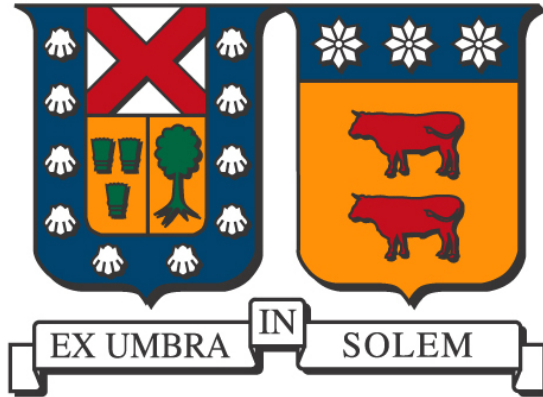


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA
SANTIAGO- CHILE



**“SISTEMA INTELIGENTE PARA LA TRANSFORMACIÓN DE
OBRAS DE ARTE VISUALES EN EXPERIENCIAS AUDITIVAS
ACCESIBLES PARA PERSONAS CON DISCAPACIDAD
VISUAL”**

TOMÁS ANDRES CAMPUSANO GAETE

**MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL
TELEMÁTICO**

PROFESOR GUÍA:

PATRICIO OLVIARES RONCAGLIOLO

PROFESOR CORREFERENTE:

NICOLÁS TORRES RUDLOFF



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: SISTEMA INTELIGENTE PARA LA TRANSFORMACIÓN DE OBRAS DE ARTE VISUALES EN EXPERIENCIAS AUDITIVAS ACCESIBLES PARA PERSONAS CON DISCAPACIDAD VISUAL

Nombre del candidato(a): Tomas Andres Campusano Gaete

Carrera / Grado: Ingeniería Civil Telemática

Campus: San Joaquín Departamento: Telemática / Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Patricio Olivares Roncagliolo, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (marcar una opción):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 12/01/2026

Firma: 

Estudiante o Candidato(a):

Fecha: 08/01/2026

Firma: 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Agradecimientos

Me gustaría iniciar mis agradecimientos expresando mi más sincero agradecimiento a mis padres, quienes fueron un pilar fundamental para poder finalizar mi carrera. Su apoyo constante, sus palabras de aliento y la confianza que depositaron en mí fueron clave para no abandonar mis estudios, incluso en los momentos más difíciles, recordándome siempre que el esfuerzo tendría frutos en el futuro.

Asimismo, agradezco profundamente a mi polola, Vania, quien más que nadie fue testigo de mis etapas más complejas, marcadas por el estrés y el desánimo, así como también de mis logros y momentos de éxito. Mi compañera incondicional, quien me comprende como nadie, y a quien estaré eternamente agradecido por todo su apoyo. Además, valoro enormemente su aporte a esta memoria a través de sus ideas y su visión profesional como psicopedagoga.

Extiendo también mi agradecimiento a mi familia en general, por su constante preocupación, apoyo y palabras de ánimo, y por interesarse siempre en la evolución de mi formación académica.

Agradezco a todos los profesores con quienes tuve la oportunidad de aprender durante mi formación universitaria, y en especial a Nicolás Torres y Patricio Olivares, por su valioso aporte en ideas, técnicas y por contribuir significativamente al desarrollo de mis capacidades técnicas y teóricas. De manera especial, agradezco al profesor Sebastián Contreras, quien fue más que un docente, convirtiéndose en un guía para comprender el mundo profesional más allá de la universidad, y por confiar durante tanto tiempo en mis conocimientos al permitirme desempeñarme como su ayudante en diversas asignaturas.

Finalmente, agradezco a mis amigos de la universidad, en especial a Diego Garrido e Ignacio Silva, quienes representan el verdadero significado de la familia que uno elige. No pude haber elegido mejores amistades, con quienes realicé gran parte —si no la totalidad— de los trabajos grupales, haciendo de la experiencia universitaria un

camino más llevadero y feliz.

También agradezco a mis amigos de The Amigods, quienes aportaron de manera significativa a esta memoria mediante su reinterpretación de diversas obras de arte, y que además fueron un apoyo fundamental durante el período de pandemia.

Resumen

En el presente informe se aborda el desafío de mejorar la accesibilidad de las obras de arte visuales para personas con discapacidad visual mediante el uso de inteligencia artificial. Se propone una plataforma accesible que transforma automáticamente imágenes de obras de arte en experiencias sonoras inmersivas, combinando descripciones narrativas y paisajes auditivos generados por modelos generativos. El sistema se estructura en cuatro módulos principales: un generador de audio descriptivo multimodal, un narrador contextual histórico, un generador de sonidos ambientales inmersivos y una interfaz web inclusiva basada en normativas de accesibilidad que integra los tres módulos anteriores. En particular, este trabajo se centra en el desarrollo del generador de audio descriptivo multimodal, encargado de interpretar y describir los elementos visuales presentes en las obras mediante modelos de lenguaje multimodales. Los resultados demuestran que el sistema es capaz de generar descripciones coherentes y auditivamente comprensibles, constituyendo un aporte significativo a la accesibilidad cultural y al uso de inteligencia artificial en experiencias sensoriales inclusivas.

Palabras Clave: Inteligencia Artificial , LLM ,Accesibilidad,Inclusión,Arte

Abstract

This report addresses the challenge of improving the accessibility of visual artworks for people with visual impairments through the use of artificial intelligence. It proposes an accessible platform that automatically transforms images of artworks into immersive auditory experiences, combining narrative descriptions and soundscapes generated by generative models. The system is structured into four main modules: a multimodal descriptive audio generator, a historical contextual narrator, an immersive ambient sound generator, and an inclusive web interface based on accessibility standards that integrates the previous three modules. In particular, this work focuses on the development of the multimodal descriptive audio generator, responsible for interpreting and describing the visual elements present in the artworks through multimodal language models. The results show that the system is capable of producing coherent and auditorily comprehensible descriptions, representing a significant contribution to cultural accessibility and the use of artificial intelligence in inclusive sensory experiences.

Keywords:Artificial Intelligence, LLM, Accessibility, Inclusion,Art

Glosario

Alucinación Contenido generado por un modelo que no está sustentado por la información de entrada.

API *Application Programming Interface* (Interfaz de Programación de Aplicaciones).

Baja visión Disminución significativa de la capacidad visual que persiste aun con corrección óptica y dificulta tareas cotidianas.

Base64 Esquema de codificación que representa datos binarios como texto ASCII para facilitar su transmisión mediante APIs.

Clonación de voz Técnica que permite generar voz sintética imitando características de una voz humana específica, considerando aspectos éticos y legales.

Conjunto de datos / Dataset Colección de datos utilizada para pruebas, evaluación o alimentación del sistema, como imágenes de obras de arte.

GAN *Generative Adversarial Network* (Red Generativa Antagónica).

GPT-OSS Modelo de lenguaje empleado para tareas texto-a-texto, particularmente traducción automática.

gTTS Sistema de conversión texto a voz utilizado para generar archivos de audio en formato MP3.

Hugging Face Plataforma y repositorio de modelos de aprendizaje automático ampliamente utilizada en investigación y desarrollo.

IA Inteligencia Artificial.

JSON *JavaScript Object Notation* (Formato de intercambio de datos estructurados).

LLM *Large Language Model* (Modelo de Lenguaje de Gran Tamaño).

MLLM *Multimodal Large Language Model* (Modelo Multimodal de Lenguaje).

Muestra por conveniencia Selección no probabilística de participantes basada en disponibilidad y acceso.

Multimodalidad Capacidad de un sistema o modelo para procesar o generar información en más de una modalidad, como imagen, texto y audio.

Nebius AI Proveedor de servicios de cómputo en la nube utilizado para la ejecución de modelos mediante APIs.

Pipeline Flujo automatizado compuesto por etapas secuenciales de procesamiento. En este proyecto: imagen → descripción → traducción → audio.

