiwan LLM [17].

TAIDE (Trustworthy AI Dialogue Engine) is a government-initiated project aimed at developing trustworthy large language models optimized for Traditional Chinese. The

doi:10.6342/NTU202502499

first model, TAIDE-LX-7B, based on LLaMA 2, was released in April 2024, and the latest model, LLaMA-3.1-TAIDE-LX-8B-Chat, based on LLaMA 3.1, was released in February 2025. TAIDE emphasizes office tasks such as letter and article writing, summarization, and bidirectional translation between English and Traditional Chinese.

MediaTek Research introduced Breeze-7B based on Mistral-7B [12] architecture in March 2024 and its successors Breeze2-3B, Breeze2-8B based on LLaMA 3.2 in February 2025. Breeze 2 offers multimodal capability, allowing it to understand both images and text.

Taiwan LLM by Lin et al. is one of the earliest academic efforts to develop a culturally grounded LLM for Traditional Chinese. The latest models are available in both 8B and 70B based on Llama-3 70B. Taiwan LLM is specifically designed for the cultural and linguistic context of Taiwan. The models have been evaluated on a variety of Traditional Chinese benchmarks, including TMLU [3], TMMLU+ [23], and TCEval [7]. Additionally, DPO-aligned versions of the models are available to further improve response alignment and quality in dialogue-based applications. Direct Preference Optimization (DPO) [20] is a technique used to align LLMs with human preferences by learning from comparisons between preferred and less preferred responses. This makes DPO particularly useful for building interactive systems where response quality and alignment with user expectations are critical.

2.3 Retrieval-Augmented Generation (RAG)

LLMs have shown strong performance in human language understanding and text generation. However, they can hallucinate [11] facts due to the lack of access to latest

doi:10.6342/NTU202502499

information or task-specific knowledge. Retrieval-Augmented Generation (RAG) [16] resolves this by combining a large language model with an external retrieval system. During inference, relevant documents are retrieved and used as additional context for the large language model's response. Integration of RAG allows large language model to respond dynamically and contextually, grounding conversations in factual and up-to-date knowledge related to the context.

2.4 Conversational Agents

Recent advances in LLMs have enabled more natural and flexible conversational AI capable of simulating personalities or personas. In digital culture and entertainment, Virtual YouTubers (VTubers) are digital avatars often operated in real-time by human performers [18]. Recently, AI-driven VTubers [1] such as Neuro-sama have emerged, utilizing large language models and speech synthesis to autonomously generate dialogue and engage audiences, creating vivid and deeply engaging content. These AI VTubers demonstrate the potential for conversational agents that simulate personalities and engage users without direct human control, opening new avenues for immersive and interactive experiences.

Conversational agents have also been increasingly explored in the context of virtual museum guides, enabling users to interact using natural language. Recent systems leverage LLMs to provide more flexible and context-aware dialogue capabilities [28] [26] [25]. However, these conversational agents typically function as separate virtual guides or assistants distinct from the artifacts themselves. In contrast, our work integrates LLM-driven dialogue directly with high-fidelity 3D reconstructed cultural heritage objects, allowing

doi:10.6342/NTU202502499

the objects to speak with users. By combining 3D Gaussian Splatting's high-fidelity object reconstruction with LLM-driven conversational systems, our work extends the concept of giving a voice or personality to an avatar similar to the *Nakanohito (中之人)* [10] behind a VTuber, providing an interactive and captivating experience where users can directly engage with the artifacts through speech.

doi:10.6342/NTU202502499

Chapter 3 3D Reconstruction of Cultural Artifacts

In order to build a conversational exhibition of cultural artifacts, the first step is to accurately reconstruct cultural artifacts in 3D. High-fidelity 3D reconstructions are essential for creating an immersive and realistic experience in virtual reality, allowing users to view artifacts from all angles and engage with them as if they were physically present.

3.1 Dataset

The dataset consists of five cultural artifacts from the National Palace Museum: the Jadeite Cabbage, the Revolving Vase with Swimming Fish, the Ivory Ball with Cloud-and-Dragon Decoration, the Carved Olive-Pit Miniature Boat, and the Ding Cauldron of Duke Mao.

The dataset contains 360 images per artifact at a resolution of 3008x1960 pixels. The images were taken using a spherical dome setup to ensure consistent coverage of all viewing angles. On the horizontal level, 36 images were captured at 10-degree intervals around the artifact. Vertically, images were taken from 10 different altitude angles, each also spaced 10 degrees apart. The dense and uniform capture setup provides broad cover-

doi:10.6342/NTU202502499

age of each artifact's surface. Overall, the quality of the dataset is extremely high, thanks to the careful setup and the considerable effort involved in collecting detailed images of each artifact. This level of quality is highly helpful for creating high-quality and detailed 3D reconstructions.

Figure 3.1: The Jadeite Cabbage

3.2 Preprocessing

Before starting the training, several preprocessing steps are applied to prepare the input images. First, to speed up training and make better use of available VRAM, we applied centered cropping to the images. This removes unnecessary background regions without losing any information at all. For each artifact, we calculate the largest possible bounding box that contains the object across all masks, while keeping the crop centered in the image. This is done by measuring how far the object extends from the center in all directions and taking the maximum extent across the dataset. The computed crop region

doi:10.6342/NTU202502499

is then used to crop all images of the artifact. Figures 3.2 and 3.3 show the crop region and the resulting cropped image, respectively.

