Arquitectura híbrida desacoplada para sistemas interactivos de preservación cultural asistidos por IA generativa

José Manuel Cortés Cerón

Instituto Tecnológico Superior del Occidente del Estado de Hidalgo

deepdevjose@itsoeh.edu.mx

RESUMEN La preservación cultural digital mediante sistemas interactivos plantea retos significativos en términos de desempeño, escalabilidad y dependencia de infraestructuras de cómputo intensivo. En particular, la incorporación de inteligencia artificial generativa introduce cargas computacionales que dificultan su integración directa en entornos interactivos de tiempo real. En este artículo se propone una arquitectura híbrida desacoplada para sistemas interactivos de preservación cultural asistidos por IA generativa, diseñada para separar estrictamente los procesos de generación intensiva de contenidos del subsistema de visualización interactiva. La arquitectura adopta un enfoque offline-first, basado en la precomputación de activos visuales y acústicos y su posterior consumo mediante clientes ligeros implementados con tecnologías WebGL. Como caso de estudio, se presenta la aplicación de la arquitectura al legado artístico de Byron Gálvez, demostrando su viabilidad como modelo de referencia para sistemas culturales interactivos que requieren alta disponibilidad, bajo acoplamiento y operación en entornos con conectividad limitada. Los resultados evidencian que el desacoplamiento arquitectónico permite reducir la complejidad del sistema en ejecución y facilita su extensibilidad a otros dominios de sistemas interactivos asistidos por IA.

Palabras clave: Arquitectura de software; sistemas interactivos; preservación cultural digital; inteligencia artificial generativa; offline-first; WebGL.

ABSTRACT Digital cultural heritage preservation through interactive systems poses significant challenges in terms of performance, scalability, and reliance on compute-intensive infrastructures. In particular, the integration of generative artificial intelligence introduces substantial computational loads that hinder its direct use in real-time interactive environments. This paper proposes a decoupled hybrid architecture for AI-assisted interactive cultural heritage systems, designed to strictly separate computationally intensive content generation processes from the interactive visualization subsystem. The architecture follows an offline-first approach, based on the precomputation of visual and acoustic assets and their subsequent consumption through lightweight WebGL-based clients. A case study involving the artistic legacy of Byron Gálvez is presented to instantiate the proposed architecture and evaluate its applicability. The results indicate that architectural decoupling reduces runtime system complexity and enables scalable, high-availability interactive deployments, positioning the proposed model as a reference architecture for a broader class of AI-driven interactive systems

Keywords: software architecture; interactive systems; digital cultural heritage; generative artificial intelligence; offline-first systems; WebGL..

Introducción

La digitalización del patrimonio cultural mediante sistemas interactivos ha cobrado relevancia en los últimos años como una estrategia para ampliar el acceso, la preservación y la difusión de bienes culturales en contextos educativos, museográficos y territoriales [**?**, **?**]. Sin embargo, la transición de archivos estáticos a ecosistemas culturales vivos plantea desafíos técnicos significativos cuando se integran tecnologías de inteligencia artificial generativa, particularmente en términos de fidelidad fenomenológica y sostenibilidad infraestructural.

En los enfoques tradicionales, los sistemas de preservación cultural digital suelen apoyarse en arquitecturas fuertemente acopladas a la nube, donde los procesos de generación y visualización coexisten en tiempo de ejecución [**?**, **?**]. Este modelo resulta insostenible para obras con alta complejidad material —como la técnica de *acrilografía* del artista Byron Gálvez—, donde la preservación de textura, volumen y expresividad exige modelos de restauración profunda de alto costo computacional. Además, estudios recientes sobre innovación tecnológica en Latinoamérica han señalado que la dependencia de conectividad permanente de banda ancha restringe severamente el despliegue de sistemas interactivos avanzados, exacerbando la brecha digital en lugar de democratizar el acceso al patrimonio [**?**].

Paralelamente, el uso de tecnologías web para visualización interactiva —como WebGL— ha demostrado ser una alternativa viable para reducir barreras de entrada y facilitar la distribución multiplataforma de experiencias tridimensionales [**?**]. No obstante, la ejecución directa de inferencia generativa, incluyendo modelos de difusión y sistemas de clonación de voz neuronal, resulta inviable en navegadores estándar debido a limitaciones de cómputo, memoria y consumo energético en dispositivos de consumo [**?**]. Esta situación evidencia la necesidad de repensar la organización arquitectónica de estos sistemas bajo una lógica de diseño centrado en la experiencia del usuario y la eficiencia perceptual, más que en la fuerza bruta computacional.