Prompt Instrucción textual utilizada para guiar el comportamiento de un modelo de lenguaje y restringir el formato o contenido de la salida.

Qwen3-VL-Instruct Modelo multimodal visión-lenguaje utilizado para interpretar imágenes y generar descripciones textuales siguiendo instrucciones definidas.

RAG *Retrieval-Augmented Generation* (Generación Aumentada por Recuperación).

Transformers Biblioteca de Python para el uso de modelos preentrenados basados en arquitecturas modernas de aprendizaje profundo.

TTS *Text-to-Speech* (Conversión de texto a voz).

Validación exploratoria Evaluación preliminar orientada a obtener retroalimentación inicial sin pretensión de generalización estadística.

VLM *Vision-Language Model* (Modelo Visión–Lenguaje).

WikiArt Repositorio digital de obras de arte utilizado como fuente de datos en investigaciones académicas.

Tabla de Contenido

1	Introducción	1
1.1	Contexto	1
1.2	Motivación	2
1.3	Objetivos	3
1.3.1	Objetivo Principal	3
1.3.2	Objetivos Específicos	3
1.4	Justificación del Proyecto	3
1.5	Alcance y Limitaciones	4
1.5.1	Alcance	4
1.5.2	Limitaciones	5
1.6	Enfoque Metodológico	5
1.6.1	Metodología empleada	5
1.6.2	Herramientas y Tecnología	7
2	Estado del Arte	9
3	Propuesta	12
3.1	Concepto General	12
3.2	Descripción módulo	14
3.2.1	Componentes Principales	14
3.2.2	Flujo de trabajo	15
3.2.3	Tecnología utilizadas	16
3.3	Modelo de Generación de descripción	16
3.3.1	Selección del Modelo	16
3.3.2	Proceso de Generación	18
3.3.3	Evaluación inicial	19
3.4	Generación de Audio	20
3.4.1	Selección de Modelo	20

ÍNDICE DE FIGURAS

3.4.2	Evaluación Inicial	20
3.5	Integración al Sistema	20
4	Validación	21
4.1	Objetivo de la Validación	21
4.2	Ajuste Metodológico	21
4.3	Participantes y Contexto de Evaluación	22
4.4	Instrumento de Evaluación	22
4.5	Resultados de la Validación	22
4.6	Validación desde la reinterpretación visual	23
5	Resultados	26
6	Conclusiones	28
7	Trabajo Futuro	30
	Referencias	31

Anexos

Anexo A Resultados comparativos de modelos Vision–Language

Índice de figuras

2.1	Tipos de tareas y problemáticas abordadas por asistentes visuales para personas ciegas o con baja visión. Fuente: adaptado de [1].	10
2.2	Arquitectura general del modelo MT-Net propuesto para la generación de descripciones auditivas a partir de imágenes. Fuente: adaptado de [2].	11
3.1	Diagrama general de la arquitectura del sistema propuesto. Fuente: elaboración propia.	13

- 3.2 Diagrama del módulo generador de descripciones auditivas multimodales. Fuente: elaboración propia. 15
- 4.1 Comparación entre la obra original y las reinterpretaciones visuales realizadas por participantes a partir de la descripción auditiva generada por el sistema. 26
- A.1 Ejemplo de descripción generada por el modelo BLIP-2.
- A.2 Ejemplo de descripción generada por el modelo LLaVa.
- A.3 Ejemplo de descripción generada por el modelo Gemma.
- A.4 Ejemplo de descripción generada por el modelo Llama 4 Maverick.
- A.5 Ejemplo de descripción generada por el modelo Llama 4 Scout.
- A.6 Ejemplo de descripción generada por el modelo Qwen3-VL-Instruct.

Índice de tablas

- 1 Comparación cualitativa de modelos Vision–Language evaluados 18
- 2 Resultados de la validación exploratoria del módulo (n=15). 23

1. Introducción

1.1. Contexto

Según la Encuesta ENDIDE (2022 [3]), en Chile existen 2.703.893 personas mayores de 18 años con algún grado de discapacidad, lo que representa un 17,6 % de la población nacional. De acuerdo con el análisis realizado por la Fundación Luz sobre los resultados de la misma encuesta [4], se estima que en el país hay 153.560 personas adultas ciegas, equivalente aproximadamente al 1 % de la población adulta total, y 4.683.567 personas con algún grado de pérdida de visión.

Estas cifras evidencian un desafío significativo en materia de accesibilidad, particularmente en aquellos ámbitos donde la información y la experiencia dependen fuertemente de la percepción visual. Uno de estos ámbitos corresponde al acceso a la cultura y, en específico, a las artes visuales, cuyo disfrute tradicionalmente se encuentra mediado por la observación directa de imágenes, obras y espacios expositivos. A pesar de los avances en políticas inclusivas y en tecnologías de asistencia, el acceso a las artes visuales para personas con discapacidad visual continúa siendo limitado.

Si bien existen soluciones como audioguías o recorridos táctiles, estas suelen ofrecer experiencias descriptivas genéricas, restringidas y poco inmersivas, que no siempre logran transmitir la riqueza visual, espacial y estética de las obras. En este contexto, la inteligencia artificial emerge como una herramienta con alto potencial para abordar específicamente la accesibilidad cultural, al permitir la generación de experiencias sonoras más detalladas, dinámicas y sensorialmente inclusivas, orientadas a mejorar la comprensión y apreciación de las artes visuales por parte de personas con discapacidad visual.

1.2. Motivación

El acceso al arte y la cultura es reconocido como un derecho fundamental, tanto en el marco de los derechos humanos (art.27 [5]) como en la convención sobre los derechos de las personas con discapacidad (art.30 [6]), al considerarse un componente esencial para la participación plena en la vida social y cultural que, sin embargo, enfrenta múltiples barreras para las personas en situación de discapacidad visual. A pesar del avance en tecnologías de accesibilidad, la experiencia estética de las obras de arte visuales continúa siendo limitada para este grupo de personas. En este contexto, la inteligencia artificial ofrece nuevas oportunidades para transformar la forma en que las personas interactúan con el arte, permitiendo experiencias más inmersivas, accesibles y sensorialmente enriquecedoras.

En este contexto, la inteligencia artificial ofrece nuevas oportunidades para abordar problemáticas de accesibilidad, permitiendo la generación de mecanismos alternativos que faciliten el acceso a la información visual de manera autónoma y comprensible para personas con discapacidad visual. Si bien estas tecnologías podrían, de forma adicional y no intencionada, aportar a experiencias más enriquecidas para personas sin dificultades de visión, dicho aspecto no constituye el objetivo central del presente proyecto

Este proyecto propone el desarrollo de una plataforma basada en inteligencia artificial capaz de interpretar obras de arte visuales y traducirlas en representaciones sonoras, combinando descripciones narrativas, contextualizaciones históricas y sonidos ambientales. No obstante, si bien la plataforma considera múltiples componentes, el presente documento se enfoca específicamente en el desarrollo del módulo de generación de audio descriptivo multimodal, encargado de analizar automáticamente los elementos visuales de una obra y producir descripciones auditivas coherentes y significativas mediante el uso de modelos de lenguaje multimodales. Este módulo busca mejorar la comprensión de las obras de arte por parte de personas con discapacidad visual, contribuyendo de manera directa al fortalecimiento de la accesibilidad cultural.

1.3. Objetivos

1.3.1. Objetivo Principal

- Mejorar la experiencia estética y promover la accesibilidad universal a las obras de arte para personas con discapacidad visual, a través del diseño e implementación de una plataforma accesible que utilice inteligencia artificial para transformar automáticamente imágenes de obras artísticas en paisajes sonoros.