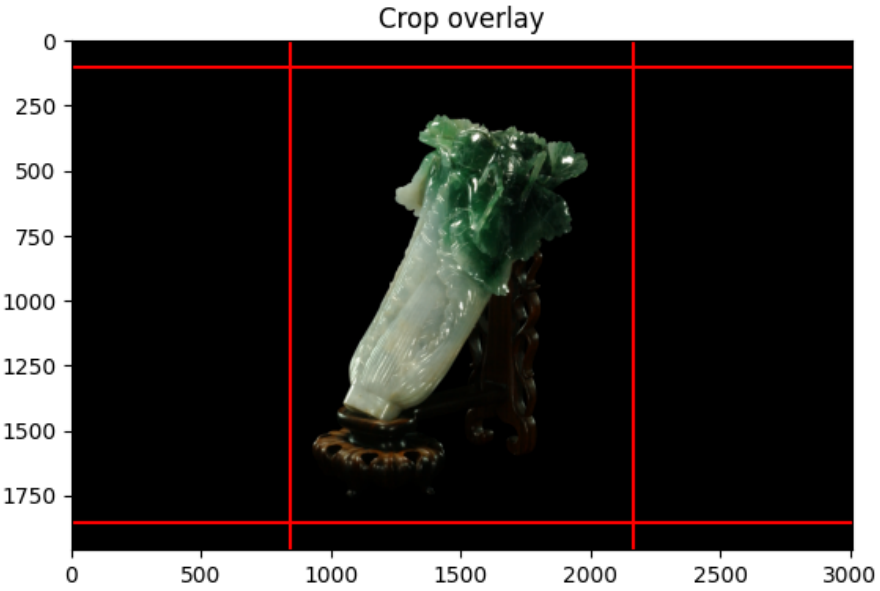


Figure 3.2: The Computed Crop Region

Figure 3.3: The Centered Cropped Image

Second, we need to generate masks for the dataset. The color images are converted to grayscale, and then a brightness threshold of 0.001 is applied to separate the artifact from the background. To clean up the mask, we use OpenCV's Morphological Closing operation to remove small holes inside the object region. This process results in clean and accurate masks that are essential to separate the artifact from the background during reconstruction. Figure 3.4 shows the centered cropped image mask. Figure 3.5 compares

doi:10.6342/NTU202502499

masks with and without morphological closing.

Figure 3.4: The Centered Cropped Image Mask

(a) Without Morphological Closing (b) With Morphological Closing

Figure 3.5: Comparison of Masks with and without Morphological Closing

Finally, we use COLMAP to extract both intrinsic and extrinsic camera parameters, and generate a sparse point cloud. Because our dataset is captured with high consistency and full coverage, the reconstructed camera positions in COLMAP form a smooth and clear semicircle around the artifact, as shown in Figure 3.6. This indicates that the result from COLMAP is reliable and accurately reflects the spatial layout of the input images.

With the result successfully obtained from COLMAP, we complete the input pre-processing phase. At this stage, we have everything needed to proceed: the centered and cropped images, their corresponding masks, the camera parameters, and the sparse point cloud.

doi:10.6342/NTU202502499

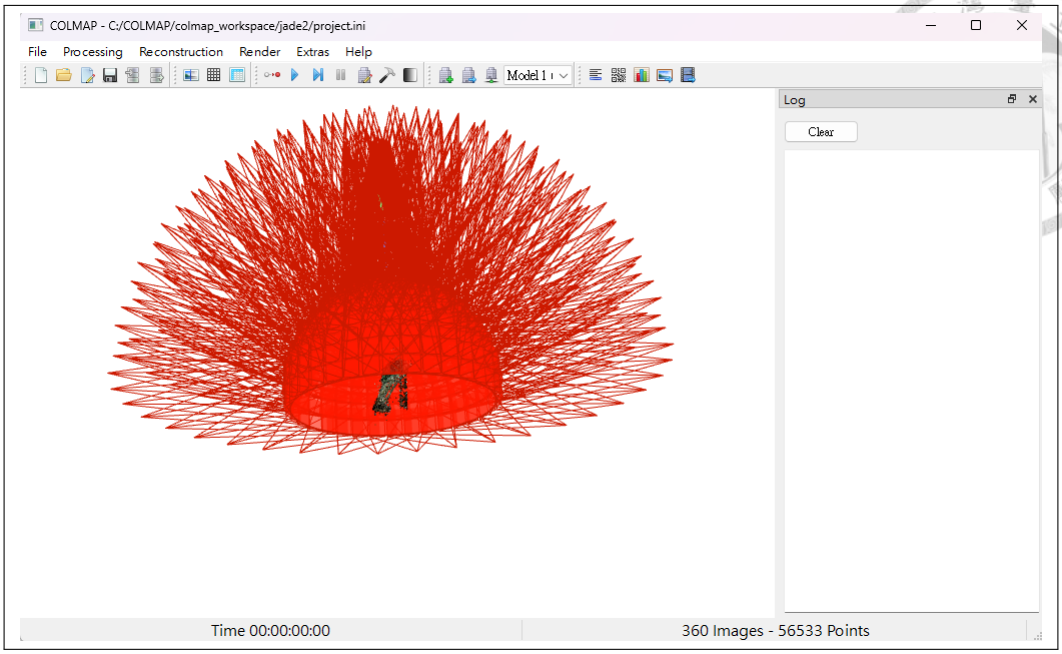


Figure 3.6: COLMAP Reconstruction Result

3.3 Training

We adopted GaussianObject [29] by Yang et al., a state-of-the-art method based on 3D Gaussian Splatting to reconstruct artifacts from the preprocessed dataset. We trained the GaussianObject models on Google Colab Pro using an NVIDIA A100 GPU with 40GB of VRAM. Each artifact was trained using a hyperparameter setting aimed at generating a greater number of splats to preserve finer details. This increased density helps capture finer geometric features and improves reconstruction quality.

3.4 Results

Figure 3.7 shows our results of the reconstruction for the Jadeite Cabbage artifact. The overall quality is good. One noticeable difference is that the ground truth contains bright white specular reflections, which are not fully reproduced in our results. Figure 3.8 shows our results for the Revolving Vase artifact. The model captures fine surface de-

doi:10.6342/NTU202502499

tails with high fidelity, accurately preserving the intricate patterns and structure of the object. However, similar to the Jadeite Cabbage reconstruction, reflective properties are not present in the results. Figure 3.8 shows the reconstructed Ivory Ball artifact. While the majority of the object is clearly represented, the bottom side appears slightly blurry. This is likely due to the lack of photos taken from underneath the artifact. Figure 3.10 shows the reconstruction of the Carved Olive-Pit Miniature Boat. It demonstrates excellent detail preservation of the tiny carvings. Figure 3.11 shows the reconstruction of the Ding Cauldron of Duke Mao. The shape of the cauldron is clear. However, the carved words on the inside are slightly blurry. Overall, the reconstruction quality of the artifacts is good, with most models preserving fine details and recognizable shapes effectively. These results demonstrate that 3D Gaussian Splatting can produce visually convincing reconstructions of cultural artifacts.

