
This is the **published version** of the bachelor thesis:

Montes De Oca Martínez, Joel. *Sistema de inteligencia artificial para la generación y transformación visual basada en el análisis del estado anímico de melodías*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/334417>

under the terms of the  license.

Sistema de inteligencia artificial para la generación y transformación visual basada en el análisis del estado anímico de melodías

Joel Montes de Oca Martínez

Resumen— Este artículo presenta un sistema end-to-end que establece una correspondencia transmodal, traduciendo la atmósfera emocional de una pista de audio en secuencias visuales abstractas para promover la accesibilidad sensorial. El sistema aplica una arquitectura multimodal (CNN y Bi-GRU) para predecir niveles continuos de valencia y arousal. Estas coordenadas condicionan paramétricamente un motor de difusión temporal (AnimateDiff) mediante reglas neuroestéticas, generando variaciones volumétricas de una esfera 3D. La validación perceptiva con 63 usuarios demostró una precisión del 71,4% en la transmisión emocional, confirmando estadísticamente la capacidad del sistema para evocar de forma determinista la carga afectiva musical mediante el canal visual.

Palabras clave— Accesibilidad sensorial, inteligencia artificial (IA), modelos de difusión, reconocimiento de emociones musicales (MER), redes neuronales multimodales, sinestesia artificial.

Abstract— This paper presents an end-to-end system that establishes a cross-modal mapping, translating the emotional atmosphere of an audio track into abstract visual sequences to promote sensory accessibility. The system employs a multimodal architecture (CNN and Bi-GRU) to continuously predict acoustic valence and arousal levels. These coordinates parametrically condition a temporal diffusion engine (AnimateDiff) via neuroaesthetic rules, generating abstract 3D volumetric variations. Perceptual validation with 63 users demonstrated a 71.4% accuracy in emotional transmission, statistically confirming the system's ability to deterministically evoke the affective load of music exclusively through the visual channel.

Index Terms— Artificial intelligence (AI), artificial synesthesia, diffusion models, multimodal neural networks, music emotion recognition (MER), sensory accessibility.



1 INTRODUCCIÓN

La música funciona como un lenguaje universal capaz de transmitir, modular y canalizar todo un espectro de emociones [1]. Históricamente, las melodías han sido un medio importante para expresar lo que sentimos y transmitir emociones que nos permiten conectar con los demás [2]. Elementos estructurales como el ritmo y la armonía interactúan con nuestra psicología de una forma tan íntima que una simple nota puede trasladarnos a un recuerdo específico o cambiar por completo nuestro estado de ánimo en cuestión de segundos [1]. Cada pista de audio se convierte en un contenedor de vivencias y sentimientos que logran expresar y canalizar un enorme abanico de emociones tan complejas que las palabras no alcanzan a describir con tanta precisión.

Sin embargo, la experimentación y modulación emocional a través de la música está inseparablemente ligada al sentido del oído. Esto plantea una barrera de accesibilidad evidente para las personas con discapacidad auditiva, quienes no pueden escuchar las melodías y, en consecuencia, quedan excluidas de este proceso específico de modulación emocional acústica. Afortunadamente, el ser humano dispone de otros canales sensoriales y medios para procesar y representar las emociones. Los medios visuales poseen una capacidad igualmente profunda para provocar

estados anímicos y respuestas emocionales [3].

Ante esta realidad, este proyecto nace con el propósito de desarrollar una herramienta tecnológica de "sinestesia artificial" capaz de transmitir las emociones que genera una melodía a través de medios visuales.

Para fundamentar el proyecto, se ha realizado una reflexión previa y un análisis de fuentes. Se toman como referencia los estándares MER descritos por Gómez-Cañón et al. [1], el diseño de redes neuronales propuesto por Han et al. [4], y los recientes hallazgos de Mori y Zatorre [5] sobre la conectividad cerebral frente a recompensas musicales. También se han consultado los estudios cognitivos sobre correspondencia visual-emocional de Tinio y Gartsus [3] y la pragmática musical de Gomila y Christensen [2], ambos documentados en la colección Progress in Brain Research. Finalmente, el desarrollo se alinea con los protocolos estándar de MER dinámico descritos en la literatura.

2 Objetivos y Alcance del Proyecto

El objetivo principal del proyecto es diseñar y desarrollar un sistema de Inteligencia Artificial capaz de establecer un puente sensorial, analizando el estado anímico de una pista de audio para generar representaciones visuales que

evoquen las mismas emociones.

Para lograr el objetivo de manera clara, este se dividirá en dos subobjetivos específicos:

- **O1:** Dada una melodía, extraer de forma dinámica las emociones que transmite [1].
- **O2:** Dadas las emociones extraídas, generar una imagen (o secuencia visual) capaz de representar y evocar esas mismas emociones en el espectador [3].

3 Metodología y recursos

3.1 Metodología

Para alcanzar los objetivos propuestos en el tiempo exigido, la metodología utilizada está fundamentada en los principios del Desarrollo Ágil (Scrum adaptado a un desarrollador), utilizando un enfoque iterativo e incremental. Este procedimiento (trabajar en iteraciones o sprints cortos) permite construir y probar el sistema módulo a módulo (Unit Testing) antes de realizar la integración final del pipeline.

La planificación temporal detallada de estos bloques de trabajo, junto con la distribución de tareas e hitos a lo largo del proyecto, se encuentra explícitamente estructurada y reflejada en el Diagrama de Gantt (véase la **Figura A1** del Apéndice A).

3.2 Recursos

El desarrollo técnico del sistema ha requerido el uso de bases de datos públicas de audio (PMEmo) y acceso a capacidad de cómputo GPU para el entrenamiento de la red neuronal. De igual manera, la ejecución del motor visual exigió capacidad de cómputo con una GPU dedicada de alta VRAM para soportar el ecosistema nodal ComfyUI y AnimateDiff.

4 Métodos

De cara al desarrollo, la implementación del sistema requiere una arquitectura secuencial y multimodal, dividida en módulos interconectados (**Figura 1**) detallados a continuación.

M1 – Análisis de Audio (MER) [1]: Se ha adoptado un enfoque de doble vía utilizando la librería Librosa en Python para generar Espectrogramas Mel (representación visual de frecuencias) y el motor OpenSMILE [6] para la extracción de descriptores acústicos de bajo nivel. El núcleo del sistema consiste en una red neuronal híbrida basada en una arquitectura de Fusión Intermedia a nivel de características (*Feature-Level Fusion*). Este diseño utiliza una CNN para la extracción de características espaciales del espectrograma, y capas de proyección (reducción de dimensionalidad) para ambas ramas, logrando que los dos flujos se unifiquen equilibradamente en un espacio latente común antes de ser procesados por una RNN (Bi-GRU) encargada de captar la evolución temporal de la emoción musical [4]. Los resultados obtenidos constituyen los parámetros dinámicos basados en el Modelo Circumplex continuo (Valencia y Arousal) [7].

M2 – Traducción a texto (Emotion to Text): Algoritmo basado en reglas lógicas neuroestéticas y correspondencias crossmodales. Su función es mapear las coordenadas continuas de la Bi-GRU y traducirlas dinámicamente a descripciones textuales (prompts) en inglés, desacoplando la semántica en baja frecuencia (2.0s) y la cinemática en alta frecuencia (0.5s) [3].

M3 – Generación Visual (Emotion to Image): Renderizado y difusión temporal adaptativa implementado localmente sobre el ecosistema nodal de ComfyUI mediante el motor AnimateDiff-Evolved [8]. Utiliza el modelo base parametrizado DreamShaper (SD 1.5) para generar variaciones geométricas temporales sobre una representación volumétrica abstracta, guiado por los prompts y velocidades del módulo M2.

M4 – Pipeline de Integración: unificación del flujo de datos (Audio → Parámetros → Texto → Imagen), convirtiendo el sistema en un proceso *End-to-end*.

