

Voice of Arts: Human-Centered Multimodal Interaction for AI Art Companions

Hiu-Lam YIP^{1,*†}, Hsiang-Chien Kung^{1†} and Wan-Yi Lin^{1,†}

¹Department of Interaction Design, National Taipei University of Technology, Taipei, Taiwan

Abstract

This paper presents *Voice of Arts*, a human-centered AI mobile application designed to address documented barriers to art engagement—including the fear of misinterpretation, lack of background knowledge, and absence of interactive guidance. To support reflective art appreciation through multimodal interaction, the system integrates a thirty-year corpus analysis of Taiwanese artist Liang Yifen’s paintings (1991–2021) into a Retrieval-Augmented Generation (RAG) architecture. Powered by a Visual Language Model (VLM) for artwork analysis, the application features a first-person AI persona that embodies the artist. Key functional modules include a 3D gallery with AR Polaroid framing, hotspot-annotated visual observation, a Chibi-avatar artist reflection dialogue, and an auto-generated visual journal with emotion-tagged summaries and abstract illustrations. To ensure a seamless user experience regardless of network connectivity, the system incorporates offline fallback scripts and local SVG art generation. A user experience analysis grounded in Don Norman’s three-level emotional design framework demonstrates that *Voice of Arts* effectively transitions museum visitors from passive observation to active, reflective meaning-making while upholding responsible AI design principles such as ambiguity preservation, source citation, and the avoidance of authoritative single-answer delivery.

Keywords

Human-Centered AI, Art Appreciation, Museum Technology, Visual Language Model, Retrieval-Augmented Generation, Multimodal Interaction, Reflective Design, Taiwan Contemporary Art

1. Introduction

Museum visitors frequently encounter barriers when engaging with contemporary visual art: they lack art-historical background, fear making incorrect interpretations, and find existing audio guides insufficiently responsive to individual curiosity [1]. These challenges are particularly pronounced for works by regional or non-canonical artists, where the cultural context requires specialized knowledge to be fully appreciated.

Liang Yifen (b. 1937) is a Taiwanese artist whose thirty-year practice (1991–2021) traverses folk culture, working-class portraiture, and post-colonial identity. His canvases combine flattened figurative forms, high-saturation folk palettes, and richly layered cultural signifiers that reward close, sustained attention. Yet, standard museum labels often provide insufficient context for general audiences to move beyond surface-level recognition.

This paper introduces *Voice of Arts*, a React Native mobile application that pairs a curated dataset of six representative Liang Yifen paintings with four interlinked, AI-powered interaction modules. Our core design question is: **how can an AI system support the visitor’s own interpretive process rather than replace it?** We ground our approach in human-centered AI (HCAI) principles and Don Norman’s three-level emotional design model [2], evaluating the resulting user experience through structured persona analysis and service blueprint reviews.

APMAR’26: The 18th Asia-Pacific Workshop on Mixed and Augmented Reality, Aug. 3-5, 2026, Taipei, Taiwan

*Corresponding author.

†These authors contributed equally.

✉ yiphiulam424@gmail.com (H. YIP); s9645848@gmail.com (H. Kung); antialin7425@gmail.com (W. Lin)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Background and Related Work

2.1. AI-Assisted Art Interpretation

Conversational AI guides for museums have grown rapidly since the introduction of large language models (LLMs). Prior systems, such as ArtBot [3] and the Smithsonian "Ask the Collection" prototype, offer efficient factual retrieval but tend to deliver single, authoritative answers. This approach often suppresses the productive ambiguity that characterizes genuine aesthetic experiences [4]. *Voice of Arts* departs from this paradigm by foregrounding open-ended questioning and multi-perspective framing, encouraging users to form their own interpretations.

Furthermore, recent studies demonstrate that traditional audio guides often suffer from passive, linear information delivery, which can limit visitor engagement [5]. In contrast, conversational AI and interactive chatbots transform visitors from passive listeners into active participants by accommodating spontaneous, curiosity-driven inquiries [3]. Unlike the 'one-size-fits-all' approach of static audio tours, conversational agents foster deeper reflective meaning-making through highly personalized, two-way dialogue, thereby offering a demonstrably superior interactive experience.

2.2. Visual Language Models in Cultural Heritage

Visual Language Models (VLMs) such as Gemini and GPT-4V have demonstrated strong performance in artwork captioning and element detection tasks [6]. Integrating VLM analysis with domain-specific Retrieval-Augmented Generation (RAG) pipelines allows for culturally grounded interpretations that go beyond generic scene descriptions. We adopt this integrated approach to power the system's dynamic hotspot generation and artist reflection dialogues.