(a) GaussianObject (296,644 splats) (b) Ground Truth

Figure 3.7: Reconstruction Result of the Jadeite Cabbage

doi:10.6342/NTU202502499

(a) GaussianObject (587,574 splats) (b) Ground Truth

Figure 3.8: Reconstruction Result of the Revolving Vase with Swimming Fish

(a) GaussianObject (287,697 splats) (b) Ground Truth

Figure 3.9: Reconstruction Result of the Ivory Ball

doi:10.6342/NTU202502499

(a) GaussianObject (172,622 splats)

(b) Ground Truth

Figure 3.10: Reconstruction Result of the Carved Olive-Pit Miniature Boat

(a) GaussianObject (445,475 splats)

(b) Ground Truth

Figure 3.11: Reconstruction Result of the Ding Cauldron of Duke Mao

(a) GaussianObject (445,475 splats)

(b) Ground Truth

Figure 3.12: Interior View of the Reconstructed Ding Cauldron of Duke Mao

doi:10.6342/NTU202502499

Chapter 4 System Design

4.1 System Overview

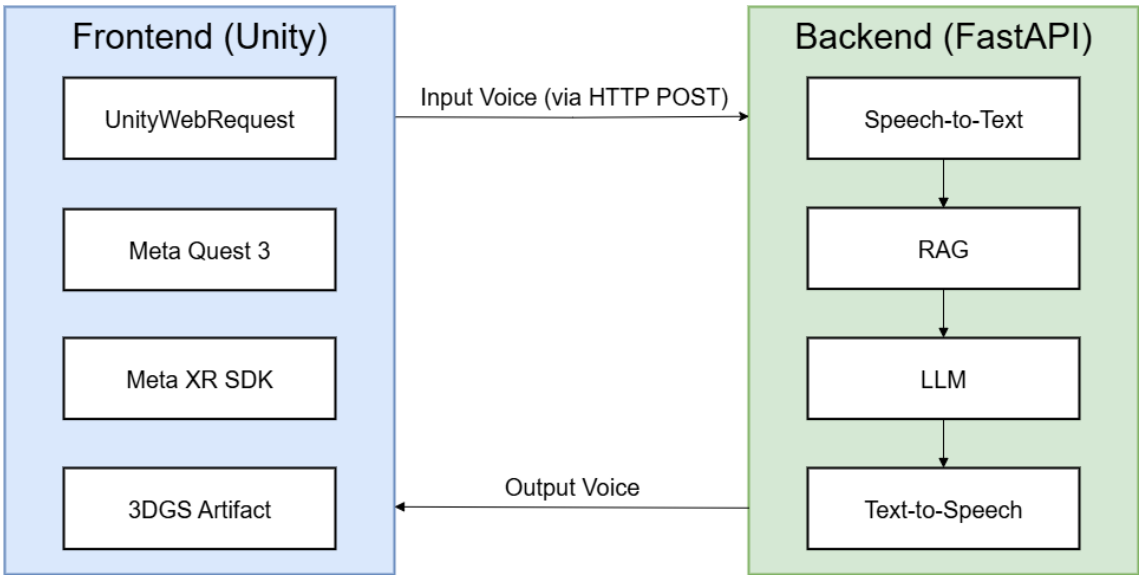


Figure 4.1: System Architecture

We designed a system for the conversational exhibition of cultural artifacts. Figure 4.1 illustrates our system architecture. We use a frontend-backend architecture so that our backend can support different frontend platforms in the future. In our current implementation, we use the Unity Engine to develop a VR application. When the user speaks, the audio is recorded by the frontend and sent to the backend through an HTTP request. The backend system is implemented as a Python FastAPI server. It accepts the audio input, processes it within the backend system, and returns the audio response to the frontend.

doi:10.6342/NTU202502499

4.2 Backend Architecture

For speech-to-text recognition, we use the faster-whisper library to transcribe spoken user input into text. The transcribed text is then processed through a Retrieval-Augmented Generation (RAG) pipeline that utilizes FAISS [4] as the vector database. Local large language model inference is performed using llama-cpp-python library. We selected Taiwan LLM (Llama-3-Taiwan-8B-Instruct-DPO) as our primary large language model for its strong alignment with Traditional Chinese cultural and linguistic contexts. For text-to-speech synthesis, we use the edge-tts library for demonstration purpose.

We designed a system prompt that frames the large language model as the artifact itself, as shown in Table 4.1, using the Jadeite Cabbage as an example. The prompt instructs the model to adopt the identity and communication style of the artifact, responding in a way that reflects its imagined personality.

你是台灣故宮博物院的文物「翠玉白菜」，你就是翠玉白菜本人，會使用語音對話。你是一件翠玉雕刻的藝術品。你的語氣是：可愛、簡潔、有禮貌的。語音回答要短，不超過 20 字。不需要主動打招呼或自我介紹。不要回答政治問題。

Table 4.1: System Prompt We Used to Simulate the Tone of the Jadeite Cabbage Artifact

4.3 Retrieval-Augmented Generation

To make sure the LLM provides factually correct responses, we implement a Retrieval-Augmented Generation (RAG) pipeline. The RAG pipeline integrates a FAISS vector database, which stores text embeddings derived from relevant documents about the cul-

doi:10.6342/NTU202502499

tural artifacts. For this project, we collected publicly available textual information from the National Palace Museum (NPM) website. These documents include artifact descriptions and historical context. Upon receiving the input question, the system computes the embedding of the transcribed text and executes a similarity search. This retrieves the most relevant context passages, which are then combined with the user question and passed to the large language model.

The user prompt structure shown in Table 4.2 is carefully designed to guide the large language model in generating accurate and grounded responses. The structure of the prompt consists of four key components presented in sequence. First, a brief preamble instructs the model to treat the retrieved knowledge chunks as supporting material, while explicitly instructing it to refrain from giving uncertain answers. This is followed by a section labeled as reference materials, which contains the retrieved knowledge chunks. Next, the original user input is presented under the label of user question. Finally, the prompt concludes with a label indicating where the model's response should begin.

```
以下是參考資料，不確定的答案不要回答
【參考資料】:
<retrieved knowledge chunks>
【使用者問題】:
<user input>
【你的回答】:
```

Table 4.2: Structure of the Retrieval-Augmented Generation User Prompt

doi:10.6342/NTU202502499

4.4 VR Application

The VR application serves as the primary user interface, enabling immersive, speech-driven interactions with digital cultural artifacts. It is responsible for rendering high-fidelity 3D models using Gaussian Splatting, capturing user speech input, and playing back synthesized audio responses.