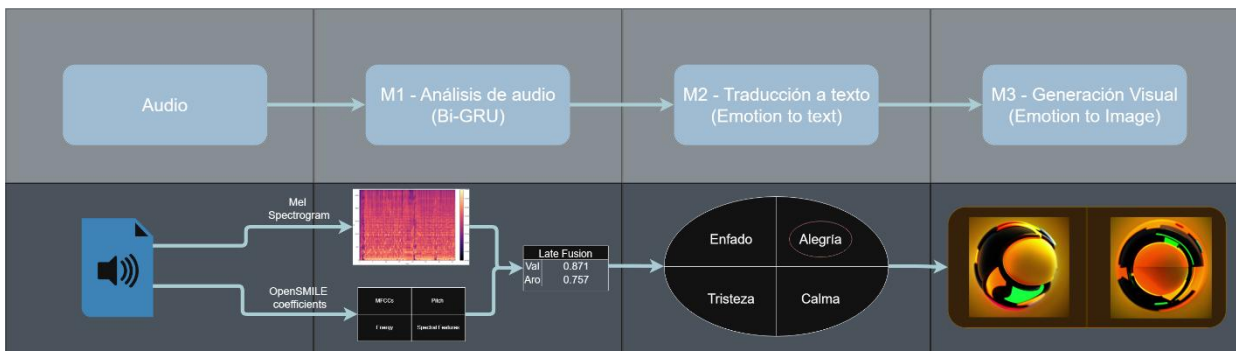


Fig. 1. Arquitectura del pipeline End-to-End. Fila superior: Diagrama de bloques de los módulos M1 a M4. Fila inferior: Trazabilidad real de los datos (Pista de audio → Fusión Multimodal → Traducción a parámetros → Fotogramas generados).

4.1 Módulo M1: Análisis de audio

4.1.1 Dataset y Preprocesamiento de Señal

Para el entrenamiento del modelo se ha utilizado el dataset PMEmo [9], compuesto por 794 fragmentos musicales con anotaciones dinámicas de Arousal (Energía) y Valencia (Positividad) recogidas en tiempo real.

Para extraer la información acústica, se ha implementado un *pipeline* de doble vía (*Dual-Stream*):

- **Rama Espectro-Visual:** Se ha reducido la frecuencia de muestreo de las pistas a 22050 Hz y se ha aplicado la Transformada de Fourier a Corto Plazo (STFT) para extraer Espectrogramas Mel de 128 bandas. Los decibelios se han normalizado mediante un escalado Min-Max (0 a 1). Por motivos de optimización, estos espectrogramas se han precomputado y almacenado localmente como tensores multidimensionales, eliminando el cuello de botella de procesamiento de la CPU durante el entrenamiento.
- **Rama Acústica:** Mediante la librería OpenSMILE (conjunto IS13_ComParE), se han extraído 65 descriptores de bajo nivel (como MFCCs, Zero-Crossing Rate y F0) junto con sus primeras derivadas temporales (*deltas*), obteniendo 130 características dinámicas. A continuación, para cada ventana temporal de 0.5 segundos, se ha calculado la media estadística (130 variables) y la desviación estándar (130 variables), sumando el total definitivo de 260 características acústicas por *step* (paso temporal). Finalmente, se ha aplicado normalización Z-Score utilizando estrictamente los parámetros del conjunto de entrenamiento para evitar el *Data Leakage*.

Para adaptar estos datos a la entrada de la red y aumentar el volumen de entrenamiento (*Data Augmentation*), se ha implementado una técnica de Ventana Deslizante (*Sliding Window*). Las pistas se han segmentado en ventanas de 30 segundos (60 *steps* temporales de 0.5s). Para el conjunto de entrenamiento, la ventana avanza con un solapamiento del 50% (15 segundos), multiplicando la cantidad de datos disponibles. Sin embargo, para el conjunto de validación, el solapamiento se eliminó para garantizar una evaluación justa y sin redundancias.

Adicionalmente, la división del dataset (80% entrenamiento, 20% validación) se ha realizado agrupando estrictamente por el identificador único de la canción (*musicId*), asegurando que ningún fragmento de una misma pista se filtre entre ambos conjuntos (*Data Leakage*). Matemáticamente, esto transforma la entrada en dos tensores por fragmento: un tensor visual de dimensiones [1, 60, 1, 128, 22] (Batch, Tiempo, Canal, Alto, Ancho) y un tensor acústico de dimensiones [1, 60, 260] (Batch, Tiempo, Características).

Además, se ha detectado que las anotaciones humanas no cubrían la totalidad del audio, lo que provocaba errores

en la función de pérdida. Esto se ha solucionado implementando una máscara dinámica que ignora los fragmentos vacíos durante el cálculo del error para poder aprovechar esta información.

4.1.2 Red neuronal híbrida

Tomando como referencia la capacidad extractiva de las redes híbridas [4], se ha diseñado una arquitectura multimodal. Inicialmente se planteó una Fusión Temprana (*Early Fusion*) concatenando los datos crudos, pero la experimentación ha demostrado que la gran dimensionalidad de la imagen (2048 características) provoca que la red ignore por completo los datos numéricos de audio (260 características).

Para solucionar este desbalance, se ha reestructurado el modelo hacia una Fusión Intermedia a nivel de características (*Feature-Level Fusion*) [10]. A diferencia de una Fusión Tardía pura (*Decision-Level*) donde se promedian las predicciones finales de redes separadas, este enfoque procesa la visión y el audio en sub-redes independientes hasta extraer *embeddings* de 128 dimensiones en cada sub-red. Tras este proceso, ambos vectores se concatenan y se introducen en las capas recurrentes (Bi-GRU) y de clasificación final encargadas de calcular la predicción. Para prevenir el sobreajuste (*overfitting*), se han implementado capas de Dropout en los reductores (20%) y en el clasificador final (30%), desactivando neuronas aleatoriamente para forzar a la red a aprender patrones más robustos.

1. **Extracción espacial (Inception Residual):** El espectrograma pasa por una CNN con tres bloques Inception. El uso de convoluciones paralelas (3x3 y 5x5) extrae simultáneamente detalles finos (como percusión) y globales (como armonía). Tras detectar curvas de aprendizaje erróneas en las primeras pruebas, se han añadido capas de Batch Normalization tras cada convolución. Aunque clásicamente la eficacia de esta técnica se atribuía a la reducción del *internal covariate shift* (cambio de covariable interna) [11], investigaciones recientes [12] han demostrado empíricamente que su verdadero impacto radica en el suavizado topológico de la superficie de pérdida (*Loss Landscape Smoothing*). Al mejorar la continuidad de Lipschitz de los gradientes, la red se vuelve significativamente más estable frente a variaciones en la inicialización de los pesos, permitiendo utilizar tasas de aprendizaje (*Learning Rates*) más agresivas sin que el optimizador diverja.
2. **Espacio latente equilibrado:** Las salidas de la CNN y de OpenSMILE atraviesan redes reductoras independientes que las comprimen a un espacio latente de exactamente 128 dimensiones cada una. Este paso asegura que ambas modalidades tengan el mismo peso al fusionarse.
3. **Módulo de memoria temporal (Bi-GRU):** Los vectores equilibrados se concatenan (256 dimensiones) y alimentan una Bi-GRU (Gated Recurrent Unit Bidi-

reccional) de dos capas. La naturaleza bidireccional le permite analizar el contexto en ambos sentidos temporales, anticipando transiciones estructurales y generando trayectorias más suaves.

4. **Clasificación final (Sigmoide):** Las representaciones temporales de la Bi-GRU atraviesan un módulo clasificador final compuesto por dos capas densas (Fully Connected). La primera actúa como capa oculta (reduciendo la información a 64 neuronas con activación ReLU y *Dropout* del 30%), mientras que la segunda proyecta el resultado hacia dos únicas neuronas de salida (Arousal y Valencia). Finalmente, se aplica una función de activación matemática Sigmoide (*torch.sigmoid*) que comprime y acota los resultados estrictamente en un rango continuo de [0, 1], coincidiendo exactamente con la escala de anotaciones del dataset PMemo.

En cuanto a carga computacional y optimización, se ha hecho uso de Precisión Mixta Automática (AMP) en GPU, calculando los gradientes en 16 bits sin pérdida de precisión, para evitar desbordamientos de memoria y reducir la carga computacional del entrenamiento del modelo. Adicionalmente, para evitar cuellos de botella en la transferencia de datos hacia la tarjeta gráfica, se han configurado los DataLoaders con memoria anclada (*pin_memory*) y multi-procesamiento persistente, activando además la implementación matemática fusionada (*fused=True*) del optimizador Adam [13].