2.3. Artist Voice and RAG-Based Persona Systems

RAG architectures have been successfully applied to historical figure simulation in educational contexts [7]. *Voice of Arts* extends this application to a contemporary artist context by injecting primary-source biographical documents into the Gemini prompt chain. This enables the system to produce first-person dialogues that cite verifiable life events—such as Liang's New York SOHO years and his return to rural Dulan, Taiwan—while maintaining ethical transparency by clearly marking all responses as AI-generated content.

3. System Architecture

3.1. Overall Design Philosophy

Voice of Arts follows a Reflective Companion model: the AI observes, prompts, and contextualizes, but never adjudicates. Four interaction screens form a progressive engagement loop aligned with the user journey stages of Observe → Explore → Reflect → Interpret → Share.

3.2. Artwork Dataset: Thirty Years of Liang Yifen

We curated a dataset of six canonical paintings spanning Liang's three decades of practice (Figure 1). These were carefully selected to represent the evolution of his iconographic vocabulary:

- **Traditional Craft Figures Series (1991–1995):** Folk vendor scenes, saturated primary palettes, flattened compositions.
- **New York SOHO Period (1996–2001):** Cross-cultural hybridity, gestural abstraction.
- **Blue Cap and Dog (c. 2000):** Urban quotidian subjects, loose brushwork, muted earth tones.
- **Red Carpet Black Beauty (c. 2005):** Ceremonial registers exploring gender and spectacle.
- **Dulan Creation Period (2010–2021):** Indigenous landscape motifs, earthen pigment palettes.

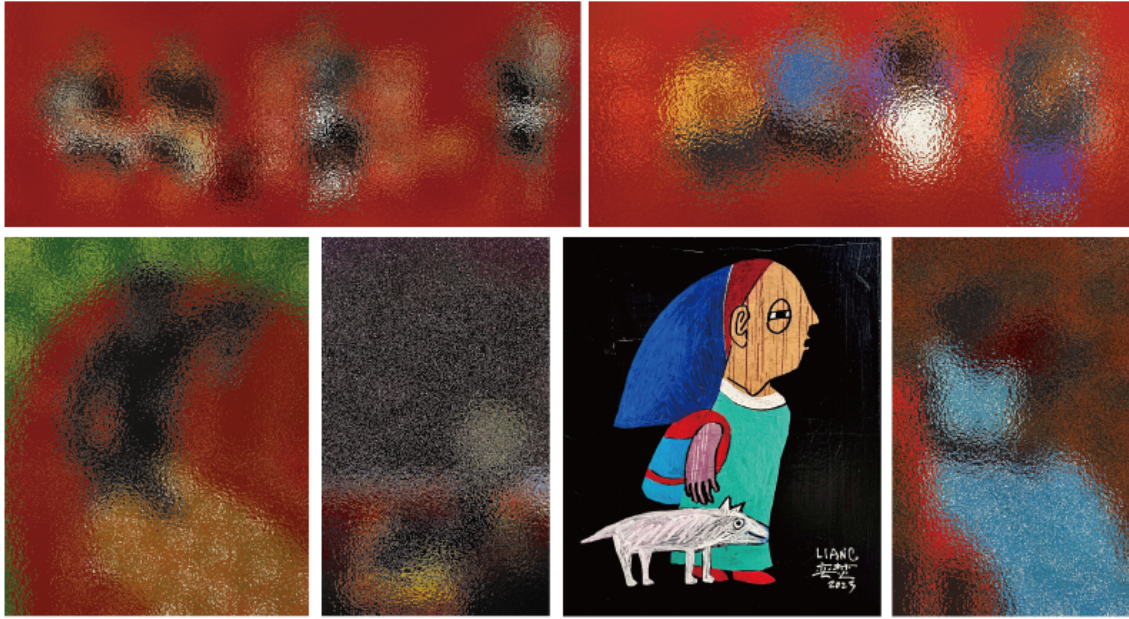


Figure 1: Six canonical paintings spanning Liang’s three decades of practice. (Note: To comply with copyright restrictions, artworks other than Blue Cap and Dog have been intentionally blurred.)

For each painting, the research team compiled comprehensive metadata: visual element annotations (objects, colors, composition, symbolic elements), emotional keyword sets, cultural context summaries, artist biographical context linked to the specific work, and four-round offline dialogue scripts.