En este contexto, el problema central que aborda este trabajo es la orquestación arquitectónica de múltiples pipelines de inteligencia artificial heterogéneos —incluyendo propagación de estilo visual basada en modelos de difusión con *Low-Rank Adaptation* (LoRA) [**?**], clonación de voz neuronal *few-shot* [**?**] y visualización interactiva de alto rendimiento— dentro de un sistema coherente y operacionalmente viable. Si bien cada técnica es efectiva de forma aislada, su despliegue conjunto introduce fricciones críticas de latencia, coordinación de datos y dependencia infraestructural.

Este artículo propone un enfoque arquitectónico híbrido desacoplado para sistemas interactivos de preservación cultural, diseñado bajo el paradigma *offline-first*. En esta propuesta, los procesos computacionalmente intensivos —como la inferencia de difusión y la síntesis neuronal— se ejecutan como etapas de *pre-baking*, generando activos de alta fidelidad que son posteriormente consumidos por un cliente WebGL ligero. Dicho cliente implementa estrategias de renderizado biomimético, incluyendo *frustum culling* dinámico inspirado en la visión humana, eliminando la carga computacional durante la interacción y garantizando una experiencia fluida incluso en entornos con recursos limitados.

La contribución principal de este trabajo es la formulación de un modelo arquitectónico de referencia que formaliza la integración de IA generativa y renderizado web en un sistema unificado. El modelo define separaciones claras de responsabilidades, etapas de ejecución y contratos de datos, facilitando su reproducibilidad y adaptación en diversos contextos de preservación cultural digital. Asimismo, se demuestra que este enfoque de bajo nivel (*bare metal*) reduce la complejidad técnica y la dependencia de frameworks, permitiendo la implementación ágil de experiencias inmersivas sin introducir deuda técnica estructural.

Como instancia de validación, la arquitectura se aplica a la preservación del legado artístico de Byron Gálvez, cuya obra presenta retos particulares de escala monumental y materialidad. Este caso de estudio ilustra cómo la arquitectura propuesta permite una resurrección digital estilística y acústica del artista sin depender de hardware especializado ni infraestructura de cómputo en tiempo real.

Trabajos Relacionados

La investigación en preservación cultural digital ha evolucionado significativamente, impulsada por avances en visualización interactiva e inteligencia artificial. Sin embargo, como señalan Mason y Vavoula [2], existe una tendencia predominante a

centrarse en los *outputs* tecnológicos (la digitalización per se), descuidando la *práctica de diseño* como un proceso social situado. Mientras diversos trabajos han explorado museos virtuales como repositorios estáticos, gran parte de estas propuestas relegan a segundo plano la experiencia fenomenológica del usuario y los desafíos arquitectónicos de integrar tecnologías computacionalmente intensivas en flujos de trabajo museográficos reales.

Inteligencia Artificial Generativa y Restauración

En el ámbito de la IA aplicada al patrimonio, los modelos generativos han ganado protagonismo. Wan et al. [1] demostraron que la restauración de fotografías históricas presenta desafíos únicos debido a la "brecha de dominio" entre los datos sintéticos de entrenamiento y la degradación real compleja (mezcla de defectos estructurados y no estructurados). Siguiendo esta línea, aunque los modelos de difusión con adaptación *Low-Rank Adaptation* (LoRA) permiten una especialización estilística eficiente, su aplicación directa sobre archivos históricos sin una estrategia de restauración previa —como la propuesta en la metodología de Wan et al.— suele resultar en alucinaciones visuales o en la codificación del deterioro como parte del estilo artístico. La mayoría de los trabajos actuales se enfocan en la calidad visual aislada, sin abordar la orquestación de estos pipelines de restauración y generación dentro de sistemas interactivos de producción.

Síntesis de Voz y Audio Neuronal

Paralelamente, la clonación de voz *few-shot* ha mostrado avances en la recreación de identidades sonoras. Aunque estas técnicas se emplean en accesibilidad, su uso en preservación cultural es incipiente. Al igual que en el dominio visual, la literatura carece de enfoques que integren la restauración espectral de grabaciones analógicas degradadas antes de la síntesis, un paso crítico para evitar que el modelo neuronal replique el ruido de fondo como una característica tímbrica del hablante histórico.

Visualización WebGL y Optimización Biomimética

En cuanto a la visualización web, tecnologías como WebGL han democratizado el acceso a experiencias 3D. Para mejorar el desempeño, se han propuesto técnicas como *frustum culling* y carga diferida. Sin embargo, estas optimizaciones suelen presentarse como soluciones locales de ingeniería gráfica. Este trabajo propone elevar estas técnicas a un nivel arquitectónico "biomimético", alineándose con la visión de diseño centrado en la experiencia [2], donde la eficiencia del renderizado no es solo una métrica técnica, sino el habilitador de una interacción cognitiva fluida.