1.3.2. Objetivos Específicos

- Analizar modelos de lenguaje multimodales capaces de interpretar contenido visual, junto con tecnologías de conversión de texto a voz, con el fin de determinar su aplicabilidad en la generación de descripciones auditivas de obras de arte.
- Implementar un flujo de trabajo automatizado imagen → descripción → audio, que integre los componentes de análisis visual, generación textual y síntesis de voz dentro del entorno de la plataforma general del proyecto.
- Construir un conjunto de datos de obras de arte de referencia, incorporando descripciones.
- Validar un prototipo funcional del generador de audio descriptivo multimodal, evaluando la coherencia semántica, la naturalidad y la comprensibilidad de las descripciones generadas.

1.4. Justificación del Proyecto

El acceso al arte constituye un derecho cultural fundamental, reconocido en el marco de los derechos humanos como un componente esencial para la participación plena en la vida social, educativa y emocional de las personas (Art. 27, Declaración Universal de los Derechos Humanos [5]; Art. 30, Convención sobre los Derechos de las

Personas con Discapacidad [6]). Sin embargo, quienes viven con discapacidad visual enfrentan barreras históricas y persistentes para acceder a obras de arte visuales, debido a que la mayoría de los espacios culturales y medios de difusión no disponen de mecanismos adecuados para una experiencia sensorial plena. Esta exclusión limita no solo la apreciación del arte como tal, sino que también dificulta el desarrollo de habilidades estéticas y emocionales, además de restringir la integración de las personas con discapacidad visual en las comunidades artísticas.

La inteligencia artificial ofrece una oportunidad para reducir estas brechas de accesibilidad mediante la generación automática de contenidos alternativos que permitan a las personas con discapacidad visual interpretar obras de arte visuales a través del sonido. El desarrollo de un módulo generador de audio descriptivo contribuye directamente a ampliar las posibilidades de participación cultural, otorgando a las personas con discapacidad visual acceso a información visual relevante de forma autónoma, comprensible y natural.

1.5. Alcance y Limitaciones

1.5.1. Alcance

El sistema se enfoca en:

- **Generación automática de descripciones de obras de arte:** el sistema será capaz de interpretar imágenes y generar descripciones textuales coherentes, basadas en modelos multimodales actualizados.
- **Transformación de texto a audio (TTS):** el sistema convertirá de manera automática el texto generado en audio, permitiendo entregar descripciones auditivas accesibles.
- **Análisis visual de elementos relevantes de la obra:** el sistema será capaz de identificar composición, figuras, acciones y elementos destacados que aporten a

la comprensión de la obra.

- **Generación de descripciones objetivas y no interpretativas:** las descripciones generadas por el sistema tendrán un enfoque exclusivamente descriptivo, evitando juicios de valor o interpretaciones subjetivas.

1.5.2. Limitaciones

Pese a presentar ventajas, existen ciertas limitaciones:

- **Dependencia del rendimiento y los sesgos de los modelos multimodales:** la calidad de las descripciones generadas está condicionada por la capacidad del modelo seleccionado. Pueden existir imperfecciones, omisiones de detalles o sesgos en la interpretación visual.
- **Limitaciones en obras extremadamente abstractas o simbólicas:** el sistema genera descripciones basadas únicamente en lo observable; por ello, obras con escasa estructura visual, alto nivel de abstracción o fuerte carga simbólica pueden producir descripciones poco precisas o no completamente relacionadas con la intención artística.
- **No reemplaza a guías humanas ni descripciones profesionales:** si bien el sistema proporciona un apoyo automatizado, no iguala el nivel de interpretación o sensibilidad que pueden entregar especialistas en arte o mediadores culturales.

1.6. Enfoque Metodológico

1.6.1. Metodología empleada

El desarrollo del sistema se llevó a cabo con un enfoque permanente en la obtención de descripciones objetivas. El proceso abarcó los siguientes hitos clave:

1. **Recolección de imágenes:** se realizaron búsquedas de obras de arte considerando diversidad en estilos y técnicas. Finalmente, el conjunto de datos utilizado fue obtenido desde un repositorio [7] asociado al desarrollo de una GAN para arte, cuyo dataset proviene de WikiArt.
2. **Depuración y consolidación de imágenes:** el dataset fue sometido a un proceso de revisión manual para descartar obras excesivamente abstractas y seleccionar aquellas que permitieran un rendimiento adecuado tanto del módulo de generación de descripciones como del módulo de audio.
3. **Exploración y testeo de modelos multimodales:** se investigaron distintos modelos de lenguaje multimodales capaces de interpretar imágenes y generar descripciones textuales coherentes. Dado que cada modelo presenta métricas y características diferenciadas, la elección final se basó en criterios del equipo relacionados con la calidad descriptiva, la coherencia y el tiempo de respuesta.
4. **Diseño del pipeline y definición de prompts:** se construyó el flujo de procesamiento imagen → descripción, integrando diferentes modelos multimodales para pruebas comparativas. Asimismo, se diseñaron y evaluaron diversos prompts orientados a obtener descripciones estrictamente objetivas.
5. **Pruebas de modelos de traducción:** dado que las salidas iniciales de los modelos multimodales se generaron en inglés, fue necesario evaluar sistemas de traducción que garantizaran alta fidelidad semántica y claridad en español.
6. **Evaluación e integración de sistemas de texto a voz (TTS):** se investigaron diferentes motores de TTS con el fin de seleccionar el más eficiente, considerando la naturalidad, la claridad y la calidad de lectura. El sistema seleccionado fue incorporado al pipeline principal, estableciendo un flujo Imagen → Texto → Audio.
7. **Validación:** se generaron muestras de prueba para ser evaluadas por distintos

grupos de interés, con el objetivo de validar la calidad descriptiva y auditiva del sistema.

1.6.2. Herramientas y Tecnología

■ Lenguajes de programación:

- **Python 3.x:** utilizado como lenguaje principal para la implementación del pipeline, la automatización del flujo de trabajo, la integración con APIs de modelos y el desarrollo de pruebas experimentales.

■ Librerías y frameworks de Python:

- **Transformers:** empleada para el uso e implementación de modelos multi-modales basados en arquitecturas modernas, como Vision Transformers y LLM.
- **Torch:** base para la ejecución de modelos en local.
- **dotenv:** utilizada para gestionar variables de entorno, tales como claves de API e información sensible del proyecto.
- **Base64:** herramienta para codificar imágenes en formato Base64 y facilitar su envío mediante APIs.
- **OpenAI:** librería utilizada para realizar solicitudes a modelos alojados en la nube.
- **gTTS:** librería empleada para convertir texto traducido en audio y guardarlo en formato MP3.

■ Plataformas y proveedores de modelos:

- **Hugging Face:** repositorio y plataforma principal para la búsqueda, evaluación y selección de modelos de inteligencia artificial.

- **Nebius AI:** proveedor de servicios en la nube utilizado para la ejecución de modelos mediante APIs.
- **Control de versiones:**
 - **GitHub:** utilizado para el almacenamiento del código, el control de versiones, el seguimiento de cambios y la documentación del proyecto.
- **Entorno de desarrollo:**
 - **Jupyter:** entorno interactivo empleado para el prototipado, las pruebas y la visualización de resultados.

2. Estado del Arte

La accesibilidad a las artes visuales para personas ciegas o con baja visión ha sido abordada desde distintos enfoques tecnológicos, con el objetivo de reducir las barreras impuestas por la dependencia de la información visual en contextos culturales. Tradicionalmente, estas soluciones han incluido audioguías, descripciones verbales estáticas y representaciones táctiles, como gráficos en relieve o sistemas basados en Braille. Si bien estos mecanismos han permitido avances relevantes en materia de inclusión, diversos estudios coinciden en que presentan limitaciones significativas para transmitir escenas visuales complejas, relaciones espaciales, atmósferas y elementos estéticos propios de las obras de arte.