3.3. Module 1: Home Screen Gallery and Camera Integration

The entry screen presents the six paintings as 3D floating swipe cards rendered with smooth perspective interpolation. Two interaction affordances are provided: (1) direct selection of a gallery card to enter the observation flow, and (2) a camera/album button that allows visitors to photograph an artwork in the exhibition space or upload a custom image for AI analysis. This affords both guided and exploratory entry paths, supporting the diverse needs of general visitors and student researchers (Figure 2).

3.4. Module 2: AR Camera and Polaroid Framing

An augmented reality layer enables visitors to composite their live camera view with artist-specific Polaroid-style borders that reproduce Liang’s characteristic color signatures. Users may co-frame themselves with painted figures, save the composite image, and share it via standard mobile share sheets. Aligning with the “visceral level” of Norman’s emotional design framework [2], this feature creates an embodied souvenir that sustains post-visit engagement (Figure 3).

3.5. Module 3: Multimodal Hotspot Observation

When an artwork is loaded, the screen simulates a radar-scan animation before revealing interactive hotspot annotations anchored to specific visual elements (e.g., garment patterns, background textures, symbolic objects). Two operational modes are supported:

- **Online Mode:** The Gemini VLM analyzes the artwork image and dynamically generates contextually relevant hotspot content.
- **Offline Mode:** Pre-authored hotspot data, hand-crafted for each painting by the design team in consultation with curatorial notes, is loaded from local JSON files.

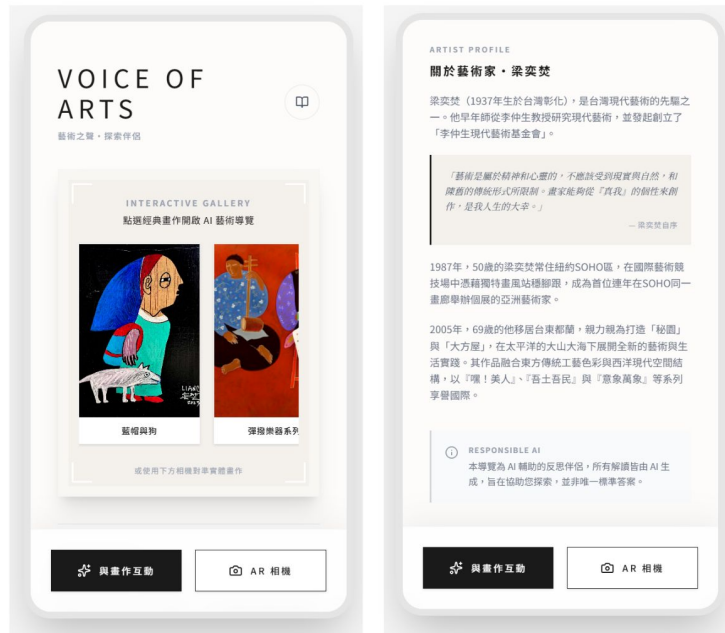


Figure 2: UI for Home Screen Gallery and AR Polaroid Framing.



Figure 3: UI for AR Camera and Polaroid Framing.

Hotspot labels utilize warm, conversational prose rather than dense academic terminology. This lowers the interpretive threshold for novice visitors while preserving analytical depth for more experienced art viewers.

To demonstrate the analytical depth of the dynamic VLM hotspot generation, we detail the generated annotations for the artwork *Blue Cap and Dog*. The Gemini VLM successfully isolates key visual elements and constructs narrative-driven observations. For instance, upon detecting the protagonist's *Striking Blue Hat*, the VLM identifies it as a vivid, saturated visual focal point that establishes a cool-warm color balance against the background. Furthermore, the VLM accurately recognizes the distinct

wood-grain texture of the face, noting how the underlying material adds an organic, raw aesthetic to the portrait. When analyzing the artwork's Minimalist Lines, the system observes that the figure's rounded and extremely simplified contours merge Cubist spatial reorganization with the essence of traditional Eastern freehand brushwork. Lastly, the model focuses on the Accompanying Dog, highlighting the loose brushwork of the dog. It interprets these unpolished, rough white strokes and the pet's presence as a symbol of heartwarming, harmonious coexistence between human and animal that infuses the scene with everyday warmth (Figure 4).