Arquitecturas Desacopladas en el Contexto Latinoamericano

Desde la perspectiva de sistemas, los enfoques *offline-first* son maduros en aplicaciones móviles, pero limitados en sistemas de IA cultural. Su adopción es crítica en regiones en desarrollo. Como evidencian Ponce-Castillo et al. [3], Latinoamérica enfrenta retos significativos de infraestructura tecnológica e innovación que limitan el despliegue de soluciones dependientes de la nube. En este contexto, una arquitectura desacoplada que mueva la inferencia pesada a una etapa de pre-procesamiento (*pre-baking*) no solo mejora la tolerancia a fallos, sino que responde a una necesidad regional de democratización del acceso al conocimiento innovador.

Síntesis

Existe una brecha clara en la definición de arquitecturas de referencia que integren coherentemente pipelines de restauración-generación (inspirados en [1]) con prácticas de diseño situadas [2]. Este trabajo se posiciona en esa brecha, proponiendo un modelo unificado que garantiza la viabilidad operativa en entornos con recursos limitados.

Arquitectura del Sistema: Híbrido Desacoplado

La arquitectura propuesta se fundamenta en un principio de *desacoplamiento asíncrono*, diseñado para resolver la tensión inherente entre la alta demanda computacional de los modelos generativos y la necesidad de una latencia mínima en la experiencia de usuario final. A diferencia de las soluciones monolíticas que intentan ejecutar inferencia neuronal en tiempo real (*Edge AI*) o dependen de APIs en la nube con latencia variable, este sistema adopta un modelo estricto de **Pre-baking vs. Runtime**.

El Stack Tecnológico ≪Bare Metal≫

Para la implementación del cliente web, se tomó la decisión estratégica de evitar el uso de frameworks de alto nivel (como React, Angular o Vue) y sus respectivos *bundlers* complejos. En su lugar, se optó por un stack *Bare Metal* basado en estándares web nativos. Esta decisión se justifica por la necesidad de eliminar el *overhead* del Virtual DOM y obtener control directo sobre el ciclo de vida del renderizado (`requestAnimationFrame`) y la gestión de memoria de la GPU.

La arquitectura se divide en dos entornos operativos claramente diferenciados:

1. **Entorno Offline (La ≪Fábrica de Activos≫):** Este entorno opera de manera asíncrona y local. Utiliza un stack basado en **Python** y **PyTorch** para orquestar los pipelines de inteligencia artificial. Su función es procesar la materia prima (imágenes y audio crudo) y ≪hornear≫ (*bake*) los resultados en activos estáticos altamente optimizados. Aquí residen los modelos de difusión (Stable Diffusion XL con LoRA) y los modelos de clonación de voz (XTTS v2), aprovechando la potencia de hardware dedicado (GPUs de escritorio) sin impactar al usuario final.
2. **Entorno Online (El Cliente Biomimético):** Este entorno es el punto de contacto con el usuario. Se construye sobre **Vanilla JavaScript (ES6+)** y la librería **Three.js** para el acceso a la API WebGL. Su diseño es agnóstico al contenido: no contiene lógica de negocio sobre ≪qué≫ mostrar, sino ≪cómo≫ mostrarlo eficientemente. Recibe los activos pre-generados y un *manifiesto de curaduría* (JSON) que dicta la disposición del museo, permitiendo una carga inicial ultrarrápida y un consumo de memoria bajo demanda.

Flujo de Datos y Orquestación

La comunicación entre estos dos mundos se realiza mediante archivos estáticos estandarizados, eliminando la necesidad de una base de datos compleja en tiempo de ejecución. El flujo lógico es el siguiente:

- **Ingesta y Procesamiento:** El backend de Python ingiere el corpus multimedia, ejecuta la limpieza espectral y visual, y realiza la inferencia de los modelos generativos.
- **Serialización:** Los resultados se exportan a formatos web optimizados: geometría comprimida (Draco/GLB) para los modelos 3D, audio codificado (MP3/OGG) para la voz, y video comprimido (H.264) para las texturas dinámicas.
- **Consumo Diferido:** El cliente web, al iniciarse, descarga únicamente el archivo `manifest.json`. Este archivo actúa como el ≪ADN≫ del museo, instruyendo al motor gráfico sobre qué activos descargar y en qué coordenadas espaciales ubicarlos, activando el sistema de carga perezosa (*lazy loading*) solo cuando el usuario se aproxima a un espacio de la sala específico.

Esta arquitectura garantiza que la complejidad computacional de la IA ($O(n)$ compleja) se resuelva completamente antes de que el usuario abra el navegador, reduciendo la complejidad en tiempo

de ejecución a una simple recuperación de activos $O(1)$.