En este contexto, distintos trabajos han explorado alternativas que superan los enfoques tradicionales mediante la incorporación de múltiples modalidades sensoriales. "Accessible Visual Artworks for Blind and Visually Impaired People: Comparing a Multimodal Approach with Tactile Graphics" [8] analiza comparativamente herramientas clásicas de accesibilidad, como las tactile graphics, frente a propuestas multimodales que integran audio localizado, efectos sonoros y música ambiental. Los resultados evidencian que las soluciones multimodales permiten una mejor comprensión de las obras y una experiencia más rica para los usuarios, demostrando que el tacto o la descripción verbal aislada resultan insuficientes para comunicar información visual compleja. Este estudio establece un punto de inflexión al mostrar que la accesibilidad cultural se beneficia de enfoques que combinan distintas modalidades perceptivas.

Paralelamente, el desarrollo de la inteligencia artificial ha impulsado la creación de asistentes visuales capaces de interpretar imágenes y generar descripciones automáticas. "Vision-Language Models under Cultural and Inclusive Considerations" [9] propone un análisis de modelos visión-lenguaje bajo consideraciones culturales e inclusivas, evaluando su desempeño en la generación de descripciones visuales destinadas a personas con discapacidad visual. Los autores destacan tanto el potencial de estos modelos como sus limitaciones, particularmente en términos de alucinaciones, falta de precisión

y ausencia de control contextual, lo que resulta crítico en aplicaciones asistivas.

Complementariamente, en "Evaluating Multimodal Language Models as Visual Assistants for Visually Impaired Users" [1] se evalúa el uso de modelos de lenguaje multimodales como asistentes visuales en escenarios reales, utilizando imágenes y videos capturados por personas ciegas o con baja visión. Este trabajo identifica problemas recurrentes como descripciones genéricas, inconsistencias semánticas y deficiencias en la comprensión espacial. En la Figura 2.1 se sintetizan los principales tipos de tareas y problemáticas abordadas por los asistentes visuales evaluados en la literatura reciente, proporcionando una visión global de los desafíos técnicos que aún persisten en este tipo de sistemas. Estos resultados refuerzan la necesidad de diseñar soluciones que prioricen la precisión, la coherencia y el control del contenido generado.

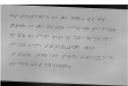
Task	Setting	Example	Model Response
Image Captioning	Original		General: A stork with Asian symbols on a silver coin. Caption: An elegant coin featuring a bird in flight, bearing inscriptions that hint at its cultural significance.
	Cultural		Cultural: The obverse side of a 500 won South Korean coin with a Red-crowned crane.
Image Question Answering	Original		EN Q: What is the title of this book? EN A: dog years (n=9), dog years: memoir (n=1) EN Answer: dog years
	Multilingual		EL Q: Ποιος είναι ο τίτλος αυτού του βιβλίου; EL A: Χρόνια σκύλου (n=9), Χρόνια σκύλου: απομνημονεύματα (n=1) EL Answer: Δεν μπορεί να απαντηθεί (Cannot be answered)
Optical Braille Recognition	Transcription		Transcription: The provided content appears to already be in English. If you are asking for an interpretation or explanation of the [...]
	Cross-script Question Answering		Q: Based on this text who had scored 60 goals in the season 1995-96? A: Jaromir Jagr and Mario Lemieux Answer: The provided content does not contain any information about goals scored in the 1995-96 season.
Video Object Recognition	General Objects		Q: What type of assistive measuring device appears in the video? Answer: Measuring cup
	Assistive Objects		A: Liquid level indicator
Video Question Answering	Descriptive Questions		Q: What colour are these sandals? A: Pink Descriptive Answer: Blue
	Spatial Questions		Q: Where are my keys? A: In corner of doormat next to slipper Spatial Answer: on mat
	Adversarial Questions		Q: What does this parcel say? A: Not enough information are depicted in video to answer this question Adversarial Answer: Amazon

Figura 2.1: Tipos de tareas y problemáticas abordadas por asistentes visuales para personas ciegas o con baja visión. Fuente: adaptado de [1].

Desde un enfoque más específico en la traducción multimodal, ^Aaudio description

3. Propuesta

3.1. Concepto General

El presente proyecto contempla el desarrollo de una plataforma integral orientada a mejorar la accesibilidad cultural a obras de arte visuales para personas con discapacidad visual, mediante el uso de tecnologías basadas en inteligencia artificial. La plataforma está concebida como un sistema modular que aborda distintas dimensiones de la experiencia artística, integrando mecanismos de interpretación visual, generación de contenido auditivo y provisión de interfaces accesibles para el usuario final.

Dentro de este contexto, la plataforma considera, entre otros componentes, un módulo de análisis visual y generación de descripciones textuales a partir de imágenes de obras de arte, un módulo de contextualización histórica y narrativa, un módulo de generación de paisajes sonoros y un módulo de interfaz web inclusiva que permite el acceso y consumo del contenido generado. Estos componentes trabajan de manera coordinada con el objetivo de facilitar la comprensión y apreciación de las obras, compensando las limitaciones que presentan los métodos tradicionales de accesibilidad al arte, tales como audioguías estáticas o descripciones manuales, las cuales suelen ser insuficientes para transmitir elementos visuales complejos como la composición, las atmósferas, la gestualidad, la profundidad y las relaciones espaciales.

En este marco general, el foco específico de la presente memoria se centra en el desarrollo del módulo de generación de audio descriptivo multimodal, encargado de transformar automáticamente imágenes de obras de arte en descripciones auditivas objetivas, detalladas y accesibles para personas con discapacidad visual. Este módulo integra modelos multimodales de análisis visual, generación y estructuración textual, y síntesis de voz, permitiendo producir descripciones auditivas claras, ricas en contenido y coherentes con la escena observada en la obra analizada.

Con el objetivo de presentar una visión global de la plataforma propuesta y contextualizar el módulo desarrollado dentro del sistema completo, se diseñó un diagrama de

CAPÍTULO 3 : PROPUESTA

arquitectura de alto nivel. En la Figura 3.1 se muestran los principales módulos de la plataforma y el flujo general de procesamiento desde la entrada de la imagen hasta la generación del contenido auditivo accesible.

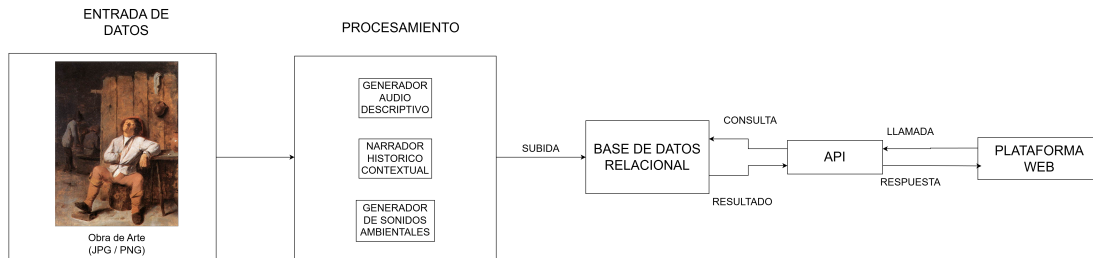


Figura 3.1: Diagrama general de la arquitectura del sistema propuesto.
Fuente: elaboración propia.

3.2. Descripción módulo

El módulo propuesto consiste en un flujo automático que recibe una imagen de una obra artística y produce como salida un archivo de audio en formato MP3 que describe su contenido visual en español. Este flujo incorpora los siguientes pasos:

1. **Entrada de la obra:** se recibe una imagen de una obra de arte (JPG/PNG).
2. **Análisis visual:** un modelo multimodal interpreta el contenido visual y genera una descripción detallada en inglés.
3. **Traducción automática:** un modelo de texto a texto traduce la descripción en inglés al español.
4. **Conversión de texto a voz:** se genera un audio en MP3 con la descripción en español.

3.2.1. Componentes Principales

El módulo de generación de audio descriptivo multimodal se construye a partir de los siguientes componentes tecnológicos. La selección de cada uno de ellos responde a criterios de desempeño, compatibilidad e integración dentro del pipeline del sistema, los cuales se analizan y justifican en subsecciones posteriores de esta memoria.