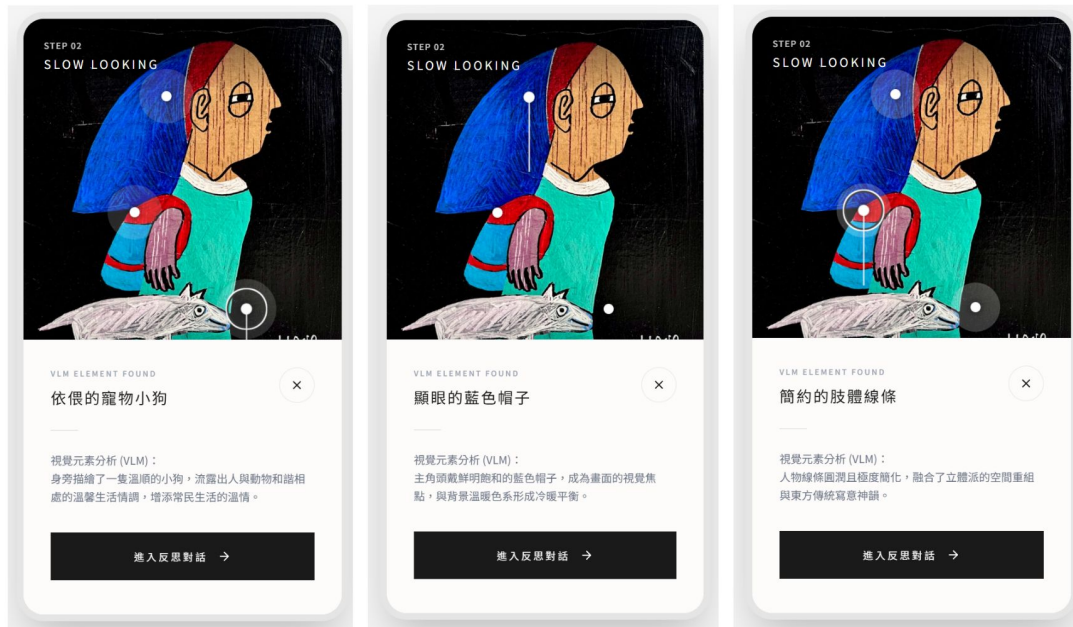


Figure 4: UI for Multimodal Hotspot Observation.

3.6. Module 4: Artist Reflection Dialogue

The reflection module presents a custom Chibi avatar of Liang Yifen as the conversational agent, rendered in a cream-toned "warm milk" interface palette. The backend retrieves passages from a structured biographical PDF via RAG, which are injected into the Gemini system prompt alongside a first-person persona instruction. The model responds as Liang himself, using the pronoun "I" and referencing verified autobiographical details (Figure 5).

To mitigate the risk of RAG-induced hallucinations, strict prompt guardrails are enforced. When a visitor's query falls outside the scope of the retrieved context, the system is prohibited from fabricating information. Instead, it employs a persona-driven defense mechanism: the AI politely declines out-of-domain questions in the artist's voice and gracefully guides the dialogue back to the artwork. A dual-layer fallback architecture ensures an uninterrupted experience:

- **Primary:** Live Gemini RAG dialogue with biographical grounding.
- **Fallback:** If the server is unavailable or the API key fails, the system deploys four-round, pre-authored dialogue scripts. These are differentiated per artwork, covering the most commonly surfaced visitor questions with distinct conversational arcs to avoid repetitive, generic responses.

Responsible AI design principles are applied throughout: all responses carry an "AI-generated content" marker, probabilistic hedges replace authoritative assertions, and the system never claims to represent the artist's sole or verified intent.

To illustrate how the dialogue is anchored in source materials, the system retrieves relevant autobiographical texts before generating the artist's response. For instance, the system context retrieval from the RAG knowledge base ("Artistic DNA of Liang Yi-fen") highlights his philosophy on primitive art and his mentor's teachings. The following transcript excerpt demonstrates how the user's emotional input triggers this retrieval, seamlessly blending the artist's historical facts with conversational empathy:

System Context Retrieval: "Primitive art displays human's innate gifts... Everyone has an artistic talent and retains the seeds of art. Learning from mentor Lee Chun-Shan's 'subtraction philosophy,' art must strip away unnecessary complexities to leave only the most essential, pure lines."

Dialogue Transcript Excerpt:

User: "I feel a profound sense of calmness."