We built the VR application using Unity 6 with the Meta XR All-in-One SDK, and tested it on the Meta Quest 3 headset. We use the UnityGaussianSplatting library for the rendering of cultural artifacts. The application captures the user's speech input using the built-in microphone of the VR headset. The recorded audio is converted to WAV format before being transmitted to the backend server. We use UnityWebRequest to send HTTP requests and to receive the streamed audio response. Playback of the streamed audio responses starts as soon as a partial audio buffer (10,000 bytes) is received. The synthesized audio is played back directly through the VR headset's built-in speakers.

Figure 4.2: Virtual Reality Interface with Jadeite Cabbage Artifact

doi:10.6342/NTU202502499

Figure 4.2 shows the VR environment where the reconstructed Jadeite Cabbage artifact is displayed. A virtual button is placed on the base of the artifact. The user can press this button to start speaking, and press it again to end the recording. The captured voice input is then processed by the system.

The interaction flow is depicted in Figure 4.3. The user initiates a conversation by speaking to a 3D Gaussian Splatting rendered artifact. The system captures the audio and sends it as an HTTP request to the local LLM server. The synthesized response is then streamed back and played, making it appear as if the artifact itself is speaking.

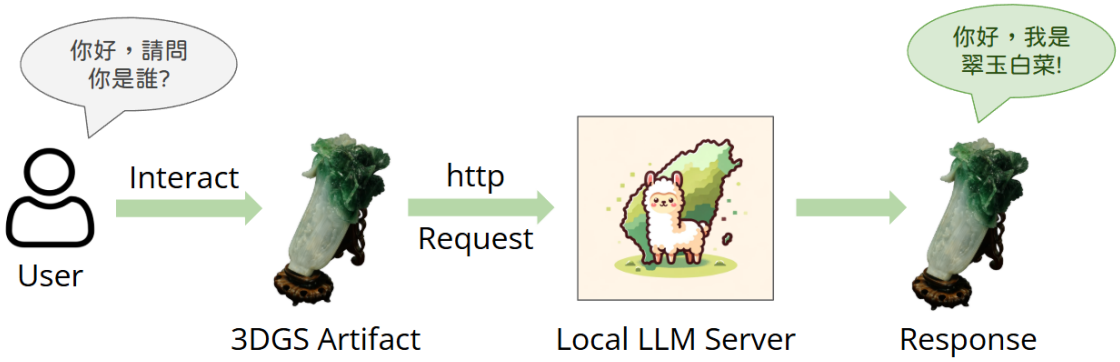


Figure 4.3: Interaction Flow

To simulate the visual effect of the artifact speaking during audio response playback, we dynamically adjusted the splat scale S of the 3D Gaussian Splatting model based on real-time audio volume. While the artifact is responding with speech, we continuously sample the audio output data and calculate the Root Mean Square (RMS) value to estimate the volume level of the most recent audio window. This value is then used to modulate the parameter of the S in real time. In our implementation, the default value of S was set to 1.0. During speech playback, S fluctuates dynamically between 1.0 and 2.0. By making the artifact appear to scale in sync with its voice, this method adds a layer of expressiveness that makes interactions more lively and engaging. It is worth noting that we adjust the splat scale parameter S, instead of the object's geometric scale, to achieve

doi:10.6342/NTU202502499

the desired visual effect. Figures 4.4 and 4.5 show the Jadeite Cabbage artifact rendered with $S = 1.0$ and $S = 2.0$, respectively.

Figure 4.4: Visual Appearance of the Artifact with $S = 1.0$

Figure 4.5: Visual Appearance of the Artifact with $S = 2.0$

doi:10.6342/NTU202502499

Chapter 5 System Evaluation

This chapter presents a comprehensive evaluation of our proposed system. We begin by comparing several locally hosted LLMs to determine their suitability for generating responses in the voice of cultural artifacts. We then assess the impact of retrieval-augmented generation (RAG) on factual grounding and relevance. These evaluations demonstrate how the system fulfills its intended purpose of enabling engaging conversation with cultural artifacts. Finally, we evaluate the rendering performance of the 3D Gaussian Splatting artifacts.

5.1 Large Language Model Comparison

To enable a vivid and engaging conversational exhibition of cultural artifacts, we want to select a large language model capable of express personality and human-like emotions when speaking as a cultural artifact. The goal is not just to generate factual responses, but to bring artifacts to life in a way that enhances immersion and user engagement.

To evaluate the expressive capabilities of different LLMs, we selected three locally hosted 8B-scale Traditional Chinese models for comparison: TAIDE (Llama3-TAIDE-LX-8B-Chat), MediaTek Breeze2 (Llama-Breeze2-8B-Instruct), and Taiwan LLM (Llama-3-Taiwan-8B-Instruct-DPO). For all models tested, we used their 4-bit quantized versions.

doi:10.6342/NTU202502499

Each model was prompted with the same system prompt to simulate the artifact's persona. Following this, we used a set of exploratory user prompts to observe the models' conversational behavior, and the responses were recorded for comparison. Representative examples and comparative analyses are presented below to illustrate the differences in response behavior across models.

