PROJECT METHODOLOGY

MelodicLens: AI-Powered Mood-Based Music Recommendation

With over 5.22 billion social media users sharing images and stories daily, finding the perfect soundtrack to match their content's mood and aesthetic can be overwhelming. MelodicLens simplifies this process by leveraging Generative AI to analyze uploaded images and recommend the perfect songs—enhancing user engagement and social media presence effortlessly. The main steps are as described below:

1) Mood Detection from Images:

A Visual Language Model (VLM) processes the user's uploaded image. We prompt it to give a textual description of the model. The image and text information is concatenated to form the query vector.

2) Playlist-Based Mood Matching

We use an embedding model to embed the audio features and lyrics of the user's playlist and store it in a vector DB for easy retrieval. We then pass the query vector and get the top 3 matches based on the user's playlist. We also add an external data of the most trending songs for RAG.

By integrating computer vision and natural language processing (NLP), MelodicLens personalizes music discovery—bridging the gap between visuals and sound effortlessly.

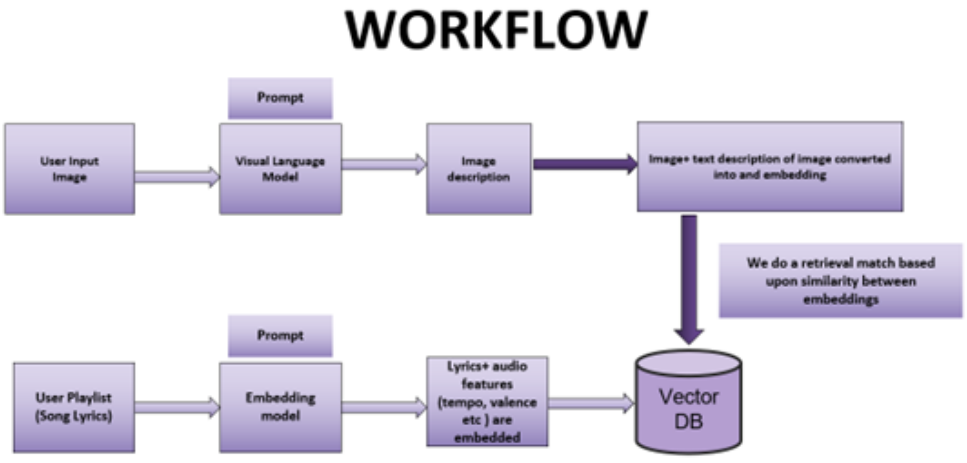

Figure 1: Project workflow

1. Purpose of the Methodology

The chosen methodologies effectively address the problem of music recommendation based on mood detection from images. The vision model accurately analyzes the visual input, identifying the mood through facial expressions or scenes. We then use an embedding model to convert the audio features + lyrics (SentenceTransformer) and an LLM-based model (BART) to extract thematic information from song lyrics, enhancing the emotional resonance of recommendations. The system combines both inputs to match detected moods with relevant songs, while user playlist data ensures personalized recommendations. This hybrid approach

`https://github.com/NitikaNahata/DS_5500`

allows for a more tailored, cross-modal mapping of visual and auditory content, providing accurate, context-aware music suggestions.

2. Problem Statement

2.1. Defining the problem

This project falls under Machine Learning and LLMs, combining classification and recommendation systems. A vision model processes an image along with a prompt to get a description from the image such as the main theme, key elements, emotional tone, music trait suggestions. While an LLM model analyses song lyrics to classify their mood, we also use embedding models to vectorise the lyrics and audio features (valence, tempo, danceability, etc.). We also embed the user input image and the description and concatenate them into a single query vector which we use to find the top 3 matches from the user's playlist or the external trending song database that we have. Finally, it integrates the user's Spotify playlist to refine suggestions, ensuring the recommended song aligns with both the image's mood and the user's music preferences.

2.2. Significance

Social media users often struggle to find the perfect music to accompany their social media posts and stories, limiting the emotional impact and engagement of their content. Music choices can greatly influence how posts are perceived, but manually selecting tracks can be time-consuming and subjective. By integrating AI-driven recommendations and music platform APIs like Spotify, this project aims to automate and personalize music suggestions based on user preferences and media content analysis. Addressing this challenge can enhance content personalization, improve social media engagement, and provide a creative edge to users. We chose this problem because it blends AI with creative media applications, offering a unique and engaging user experience while challenging us to explore both technical and design aspects.

3. Data Collection and Preparation

3.1. Data Sources

Our project utilizes a combination of public datasets, APIs, and user-generated data to build a robust mood-based music recommendation system. The primary data sources include publicly available Kaggle datasets that provide extensive song metadata, including audio features, valence, energy, and classification labels. Specifically, we use:

- Spotify Top Songs and Audio Features – Contains song metadata with detailed audio attributes.
- Spotify Classification Dataset – Provides labeled song features for classification tasks.
- TikTok Trending Tracks – Includes trending music data from TikTok for real-world popularity insights.

Additionally, we integrate Spotify's Web API to retrieve user-specific listening history, playlists, and real-time song features, allowing for personalized recommendations. Our vision model analyzes user-uploaded images to classify moods, while LLMs process song lyrics to extract mood-based labels.

`https://www.kaggle.com/datasets/julianoorlandi/spotify-top-songs-and-audio-features`

`https://www.kaggle.com/datasets/geomack/spotifyclassification`

`https://www.kaggle.com/datasets/yamqwe/tiktok-trending-tracks`

3.2. Data Description

Our final dataset consists of 710 unique rows, with each row representing a song selected from a user's playlist, matched based on audio features and lyrics fetched from the Genius API. This dataset is built by combining the datasets from Kaggle containing 28,451 unique tracks and the user-specific data obtained through the Spotify API. The Kaggle dataset includes key attributes such as danceability, energy, key, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time_signature, and popularity. The user-specific data, extracted from Spotify, included these same features and matched them with the relevant songs that appear in the user's playlist.