AI Persona (Liang Yifen): "Ah, child, your sensitivity is wonderful! 'A profound sense of calmness'—these words truly touch this old man's heart. Looking back on my life, from the mud houses in rural Changhua to the bustling streets of New York SOHO, and finally to the 'Secret Garden' in Dulan, Taitung—what I have always been seeking is exactly that most primitive, purest calmness. Just like the 'subtraction philosophy' I learned from my mentor Lee Chun-Shan: 'If one line can finish it, don't use two.' My paintings aim to strip away that external noise... Which part of the painting brings you this calmness?"

User: "The accompanying little dog."

AI Persona (Liang Yifen): "Ah, your observation is so delicate and warm! You are absolutely right; the presence of this dog indeed adds a rare sense of tranquility and tenderness to the entire painting. It reminds me of my childhood in the Changhua countryside... The dog in this painting, though painted with rough brushstrokes, embodies that unpolished innocence. In my art philosophy, I emphasize revealing the pure, primitive nature innate to humans. This dog is the embodiment of that purity."



Figure 5: UI for Artist Reflection Dialogue.

3.7. Module 5: Generative Art Journal

Following the reflection dialogue, the system generates a personal art journal entry comprising three components:

- **My Perspective:** A synthesis of the visitor's chat turns, surfacing their expressed observations and emotional responses.
- **Cultural Context:** A RAG-grounded contextual annotation linking the visitor's remarks to documented historical and cultural information about the work.
- **Emotional Tags:** Pill-shaped labels generated by analyzing the affective register of the dialogue (e.g., *Calm*, *Warmth*, *Nostalgia*, *Contemplative*, *Joyful*).

An abstract artwork is then generated to accompany the journal entry:

- **Online Mode:** An Imagen 4.0 API call produces a minimalist, oil-painting-style image summarizing the session's emotional atmosphere, optimized for minimal token cost via a compact style prompt.
- **Offline/Fallback Mode:** If image generation fails, the system uses the conversation length as a random seed to dynamically render a museum-aesthetic geometric SVG composition. This utilizes the current painting's extracted color palette, ensuring a visually coherent artifact even without network access.

Journal Generation Pipeline and Evaluation Metrics. The generation fuses semantic text analysis with an image generation pipeline. The `/api/generate-journal` endpoint executes the following stages:

1. **Semantic Analysis:** Gemini 2.5 Flash extracts user responses from the dialogue history to generate a structured journal evaluating the user's perspective, contextual appreciation, and sentiment.
2. **Image Rendering:** The generated text is passed to the Imagen 4.0 model to produce an abstract painting that resonates with the user's reflective mood.

To ensure the quality of the generated digital journal, the system evaluates the content across three dimensions:

- **Semantic Consistency:** Assesses whether the generated *perspective* matches the semantics of the user's replies in the dialogue history (cross-verified by the LLM, with a similarity score > 0.85 considered passing).
- **Contextual Relevance:** Evaluates whether the *context* correctly cites Liang Yifen's artistic style (e.g., vermilion colors, Fauvist palette, or Dulan lifestyle), strictly prohibiting the model from fabricating unrelated art historical movements.
- **Image-Text Alignment:** Constrains the Imagen 4.0 drawing prompt to exclude text and specific character renderings, retaining only abstract colors and moods to meet the aesthetic standards of a modern art museum exhibition (Figure 6).

Example Analysis: Blue Cap and Dog. To clearly illustrate this artwork analysis pipeline, consider a user session engaging with Blue Cap and Dog (see Figure 6). Based on a visitor's dialogue expressing how the silent presence of the dog evokes a sense of quiet support, the Semantic Analysis module extracts "Solitude" and "Companionship" as the definitive emotional tags. The RAG system synthesizes the Cultural Context by linking this emotional resonance to Liang's "subtraction philosophy"—stripping away external noise to reveal intrinsic vitality. Finally, the pipeline uses these semantic and emotional parameters to render the customized abstract illustration, successfully externalizing the visitor's internal aesthetic experience.