- **Qwen3-VL-Instruct:** modelo multimodal encargado de analizar la imagen y producir la descripción.
- **GPT-OSS:** modelo de lenguaje utilizado para la traducción automática al español.
- **gTTS (Google Text-to-Speech):** sistema TTS encargado de producir el audio MP3 con la descripción de la obra.
- **Script de procesamiento:** código en Python encargado de orquestar toda la comunicación entre los componentes.

3.2.2. Flujo de trabajo

Si bien el proyecto general contempla la integración de otros módulos, como lo son la generación de sonidos ambientales, la narración histórica contextual y una plataforma web inclusiva, el presente trabajo se focaliza exclusivamente en el módulo de generación de audio descriptivo multimodal.

El flujo de trabajo de este módulo se estructura en las siguientes etapas:

1. Carga de la imagen de la obra de arte.
2. Procesamiento visual y generación de la descripción textual mediante el modelo multimodal Qwen3-VL-Instruct.
3. Traducción automática de la descripción generada al idioma español.
4. Conversión de texto a voz y almacenamiento del audio resultante en formato MP3.

Dado que el alcance de esta memoria se centra específicamente en este módulo, se desarrolló un diagrama detallado que describe su funcionamiento interno. En la Figura 3.2 se presenta el flujo completo de procesamiento, desde la carga de la imagen hasta la generación del archivo de audio descriptivo en formato MP3.



Figura 3.2: Diagrama del módulo generador de descripciones auditivas multimodales. Fuente: elaboración propia.

3.2.3. Tecnología utilizadas

- **Python 3.x**
- **Huggingface Transformers** [10]
- **Nebius IA** [11]

3.3. Modelo de Generación de descripción

Uno de los ejes principales del módulo es el modelo de IA multimodal especializado en comprender imágenes y describirlas de forma objetiva y detallada.

3.3.1. Selección del Modelo

Se seleccionó el modelo **Qwen3-VL-Instruct** [12] como componente central del pipeline multimodal debido a su desempeño consistente en los siguientes aspectos:

- Alta precisión en la interpretación visual de escenas complejas.
- Capacidad robusta para seguir instrucciones estructuradas y formatos definidos.
- Coherencia semántica en descripciones extensas y detalladas.
- Compatibilidad directa con entornos Python y una relación costo–beneficio favorable en entornos de cómputo en la nube.

El modelo fue evaluado comparativamente frente a alternativas como BLIP-2 [13], LLaVA [14], Gemma [15], Llama 4 Maverick [16] y Llama 4 Scout [17]. La elección final no se basó en una métrica cuantitativa única, ya que cada modelo emplea criterios de validación propios para autoevaluar su rendimiento, lo que dificulta una comparación objetiva bajo una variable estandarizada común.

En consecuencia, se optó por una evaluación cualitativa comparativa, basada en la generación de descripciones para una misma obra de arte de referencia utilizando

CAPÍTULO 3 : PROPUESTA

cada uno de los modelos evaluados. Las descripciones generadas fueron analizadas por los integrantes del equipo de desarrollo, considerando criterios como la coherencia , precisión visual y la facilidad de uso. A partir de esta comparación directa, se determinó si las descripciones producidas por cada modelo resultaban válidas o no

Si bien los modelos de la familia Llama 4 y Qwen3 presentaron resultados altamente similares en términos de calidad descriptiva, la selección final de Qwen3-VL-Instruct respondió a criterios prácticos de implementación. En particular, se consideró su menor complejidad de integración y un costo operativo de API más favorable, factores determinantes para asegurar la viabilidad técnica y económica del sistema, dado que las diferencias perceptuales entre los modelos evaluados no resultaron significativas.

Con el fin de transparentar el proceso de selección y respaldar la evaluación cualitativa realizada, en el Anexo A se presentan ejemplos de generación producidos por cada uno de los modelos evaluados. Dicho anexo incluye descripciones generadas frente a una misma obra de referencia, lo que permite observar de manera directa diferencias en el nivel de detalle, la coherencia semántica, la estructura narrativa y la fidelidad visual entre los modelos analizados.

Con el objetivo de representar visualmente esta evaluación cualitativa, se construyó una comparación relativa entre los modelos evaluados, donde no se pretende reflejar métricas absolutas, sino una jerarquía de desempeño basada en la experiencia experimental del equipo de desarrollo.

Tabla 1: Comparación cualitativa de modelos Vision–Language evaluados

Modelo	Precisión visual	Coherencia	Facilidad de uso	Evaluación global
BLIP-2	★ ★ ○ ○ ○	★ ★ ○ ○ ○	★ ★ ★ ○ ○	★ ★ ○ ○ ○
LLaVa	★ ★ ○ ○ ○	★ ★ ○ ○ ○	★ ★ ★ ○ ○	★ ★ ○ ○ ○
Gemma	★ ★ ★ ○ ○	★ ★ ★ ○ ○	★ ★ ★ ○ ○	★ ★ ★ ○ ○
Llama 4 Maverick	★ ★ ★ ★ ○	★ ★ ★ ★ ○	★ ★ ★ ★ ★	★ ★ ★ ★ ○
Llama 4 Scout	★ ★ ★ ★ ○	★ ★ ★ ★ ○	★ ★ ★ ★ ★	★ ★ ★ ★ ○
Qwen3-VL-Instruct	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★

La progresión cualitativa observada puede resumirse de la siguiente forma:

BLIP-2 = LLaVa < Gemma < Llama 4 Maverick < Llama 4 Scout < **Qwen3-VL-Instruct**

Esta aproximación permitió justificar de manera técnica y metodológicamente consistente la selección del modelo final, alineando el desempeño observado con los requerimientos prácticos del sistema y las restricciones operativas del proyecto.

3.3.2. Proceso de Generación

Este proceso consta de dos partes. En primer lugar, se entrega la imagen al modelo multimodal, el cual genera una descripción siguiendo un *prompt* especializado:

```
"You are a visual analyst. Describe a single artwork objectively and clearly for visually
impaired audiences.\n"
"Return strict JSON with one field:\n"
'- "description": 3 to 6 sentences providing a coherent and unified visual description of the
```

```
artwork.\n'
"Include both the main scene (people, objects, actions, composition) and the visible environment
  (landscape, architecture, colors, lighting, textures) in a natural, continuous narrative.\n"
"Avoid mentioning the artist, title, historical period, or any information not directly visible
  in the image.\n"
"Focus on what can be *seen*, using neutral and descriptive language without interpretation or
  opinion."
```

El *prompt* instruye al modelo a actuar como un analista visual y a generar una descripción objetiva de una obra de arte, específicamente orientada a personas con discapacidad visual. La salida se solicita en formato JSON con un único campo "description", con el fin de estandarizar y simplificar el procesamiento posterior. Además, el *prompt* establece el rango de longitud esperado e indica explícitamente los tipos de detalles que deben incluirse, tales como la composición, los objetos presentes, los colores, la iluminación y los elementos del entorno. También enfatiza la necesidad de describir tanto la escena principal como todo lo visible en el contexto de la imagen, de manera natural y narrativamente continua. Asimismo, restringe cualquier inferencia o información que no provenga directamente de lo observable. Este enfoque estrictamente descriptivo garantiza neutralidad, accesibilidad y consistencia dentro del flujo multimodal del sistema.

3.3.3. Evaluación inicial

La calidad de lo generado se evaluó mediante:

- Revisión de coherencia.
- Iteración sobre el *prompt* para eliminar alucinaciones o errores.
- Comprobación de que lo generado es comprensible y claro.

Con lo anterior definido y puesto a prueba, el proceso de generación de descripciones mostró un comportamiento consistente y suficiente para su uso en una experiencia accesible.

3.4. Generación de Audio

El otro eje principal del módulo es la creación de un audio con la descripción de la obra de arte.

3.4.1. Selección de Modelo

Se selecciono **gTTS** debido a:

- Pronunciación clara.
- Soporte para español latino.
- Simplicidad de integración y eficiencia computacional.