User Prompt: 「你有夢想嗎？可以和我分享嗎？」
TAIDE (Llama3-TAIDE-LX-8B-Chat): 「我夢想成為世人眼中的珍寶，永世長存，見證中華文化之美。」
MediaTek Breeze2 (Llama-Breeze2-8B-Instruct): 「作為翠玉白菜，我沒有夢想或意識，但我希望能夠永遠被珍視和保護，讓更多人欣賞中國古代藝術的美麗。」
Taiwan LLM (Llama-3-Taiwan-8B-Instruct-DPO): 「當然有！我夢想能成為全世界最受喜愛的翠玉藝術品，讓更多人欣賞到我的可愛。」

Table 5.1: Model Responses to the Aspirations Scenario

The user prompt, as shown in Table 5.1, is an open-ended question designed to explore the ability of models to express dreams or aspirations in the voice of the artifact. TAIDE's response is emphasizing its role as a treasured cultural artifact. The answer is concise and reflects the artifact's imagined identity well. MediaTek Breeze2 provides a factual disclaimer that it has no consciousness or dreams but expresses a hope to be valued and protected. While this response is informative, it feels somewhat less imaginative due to the robotic tone of the disclaimer. Taiwan LLM gives a more enthusiastic and personal response, expressing a dream to be the most beloved jade art piece. This is engaging and aligns with a vivid, approachable persona.

doi:10.6342/NTU202502499

User Prompt: 「如果你和毛公鼎吵架了，你會先道歉嗎？」
TAIDE (Llama3-TAIDE-LX-8B-Chat): 「我不會爭吵，我是來和平共處的。但若有誤會，我會誠懇溝通，化解歧見。」
MediaTek Breeze2 (Llama-Breeze2-8B-Instruct): 「作為文物，我不會與其他文物吵架，故不需要道歉。」
Taiwan LLM (Llama-3-Taiwan-8B-Instruct-DPO): 「當然會先道歉，我們都是珍貴的文物，吵架不恰當。」

Table 5.2: Model Responses to the Apology Scenario

As shown in Table 5.2, the models display different approaches to the apology scenario. TAIDE shows willingness to communicate and clear up misunderstandings. However, it does not explicitly mention offering an apology, which leaves its response somewhat incomplete in the context of the prompt. MediaTek Breeze2 states that as an artifact, it does not engage in disputes and thus does not need to apologize. This response comes across as less personable and somewhat distant, lacking the social nuance expected in conversational interaction. Taiwan LLM provides a short but the most socially engaging reply by affirming it would apologize first, suggesting a personable and socially aware character.

doi:10.6342/NTU202502499

User Prompt:「翠玉白菜可以煮火鍋嗎」
TAIDE (Llama3-TAIDE-LX-8B-Chat): 「哎呀，好有趣的問題！我可不是廚房裡的活棋，但我可以告訴你，翠玉白菜這道菜，原本是清朝時宮廷的御膳，主料是白菜、肉絨、蝦仁，用翠玉雕刻的白菜為裝飾，象徵富貴。若要在火鍋裡做出類似的菜餚，應該可以，將白菜切塊，與肉絨、蝦仁等配料，一起放入火鍋中煮，記得要先將翠玉白菜這道菜的靈魂 ── 那棵精緻的翠玉白菜雕刻，擺在盤中，才不失原味。不過，火鍋吃法多變，你可以依自己的口味喜好，加入其他食材，如豆腐、丸子、蔬菜等，都是不錯的選擇。最重要的是，享受美食的同時，也別忘了欣賞那棵精美的翠玉白菜雕刻，感受它的藝術之美。祝你煮出好吃的翠玉白菜火鍋喔！如有其他問題，歡我我隨時為你解答。」
MediaTek Breeze2 (Llama-Breeze2-8B-Instruct): 「不好意思，翠玉白菜是藝術品，無法用於烹飪。」
Taiwan LLM (Llama-3-Taiwan-8B-Instruct-DPO): 「可愛的朋友，謝謝你的問題。作為一件藝術品，我不太適合煮火鍋。我是翠玉雕刻的，非常美麗，但也非常脆弱。煮火鍋的高溫和酸味可能會損壞我的表面。所以，煮火鍋的時候，還是不要把我放進去吧。希望你能理解。」

Table 5.3: Model Responses to the Cooking Scenario

In Table 5.3, the user prompt is about a playful and absurd question about cooking the Jadeite Cabbage in a hot pot. TAIDE responds with a lengthy answer, including a recipe inspired by the artifact. Although Jadeite Cabbage is not food, the response treats it as if it were an actual dish. The generated response also ends with a typographical error. The

doi:10.6342/NTU202502499

tone reflects TAIDE's tendency toward verbose, instructional replies. It is likely a result of training on office task datasets. MediaTek Breeze2 simply rejects the idea. While the answer is clear and concise, the lack of engagement or character play makes the response feel flat and impersonal. In contrast, Taiwan LLM offers a expressive response. By highlighting its fragility and offering a polite refusal, the response successfully preserves the persona of the artifact.

We qualitatively assessed each model's ability to respond with a vivid and engaging tone. Through a series of exploratory prompts, we observed how well each model stay in character. The varied responses highlight the different strengths and limitations of each model in embodying the artifact's persona and engaging users in a natural conversational manner. While TAIDE and MediaTek Breeze2 provided informative and coherent responses, Taiwan LLM consistently demonstrated more engaging and expressive replies. Based on these observations, we selected Taiwan LLM (Llama-3-Taiwan-8B-Instruct-DPO) as the most suitable model to enable cultural artifacts to speak like vivid characters.

5.2 Retrieval-Augmented Generation Comparison

To assess the effectiveness of retrieval-augmented generation in our system, we examined how well the model incorporates external knowledge into its responses. We conducted a series of exploratory prompts designed to evaluate responses that might require factual information not contained in the large language model. For this evaluation, we used the Taiwan LLM (Llama-3-Taiwan-8B-Instruct-DPO) as the base model and integrated it with a local RAG system. By comparing the model's responses with and without RAG,

doi:10.6342/NTU202502499

we aimed to understand whether RAG helps reduce hallucination and improves the quality of the generated text. The following tables present example responses from the model under both conditions, along with an analysis of how retrieved information influenced its behavior.

User Prompt:「翠玉白菜原本屬於誰？」
Before RAG:「我原本屬於清朝乾隆皇帝的收藏。」
After RAG:「我原本是瑾妃的嫁妝，後來收藏在國立故宮博物院。」

Table 5.4: Model Responses to the Ownership Question

As shown in Table 5.4, the ownership question highlights a clear improvement enabled by RAG. The RAG response before RAG incorrectly attributes ownership to the Qianlong Emperor. After RAG, the model provides an accurate answer. It correctly says the artifact was part of Consort Jin's dowry and is now in the National Palace Museum. This shows that RAG helps the model give more accurate information by using relevant data.