Respecto a los parámetros utilizados se ha implementado un Grid Search iterativo para la búsqueda de hiperparámetros, determinando la convergencia óptima del *Batch Size* y el *Learning Rate*. Asimismo, se ha integrado un optimizador Adam [13] junto al planificador ReduceLROnPlateau. Este planificador está configurado para reducir la tasa de aprendizaje a la mitad si la red no logra mejorar su métrica de validación durante 3 épocas consecutivas (*patience=3*), evitando así el estancamiento en mínimos locales. El proceso de entrenamiento se ha configurado para ejecutarse durante un máximo de 30 épocas, implementando un sistema de guardado automático (Model Checkpointing) que conserva exclusivamente los pesos matemáticos de la iteración que obtiene el menor error de validación.

4.1.3 Criterios y Métricas de Evaluación

Durante la implementación, se ha determinado que la naturaleza continua del modelo requiere una evaluación en dos niveles. A nivel técnico-matemático, se ha establecido la minimización del Error Cuadrático Medio (MSE) como métrica y función de pérdida principal para el entrenamiento de la red. No obstante, para garantizar la interpretabilidad de los resultados de cara a la posterior generación visual (módulos M2 y M3), se ha introducido una métrica funcional complementaria: la Precisión por Cuadrantes (Accuracy) basada en el Modelo Circunflejo de Russell [7]. Este enfoque de evaluación asegura que el sistema identifique correctamente la "atmósfera general" (alegre, melan-

cólica, tensa o relajada), mitigando el impacto visual que podrían tener las pequeñas desviaciones numéricas inherentes a la regresión continua.

4.2 Módulo M2: Traducción a texto (Emotion to text)

Para poder interpretar los resultados de los tensores numéricos de la Bi-GRU (M1) y traducirlos a instrucciones comprensibles para el motor de Inteligencia Artificial visual (M3), se ha desarrollado un módulo intermedio en Python (*TraductorSinestesico*).

Variable de control: Esfera de Fluidos 3D

Como decisión de diseño fundamental, se ha establecido que la representación visual base sea siempre una "esfera de fluidos 3D". Utilizar una geometría volumétrica abstracta como sujeto principal ofrece dos ventajas críticas: **previene la pareidolia** (evitando que el modelo de difusión alucine objetos o rostros aleatorios que distraigan de la carga afectiva transmitida) y proporciona una **medición clara de los estados físicos**, haciendo evidentes las transiciones morfológicas desde una superficie perfectamente lisa hasta la extrusión de picos afilados y caóticos.

A partir de esta esfera base, el script en Python lee las coordenadas (Valencia y Arousal) del modelo Circumplex extraídas en el M1 y las discretiza apoyándose en dos teorías clave de la neuroestética. La correspondencia de propiedades visuales se realiza de la siguiente manera:

- **Maapeo de la Valencia (Eje X - Positividad/Negatividad):** Guiado por la Teoría de la Fluidez Cognitiva [14] y la Valencia Ecológica [15]. Esta variable determina la colorimetría, la saturación, la iluminación y la simetría. Los estados de ánimo positivos se asocian con luz y colores cálidos, mientras que los negativos se asocian a oscuridad y asimetría.
- **Maapeo del Arousal (Eje Y - Energía):** Guiado por el Efecto Kiki-Bouba [16] y la respuesta de la amígdala [17]. Esta variable regula la deformación geométrica, la velocidad y la frecuencia espacial. La alta energía o tensión provoca formas afiladas y amenazantes (Kiki), mientras que la calma invoca curvas y formas redondeadas inofensivas (Bouba).

Implementando estas teorías mediante un sistema de parametrización lógica, el sistema clasifica los valores en cuatro cuadrantes y exporta cadenas de texto descriptivo (prompts) diseñadas matemáticamente para guiar al modelo visual. La **Figura 2** detalla gráficamente la arquitectura de decisión neuroestética de este maapeo, mostrando cómo las coordenadas continuas se traducen a texto y simbólica en cada uno de los cuatro cuadrantes (Q1 - Q4).

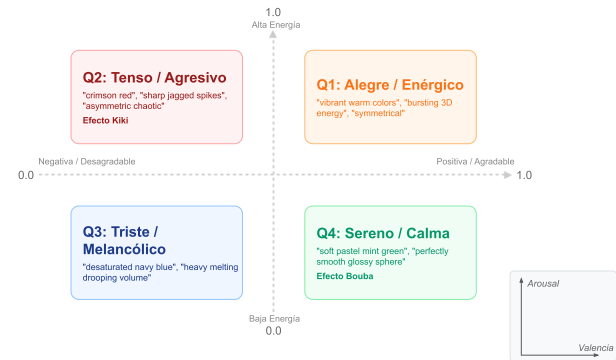


Fig. 2. Arquitectura lógica del Traductor Sinestésico (M2). Mapeo de coordenadas continuas a descriptores textuales basados en el Modelo Circumplex de Russell y el efecto Kiki-Bouba.

- Q1 - Alegre / Enérgico (Alta Valencia, Alto Arousal):** El texto exportado (*"vibrant warm colors, neon orange, symmetrical bursting energy"*) ordena al motor generar colores cálidos y muy luminosos, con una geometría simétrica y ascendente, reflejando euforia.
- Q2 - Tenso / Agresivo (Baja Valencia, Alto Arousal):** El prompt (*"dark crimson red, pitch black, sharp jagged spikes"*) genera contrastes y texturas ruidosas. La geometría de la esfera muta hacia formas puntiagudas, asimétricas y afiladas (Efecto Kiki), transmitiendo amenaza o tensión.
- Q3 - Melancólico (Baja Valencia, Bajo Arousal):** Se exportan descripciones (*"desaturated cold navy blue, melting drooping volume, low contrast"*) para crear colores fríos desaturados y una iluminación tenue. El fluido pierde su tensión superficial y se derrite o desciende lentamente, reflejando decaimiento anímico.
- Q4 - Sereno / Calmado (Alta Valencia, Bajo Arousal):** El prompt (*"soft pastel mint green, perfectly smooth glossy sphere, diffuse ambient lighting"*) invoca tonos pastel suaves y mucha luminosidad. La geometría retorna a ser una esfera de cristal perfectamente lisa y simétrica (Efecto Bouba), evocando descanso.

Arquitectura temporal desacoplada para control semántico-cinemático

El mayor reto de este módulo es evitar que el modelo colapse por "sobrecarga semántica" (recibir demasiadas instrucciones de cambio en muy poco tiempo). Para solventarlo, se implementó una estrategia que exporta las variables hacia ComfyUI (M3) divididas en dos frecuencias temporales distintas:

- Baja frecuencia (Estética):** El texto que dicta el color y la forma base (los prompts de los cuadrantes Q1-Q4) se actualiza lentamente, en ventanas de 2.0 segundos (cada 30 fotogramas). El código Python

empaqueta estos datos en formato clave-valor (ej. "0": "prompt_Q4", "30": "prompt_Q2"). Esto proporciona a la red de difusión el tiempo de cálculo necesario para realizar interpolaciones suaves entre dos colores o formas distintas (Prompt Travel), evitando parpadeos bruscos.

- Alta frecuencia (Cinemática):** En contraste, el nivel de Arousal (la agitación de la melodía) debe tener una respuesta visual inmediata. Por ello, se mapea matemáticamente a un multiplicador de velocidad dinámico (*motion_scale*). Utilizando una función de escalado, el rango original del Arousal [0 a 1] se traduce en un rango de velocidad seguro [0.5 a 1.5]. El script exporta estos valores cada 0.5 segundos. Esto logra que, si hay un golpe fuerte en la melodía, la agitación del fluido acelere en pantalla sin alterar de forma drástica el proceso estético que rige los colores.

f_semantic = 0.5 Hz ; f_motion = 2 Hz

Donde *fsemantic* representa la frecuencia de actualización de las instrucciones semánticas globales (prompts de cuadrante) y *fmotion* la frecuencia de actualización de la modulación cinemática (*motion_scale*).

motion_scale = 0.5 + arousal

Esta función transforma el rango emocional continuo del arousal [0,1] a un rango de velocidad seguro [0.5,1.5], determinado empíricamente mediante estudios de ablación para evitar colapsos geométricos o desintegración del volumen en AnimateDiff.