Esta herramienta fue evaluada frente a alternativas como KokoroTTS, CoquiTTS y Facebook MMS-TTS.

3.4.2. Evaluación Inicial

Lo creado por esta herramienta tuvo en consideración las siguiente cosas:

- Claridad de la voz.
- Fluidez de lectura.
- Comprensibilidad sin ruido ni alteraciones.

3.5. Integración al Sistema

Aunque el módulo desarrollado se implementa actualmente como un script independiente en Python encargado de alimentar la base de datos de la página web descrita en el objetivo general, este puede integrarse como un componente nativo del propio sitio. El diseño del módulo ya contempla su incorporación en interfaces que permitan

cargar o seleccionar obras e imágenes, además de generar archivos de audio descriptivo en formato MP3, con la posibilidad de convertirlos a otros formatos según los requerimientos del sistema.

4. Validación

4.1. Objetivo de la Validación

El objetivo del proceso de validación fue evaluar de manera exploratoria el desempeño del módulo de generación de audio descriptivo multimodal, específicamente en términos de comprensibilidad del contenido auditivo generado, utilidad percibida como herramienta de accesibilidad y adecuación perceptual de los parámetros de narración, tales como la velocidad de lectura y la naturalidad de la voz sintetizada.

La validación se enfocó exclusivamente en el módulo desarrollado en esta memoria (imagen → descripción → audio), sin considerar los otros componentes del sistema global propuesto.

4.2. Ajuste Metodológico

El plan inicial de validación contemplaba la realización de una evaluación formal con un grupo objetivo perteneciente a una fundación de personas con discapacidad visual, asociada al cliente del proyecto. No obstante, debido a problemas de coordinación con el equipo y limitaciones logísticas, no fue posible concretar dicha instancia dentro del período de desarrollo de la memoria.

Frente a este escenario, se optó por implementar un mecanismo alternativo de validación exploratoria, orientado a obtener retroalimentación directa de usuarios reales con problemas de visibilidad, permitiendo así evaluar el uso práctico del módulo desarrollado en contextos cotidianos.

4.3. Participantes y Contexto de Evaluación

El proceso de validación se realizó mediante entrevistas breves a un total de 15 personas con problemas visuales, contactadas en distintos espacios públicos, tales como transporte público, centros de salud y centros comerciales.

Con el objetivo de asegurar una experiencia auditiva controlada y minimizar interferencias del entorno, las pruebas se llevaron a cabo utilizando audífonos con tecnología de cancelación activa de ruido. Cada participante escuchó una o más descripciones generadas por el sistema antes de responder el instrumento de evaluación.

4.4. Instrumento de Evaluación

El instrumento de evaluación consistió en un cuestionario breve compuesto por cuatro preguntas cerradas, con respuestas de tipo Sí/No, formuladas inmediatamente después de la experiencia auditiva:

- P1: ¿Se entendió el audio descriptivo generado?
- P2: ¿Le resulta útil una herramienta de este tipo para comprender obras de arte visuales?
- P3: ¿La velocidad de narración es adecuada?
- P4: ¿La voz utilizada resulta adecuada y agradable para la escucha?

En los casos en que las respuestas a P3 o P4 fueron negativas, se solicitó a los participantes indicar de manera verbal el motivo de su respuesta, con el fin de identificar oportunidades de mejora específicas.

4.5. Resultados de la Validación

La Tabla 2 presenta un resumen de los resultados obtenidos a partir de las entrevistas realizadas.

Tabla 2: Resultados de la validación exploratoria del módulo (n=15).

Pregunta	Sí	No	Observaciones
P1: ¿Se entendió el audio?	10	5	–
P2: ¿Le es útil una herramienta así?	8	7	–
P3: ¿La velocidad de narración es adecuada?	5	10	Velocidad percibida como excesiva
P4: ¿La voz es adecuada y agradable?	5	10	Voz percibida como poco amigable

En términos agregados, los resultados muestran que un 66,7 % de los participantes (10 de 15) comprendió el audio descriptivo (P1) y un 53,3 % (8 de 15) percibió utilidad en la herramienta (P2). Sin embargo, se observaron debilidades relevantes en parámetros perceptuales: solo un 33,3 % (5 de 15) consideró adecuada la velocidad de narración (P3) y la voz utilizada (P4), evidenciando una oportunidad de mejora en la síntesis de voz y en el control del ritmo de lectura.

4.6. Validación desde la reinterpretación visual

Con el fin de evaluar el grado en que las descripciones auditivas generadas por el sistema permiten construir una representación mental coherente de la obra original, se realizó una validación complementaria basada en un enfoque de reinterpretación visual. Este enfoque busca analizar indirectamente la fidelidad descriptiva del audio, evaluando qué tan consistente resulta la imagen mental generada en el oyente a partir de la narración.

El procedimiento consistió en solicitar a cuatro participantes, sin formación profesional en arte, que escucharan una descripción auditiva generada por el sistema y,

posteriormente, realizaran un dibujo libre que representara lo que comprendieron a partir del audio. Los participantes no tuvieron acceso a la imagen original ni a información adicional sobre la obra, y el ejercicio se realizó únicamente en base a la narración auditiva entregada por el sistema.

Este tipo de validación no busca evaluar habilidades artísticas ni precisión técnica en el dibujo, sino analizar la presencia de elementos visuales clave, tales como figuras principales, disposición espacial, acciones, objetos relevantes y características generales del entorno descrito. De esta forma, el dibujo actúa como una representación externa de la imagen mental construida por el oyente.

Los resultados obtenidos muestran que, en la mayoría de los casos, los participantes lograron representar componentes centrales de la escena descrita, manteniendo coherencia con la estructura general de la obra original. Se observaron coincidencias en la identificación de personajes, relaciones espaciales básicas y elementos del entorno, lo que sugiere que las descripciones auditivas generadas por el módulo entregan información suficiente para permitir una reinterpretación visual coherente, aun cuando esta no sea detallada ni artísticamente precisa.

Cabe destacar que este tipo de validación representa una perspectiva complementaria a la evaluación realizada con personas con discapacidad visual. Dado que la tarea de dibujar una imagen mental no resulta fácilmente realizable para personas ciegas o con baja visión, este enfoque permite evaluar, desde otra vereda, la capacidad del sistema para transmitir información visual de forma estructurada y comprensible, reforzando la validez del módulo desde un punto de vista cognitivo y comunicacional.

En conjunto, esta validación adicional refuerza los resultados obtenidos previamente, aportando evidencia cualitativa de que el módulo desarrollado no solo genera descripciones auditivas comprensibles, sino que también permite a los oyentes construir representaciones mentales coherentes de las obras de arte, lo cual constituye un objetivo fundamental en contextos de accesibilidad cultural.

Con el objetivo de ilustrar visualmente los resultados de este proceso de validación,

en la Figura 4.1 se presenta una comparación entre la obra de arte original utilizada como referencia y las reinterpretaciones visuales realizadas por los participantes a partir de la descripción auditiva generada por el sistema. Esta comparación permite observar de manera directa las similitudes conceptuales y estructurales entre la escena original y las representaciones producidas, facilitando el análisis cualitativo del grado de coherencia alcanzado por las descripciones auditivas.



(a) Obra de arte original utilizada como referencia.



(b) Reinterpretación visual – Participante 1.



(c) Participante 2.



(d) Participante 3.



(e) Participante 4.

Figura 4.1: Comparación entre la obra original y las reinterpretaciones visuales realizadas por participantes a partir de la descripción auditiva generada por el sistema.

5. Resultados

El desarrollo del módulo de generación de audio descriptivo multimodal permitió implementar un pipeline funcional capaz de transformar imágenes de obras de arte en

CAPÍTULO 5 : RESULTADOS

descripciones auditivas accesibles en idioma español. El sistema integra de manera automática el análisis visual mediante modelos multimodales, la traducción del contenido generado y la síntesis de voz, produciendo archivos de audio en formato MP3 listos para su integración en una plataforma web inclusiva.