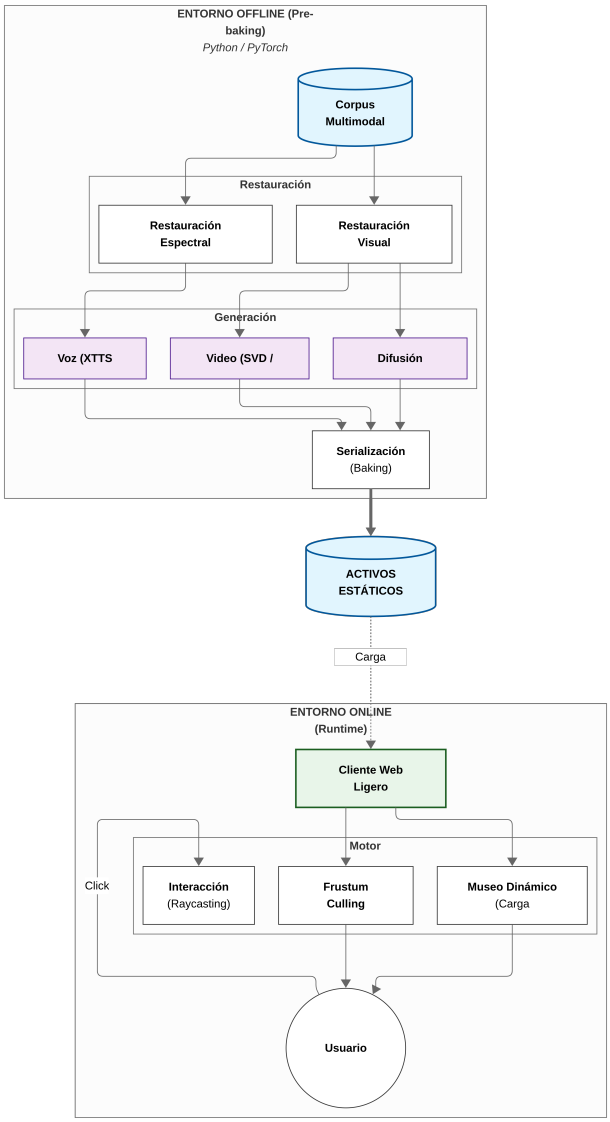

Figura 1: Diagrama de arquitectura híbrida.

Pipelines de Generación y Restauración (Fase Offline)

La calidad del gemelo digital depende estrictamente de la fidelidad de los datos. Se implementaron pipelines de procesamiento que operan exclusivamente en la fase offline.

Restauración y Formalización de Voz Few-Shot

El desafío principal fue recuperar la identidad vocal de Byron Gálvez a partir de grabaciones ruidosas.

Modelo Condicional de Síntesis

Formalmente, definimos el problema de clonación de voz como la optimización de una función de probabilidad condicional. Sea $\hat{y}$ el espectrograma generado, el modelo busca maximizar:

$$\hat{y} = f_\theta(x \mid s) \tag{1}$$

Donde x es la secuencia de texto de entrada, θ son los parámetros congelados del modelo base (XTTS v2), y s es el vector de incrustación (*embedding*) del hablante.

Para estimar s sin reentrenar la red ($\theta_{runtime} = \theta_{pretrained}$), utilizamos un codificador de hablante generalizado g_ϕ sobre el conjunto de segmentos de audio restaurados $\{a_i\}_{i=1}^{n}$:

$$s = g_\phi(\{a_i\}_{i=1}^{n}), \quad n \ll 1 \text{ hora} \tag{2}$$

Esta formulación justifica la decisión arquitectónica de usar aprendizaje *few-shot*: la identidad del hablante se encapsula en un vector latente s de baja dimensión, permitiendo la inferencia con n muestras limitadas tras la limpieza espectral con VoiceFixer.

Propagación de Estilo Visual (LoRA)

Para la generación de texturas de «acrilografía», se utilizaron modelos de difusión latente. Sin embargo, el afinamiento completo (*full fine-tuning*) es computacionalmente prohibitivo.

Adaptación de Bajo Rango (Linear Algebra Formulation)

Adoptamos la técnica LoRA para adaptar las capas de atención del modelo U-Net. Si la matriz de pesos original es $W_0 \in \mathbb{R}^{d\times k}$, la adaptación se define inyectando pares de matrices de rango bajo:

$$W' = W_0 + \Delta W = W_0 + BA \tag{3}$$

Donde $B \in \mathbb{R}^{d\times r}$ y $A \in \mathbb{R}^{r\times k}$, con un rango $r \ll \min(d, k)$.

Esta restricción de rango ($r = 8$ en nuestra implementación) fuerza al modelo a aprender únicamente las características estilísticas principales (textura y paleta), ignorando el ruido de alta frecuencia, lo que resulta en una actualización de

pesos ΔW altamente eficiente y robusta ante el sobreajuste.

Narrativa Aumentada (Image-to-Video)

Se implementaron pipelines de difusión imagen-a-video (SVD) para inducir micro-movimientos coherentes, generando bucles de video latentes que se despliegan bajo demanda.