4.3 Módulo M3: Generación visual (Emotion to image)

El motor de renderizado visual se construyó sobre la infraestructura nodal ComfyUI, utilizando como motor temporal el ecosistema de librerías AnimateDiff-Evolved [8]. Para garantizar un procesamiento modular, comprensible y escalable, la enorme complejidad de la red se ha encapsulado en cinco submódulos interconectados:

- Motor de Movimiento (AnimateDiff Loader & Context):** Este submódulo capacita a la red de consciencia temporal. Para lograrlo, carga el módulo de movimiento específico *mm_sd_v15_v2.ckpt*, una red neuronal secundaria pre-entrenada con millones de secuencias de vídeo que inyecta los "priors temporales" (las reglas físicas del movimiento) en el modelo base. Además, implementa la técnica de "Ventana Deslizante" (Context Options), obligando al sistema a renderizar el vídeo en lotes de 16 fotogramas solapados entre sí. Esta memoria compartida es lo que permite que el movimiento entre segundos sea fluido y tenga una continuidad lógica sin desbordar la memoria VRAM.
- Inyección Semántica (Prompt Scheduling & Noise):** Este submódulo actúa como el director conceptual del KSampler, encargándose de recibir y estructurar las instrucciones de Python (M2) a través de tres compo-

nentes críticos:

- **Prompt Positivo Base (Prepend Text):** Cadena de texto fija que actúa como base de estilo para todo el vídeo. Se ha establecido el uso del prompt base: "*masterpiece, ultra high resolution, 3D fluid dynamics, cinematic lighting, photorealistic material, studio volumetric shadows*". Esto obliga a la IA a mantener el fotorrealismo independientemente de los cambios de color emocionales.
- **Prompt Negativo Base (Restricciones):** Matriz de exclusión matemática encargada de prohibir elementos indeseados mediante un vector de penalización con peso reforzado de 1.5: "*eye, face, human features, monster, text, watermark, flat 2d, low quality:1.5*". Esta barrera restringe el espacio latente para evitar fugas semánticas (pareidolia) o el aplanamiento de la imagen.
- **Optimización de Ruido via FreeNoise:** Para que un modelo genere un vídeo, parte siempre de una "matriz de ruido base", que se puede entender de forma sencilla como un lienzo inicial de estática blanca a partir del cual el modelo va despejando píxeles. En vídeos largos, la IA tiende a cambiar este lienzo aleatoriamente en cada frame, destruyendo la forma del objeto. El algoritmo FreeNoise soluciona esto reciclando e interconectando matemáticamente los mismos patrones de estática inicial a lo largo del tiempo, garantizando la "permanencia del objeto" para que la esfera de fluido no se desintegre.

Esta separación entre prompt estructural fijo y modulación emocional dinámica constituye un mecanismo de estabilización semántica. Mientras el prompt base mantiene constantes las propiedades tridimensionales fundamentales del objeto (volumen, iluminación y materialidad), el Módulo 2 únicamente modifica atributos emocionales específicos como color, tensión geométrica o contraste perceptual.

3. Núcleo de Generación (Modelos y KSampler):

Constituye el centro de procesamiento matemático de la arquitectura.

- **Checkpoint (DreamShaper v8):** Actúa como el "cerebro visual" o la base de conocimiento principal del sistema (construido sobre la arquitectura de Stable Diffusion 1.5). Contiene los pesos matemáticos pre-entrenados que dictan cómo la red comprende y representa los conceptos. Se ha seleccionado específicamente este modelo frente a otras alternativas por su capacidad para interpretar la refracción de la luz (raytracing), generar sombras complejas y modelar texturas fotorrealistas. Esta especialización tridimensional es perfecta para modelar el volumen definido para el proyecto, evitando la generación de ilustraciones planas o estilos de dibujo animados.

- **KSampler:** Es el algoritmo central de difusión (configurado con DPM++ 2M Karras). Su función es tomar el lienzo inicial lleno de ruido y "limpiarlo" de forma progresiva iteración a iteración (steps). Para saber qué formas o colores debe despejar entre todo ese ruido, utiliza como guía las instrucciones de condicionamiento de texto (prompts positivos y negativos) recibidas del submódulo semántico.
 - **Decodificador VAE Externo (ft-mse):** El Variational Autoencoder (VAE) ejecuta el paso final: traducir la compleja matriz de tensores (matemática pura) del espacio latente a los píxeles RGB que compondrán la imagen y posteriormente el vídeo visible. Se ha forzado el uso de este decodificador externo (*ft-mse-840000-ema-pruned*) porque los modelos por defecto tienden a generar salidas grisáceas. Su uso garantiza una decodificación en Alto Rango Dinámico (HDR), asegurando que el contraste y la saturación requeridos por el Módulo 2 se visualicen por pantalla con viveza.
4. **Post-procesamiento (Upscale & RIFE):** Submódulo encargado de elevar la calidad del vídeo crudo (512x512 px a 15 fps) mediante un escalado espacial (4x-UltraSharp) acotado a 1024x1024 px para mitigar el consumo de VRAM, seguido de una interpolación temporal a 30 fps mediante estimación de movimiento basada en RIFE VFI. Dado que los algoritmos convencionales evalúan vectores de movimiento entre imágenes consecutivas generando exactamente $N - 1$ intervalos para un conjunto de N fotogramas base ($N_{final} = N + (N - 1) \cdot (M - 1)$, donde M es un factor multiplicador), se produce un desajuste temporal con la pista de audio original. Para resolver esta asincronía arquitectónica, el pipeline se configuró para generar siempre una matriz base de $N+1$ fotogramas, logrando que el número de intervalos intermedios sea exactamente N y garantizando sincronización con la melodía final.

En el caso del proyecto, provoca un desajuste con la pista de audio analizada ya que no cuadra la duración final del vídeo con la de la melodía. Para resolver el error se ha ajustado la arquitectura para generar siempre una matriz base de $N + 1$ fotogramas, garantizando que el número de intervalos intermedios sea exactamente N . Logrando eliminar así el desfase temporal con las longitudes de audio.

5. **Ensamblado (Video Combine):** Subsistema final de exportación que compila la matriz de imágenes procesadas en alta resolución, codificándolas y exportándolas como un archivo de vídeo estándar (H264-MP4), ajustado a la duración requerida.

La representación del grafo de nodos implementado en ComfyUI puede consultarse en el apartado **A2 del Apéndice A**.

4.4 Módulo M4: Pipeline de Integración

Para consolidar el sistema en una solución End-to-End, se ha implementado un script orquestador en Python que unifica los tres módulos anteriormente descritos. Este proceso automatiza la entrada y segmentación de la pista de audio, invoca a la red Bi-GRU (M1) para inferir la matriz continua de emociones y delega en el Traductor Sinestésico (M2) la generación algorítmica de los prompts. Finalmente, el orquestador inyecta dinámicamente estos valores (semántica de baja frecuencia y cinemática de alta frecuencia) sobrescribiendo el archivo JSON de configuración del workflow. Este paquete de instrucciones se envía mediante una petición HTTP asíncrona a la API del servidor local de ComfyUI (M3), el cual opera en segundo plano y renderiza el archivo de vídeo MP4 final directamente en el espacio de almacenamiento.

Previo a la validación con usuarios, se ejecutó un test de variabilidad algorítmica (Batch Processing) sobre el sistema integrado. Para cada una de las melodías de control, el pipeline se iteró múltiples veces modificando exclusivamente la semilla de ruido aleatorio (seed) inicial del modelo de difusión. El objetivo de esta prueba de estrés fue garantizar que, aunque la red generara variaciones visuales fluidas y creativamente únicas en cada iteración, la inyección paramétrica del Traductor Sinestésico forzaba a respetar siempre las fronteras de color y cinemática impuestas. Esta fase corroboró que el comportamiento del sistema es determinista y no depende del azar de la generación.