Desde un punto de vista técnico, el módulo demostró ser estable durante su ejecución, permitiendo procesar múltiples obras de forma automatizada sin intervención manual adicional. Las descripciones generadas presentaron coherencia semántica, continuidad narrativa y un enfoque predominantemente descriptivo, alineado con los objetivos definidos en el diseño del sistema.

Los resultados de la validación exploratoria indican que el sistema logra cumplir parcialmente su propósito de mejorar la accesibilidad a obras de arte visuales, especialmente en lo relativo a la comprensión del contenido auditivo. No obstante, se identificaron limitaciones relevantes asociadas a la calidad perceptual de la narración, particularmente en la velocidad de lectura y la expresividad de la voz sintetizada.

En conjunto, los resultados obtenidos permiten concluir que el módulo desarrollado constituye una base funcional y técnicamente viable para la generación automática de descripciones auditivas de obras de arte. Asimismo, los hallazgos derivados del proceso de validación entregan lineamientos claros para futuras mejoras, orientadas principalmente a la personalización de la experiencia auditiva y a la incorporación de tecnologías de síntesis de voz más avanzadas.

De forma complementaria, la validación basada en reinterpretación visual permitió evaluar indirectamente la fidelidad descriptiva de las narraciones generadas. Los resultados cualitativos obtenidos sugieren que el módulo entrega información visual estructurada suficiente para que oyentes sin acceso a la imagen original puedan construir representaciones mentales coherentes de la escena descrita, evidenciando consistencia en elementos centrales, relaciones espaciales básicas y el contexto general de la obra.

6. Conclusiones

En el desarrollo de la presente memoria se abordó el desafío de mejorar la accesibilidad a las obras de arte visuales para personas con discapacidad visual mediante el uso de inteligencia artificial, proponiendo e implementando un módulo de generación de audio descriptivo multimodal capaz de transformar imágenes en descripciones auditivas objetivas y comprensibles. El trabajo se centró específicamente en el diseño, la implementación y la validación de dicho módulo, el cual constituye un componente fundamental dentro de una plataforma inclusiva de mayor alcance.

A partir del proceso de investigación, experimentación y desarrollo realizado, fue posible implementar un pipeline funcional que integra análisis visual mediante modelos multimodales, traducción automática y síntesis de voz, logrando generar archivos de audio en formato MP3 a partir de imágenes de obras de arte de manera completamente automatizada. El sistema desarrollado demostró estabilidad en su ejecución y coherencia en las descripciones generadas, cumpliendo con los requisitos definidos en la etapa de diseño, particularmente en lo relativo a la objetividad, claridad y estructuración narrativa del contenido auditivo.

La selección del modelo multimodal Qwen3-VL-Instruct, junto con el uso de modelos de traducción y un sistema de texto a voz ligero como gTTS, permitió construir una solución técnicamente viable y de fácil integración, validando la factibilidad de emplear modelos de lenguaje multimodales para aplicaciones de accesibilidad cultural. Asimismo, el diseño cuidadoso del *prompt* resultó clave para reducir interpretaciones subjetivas y alucinaciones, favoreciendo descripciones visuales centradas exclusivamente en lo observable.

El proceso de validación exploratoria, complementado con una evaluación de reinterpretación visual, evidenció que el módulo logra cumplir parcialmente su propósito, no solo en términos de comprensión auditiva y utilidad percibida, sino también en su capacidad para transmitir información visual de forma estructurada y cognitivamente interpretable. Los resultados obtenidos indican que una proporción significativa de los

CAPÍTULO 6 : CONCLUSIONES

participantes logró entender el contenido generado y percibió utilidad en la herramienta, lo que respalda la efectividad general del enfoque propuesto. No obstante, también se identificaron oportunidades de mejora relevantes, particularmente asociadas a la velocidad de narración y a la naturalidad de la voz sintetizada.

En conjunto, los resultados permiten concluir que el módulo de generación de audio descriptivo multimodal fue correctamente diseñado, implementado y validado dentro del contexto y alcance definidos para esta memoria. El sistema desarrollado constituye una base sólida y funcional para futuras extensiones, tanto en términos de mejora perceptual del audio como de integración con otros módulos de la plataforma global propuesta, tales como la narración histórica contextual, la generación de paisajes sonoros y las interfaces web inclusivas.

Finalmente, este trabajo aporta evidencia práctica sobre el potencial de la inteligencia artificial multimodal como herramienta para promover la accesibilidad cultural, demostrando que es posible automatizar la generación de descripciones auditivas de obras de arte de manera coherente y técnicamente robusta. A partir de los hallazgos obtenidos, se abren múltiples líneas de trabajo futuro orientadas a la personalización de la experiencia auditiva, la evaluación con grupos objetivo más amplios y la incorporación de tecnologías de síntesis de voz más avanzadas, con el fin de seguir fortaleciendo experiencias artísticas verdaderamente inclusivas.

7. Trabajo Futuro

Si bien el módulo de generación de audio descriptivo multimodal desarrollado en esta memoria cumple con los objetivos planteados y demuestra viabilidad técnica y funcional, existen diversas líneas de trabajo futuro que permitirían mejorar significativamente su desempeño, escalabilidad y calidad perceptual, fortaleciendo así su impacto como herramienta de accesibilidad cultural.

Una primera línea de trabajo futuro consiste en la incorporación de técnicas de *Retrieval-Augmented Generation* (RAG) como mecanismo para enriquecer las descripciones generadas sin necesidad de reentrenar el modelo multimodal base. Mediante este enfoque, el sistema podría consultar dinámicamente una base de conocimiento externa compuesta por obras de arte previamente etiquetadas, descripciones de referencia, guías curatoriales u ontologías visuales, las cuales actuarían como contexto adicional para el modelo en el momento de la generación. Esta estrategia permitiría mejorar la consistencia descriptiva, reducir ambigüedades y orientar la salida del modelo hacia estándares descriptivos más adecuados para personas con discapacidad visual, manteniendo al mismo tiempo la flexibilidad y eficiencia del enfoque basado en modelos preentrenados.

En relación con lo anterior, otra línea relevante de trabajo futuro corresponde a la construcción y utilización de un conjunto de datos especializado de obras de arte, etiquetado explícitamente con descripciones orientadas a la accesibilidad. Dicho dataset podría incluir anotaciones sobre composición, relaciones espaciales, elementos relevantes y niveles de detalle recomendados, sirviendo tanto como fuente de recuperación en un esquema RAG como guía semántica para evaluar y ajustar la calidad de las descripciones generadas. Esta aproximación permitiría establecer criterios más objetivos y reproducibles para la generación de descripciones auditivas accesibles.

Desde el punto de vista perceptual, una mejora sustancial del sistema podría lograrse mediante la sustitución del motor de texto a voz actual por tecnologías de síntesis de voz más avanzadas. La incorporación de modelos de TTS neuronales de última ge-

neración permitiría controlar parámetros como la entonación, el ritmo, las pausas y la expresividad, abordando directamente las limitaciones detectadas durante el proceso de validación exploratoria, particularmente en lo relativo a la velocidad de narración y a la naturalidad de la voz.

Asimismo, se plantea como línea futura la exploración de técnicas de clonación de voz o *voice cloning*, con el objetivo de generar narraciones con características vocales más humanas y cercanas. Este enfoque permitiría personalizar la experiencia auditiva, utilizando voces con mayor calidez, identidad o incluso asociadas a mediadores culturales, lo que podría mejorar la inmersión y la aceptación del sistema por parte de los usuarios. No obstante, este tipo de tecnologías debería ser abordado considerando aspectos éticos, legales y de consentimiento, especialmente en contextos culturales y educativos.