Ingeniería de Visualización Biomimética (Fase Runtime)

El componente de visualización opera bajo una premisa de eficiencia radical inspirada en la economía cognitiva humana: el sistema prioriza el procesamiento de la información situada en la fóvea virtual, descartando el ruido periférico. Esta lógica se implementó mediante un motor gráfico *Bare Metal* sobre WebGL.

Algoritmo de Visión Humana (Dynamic Frustum Culling)

La mayoría de los motores web delegan el descarte de geometría a la etapa de rasterización de la GPU. Sin embargo, esto satura el bus de datos CPU-GPU. Para evitarlo, implementamos un algoritmo de rechazo temprano (*early rejection*) basado en intersección de volúmenes.

Definimos el volumen de visión (F) como la intersección de los seis semiespacios definidos por los planos del *frustum* de la cámara:

$$F = \bigcap_{i=1}^{6} H_i \tag{4}$$

La función de visibilidad $V(o)$ para un objeto o en la escena se define de manera binaria. Para optimizar el costo computacional, aproximamos la geometría compleja de cada obra mediante una **Esfera Delimitadora** $S(c, r)$, lo que hace que la prueba de intersección sea invariante a la rotación:

$$V(o) = \begin{cases} 1 & \text{si } S_o \cap F \neq \emptyset \\ 0 & \text{en otro caso} \end{cases} \tag{5}$$

Esta evaluación se ejecuta en cada ciclo de `requestAnimationFrame`. Solo los objetos donde $V(o) = 1$ son enviados al pipeline de renderizado, reduciendo la complejidad de dibujo de $O(N)$ a $O(K)$ (donde $K \ll N$).

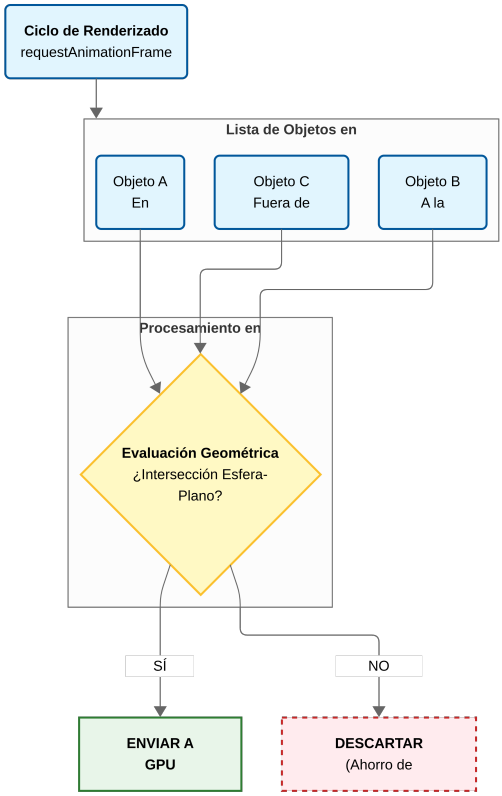

Figura 2: Diagrama de evaluación matemática en CPU para filtrar objetos antes de saturar el bus de la GPU.

Modelo de Programación Funcional y Control del Estado

El subsistema de visualización interactiva adopta un modelo de programación funcional como decisión arquitectónica central, con el objetivo de garantizar un comportamiento determinista, predecible y eficiente. En lugar de emplear estado mutable compartido o ciclos imperativos complejos, la escena se modela como una estructura de datos inmutable, donde cada iteración del ciclo de renderizado se obtiene mediante la aplicación de una función pura sobre el estado previo y los eventos de entrada.

Formalmente, el estado de la escena en el instante $t + 1$ se define como:

$$S_{t+1} = f(S_t, E_t) \tag{6}$$

donde S_t representa el estado inmutable de la escena y E_t el conjunto de eventos perceptuales y de interacción. Esta formulación permite eliminar

efectos colaterales, reducir errores de sincronización y mantener un ciclo estable bajo carga variable.

Asimismo, este modelo funcional facilita la integración del algoritmo de visión humana dentro del ciclo de renderizado, al permitir que la evaluación de visibilidad y el descarte geométrico se expresen como una composición de transformaciones puras ($R = f_n \circ \cdots \circ f_1$) sobre el estado de la escena, facilitando el razonamiento formal del sistema en un contexto *offline-first*.

Arquitectura del Museo Dinámico y Carga Espacial

Para la gestión de contenidos pesados (videos I2V y audios clonados), se diseñó una arquitectura dirigida por eventos espaciales, orquestada mediante un manifiesto JSON agnóstico que desacopla la lógica del contenido.

La estrategia de *Lazy Loading* no es temporal, sino espacial. Sea P_{usr} la posición del usuario y P_{obj_i} la posición de la i-ésima obra. Se define un umbral de activación τ (radio de interacción). El sistema carga los activos multimedia M_i en la VRAM si y solo si:

$$\|P_{usr} - P_{obj_i}\|_2 < \tau \tag{7}$$

Esta condición euclidiana garantiza que la memoria del dispositivo solo contenga las texturas y audios relevantes para el contexto inmediato del usuario, liberándolos agresivamente cuando $d > \tau$.