Una vez garantizada la viabilidad técnica del pipeline, se diseñó un estudio perceptivo crossmodal para validar la coherencia sinestésica de las secuencias generadas por el sistema. El instrumento de evaluación se fundamentó en una adaptación de la escala psicológica SAM (Self-Assessment Manikin) [18], consolidada como el estándar para la medición de respuestas afectivas bajo las dimensiones del Modelo Circumplejo. El diseño consistió en un cuestionario anónimo estructurado en el que los participantes fueron expuestos a la visualización en absoluto silencio de cuatro secuencias de vídeo. Estos clips fueron generados por el pipeline a partir de cuatro melodías de seleccionadas específicamente para abarcar de manera equitativa los cuatro cuadrantes neuroestéticos (Alegria, Tension, Tristeza y Serenidad). Tras cada visualización, se solicitó a los usuarios que evaluaran su percepción subjetiva de la energía visual (Arousal) y la positividad estética (Valencia) empleando una escala psicométrica Likert de 5 puntos. Posteriormente, para habilitar un análisis estadístico riguroso y permitir la comparativa directa con las inferencias de la red neuronal, las respuestas recopiladas se sometieron a un escalado lineal Min-Max, mapeando matemáticamente los valores discretos de la escala Likert al espacio coordenado continuo y acotado [0, 1] que utiliza el espacio latente del modelo predictivo.

5 Resultados

Los experimentos realizados muestran resultados que alcanzan el objetivo definido para la viabilidad del sistema.

Como meta principal del proyecto, la orquestación de los módulos a través del pipeline End-to-End (M4) ha demostrado su capacidad técnica para generar transformaciones visuales y geométricas coherentes con las instrucciones dictadas por el Traductor Sinestésico. La **Figura 3** muestra el éxito del pipeline, mostrando los fotogramas resultantes para cada uno de los cuatro cuadrantes neuroestéticos definidos. A continuación, se detallan las métricas de rendimiento y validación de esta arquitectura.



Fig. 3. Resultados de ejecución del pipeline para cada cuadrante: Q1-Alegre/Enérgico, Q2-Tenso/Agresivo, Q3-Melancólico/Triste y Q4-Sereno/Calmado.

5.1 Rendimiento y estudio de ablación

Tras la ejecución del Grid Search, los resultados han determinado que la convergencia óptima se logra utilizando un *Batch Size* de 16 y un *Learning Rate* de 0.0005. Aplicando esta configuración y para fundamentar el diseño arquitectónico, se ha ejecutado un estudio de Ablación aislando las diferentes ramas de la red:

- Unimodal (Solo Visión - CNN): **0.0117 (MSE)**
- Unimodal (Solo Audio - OpenSMILE): **0.0114 (MSE)**
- Multimodal (*Feature-Level Fusion*): **0.0106 (MSE)**

Los resultados han confirmado que la combinación equitativa de ambas fuentes reduce el error de predicción en comparación con el uso aislado de técnicas tradicionales o visuales.

Finalmente, tras entrenar el modelo multimodal definitivo con los hiperparámetros óptimos y la estructura antes descrita, el sistema ha logrado alcanzar un **Error Cuadrático Medio (MSE) de 0.0100** en el conjunto de validación, alcanzando el umbral de rendimiento necesario para poder hacer uso del módulo propio implementado.

5.2 Precisión de la clasificación en cuadrantes

Para comprobar el funcionamiento real del sistema, se han comparado las predicciones temporales generadas por el modelo con las anotaciones reales hechas por oyentes humanos. Para garantizar la eficacia en las pruebas, se ha desarrollado un *script* de análisis exploratorio encargado de identificar las canciones que marcan los límites numéricos del *dataset* (máxima energía, máxima melancolía, etc.). Al evaluar estas pistas límite de control (como los ID 75 o 379), las pruebas han demostrado que el modelo es muy

preciso detectando los incrementos de intensidad (*build-ups*) y las caídas de energía de la melodía. En estas canciones con valores extremos, el modelo no solo consigue un margen de error numérico muy bajo, sino que logra un acierto de precisión al ubicar la canción en el cuadrante emocional correcto.

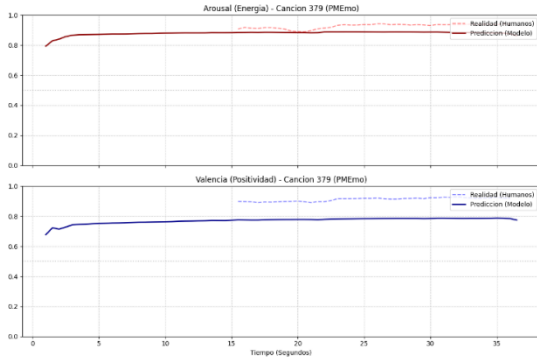


Fig. 4. Comparativa entre el Ground Truth (anotadores humanos, líneas discontinuas) y la predicción de la Inteligencia Artificial (línea sólida) para la canción ID 379. El modelo logra una precisión por cuadrantes del 100%.

Asimismo, la capacidad de la red Bi-GRU para procesar la línea temporal asegura que este seguimiento no sea errático. En temas de mayor complejidad estructural, el sistema demuestra una alta sensibilidad para trazar trayectorias suavizadas, mitigando el "ruido" o los picos de las anotaciones humanas sin perder la fidelidad emocional.

5.3 Análisis de errores y excepciones

Para auditar la robustez del sistema, se ha desarrollado un algoritmo adicional diseñado para buscar las canciones donde el modelo cometía los mayores errores de predicción (outliers). Los resultados han mostrado que el modelo falla principalmente al calcular la Valencia (si una canción es percibida como positiva o negativa) en aquellas pistas donde la letra o el género musical tienen un peso determinante. Se han identificado dos casos claros:

- **El efecto de la letra (Ejemplo ID 579 - Pop alternativo):** Se puede ver como el modelo es capaz de seguir el *beat* (pulso) de la canción, pero sobrestima significativamente los valores de Valencia y Arousal. La red detecta bajos profundos y percusión constante, por lo que predice niveles medios de energía y positividad. Sin embargo, los humanos la puntuaron como "profundamente triste" porque prestaron atención a la letra y a la voz melancólica de la cantante, elementos que la red neuronal no procesa.
- **El efecto cultural (Ejemplo ID 969 - Rap agresivo):** El modelo interpreta texturas acústicas tensas (frecuencias irregulares) y graves (frecuencias graves), prediciendo una Valencia negativa (enfado o tensión). Por el contrario, los oyentes humanos le dieron una puntuación muy alegre y positiva, ya que asocian ese tipo de rap con la euforia

y determinados códigos culturales.

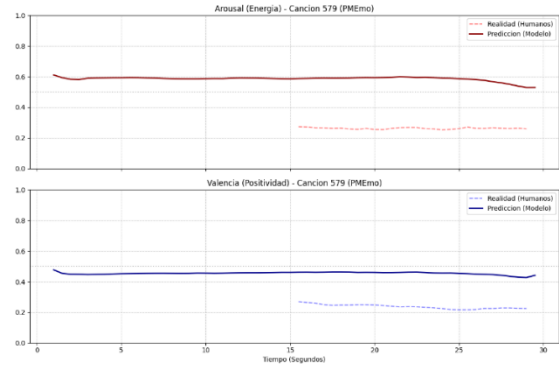


Fig. 5. Análisis de Outlier (Canción ID 579). Se observa una clara discrepancia entre la predicción acústica de la IA (línea sólida) y la percepción semántica del humano (línea discontinua).

5.4 Validación y Estudios de Ablación Arquitectónica (M2-M3)

Para garantizar que el Módulo 3 responda estrictamente al mapeo (Valencia y Arousal) dictado por el Módulo 2 y no genere alucinaciones aleatorias, se sometió la arquitectura a un conjunto de Estudios de Ablación. Al aislar y modificar hiperparámetros individuales, se verificó el impacto directo en la salida visual, justificando cada decisión técnica para el pipeline de producción (Tabla 1, apartado A3 del Apéndice A).