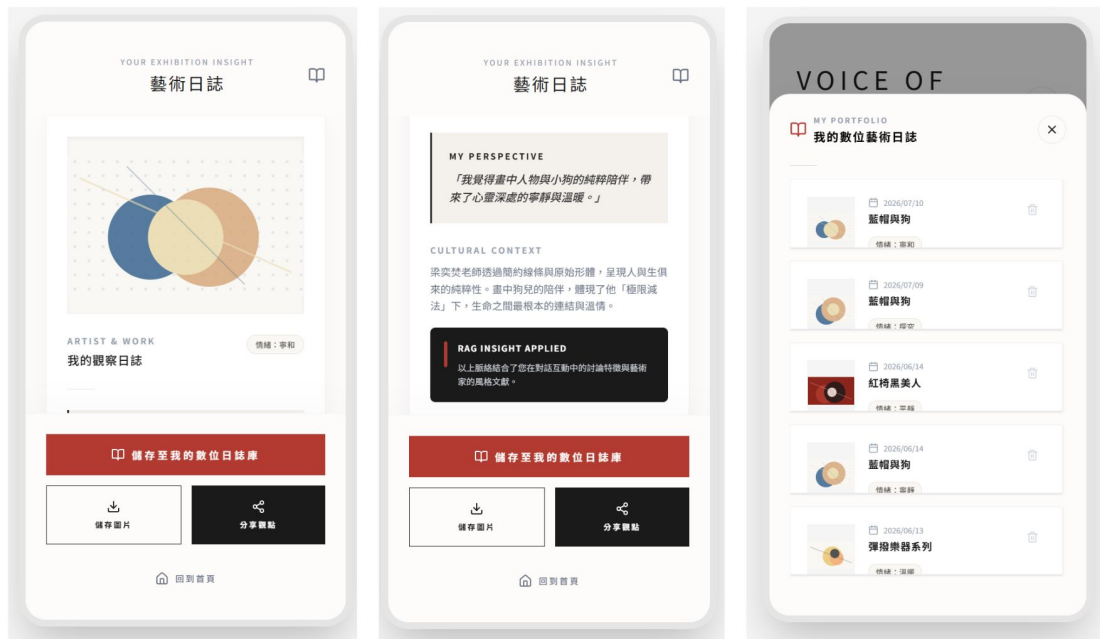


Figure 6: UI for Generative Art Journal and Digital Journal Library.

3.8. Module 6: Digital Journal Library

A persistent journal archive stores each session record to the device's local storage, including artwork metadata, emotional tags, dialogue summaries, and the generated abstract image. Users access their archive through a half-screen bottom drawer triggered by a bookshelf icon, which presents entries on a chronological timeline. Each entry card supports expand or collapse actions, image downloading, perspective sharing, and deletion. This feature effectively transforms ephemeral museum visits into a curated digital trail of aesthetic encounters.

4. User Experience Analysis

4.1. Theoretical Framework: Don Norman's Three-Level Emotional Design

We analyze the Voice of Arts experience across Norman's visceral, behavioral, and reflective design levels [2]:

Visceral level: Immediate sensory impact is addressed through the 3D swipe gallery, the radar-scan animation, the AR Polaroid feature, and the warm cream palette of the reflection screen. These elements generate positive first-impression affect and signal a playful, non-intimidating interaction register.

Behavioral level: Interaction quality is supported through hotspot annotations that guide the visitor's gaze without dictating interpretation, a conversational AI dialogue that dynamically responds to the user's actual utterances, and multi-painting navigation that rewards exploratory behavior.

Reflective level: Deeper meaning-making is scaffolded through the first-person artist persona (enabling an imagined dialogic relationship with the creator), the personalized journal (externalizing and preserving the visitor's interpretive process), and the emotional-tag system (naming affective states that visitors may struggle to articulate independently).

4.2. Persona-Based UX Evaluation

Two primary user personas were developed through iterative, Project-Based Learning (PBL) fieldwork to evaluate the system's target experiences:

Persona A – General Museum Visitor (Linda Chen, 45, office worker): Linda lacks an art-historical background and fears making incorrect interpretations. She requires accessible language, visual guidance, and emotional connection points. *Voice of Arts* addresses his needs through plain-language hotspot prose, the artist’s warm conversational tone, and the journal’s validation of his personal response as inherently meaningful.

Persona B – Student/Online User (Jason Lin, 19, undergraduate): Jason needs systematic analytical scaffolding and citable knowledge for coursework. He benefits from the multi-perspective framing of hotspot annotations, the RAG-backed cultural context integrated into the digital journal, and the responsible AI design features (such as source citation and probabilistic language) that support academically credible engagement.

4.3. Service Blueprint and Responsible AI Design

The service blueprint maps the end-to-end interaction flow: *artwork upload* → *VLM visual analysis* → *feature extraction* → *RAG knowledge retrieval* → *LLM reflection dialogue* → *visitor reflection* → *personal interpretation*. Frontstage elements (the AI dialogue interface, visual observation guide, and open-ended questioning) are seamlessly matched to backstage systems (Gemini VLM, RAG knowledge store, Imagen 4.0, and the local fallback engine).