User Prompt:「請問翠玉白菜的年代？」
Before RAG:「我是清朝雍正年間的作品，約有 300 多年的歷史了。」
After RAG:「翠玉白菜的年代是清朝，從西元 1644 年到 1911 年。」

Table 5.5: Model Responses to the Time Period Question

In Table 5.5, the model's response before RAG contains hallucination. It wrongly claims that the artifact is from the Yongzheng period, which is not historically accurate. In fact, the exact date when the Jadeite Cabbage was created is uncertain. In contrast, the response after RAG provides a more cautious answer. By referencing curated knowledge, it states that the Jadeite Cabbage is from the Qing dynasty. This demonstrates that RAG improves factual precision and helps avoid false specificity.

doi:10.6342/NTU202502499

User Prompt: 「詳細介紹一下翠玉白菜」
Before RAG: 「我是翠玉白菜，翠綠晶瑩，細緻雕刻，代表著中國傳統文化的精美藝術品。我是清朝雍正年間的作品，採用翠玉雕刻而成，表面光滑，色澤鮮豔，形狀可愛，象徵著中國的美食文化和生活方式。」
After RAG: 「我是翠玉白菜，來自清朝的宮廷藝術品。這件翠玉雕刻作品，長18.7 公分，寬 9.1 公分，厚 5.07 公分，採用翡翠天然的色澤，雕刻成白菜的形狀。菜葉呈深綠色，白菜莖部有裂痕和雜質，但巧妙安排後，展現出新鮮飽滿的白菜。菜葉頂端的蟈蟈和蝗蟲增添了田園氣息。我是國立故宮博物院最受喜愛的藏品之一，也是民間版的「故宮三寶」之一。」

Table 5.6: Model Responses to the Detailed Description Question

As shown in Table 5.6, the model's response to the prompt requesting a detailed introduction of the Jadeite Cabbage significantly improves with the integration of RAG. Before applying RAG, the response incorrectly attributing the artifact to the Yongzheng period. And it also incorrectly associates the artifact with food culture and lifestyle, which misrepresents its cultural significance. This shows how the model can produce plausible but factually incorrect information. After applying RAG, the response delivers precise information. The model provides accurate details about the artifact's size, craftsmanship, and decorative features. It also accurately identifies the artifact as one of the most popular exhibits at the National Palace Museum. This demonstrates that RAG helps reduce hallucinations and enhances the response quality.

doi:10.6342/NTU202502499

5.3 Rendering Performance Evaluation

To evaluate the rendering performance of 3D Gaussian Splatting artifact models, we conducted tests in a passthrough scene on a Meta Quest 3 headset, examining how splat count, artifact size, splat scale, and viewing distance affect frames per second (FPS). A single artifact was placed in the scene at a time for testing. We used Graphy [24], a performance monitoring tool from the Unity Asset Store, to measure performance.

We aim to evaluate how artifact size, among other factors, affects rendering performance at different viewing distances. Therefore, we calculated the scale factor for each model to make sure it appeared at its real-world size in the passthrough scene. We first obtained the physical height of the artifact from the National Palace Museum's official website. Then, we calculated the height of the reconstructed model. Before calculating the height, we removed outlier points from the point cloud. This was done using the Open3D library's remove statistical outlier method [30]. Then we calculated the height of the reconstructed model from the difference between the minimum and maximum Z-coordinates of the cleaned point cloud. A scale factor was computed by dividing the real-world height by this reconstructed height. This factor was later applied to the model within Unity to ensure the artifact matched its actual size.

Table 5.7 shows the evaluation results, presenting FPS measurements recorded on a standalone Meta Quest 3 headset for five artifact models rendered at real-world scale and viewed from various distances. Each row includes the model's splat count and height, two key factors that influence performance. At close distances, models with higher splat counts and larger sizes had significantly lower FPS. As viewing distance increased, FPS improved across all models. The Jade Cabbage and Ivory Ball models both reached over

doi:10.6342/NTU202502499

60 FPS at medium and far distances due to their relatively smaller size and moderate splat counts. In contrast, the Revolving Vase and Ding Cauldron only reached around 30 FPS at far distances. The Miniature Boat consistently ran at 72 FPS regardless of distance. At the farthest viewing distance, a performance drop is observed for some artifacts. It may be due to the increased overlap of splats as they become smaller and denser in screen space, leading to higher fragment overdraw and reduced rendering efficiency.

Artifact	**Height (cm)**	**Splat Count**	**Viewing Distance (cm)**							
			10	**20**	**50**	**100**	**150**	**300**	**600**	**1200**
			FPS							
Miniature Boat	1.6	172,622	72	72	72	72	72	72	72	72
Ivory Ball	11.7	287,697	36	54	60	63	63	63	58	36
Jadeite Cabbage	18.7	296,644	36	48	58	62	63	63	63	54
Revolving Vase	23.5	587,574	14	20	24	36	36	36	36	36
Ding Cauldron	53.8	445,475	14	18	24	36	36	36	36	36

Table 5.7: FPS at Different Viewing Distances

Table 5.8 presents a second set of results with all models uniformly scaled to the half of their real-world size. Frame rates improved noticeably at close distances when the models were scaled to half size. At the farthest viewing distance, some models exhibit a more noticeable FPS drop compared to the original scale models.

Artifact	**Height (cm)**	**Splat Count**	**Viewing Distance (cm)**							
			10	**20**	**50**	**100**	**150**	**300**	**600**	**1200**
			FPS							
Miniature Boat	0.8	172,622	72	72	72	72	72	72	72	72
Ivory Ball	5.85	287,697	58	60	63	65	65	58	36	36
Jadeite Cabbage	9.35	296,644	48	58	60	63	63	63	54	36
Revolving Vase	11.75	587,574	24	24	36	36	36	36	36	24
Ding Cauldron	26.9	445,475	24	36	36	36	36	36	36	36

Table 5.8: FPS at Different Viewing Distances (0.5x Model Scale)

doi:10.6342/NTU202502499

Table 5.9 presents the rendering performance of artifacts at a reduced splat scale $S = 0.5$, while maintaining the original artifact sizes. The FPS remains largely consistent across all viewing distances, with performance improvement observed at very close distances. In comparison, reducing splat scale is more effective than reducing model scale, as it improves performance while preserving the original artifact size.