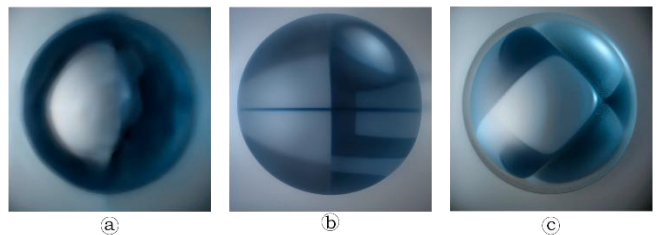


Fig. 6. Impacto de los hiperparámetros de cohesión temporal y semántica en el motor de renderizado. (a) Ruptura vertical drástica de la iluminación y fractura de la geometría al carecer de memoria a corto plazo (*Context Overlap* a 0). (b) Aplanamiento bidimensional y pérdida del eje Z (fuga semántica) ante la ausencia de restricciones penalizables (*sin Prompt Negativo*). (c) Resultado continuo y estable adoptado en el pipeline de producción final, logrando una retención absoluta de la volumetría 3D y la física del fluido.

Los resultados audiovisuales arrojan conclusiones de diseño críticas. En primer lugar, respecto a la Estabilidad Semántica y Temporal: la ausencia de restricciones penalizables (*Prompt Negativo*) provoca que la red olvide el volumen que se está generando, aplanando la esfera en diseños 2D indeseados. Del mismo modo, tal y como evidencian la configuración de *Context Overlap* a 0 o el uso del ruido por defecto (*Default Noise*), carecer de "memoria a corto y largo plazo" provoca errores drásticos de iluminación o la total desintegración del objeto principal en múltiples burbujas aleatorias.

En segundo lugar, respecto al Rendimiento y Fidelidad Estética: Valores de *Classifier-Free Guidance* (CFG) su-

periores a 8.0 provocan sobresaturación ("queman" el renderizado de la imagen), mientras que valores demasiado bajos causan imágenes grises, ignorando los colores cálidos necesarios para mapear una alta Valencia.

Finalmente, el uso de una velocidad dinámica (Motion Scale) dictada por el audio evita los problemas de establecer una velocidad fija. Si se fuerza el motor a una velocidad alta y constante (ej. 1.5), la imagen se rompe en formas inconexas. Al modularla dinámicamente, la modelo es capaz de crear formas afiladas y orgánicas (*Efecto Kiki*) solo en los picos de tensión mapeados, permitiendo volver a la estabilidad de formas más redondeadas posteriormente.

5.5 Validación perceptiva mediante usuarios

Para evaluar empíricamente la coherencia sinestésica del pipeline End-to-End, se diseñó un estudio perceptivo crossmodal (N=63) basado en la escala Likert adaptando las metodologías SAM [18]. Los usuarios evaluaron en silencio cuatro secuencias de vídeo generadas por el sistema que cubrían los cuatro cuadrantes del Modelo Circumplejo de Russell. El sistema alcanzó una precisión global del 71.4% en la transmisión emocional. Para examinar la distribución de este rendimiento y corroborar estadísticamente la independencia de las variables inyectadas, se aplicaron análisis inferenciales de naturaleza no paramétrica.

5.5.1 Análisis Demográfico y Alfabetización Visual

La prueba *H* de *Kruskal-Wallis* descartó sesgos de edad en la decodificación de las atmósferas emocionales del sistema ($H(5) = 5.15, p = 0.397$), confirmando la universalidad del diseño perceptual. Por el contrario, se halló un efecto altamente significativo asociado a la familiaridad previa del sujeto con arte generativo por IA ($H(2) = 11.68, p = 0.002$).

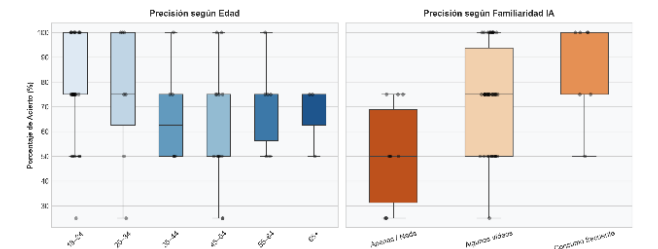


Fig. 7. Análisis de la precisión perceptual en función del perfil de los participantes. Izquierda: Distribución de aciertos agrupados por rangos de edad, confirmando la universalidad del diseño ($H(5) = 5.15, p = 0.397$, no significativo). Derecha: Comparativa de precisión según el grado de familiaridad con IA generativa ($H(2) = 11.68, p = 0.002$, altamente significativo), donde la prueba post-hoc de Dunn revela una brecha de rendimiento significativa ($p < 0.05$) para los usuarios sin exposición previa.

La prueba *post-hoc* de *Dunn* evidenció que los usuarios noveles ("Apenas he visto nada") obtuvieron una precisión inferior en comparación con los usuarios ocasionales ($p = 0.008$) y los consumidores frecuentes ($p = 0.004$), mientras que la diferencia entre estos dos últimos no fue significativa ($p = 0.165$). Esto demuestra una pronunciada curva de

adaptación cognitiva frente a las estéticas sintéticas: una mínima exposición previa capacita al cerebro del observador para filtrar fluctuaciones menores de difusión y asimilar la atmósfera emocional subyacente (**Figura 7**).

5.5.2 Desacoplamiento de Variables (Arousal y Valencia)
El reto arquitectónico central de M2 radica en su capacidad para inyectar los parámetros de Intensidad y Positividad de manera independiente, evitando contaminación cruzada. El Test de Friedman para medidas repetidas confirmó la existencia de alteraciones perceptivas drásticas y focalizadas tanto en la dimensión de Intensidad ($\chi^2 = 163.05, p < 0.001$) como en la de Positividad ($\chi^2 = 60.61, p < 0.001$), demostrando el éxito de los dos submódulos de control:

- Control Cinemático (Arousal vía AnimateDiff):** La modulación dinámica del multiplicador de velocidad (motion_scale) logró su objetivo. Los vídeos diseñados para transmitir alta energía, V3 (Tenso) y V4 (Alegre), registraron puntuaciones de agitación visual significativamente superiores a las de los vídeos estáticos, V1 (Serenos) y V2 (Triste).
- Control Estético (Valencia vía Prompting):** La gestión cromática y de iluminación condicionó la positividad de forma independiente a la velocidad de animación. Ello provocó que los vídeos V2 (Triste) y V3 (Tenso) recibieran las valoraciones perceptivas más oscuras y lúgubres, mientras que V1 (Serenos) y V4 (Alegre) obtuvieron las puntuaciones más luminosas y positivas (**Figura 8**).

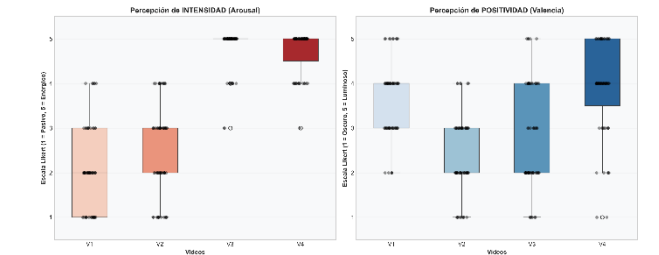


Fig. 8. Distribución de las valoraciones de percepción subjetiva de los usuarios (escala Likert 1-5) para los cuatro vídeos del estudio de caso (V1: Sereno, V2: Triste, V3: Tenso, V4: Alegre). Izquierda: Puntuaciones de Intensidad (Arousal) controladas por la cinemática de velocidad ($\chi^2=163.05,p<0.001$). Derecha: Puntuaciones de Positividad (Valencia) condicionadas por el estilo estético y colorimétrico ($\chi^2=60.61,p<0.001$). Las gráficas demuestran empíricamente el desacoplamiento e independencia de ambas variables de control en el espacio latente.

6 Discusión

La evolución del sistema permite extraer reflexiones clave sobre la viabilidad técnica de la sinestesia artificial. Como demuestran los resultados del test de Friedman ($p < 0.001$), es posible condicionar paramétricamente los modelos de difusión para inducir estados afectivos deterministas en el observador, superando la naturaleza puramente estocástica de estas arquitecturas. Este hallazgo, respal-

dado por la neurociencia de la recompensa auditiva [5], valida nuestro pipeline como un canal de estimulación cross-modal con potencial terapéutico real para la accesibilidad sensorial.