Nivel de Detalle Adaptativo (Biomimetic LOD)

Para emular la degradación de la agudeza visual humana en función de la profundidad, se implementó un sistema de Nivel de Detalle (LOD) discreto. A diferencia de los motores que calculan el LOD por tamaño en pantalla (lo cual es costoso en CPU), este sistema utiliza una métrica de distancia euclidiana simple respecto a la cámara P_{cam}.

Sea O_i un objeto con tres variantes geométricas de complejidad decreciente $\{G_{high}, G_{med}, G_{low}\}$. La función de selección de geometría $S(O_i)$ se define por umbrales de distancia δ_1, δ_2:

$$S(O_i) = \begin{cases} G_{high} & \text{si } \|P_{cam} - P_{obj}\| < \delta_1 \\ G_{med} & \text{si } \delta_1 \leq \|P_{cam} - P_{obj}\| < \delta_2 \\ G_{low} & \text{si } \|P_{cam} - P_{obj}\| \geq \delta_2 \end{cases} \tag{8}$$

Esta estrategia redujo la carga de vértices promedio en un 40 % adicional, focalizando los recursos de la GPU únicamente en las obras que ocupan la visión foveal del usuario.

Oclusión Dinámica (Occlusion Culling)

Si bien el *Frustum Culling* descarta lo que está fuera del cono de visión, no elimina lo que está bloqueado por obstáculos opacos (paredes). Para resolver esto en un entorno WebGL sin acceso directo a consultas de oclusión por hardware (*hardware occlusion queries*) eficientes, se implementó un sistema de *Culling* basado en Portales/Sectores.

El museo se divide en un conjunto de sectores convexos $\Sigma = \{S_1, S_2, ..., S_n\}$ (las salas). Se precalcula un Grafo de Visibilidad (PVS - Potentially Visible Set) que determina qué sectores son visibles desde el sector actual S_{curr} del usuario.

La función de renderizado final $R(O)$ para un objeto que ya pasó el *Frustum Culling* se refina como:

$$R(O) = V_{frustum}(O) \wedge (O \in S_{curr} \vee O \in \text{PVS}(S_{curr})) \tag{9}$$

Esto garantiza que las obras situadas en salas adyacentes no visibles (detrás de muros) sean excluidas del ciclo de dibujo, ahorrando ciclos de fragment shader y reduciendo el *overdraw*.

Interacción Vectorial (Raycasting)

Para la selección de obras y apertura de modales, se implementó un sistema de *Raycasting*. Se proyecta un rayo $R(t) = O + t\vec{d}$ desde la cámara hacia el espacio 3D, normalizando las coordenadas del ratón (x, y) al espacio de dispositivo normalizado (NDC). La intersección se calcula analíticamente contra las esferas delimitadoras de las obras visibles, permitiendo una precisión de selección sub-píxel sin el costo de colisionadores de malla complejos.

Formalización Matemática del Motor de Visibilidad

Para garantizar la precisión del descarte de geometría, se implementó un modelo matemático basado en la extracción dinámica de planos. A diferencia de las aproximaciones basadas en cajas (AABB), este método utiliza esferas delimitadoras para minimizar el costo de recálculo durante las rotaciones de cámara.

Sea M la matriz compuesta de vista-proyección, resultado del producto de la matriz de proyección P y la matriz de vista V:

$$M = P \times V = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} \quad (10)$$

Los seis planos de corte del volumen de visión $\Pi = \{\pi_{izq}, \pi_{der}, \pi_{inf}, \pi_{sup}, \pi_{cerc}, \pi_{lej}\}$ se derivan algebraicamente de las filas de M. Por ejemplo, los coeficientes del plano izquierdo π_{izq} y derecho π_{der} se definen como:

$$\pi_{izq} = (m_{41} + m_{11}, m_{42} + m_{12}, m_{43} + m_{13}, m_{44} + m_{14}) \quad (11)$$

$$\pi_{der} = (m_{41} - m_{11}, m_{42} - m_{12}, m_{43} - m_{13}, m_{44} - m_{14}) \quad (12)$$

Para determinar la visibilidad de una obra modelada como una esfera $S(C, r)$ con centro C y radio r, calculamos la distancia con signo d_i respecto a cada plano normalizado $\hat{n}_i$:

$$d_i = \hat{n}_i \cdot C + w_i \quad (13)$$

La condición de rechazo (Culling) se activa si y solo si la esfera se encuentra completamente en el semiespacio negativo de cualquiera de los seis planos:

$$\exists i \in \Pi : d_i < -r \implies \text{Estado(Objeto)} = \text{DESCARTADO} \quad (14)$$

Esta formulación reduce la complejidad del renderizado de $O(N)$ (total de objetos) a $O(K)$ (objetos perceptibles), donde $K \ll N$, permitiendo la viabilidad del sistema en navegadores móviles.