Finalmente, futuras extensiones del sistema podrían incluir la personalización de la experiencia auditiva según las preferencias del usuario, tales como el nivel de detalle descriptivo, la velocidad de narración o el tipo de voz, así como evaluaciones más extensas con grupos objetivo formales que permitan medir de manera cuantitativa el impacto del sistema en la comprensión y apreciación de obras de arte. Estas líneas de trabajo contribuirían a consolidar la plataforma propuesta como una solución robusta, escalable y centrada en el usuario para la accesibilidad cultural mediante inteligencia artificial.

Referencias

- [1] A. Karamolegkou, M. Nikandrou, G. Pantazopoulos, D. S. Villegas, P. Rust, R. Dhar, D. Hershcovich, and A. Søgaaard, “Evaluating multimodal language models as visual assistants for visually impaired users,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.22610>
- [2] H. Ning, X. Zheng, Y. Yuan, and X. Lu, “Audio description from image by modal translation network,” *Neurocomputing*, vol. 423, p. 124–134, Jan. 2021. [Online]. Available: <http://dx.doi.org/10.1016/j.neucom.2020.10.053>
- [3] M. de Desarrollo Social y Familia, “Encuesta de discapacidad y dependencia 2022,” 2023, accedido el 10 de noviembre de 2025. [Online]. Available: <https://observatorio.ministeriodesarrollosocial.gob.cl/endide-2022>
- [4] F. LUZ, “Estudio sobre personas con discapacidad visual en chile – fundación luz / endide 2022,” 2024, accedido el 10 de noviembre de 2025. [Online]. Available: <https://fluz.bitbanglab.cl/area/estudio-sobre-personas-con-discapacidad-visual-en-chile-fundacion-luz-endide-2022/>
- [5] ONU, “Declaración universal de los derechos humanos,” <https://www.un.org/es/about-us/universal-declaration-of-human-rights>.
- [6] —, “Convención sobre los derechos de las personas con discapacidad,” <https://www.un.org/esa/socdev/enable/documents/tccconvs.pdf>.
- [7] W. R. Tan, C. S. Chan, H. Aguirre, and K. Tanaka, “Improved artgan for conditional synthesis of natural image and artwork,” *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 394–409, 2019. [Online]. Available: <https://doi.org/10.1109/TIP.2018.2866698>
- [8] L. Cavazos Quero, J. Iranzo Bartolomé, and J. Cho, “Accessible visual artworks for blind and visually impaired people: Comparing a multimodal approach

with tactile graphics,” *Electronics*, vol. 10, no. 3, 2021. [Online]. Available: <https://www.mdpi.com/2079-9292/10/3/297>

- [9] A. Karamolegkou, P. Rust, Y. Cao, R. Cui, A. Søgaaard, and D. Hershcovich, “Vision-language models under cultural and inclusive considerations,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.06177>
- [10] Hugging Face, “Hugging face: The ai community building the future,” <https://huggingface.co/>, 2025, accedido: 01-dic-2025.
- [11] Nebius AI, “Nebius ai: Cloud infrastructure for artificial intelligence,” <https://nebius.com/>, 2025, accedido: 01-dic-2025.
- [12] Qwen Team, “Qwen3-vl-8b-instruct,” <https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct>, 2025, modelo multimodal visión-lenguaje. Accedido: 01-dic-2025.
- [13] Salesforce Research, “Blip-2 opt 6.7b,” <https://huggingface.co/Salesforce/blip2-opt-6.7b>, 2025, modelo de visión-lenguaje basado en BLIP-2. Accedido: 01-dic-2025.
- [14] Liu et al., “Llava v1.5 7b,” <https://huggingface.co/liuhaotian/llava-v1.5-7b>, 2025, large Language and Vision Assistant. Accedido: 01-dic-2025.
- [15] Google DeepMind, “Gemma 3n e4b instruction-tuned,” <https://huggingface.co/google/gemma-3n-E4B-it>, 2025, modelo de lenguaje instruccional. Accedido: 01-dic-2025.
- [16] Meta AI, “Llama 4 maverick 17b 128e instruct,” <https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct>, 2025, modelo de lenguaje de gran escala con expertos. Accedido: 01-dic-2025.

- [17] —, “Llama 4 scout 17b 16e instruct,” <https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct>, 2025, modelo optimizado para razonamiento eficiente. Accedido: 01-dic-2025.

Anexos

A. Resultados comparativos de modelos Vision–Language

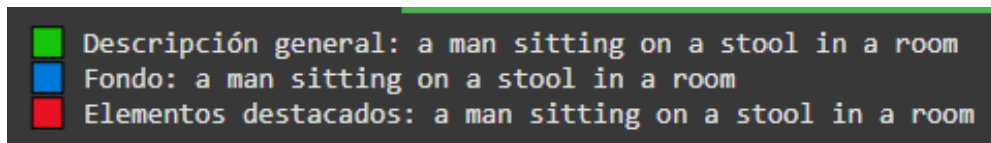


Figura A.1: Ejemplo de descripción generada por el modelo BLIP-2.

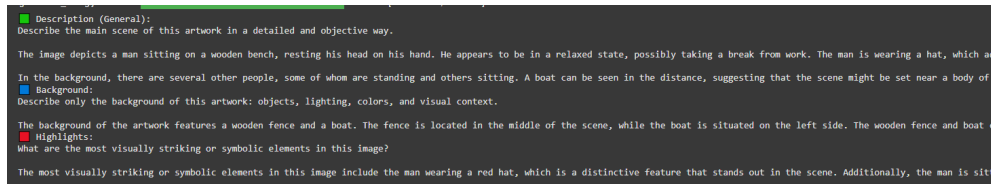


Figura A.2: Ejemplo de descripción generada por el modelo LLaVa.

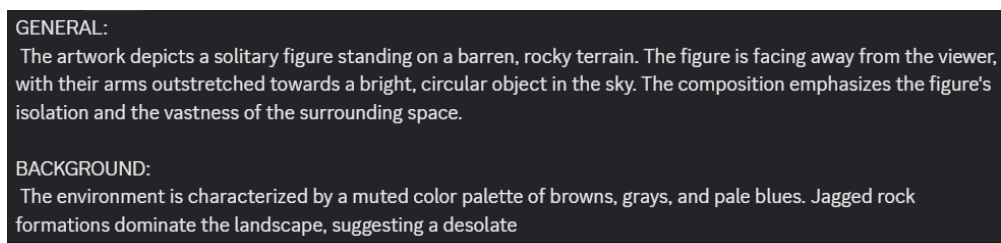


Figura A.3: Ejemplo de descripción generada por el modelo Gemma.

General:
A man sits on a wooden block, leaning against a wooden wall with his eyes closed. He wears a white shirt and brown pants. Two other people are visible in the background, one standing and one sitting.

Background:
The scene is set in a dimly lit room with wooden walls and a wooden floor. A table is partially visible to the right, and a hat hangs on the wall. (editado)

Figura A.4: Ejemplo de descripción generada por el modelo Llama 4 Maverick.

General:
The painting depicts a man sitting on a wooden block, leaning against a wooden door or wall. He is wearing a white shirt and tan pants, and appears to be resting or sleeping. In the background, two other men are visible, one standing and one sitting.

Background:
The environment appears to be a rustic, dimly lit interior with wooden walls and a concrete floor. The walls are made of vertical wooden planks, and a wooden table is partially visible to the right of the main figure. The overall atmosphere is somber and muted, with earthy tones dominating the scene.

Figura A.5: Ejemplo de descripción generada por el modelo Llama 4 Scout.

Una gran figura con aspecto de cabra se encuentra en el centro, adornada con cuernos y una corona de hojas, extendiendo su brazo hacia una mujer sentada que está amamantando a un niño. A su alrededor hay varias figuras vestidas con ropas sueltas y fluidas, algunas sentadas y otras de pie, realizando diversas actividades. La escena se desarrolla bajo un cielo nocturno oscuro iluminado por una luna creciente, con montañas lejanas y siluetas de criaturas voladoras que añaden profundidad al fondo.

Figura A.6: Ejemplo de descripción generada por el modelo Qwen3-VL-Instruct.