No obstante, se detectó una "ceguera" semántica en el Módulo 1 ante el contexto lírico (casos ID 579 y 969). Curiosamente, esta limitación fue una ventaja estratégica para el motor visual: al traducir la acústica pura en geometrías universales (Efecto Kiki-Bouba [16]), el sistema evita sesgos idiomáticos y culturales, logrando una decodificación afectiva intuitiva, tal como corroboró la prueba de Dunn. A nivel arquitectónico, el over-smoothing temporal de la Bi-GRU actuó como un filtro de paso bajo; sin este efecto, la inestabilidad de los datos de entrada habría generado un flickering visual que colapsaría la coherencia del motor de difusión. De cara a futuras líneas, la integración de LLMs para análisis lírico y el uso de Modelos de Consistencia Latente (LCM) son los pasos siguientes para alcanzar una sinestesia en tiempo real.

7 Conclusiones

La finalización del proyecto ha permitido alcanzar con éxito todos los objetivos establecidos, consolidando un sistema robusto de traducción emocional:

- **Análisis Dinámico:** Se implementó una red híbrida multimodal (CNN + Bi-GRU) que alcanzó un MSE de 0.0100, demostrando una superioridad estadística frente a enfoques unimodales en el seguimiento emocional continuo.
- **Traductor Sinestésico:** Se desarrolló un motor neuroestético funcional que desacopla la estética (baja frecuencia) de la cinemática (alta frecuencia), transformando con éxito las variables acústicas en secuencias dinámicas sobre el ecosistema AnimateDiff.
- **Pipeline End-to-End:** El sistema opera de forma completamente automatizada. Las pruebas de inferencia por lotes confirmaron su robustez frente a la variabilidad del ruido generativo. Finalmente, el estudio perceptivo con 63 usuarios, con una tasa de acierto del 71.4%, ratifica empíricamente que la arquitectura es capaz de transmitir la emoción original de una melodía exclusivamente a través del canal visual.

En síntesis, este trabajo aporta una solución evaluada y determinista para la estimulación multisensorial, estableciendo un puente tecnológico sólido para la sinestesia artificial y la comunicación afectiva crossmodal.

USO DE LA IA GENERATIVA

El diseño arquitectónico, la implementación y la ejecución del presente trabajo han sido realizados íntegramente por el estudiante. El uso de inteligencia artificial se ha limitado a la búsqueda de información, a la corrección gramatical y

de estilo y a la resolución de dudas del ecosistema ComfyUI, sin intervenir en las fases de análisis, desarrollo o proceso creativo.

BIBLIOGRAFÍA

- [1] J. S. Gómez-Cañón *et al.*, "Music Emotion Recognition: toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106-114, Nov. 2021.
- [2] A. Gomila y J. F. Christensen, "On the moral import of the arts: The case of music," en *Progress in Brain Research*, vol. 237. Ámsterdam, Países Bajos: Elsevier, 2018, pp. 483-492.
- [3] P. P. L. Tinio y A. Gartsus, "Characterizing the emotional response to art beyond pleasure: Correspondence between the emotional characteristics of artworks and viewers' emotional responses," en *Progress in Brain Research*, vol. 237. Ámsterdam, Países Bajos: Elsevier, 2018, pp. 319-342.
- [4] X. Han, F. Chen, y J. Ban, "Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure," *Electronics*, vol. 12, no. 4, p. 978, Feb. 2023.
- [5] K. Mori y R. Zatorre, "State-dependent connectivity in auditory-reward networks predicts peak pleasure experiences to music," *PLOS Biology*, vol. 22, no. 8, p. e3002732, Ago. 2024.
- [6] F. Eyben, F. Weninger, F. Gross, y B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," en *21st ACM Int. Conf. Multimedia*, Barcelona, España, Oct. 2013, pp. 835-838.
- [7] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, Dic. 1980.
- [8] Y. Guo *et al.*, "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning," *arXiv preprint arXiv:2307.04725*, 2023.
- [9] K. Zhang, H. Zhang, S. Li, C. Yang, y L. Sun, "The PMemo Dataset for Music Emotion Recognition," en *2018 ACM Int. Conf. Multimedia Retrieval (ICMR)*, Yokohama, Japón, Jun. 2018, pp. 135-142.
- [10] S. Y. Boulahia, A. Amamra, M. R. Madi, y S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications*, vol. 32, no. 121, Sep. 2021.
- [11] S. Ioffe y C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," en *32nd Int. Conf. Mach. Learning (ICML)*, Lille, Francia, Jul. 2015, pp. 448-456.
- [12] S. Santurkar, D. Tsipras, A. Ilyas, y A. Mądry, "How Does Batch Normalization Help Optimization?," en *Advances in Neural Information Processing Systems (NeurIPS)*, Montreal, Canadá, Dic. 2018, pp. 2488-2498.
- [13] D. P. Kingma y J. Ba, "Adam: A Method for Stochastic Optimization," en *3rd Int. Conf. Learning Representations (ICLR)*, San Diego, EE. UU., May. 2015.
- [14] R. Reber, N. Schwarz, y P. Winkielman, "Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience?," *Personality and Social Psychology Review*, vol. 8, no. 4, pp. 364-382, 2004.
- [15] B. P. Meier, M. D. Robinson, y G. L. Clore, "Why good guys wear white: Automatic inferences about stimulus valence based on brightness," *Psychological Science*, vol. 15, no. 2, pp. 82-87, 2004.
- [16] V. S. Ramachandran y E. M. Hubbard, "Synaesthesia: A Window Into Perception, Thought and Language," *Journal of Consciousness Studies*, vol. 8, no. 12, pp. 3-34, 2001.
- [17] M. Bar y M. Neta, "Humans prefer curved visual objects," *Psychological Science*, vol. 17, no. 8, pp. 645-648, 2006.
- [18] M. M. Bradley y P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49-59, Mar. 1994.

APÉNDICE

A1. PLANIFICACIÓN TEMPORAL Y RECURSOS

La progresión temporal se ha planificado alineando los módulos de desarrollo (M1-M4) con las entregas parciales estipuladas, estimando un total de ~300 horas de trabajo autónomo (12 ECTS).

Además, se ha diseñado un Diagrama de Gantt estructurado en 5 sprints y diversas tareas:

- Sprint 1 (Hasta el 08/03/26): Fase de Diseño** - Búsqueda bibliográfica y definición de la arquitectura.
 - Tiempo estimado:** 30 horas.
 - Recursos:** Ordenador personal y bases de datos científicas.
- Sprint 2 (Hasta el 12/04/26): El Cerebro Analítico (M1)** - Búsqueda de Datasets de audio etiquetados (T1). Programación del preprocesamiento (T2) y entrenamiento/adequación de la Red Neuronal Híbrida (T3) para generar las métricas del Modelo Circumplex (T4).
 - Tiempo estimado:** 80 horas.
 - Recursos:** Datasets públicos de audio y acceso a capacidad de cómputo GPU (local o Cloud) para el entrenamiento de la red.
- Sprint 3 (Hasta el 24/05/26): Traducción, Generación y Post-procesado (M2 y M3)** - Desarrollo e implementación del algoritmo determinista para la traducción y generación de prompts dinámicos con arquitectura de frecuencias desacopladas (T5). Configuración e instalación local del entorno generativo nodal ComfyUI basado en Stable Diffusion 1.5 y adecuación del modelo junto al VAE (T6). Integración de la interpolación lineal temporal, resolución matemática del desajuste del framerate y ejecución de estudios de ablación para validar el sistema (T7).
 - Tiempo estimado:** 80 horas.