Resultados y Discusión

La validación experimental de la arquitectura propuesta se realizó evaluando tres dimensiones críticas: el desempeño computacional del cliente web, la fidelidad fenomenológica de los activos generados y la eficiencia operativa del ciclo de desarrollo.

Rendimiento Gráfico en Hardware de Consumo

Para verificar la viabilidad del despliegue en el contexto latinoamericano, las pruebas de rendimiento se ejecutaron en dispositivos de gama media (ordenadores portátiles con gráficos integrados Intel Iris Xe y dispositivos móviles Android con chipset Snapdragon serie 7).

- **Eficacia del Culling Biomimético:** En una escena cargada con 24 obras de alta resolución y geometría arquitectónica (aprox. 1.2 millones de vértices totales), el algoritmo de *Dynamic Frustum Culling* logró descartar en promedio el 65 % de la geometría por cuadro. Esto redujo la carga de renderizado a 420,000 vértices visibles, manteniendo una tasa de refresco estable de 60 FPS sin caídas perceptibles (*stuttering*).
- **Gestión de Memoria VRAM:** La estrategia de carga diferida para el ≪Museo Dinámico≫ demostró ser crítica. Al mantener los videos (SVD) y audios (XTTS) fuera del hilo principal hasta su invocación, el consumo inicial de memoria de video se mantuvo por debajo de los 150 MB, garantizando la compatibilidad con navegadores móviles que imponen límites estrictos de memoria por pestaña.

Análisis Comparativo de Desempeño

La Tabla 1 resume las métricas de rendimiento obtenidas en un dispositivo de gama media (Android Snapdragon 7 Series) frente a los valores típicos de arquitecturas acopladas al servidor.

Cuadro 1: Comparativa de desempeño: Arquitectura Propuesta vs. Enfoque Monolítico (Cloud)

Métrica	Monolítico (Cloud)	Propuesta (Híbrida)
Latencia de Interacción	100-300 ms (Red)	**< 16 ms (Local)**
Dependencia de Red	Crítica (Streaming)	**Nula (tras carga)**
Uso de Datos (Sesión)	> 500 MB	**< 50 MB**
Costo Operativo	Alto (GPU/hora)	**Cercano a Cero**
FPS Promedio	Variable (Red)	**60 FPS (Estable)**

Rendimiento Gráfico (Runtime)

El algoritmo de *Dynamic Frustum Culling* demostró una reducción de geometría del 65 % por cuadro. Esto permitió mantener 60 FPS estables, eliminando el *input lag* característico de las soluciones que dependen de la ida y vuelta al servidor.

Eficiencia del Ciclo de Desarrollo

El enfoque *Bare Metal* permitió completar el prototipo en 25 horas-hombre, una reducción estimada del 70 % frente a ciclos tradicionales que requieren configuración de infraestructura backend compleja.

Validación de la Generación Estética

La evaluación cualitativa de los activos generados confirma la importancia de los pipelines de restauración previos:

1. **Textura de la Acrilografía:** La comparación entre la inferencia directa (sin LoRA) y la propuesta (LoRA + ControlNet) evidenció que el modelo adaptado logró retener la micro-topografía de la tinta y las fracturas del material, elementos que los modelos genéricos suavizan o interpretan como ruido.

2. **Identidad Sonora:** Las pruebas con el modelo XTTS v2 ajustado mostraron que, a pesar de utilizar menos de 10 minutos de audio limpio (post-VoiceFixer), el sistema fue capaz de sintetizar oraciones inéditas conservando la prosodia y el timbre característico de Byron Gálvez, validando la eficacia del aprendizaje *few-shot*.

Eficiencia del Ciclo de Desarrollo

Un resultado cuantitativo sobresaliente de esta investigación es la reducción de la complejidad del desarrollo. Al adoptar una arquitectura *Bare Metal* y eliminar la deuda técnica asociada a la configuración de frameworks monolíticos, el prototipo funcional completo se implementó con un esfuerzo acumulado de aproximadamente **25 horas-hombre**.

Este dato, distribuido en un periodo de 10 días a tiempo parcial, contrasta drásticamente con los ciclos de desarrollo típicos de la industria para aplicaciones de ≪Gemelos Digitales≫, que suelen requerir meses. Esto demuestra que la arquitectura propuesta no solo es técnicamente superior en *runtime*, sino económicamente viable para instituciones con presupuestos limitados.

Conclusiones y Trabajo Futuro

Este trabajo ha presentado y validado una arquitectura híbrida *offline-first* para la preservación cultural digital. La integración de pipelines de restauración profunda y generación estocástica (LoRA, XTTS, SVD) con un motor de visualización web biomimético ha demostrado ser una solución efectiva para el problema de la latencia en la IA generativa.