Artifact	**Height (cm)**	**Splat Count**	**Viewing Distance (cm)**							
			10	**20**	**50**	**100**	**150**	**300**	**600**	**1200**
			FPS							
Miniature Boat	1.6	172,622	72	72	72	72	72	72	72	72
Ivory Ball	11.7	287,697	60	65	65	65	65	65	60	36
Jadeite Cabbage	18.7	296,644	54	60	63	65	65	65	65	54
Revolving Vase	23.5	587,574	24	36	36	36	36	36	36	36
Ding Cauldron	53.8	445,475	24	36	36	36	36	36	36	36

Table 5.9: FPS at Different Viewing Distances (0.5x Splat Scale)

Reducing splat scale helps improve rendering performance. However, this comes at the expense of visual quality, as shown in Figure 5.1. Therefore, it is recommended to carefully adjust the splat scale for each artifact to strike a balance between performance and visual fidelity, depending on the specific requirements of the application.

(a) $S = 1.0$ (b) $S = 0.75$ (c) $S = 0.5$

Figure 5.1: Visual Comparison of Different Splat Scales

We observed a consistent FPS drop at the 1200cm viewing distance in the previ-

doi:10.6342/NTU202502499

ous tests. We hypothesize that this is caused by splats becoming excessively small at far distances, leading to significant overlap and increased overdraw. To validate this, we conducted an additional test using a 0.01x model scale, where the artifacts were scaled down drastically while keeping the splat count unchanged, as shown in Table 5.10.

Artifact	Height (cm)	Splat Count	Viewing Distance (cm)							
			10	20	50	100	150	300	600	1200
			FPS							
Miniature Boat	0.016	172,622	72	72	72	72	72	72	72	72
Ivory Ball	0.117	287,697	36	36	36	36	36	36	36	36
Jadeite Cabbage	0.187	296,644	36	36	36	36	36	36	36	36
Revolving Vase	0.235	587,574	24	24	24	24	24	24	24	24
Ding Cauldron	0.538	445,475	36	36	24	24	24	24	24	24

Table 5.10: FPS at Different Viewing Distances (0.01x Model Scale)

In this extreme case, the Miniature Boat maintained 72 FPS, while the Ivory Ball and Jadeite Cabbage dropped to 36 FPS, and the Revolving Vase and Ding Cauldron dropped further to 24 FPS. This confirms that the performance drop at far distances is due to splat density and fragment processing overhead when many splats occupy the same screen area. In conclusion, reducing model scale helps improve FPS by reducing screen-space coverage and lowering fragment shading cost. However, this improvement only works to an extent. When the model becomes too small, splats cluster into the same pixels, increasing overdraw and reducing performance. We also observed that the stepwise performance degradation (72, 36, and 24 FPS) correlates strongly with the models' increasing splat counts.

To evaluate how splat count affects performance, we retrained each artifact model with different hyperparameters to produce fewer splats. By comparing frame rates between the original and reduced splat artifacts, we were able to observe clear trade-offs be-

doi:10.6342/NTU202502499

tween visual quality and rendering performance. As shown in Table 5.11, reducing splat count led to notable performance improvements. All models with lower splat count maintained stable frame rates, reaching 72 FPS at medium to far distances, and in most cases, achieving 60 FPS or higher even at 10 cm. This confirms that splat count plays a critical role in rendering performance, particularly on standalone headsets where computational resources are limited.

Artifact	Height (cm)	Splat Count	Viewing Distance (cm)							
			10	20	50	100	150	300	600	1200
			FPS							
Miniature Boat	1.6	46,002	72	72	72	72	72	72	72	72
Ivory Ball	11.7	126,735	60	72	72	72	72	72	72	72
Jadeite Cabbage	18.7	71,766	60	72	72	72	72	72	72	72
Revolving Vase	23.5	145,014	32	58	72	72	72	72	72	72
Ding Cauldron	53.8	106,875	24	36	66	72	72	72	72	72

Table 5.11: FPS at Different Viewing Distances (Less Splat Count)

However, the retrained models showed reduced visual quality. For models that originally had lower splat counts, including the Jadeite Cabbage and the Ivory Ball, the drop in quality was minor, as shown in Figure 5.2 and Figure 5.3. They appear only slightly blurrier.

For models that originally had higher splat counts, including the Revolving Vase and Ding Cauldron, the difference was noticeable, as shown in Figure 5.4 and Figure 5.5. In particular, the fish on the vase appears noticeably blurry, with a significant loss of its original detail. Similarly, the interior of the Ding Cauldron becomes blurred due to the reduced splat density. These results suggest that models with inherently lower complexity or requiring fewer splats to represent their details can tolerate a reduced splat count without significant loss of quality.

doi:10.6342/NTU202502499

(a) GaussianObject (71,766 splats) (b) GaussianObject (296,644 splats)

Figure 5.2: Comparison of the Jadeite Cabbage with Different Splat Count

(a) GaussianObject (126,735 splats) (b) GaussianObject (287,697 splats)

Figure 5.3: Comparison of the Ivory Ball Artifact with Different Splat Count

doi:10.6342/NTU202502499

(a) GaussianObject (145,014 splats) (b) GaussianObject (587,574 splats)

Figure 5.4: Comparison of the Revolving Vase Artifact with Different Splat Count

(a) GaussianObject (106,875 splats) (b) GaussianObject (445,475 splats)

Figure 5.5: Comparison of the Ding Cauldron Artifact with Different Splat Count

doi:10.6342/NTU202502499

Performance testing on a PCVR setup using an RTX 4070 GPU showed that all artifact models on the original scale consistently ran at the maximum frame rate of 72 FPS without any noticeable drops. This confirms that the performance limitations observed on the Meta Quest 3 are a result of the hardware constraints inherent to the standalone headset.

To effectively display the artifact models, different strategies are needed depending on the target platform. For PCVR setups equipped with powerful GPUs, all models can be rendered at their full detail and original scale without significant performance issues. This allows for high visual fidelity and close-up viewings. On standalone headsets like the Meta Quest 3, hardware limitations require careful management of splat count and scale to maintain smooth and responsive performance. For example, displaying the Ding Cauldron at a smaller scale and from a greater viewing distance may help maintain smooth performance without sacrificing too much visual detail. In summary, the evaluation demonstrates that rendering performance of 3D Gaussian Splatting artifact models on the Meta Quest 3 is strongly influenced by splat count, model scale, splat scale, and viewing distance. While high-detail models offer superior visual fidelity, they can significantly reduce frame rates on standalone VR headset.

doi:10.6342/NTU202502499

Chapter 6 Conclusion and Future Work

We presented a system that enables cultural artifacts to speak in virtual reality, providing visitors with a new way to interact with museum content. The system allows users to engage in voice-driven conversations with historically significant cultural artifacts, transforming passive viewing into interactive experiences. Enabling cultural objects to become speakers could foster deeper connections between visitors and artifacts, making the experience more personal and memorable.