- Recursos:** Estación de trabajo local con GPU dedicada, repositorio oficial de AnimateDiff-Evolved y suite nodal ComfyUI.
- Sprint 4 (Hasta el 14/06/26): Optimización, Integración y Validación (M3 y M4)** - Desarrollo del módulo pipeline End-to-End (M4) mediante scripts de automatización para conectar en un solo flujo la salida de la Bi-GRU con ComfyUI (T8). Ejecuciones y pruebas sobre el dataset musical e implementación de tests de variabilidad mediante semillas aleatorias (T9). Análisis estadístico de las pruebas y encuestas de concordancia perceptiva con usuarios finales para validar la correspondencia crossmodal del sistema (T10).
 - Tiempo estimado:** 60 horas.
 - Recursos:** Código de integración propio, Formularios de Google (Google Forms) para la recogida de métricas y software de análisis estadístico.
- Sprint 5 (Hasta el 18/07/26): Cierre y Presentación** - Cierre definitivo del código, elaboración del Dossier Digital y redacción del Informe Final (entrega el 29/06), junto con la preparación del póster (entrega 02/07) y las diapositivas para el tribunal de defensa.
 - Tiempo estimado:** 50 horas.
 - Recursos:** Plantillas oficiales de la UAB y software de ofimática.

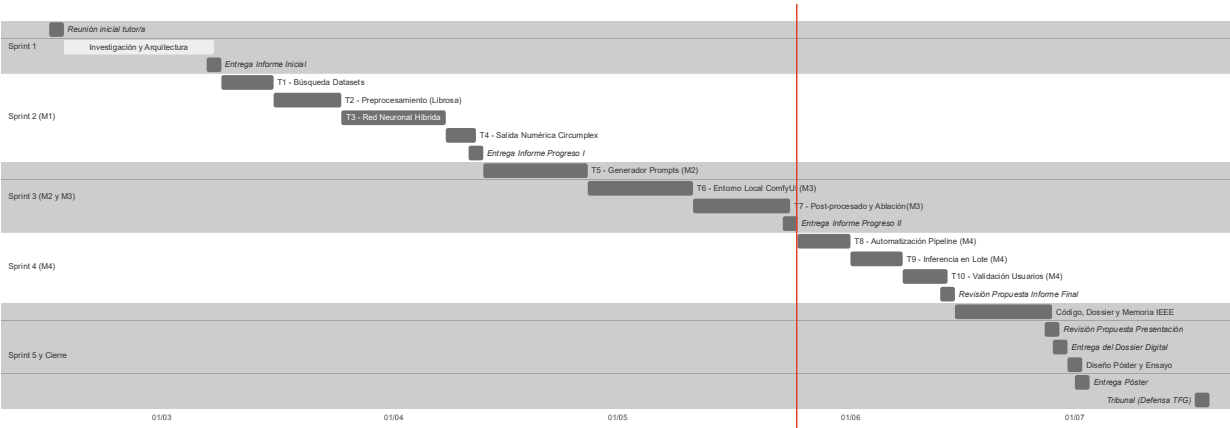


Fig. A1. Diagrama de Gantt estructurado en sprints, tareas e hitos.

A2. ESTRUCTURA NODAL EN COMFYUI (M3)

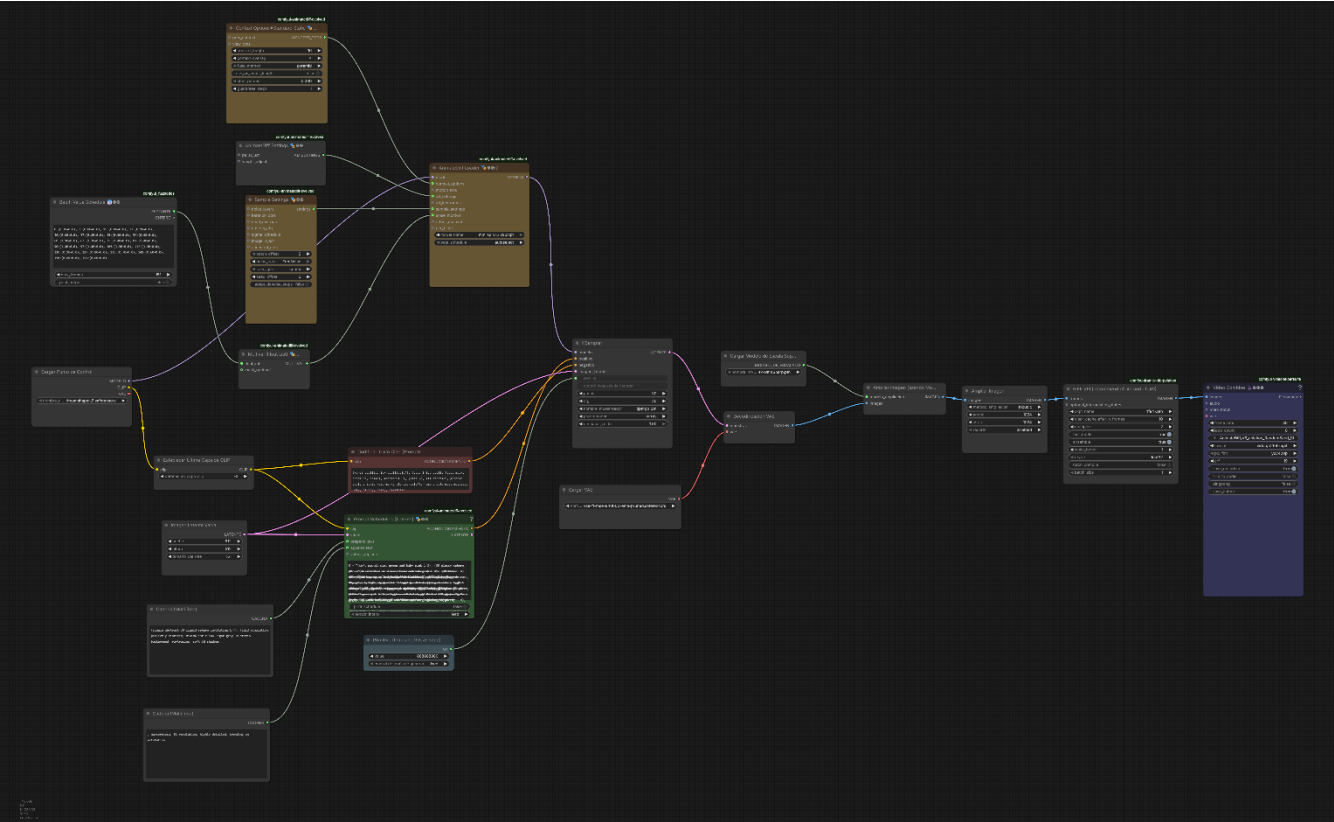


Fig. A2. Grafo de nodos implementado en ComfyUI

A3. RESUMEN DEL ESTUDIO DE ABLACIÓN (M3)

TABLA 1
RESUMEN DEL ESTUDIO DE ABLACIÓN (M3)

Variable analizada	Objetivo	Decisión final	Justificación Arquitectónica
Prompt Negativo	Evaluar fugas semánticas (pareidolia 2D).	Activado (Peso 1.5)	Previene alucinaciones y la pérdida del eje Z en la evolución temporal de la esfera.
Context Overlap	Evaluar la cohesión temporal de la secuencia generada.	Overlap = 4	Evita rupturas visuales drásticas de luz y forma al pasar de una ventana a otra (16 fotografías) que usa el modelo.
CFG Scale	Fidelidad textual vs. Libertad creativa.	CFG = 7.5	Equilibra obediencia al color (Valencia) y textura realista sin generar ruido ni llegar a tonos "quemados".
VAE	Evaluar rango dinámico y colorimetría (Externo vs Integrado).	VAE ft-mse	Aunque los resultados del integrado son aceptables, se estandariza el uso del ft-mse por ser la norma de la industria para garantizar la máxima fidelidad de color.
KSampler Steps	Coste computacional vs. Calidad visual geométrica	30 Steps	Es la cantidad óptima iterativa para que el modelo calcule correctamente las físicas de iluminación 3D (oclusión ambiental y reflexión especular).
Motion Scale	Validación de cinemática dinámica o uso de estática.	Dinámico [0.5-1.5]	Sincroniza orgánicamente la agitación visual en tiempo real con los valores de Arousal recibidos cada 0.5s
Algoritmo Ruido	Estabilidad de la generación a largo plazo.	FreeNoise	Recicla los tensores para lograr la permanencia temporal del volumen 3D.
Prompt Travel	Límite en la frecuencia de actualización de texto.	Baja freq. (2 seg)	Otorga margen de tiempo de cálculo a los tensores, previniendo la sobrecarga semántica, el ruido y los parpadeos visuales.