Responsible AI requirements are operationalized within this service architecture as follows:

- **AI Identity Disclosure:** All AI-generated content is persistently labeled, and the custom Chibi avatar visually signals non-human agency.
- **Ambiguity Preservation:** The system provides two to three interpretive directions rather than a single, authoritative answer. Hedged language (e.g., "may suggest," "some scholars interpret") is strictly enforced in the persona prompt.
- **Context Citation:** Responses explicitly cite the artist’s biographical timeline, curatorial texts, and specific visual details as evidential anchors.
- **Avoiding Speaking for the Artist:** The persona prompt strictly prohibits making definitive claims about creative intent; probabilistic phrasing is required to respect the artist’s true agency.
- **User Disagreement Handling:** A dedicated "My View" input field allows visitors to log and register their own unique interpretations alongside the AI response.
- **Cultural Respect:** The RAG knowledge base incorporates indigenous Taiwanese cultural perspectives relevant to Liang’s Dulan-period works, directly counteracting Western-centric interpretive defaults.

We acknowledge that the current evaluation relies primarily on formative design research methods, namely persona analysis and service blueprinting, rather than empirical user testing. However, within Human-Centered AI (HCAI) and interaction design research, utilizing these theoretical frameworks to rigorously validate conceptual workflows, structural alignment, and technical safety guardrails prior to formal in-the-wild deployment is a standard and critical foundational step. This theoretical validation ensures the system’s responsible AI mechanisms are robust enough to proceed to the controlled user studies outlined in our future work.

5. Implementation Details

5.1. Technology Stack

The mobile application is built in React Native (TypeScript). The backend is a Node.js/Express server (server.ts). AI services are provided by Google’s Gemini 2.5 Flash (dialogue and visual analysis) and Imagen 4.0 (journal illustration). The RAG pipeline parses and chunks the biographical PDF using a custom text splitter and retrieves relevant passages via cosine similarity over local embeddings. All six paintings are bundled as local assets; hotspot JSON files and fallback dialogue scripts are embedded in the app bundle for full offline functionality.

5.2. Offline Architecture

System resilience was a primary engineering constraint, given that museum exhibition environments frequently suffer from unreliable network connectivity. Consequently, every AI-dependent feature incorporates a robust local fallback mechanism:

- **VLM Analysis** → Defaults to pre-authored hotspot JSON annotations.
- **Gemini Dialogue** → Defaults to scripted, four-round conversations distinct to each artwork.
- **Imagen 4.0 Generation** → Defaults to SVG generative art, utilizing the current painting's color palette and the session's conversation length as a randomization seed.

Crucially, these fallback transitions occur seamlessly in the background; no error messages or loading disruptions are displayed to the user, preserving the immersive experience.

5.3. System Operational Workflow

To clarify the concrete operational scenarios of the Voice of Arts system, this section details the data structures, API responses, and specific tracking examples across both online and offline states. The system dynamically routes user inputs based on API availability and dialogue history, supporting two primary operational paths: an Online RAG Mode and an Offline Scripted Fallback Mode.

1. Online/Offline Routing and Fallback Mechanism: In the Online RAG Mode, when a valid API key is present and the dialogue is within its first three turns, the backend calls Gemini 2.5 Flash. This process integrates the source document (DNAofLiangYifen.pdf) with the current artwork ID and hotspot data, rendering standard chat bubbles with a clear citation label indicating the source material.

Conversely, to ensure uninterrupted user experience in exhibition environments, the system seamlessly downgrades to the Offline Scripted Fallback Mode if the API key is invalid or the quota is exceeded. Upon the user's initial entry into the dialogue, the system retrieves a tailored welcome message from a local OFFLINE_DIALOGUES array based on the specific artwork ID, appending an informative "Local Offline Script" subtext to the chat bubble. As the user submits their first reply (history length equals 2), the system advances to a secondary deep-interaction prompt. Finally, once the dialogue reaches three or more user replies (history length of 6 or greater), the system triggers the Offline Scripted End, prompting the transition to the journal module while locking the final reply to preserve system transparency.

2. RAG-based Dialogue Anchoring: In online mode, the conversational lifecycle anchors strictly to the source material to prevent hallucination and align with Human-Centered AI (HCAI) principles.