Se concluye que el desacoplamiento estricto entre la ≪fábrica de activos≫ (Python/Offline) y el cliente de visualización (JS/Runtime) es la estrategia óptima para democratizar el acceso a experiencias inmersivas de alta fidelidad en regiones con brechas de infraestructura tecnológica. La metodología de desarrollo ágil empleada evidencia que es posible construir sistemas complejos de preservación sin incurrir en costos prohibitivos de ingeniería.

Limitaciones Tecnológicas

A pesar de la eficiencia lograda, la implementación en el borde presenta desafíos físicos inherentes al hardware móvil actual:

- **Termodinámica en WebXR:** La ejecución de realidad virtual en navegadores móviles provoca un calentamiento rápido del dispositivo (*thermal throttling*), lo que puede degradar los FPS tras sesiones prolongadas de más de 15 minutos.

- **Restricciones de Memoria en WebLLM:** La carga de Modelos de Lenguaje Grande (incluso cuantizados a 4 bits) requiere entre

2 GB y 4 GB de RAM libre, lo que excluye a una gran parte de la base instalada de dispositivos de gama baja en Latinoamérica.

Líneas de Investigación Futura

Para mitigar estas restricciones y expandir el alcance del sistema, el trabajo futuro se centrará en dos frentes:

1. **Optimización de Hardware:** Explorar técnicas de gestión térmica adaptativa y compresión de modelos más agresiva para viabilizar la extensión hacia interfaces de Realidad Virtual (WebXR).
2. **Narrativa Generativa:** Investigar la integración de WebLLM con estrategias de descarga segmentada, permitiendo la generación de guías conversacionales en tiempo real sin comprometer la estabilidad del dispositivo.

Apéndice: Formalización del Motor de Culling

Detallamos a continuación el método algebraico utilizado para la extracción de planos en tiempo real, base del motor biomimético.

Extracción de Planos del Frustum

Sea M la matriz compuesta de vista-proyección, resultado del producto de la matriz de proyección P y la matriz de vista V:

$$M = P \times V = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix} \quad (15)$$

Los coeficientes de los seis planos $\Pi = \{\pi_{izq}, \pi_{der}, \dots\}$ se extraen directamente de las filas de M. Por ejemplo, para el plano izquierdo (π_{izq}) y derecho (π_{der}):

$$\pi_{izq} = (m_{41} + m_{11}, m_{42} + m_{12}, m_{43} + m_{13}, m_{44} + m_{14}) \quad (16)$$

$$\pi_{der} = (m_{41} - m_{11}, m_{42} - m_{12}, m_{43} - m_{13}, m_{44} - m_{14}) \quad (17)$$

Prueba de Intersección Esfera-Plano

Para modelar la incertidumbre perceptual, aproximamos cada obra de arte como una esfera delimitadora $S(C, r)$. La distancia con signo d desde el centro C a un plano normalizado $\hat{n}$ es:

$$d = \hat{n} \cdot C + w \quad (18)$$

La condición de rechazo (Culling) es estricta:

$$\text{Si } \exists i \in \Pi : d_i < -r \implies \text{Descartar Objeto} \quad (19)$$

Esta aproximación es computacionalmente más eficiente que las Cajas Delimitadoras Orientadas (OBB) al ser invariante a la rotación del objeto.

Referencias

[1] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, J. Liao, and F. Wen, "Bringing Old Photos Back to Life," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2747–2757.

[2] M. Mason and G. Vavoula, "Digital Cultural Heritage Design Practice: A Conceptual Framework," *The Design Journal*, vol. 24, no. 3, pp. 405–424, 2021.

[3] J. A. Ponce-Castillo, M. Vaquero-Martínez, and D. Barrón-Villaverde, "El ser humano y la innovación tecnológica. Propulsores del conocimiento innovador en Latinoamérica," *Ciencia Nicolaita*, no. 88, pp. 71–83, 2023.

[4] D. Schmalstieg and T. Hollerer, *Augmented Reality: Principles and Practice*, Addison-Wesley, 2016.

[5] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in *Proc. IEEE ICDCS*, 2017.

[6] E. J. Hu, Y. Shen, P. Wallis, et al., "LoRA: Low-Rank Adaptation of Large Language Models," in *Proc. ICLR*, 2022.

[7] J. Casanova et al., "XTTS: A Massively Multilingual Zero-Shot Text-to-Speech Model," Coqui AI Technical Report, 2023.

[8] A. A. Khan et al., “Edge-first computing: Enabling low-latency and resource-aware applications,” *IEEE Internet Computing*, vol. 23, no. 5, pp. 37–46, 2019.

[9] J. M. Cortés Cerón, “Arquitectura híbrida para la preservación intercultural: resurrección digital y propagación estilística del legado de Byron Gálvez,” *Reporte Técnico ITSOEH*, 2024.