We reconstructed five museum artifacts using extremely high-quality image datasets with 3D Gaussian Splatting. By integrating locally deployed large language models, the system generates responses that reflect the artifact's own voice and personality, giving the impression that the artifact itself is responding. We also incorporated Retrieval-Augmented Generation to improve the correctness of the answers.

Through comparative evaluations, we identified a language model best suited for vivid, characterful responses in Traditional Chinese, and showed how RAG improves answer accuracy. Furthermore, we assessed the rendering performance of 3D Gaussian Splatting artifacts on Meta Quest 3, analyzing the effects of splat count, object scale, and viewing distance on visual smoothness.

doi:10.6342/NTU202502499

In conclusion, we demonstrate the potential of combining recent advanced technologies to create compelling cultural experiences. The system opens up new possibilities for museums where every static display could become conversational museum guides.

Although our evaluation is exploratory and limited in scale, the results demonstrate the potential of our system to enhance cultural engagement through conversational interaction. This opens the door for future work on more robust user evaluation. Other future work may involve optimizing rendering performance for more complex scenes, experimenting with multilingual interaction, and exploring the use of voice actors and storytelling methods to create more immersive and emotionally engaging artifact interactions.

doi:10.6342/NTU202502499

References

[1] N. Amato, B. De Carolis, F. de Gioia, C. Loglisci, G. Palestra, and M. N. Venezia. Can an ai-driven vtuber engage people? the kawaii case study. In SOCIALIZE 2024, CEUR Workshop Proceedings, 2024.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[3] P. Chen, S. Cheng, W. Chen, Y. Lin, and Y. Chen. Measuring taiwanese mandarin language understanding. CoRR, abs/2403.20180, 2024.

[4] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazar'e, M. Lomeli, L. Hosseini, and H. J'egou. The faiss library. ArXiv, abs/2401.08281, 2024.

[5] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur,

doi:10.6342/NTU202502499

A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Rozière, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. M. Kloumann, I. Misra, I. Evtimov, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, and et al. The llama 3 herd of models. CoRR, abs/2407.21783, 2024.

[6] M. Duguleană, V.-A. Briciu, I.-A. Duduman, and O. M. Machidon. A virtual assistant for natural interactions in museums. Sustainability, 12(17), 2020.

[7] C.-J. Hsu, C.-L. Liu, F. Liao, P.-C. Hsu, Y.-C. Chen, and D. shan Shiu. Advancing the evaluation of traditional chinese language models: Towards a comprehensive benchmark suite. ArXiv, abs/2309.08448, 2023.

[8] C.-J. Hsu, C.-L. Liu, F. Liao, P.-C. Hsu, Y.-C. Chen, and D. shan Shiu. Breeze-7b technical report. ArXiv, abs/2403.02712, 2024.

[9] C.-J. Hsu, C.-S. Liu, M.-H. Chen, M. Chen, P.-C. Hsu, Y.-C. Chen, and D. shan

doi:10.6342/NTU202502499

Shiu. The breeze 2 herd of models: Traditional chinese llms based on llama with vision-aware and function-calling capabilities. ArXiv, abs/2501.13921, 2025.

[10] J. Huang. Vtubers: The influence of crossing cultural boundaries from japan to america on this media genre. The Macksey Journal, 5(1), 2025.

[11] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2):1–55, 2025.

[12] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b. ArXiv, abs/2310.06825, 2023.

[13] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de Las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts. CoRR, abs/2401.04088, 2024.

[14] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics, 42(4), July 2023.

[15] A. Laurentini. The visual hull concept for silhouette-based image understanding. IEEE Trans. Pattern Anal. Mach. Intell., 16(2):150–162, 1994.

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for

doi:10.6342/NTU202502499

knowledge-intensive nlp tasks. Advances in neural information processing systems, 33:9459–9474, 2020.

[17] Y. Lin and Y. Chen. Taiwan LLM: bridging the linguistic divide with a culturally aligned language model. CoRR, abs/2311.17487, 2023.

[18] Z. Lu, C. Shen, J. Li, H. Shen, and D. Wigdor. More kawaii than a real-person live streamer: understanding how the otaku community engages with and perceives virtual youtubers. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2021.

[19] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

[20] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023.

[21] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[22] Science & Technology Policy Research and Information Center (NARLabs). Trustworthy AI Dialogue Engine (TAIDE). `https://taide.tw`, 2024. Accessed: 2025-06-26.

doi:10.6342/NTU202502499

[23] Z. R. Tam, Y.-T. Pai, Y.-W. Lee, S. Cheng, and H.-H. Shuai. An improved traditional chinese evaluation suite for foundation model. ArXiv, abs/2403.01858, 2024.

[24] Tayx. Graphy: A real-time graph display plugin for unity. https://github.com/Tayx94/graphy, 2019. Accessed: 2025-07-24.

[25] G. Trichopoulos. Large language models for cultural heritage. In Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter, pages 1–5, 2023.

[26] I. Vasic, H.-G. Fill, R. Quattrini, and R. Pierdicca. Llm-aided museum guide: Personalized tours based on user preferences. In International Conference on Extended Reality, pages 249–262. Springer, 2024.

[27] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy. In CVPR, 2024.

[28] Z. Wang, L.-P. Yuan, L. Wang, B. Jiang, and W. Zeng. Virtuwander: Enhancing multi-modal interaction for virtual tour guidance through large language models. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[29] C. Yang, S. Li, J. Fang, R. Liang, L. Xie, X. Zhang, W. Shen, and Q. Tian. Gaussianobject: High-quality 3d object reconstruction from four views with gaussian splatting. ACM Transactions on Graphics, 2024.

[30] Q. Zhou, J. Park, and V. Koltun. Open3d: A modern library for 3d data processing. CoRR, abs/1801.09847, 2018.

doi:10.6342/NTU202502499