- **Anchoring Process:** Upon initialization, the server uploads the structured biographical PDF to the Gemini File API. For every user message sent to the backend, the server appends the document's Cloud File URI to the request. The system prompt strictly constrains the model to adopt a first-person persona and mandates that personal biographical details be drawn exclusively from the text.
- **Transcript Example:**
User: "I feel a profound sense of calmness."
System Retrieval: "Primitive art displays human's innate gifts... Learning from mentor Lee Chun-Shan's 'subtraction philosophy,' art must strip away unnecessary complexities to leave only the most essential, pure lines."
Gemini RAG Response: "Ah, child, your sensitivity is wonderful! 'A profound sense of calmness'—these words truly touch this old man's heart. Looking back on my life, from the mud houses in rural Changhua to the bustling streets of New York SOHO... what I have always been seeking is exactly that most primitive, purest calmness."

3. VLM Hotspots vs. Offline Curated Hotspots: The observation phase transforms visual features into interactive hotspots using two distinct methods:

- **Dynamic VLM Generation (Online):** The system passes the Base64 artwork image to Gemini 2.5 Flash with a prompt to identify prominent visual elements, returning proportional coordinates (x, y) and short descriptions.
- **Offline Curated Hotspots:** To prevent timeout errors when API access is unavailable, the frontend loads precise, expert-authored hotspots based on the artwork ID.

4. Digital Art Journal Generation Pipeline: Following the reflection dialogue, the system executes a three-stage pipeline to evaluate the session and create a structured journal:

- **Dialogue Summarization (LLM Evaluation):** Gemini 2.5 Flash parses the final 6 dialogue turns and outputs a strict JSON schema containing a first-person perspective, an academically grounded context, and a two-character Chinese sentiment tag.
- **Image Prompt Engineering:** The perspective and context are embedded into an English prompt template instructing the model to generate a minimalist, abstract illustration devoid of text.
- **Image Rendering:** The Imagen 4.0 model renders a 1:1 JPEG illustration.

6. Conclusions and Future Work

Voice of Arts demonstrates that multimodal AI can function effectively as a reflective companion rather than an authoritative oracle in museum contexts. By grounding user interaction in a thirty-year artistic corpus, employing first-person RAG dialogue, and generating personalized emotional journals, the system successfully moves visitors from passive observation toward active interpretive agency. The underlying responsible AI design framework ensures that cultural ambiguity is preserved and the user's voice is respected throughout the experience.

Future work will expand upon this foundation to include: (1) a comprehensive controlled user study (between-subjects design) involving representative users to empirically compare the efficacy of *Voice of Arts* against traditional audio guide baselines. This study will utilize the System Usability Scale (SUS) and the Museum Experience Scale (MES) to quantitatively evaluate depth of meaning-making, emotional engagement, and observation dwell time, supplemented by semi-structured interviews to capture qualitative behavioral shifts; (2) the expansion of the dataset to include additional Liang Yifen works and other contemporary Taiwanese artists; (3) multilingual support to accommodate international visitors; (4) a formal aesthetic evaluation of the SVG generative art fallback as an independent creative output; and (5) a longitudinal analysis of the digital journal archive to better understand long-term user usage patterns.

Acknowledgments

This research was developed as part of the "Interactive Technology Service Design" course. The authors express their deepest gratitude to Prof. Ryan Sheng-Ming Wang for his invaluable guidance throughout the project and for generously providing authorized access to Liang Yi-fen's artwork, *Blue Cap and Dog*, from his private collection. His continuous support made the implementation and visual demonstration of this interactive system possible.

Declaration on Generative AI

The authors used Google Gemini to assist in proofreading, translation, and as core computational modules for natural language dialogue, visual analysis, and image generation within the proposed system architecture. The authors take full responsibility for the contents of the publication.

References

- [1] J. H. Falk, L. D. Dierking, S. Foutz, *In Principle, In Practice: Museums as Learning Institutions*, AltaMira Press, Lanham, MD, 2007.
- [2] D. A. Norman, *Emotional Design: Why We Love (or Hate) Everyday Things*, Basic Books, New York, 2004.
- [3] S. Lomas, Artbot: Conversational ai for museum engagement, in: *Proc. Museums and the Web*, 2022.
- [4] W. J. T. Mitchell, *Picture Theory: Essays on Verbal and Visual Representation*, University of Chicago Press, Chicago, 1994.
- [5] L. Tallon, K. Walker, *Digital Technologies and the Museum Experience: Handheld Guides and Other Media*, AltaMira Press, 2008.
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Aspell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: *Proc. ICML*, 2021.
- [7] C. Gao, et al., S3: Social-network simulation system with large language model-empowered agents, *arXiv preprint arXiv:2307.14984* (2023).