In terms of class imbalances, there are no major issues observed in the distribution of features across the dataset, as it consists of diverse songs from both user and general data. The text corpus, which consists of song lyrics, was preprocessed by removing stopwords, applying lemmatization, and performing tokenization before generating embeddings.

3.3. Preprocessing steps

For each song in the 710 tracks, the following preprocessing steps were performed:

- **Text Cleaning:** Cleaned the song lyrics by removing unwanted contributor info, special characters, and extra spaces. The text was then converted to lowercase, and newlines and tabs were removed to ensure a uniform format for embedding generation.
- **Imputing Values:** We imputed the missing values for the tempo and time signature by analyzing their distribution.
- **Embeddings:** Using the SentenceTransformer (`all-MiniLM-L6-v2`), we generated 384-dimensional embeddings for each song's lyrics, which represent the semantic content of the lyrics in vector form.
- **Normalization:** Numerical features like danceability, energy, tempo, loudness, etc. were normalized using MinMaxScaler, scaling them to a range between 0 and 1 to ensure consistency across features. We use the same model as mentioned above for the audio features as well.

4. Model Consideration

Large Language Vision Model for Image Captioning task:

For the visual-emotional understanding component of the music recommendation system, the primary focus was on Large Language Vision Models (LLVMs) that are readily accessible for local deployment through Ollama. Ollama provides a convenient local deployment environment, enabling rapid experimentation and easy integration without incurring significant costs or requiring extensive computational resources.

The following LLVMs were evaluated:

- **LLaVA (1.6):** Combines a vision encoder with Vicuna, optimized for general-purpose visual-language understanding, strong visual reasoning, and OCR capabilities.
- **Llama 3.2 Vision:** Built upon the Llama 3.1 language model with a vision adapter; excels in document-level understanding, OCR, chart analysis, and image captioning.
- **MiniCPM-V 2.6:** An 8B parameter multimodal model based on SigLip and Qwen architectures, known for superior single-image and multi-image understanding, multilingual support, strong OCR, and efficient deployment.
- **BakLLaVA:** Built upon Mistral 7B with LLaVA architecture, offering strong multimodal reasoning and efficient inference suitable for resource-constrained environments.
- **Moondream2:** A compact VLM (1.86B parameters) optimized for edge devices, providing good performance in basic visual-language tasks despite small model size.
- **Granite3.2 Vision:** A compact model specifically optimized for visual document understanding tasks such as charts, tables, diagrams, and infographics extraction.

Paid Proprietary LLVMs (Not Included due to High Cost):

- OpenAI GPT-4 Turbo with Vision / GPT-4o

- Google Gemini 1.5 Pro / Gemini Ultra

These proprietary models offer excellent multimodal performance but were excluded due to high operational expenses (per-token or subscription-based pricing), making them unsuitable for cost-effective local deployment scenarios.

Transformer-Based Models for song analysis:

Given the importance of textual data (song lyrics) in this project, we considered transformer-based models like BART. These models are capable of capturing complex semantic relationships and understanding long-range dependencies in text. They were chosen for the following reasons:

- State-of-the-art performance on NLP tasks: Transformer-based models like BART (Bidirectional and Auto-Regressive Transformers) have shown superior performance on a variety of NLP tasks, including text classification, summarization, and theme identification.
- Handling long sequences: BART and similar models are robust at processing long sequences of text (like song lyrics) without losing context, which is essential for understanding the themes of the songs.
- Pre-trained models: Leveraging pre-trained models like `facebook/bart-large-mnli` on Hugging Face allowed us to save time and computational resources, as these models are already fine-tuned for a variety of tasks, including sentence classification and textual entailment.

Thus, for the theme identification task, we used the `facebook/bart-large-mnli` model. This model was particularly chosen because it is designed for multi-class classification tasks and performs well in identifying themes based on textual input. We used it to classify songs into themes such as romance, sadness, nature, friendship, etc.

Final Model Selection

- **Open-source Availability:** Llama 3.2 Vision is available via Ollama, facilitating straightforward local deployment without licensing concerns. Its open-source nature ensures flexibility and cost-effectiveness for academic and personal projects.
- **Multimodal Capabilities:**
  - **Visual Reasoning:** Llama 3.2 Vision demonstrated exceptional performance in understanding and interpreting complex visual inputs, such as extracting emotional context from images.
  - **OCR Accuracy:** The model excels in text extraction from images, making it suitable for scenarios involving document or infographic analysis.
  - **VQA Accuracy:** It achieved "Very High" ratings in Visual Question Answering tasks, showcasing its ability to provide accurate and detailed responses.

- **Computational Efficiency:** While the model has a moderate computational efficiency due to its larger size (11B parameters), it remains manageable for local deployment with adequate hardware resources.

Ultimately, Llama 3.2 Vision (11B) was selected due to its robust performance across all metrics, making it an optimal choice for this project's visual mood-based music recommendation system.

Model	VQA Accuracy	Visual Reasoning	OCR Capability	Multilingual Support	Computational Efficiency
LLaVA 1.6	High	High	High	Moderate	Moderate (7B–34B)
Llama 3.2 Vision (11B)	Very High	Very High	Very High	Moderate	Moderate–Low (11B)
MiniCPM-V 2.6	Very High	Very High	Very High	Excellent	High (8B)
BakLLaVA	Moderate	Moderate	Moderate–High	Limited	High (7B)
Moondream 2	Moderate	Moderate–High	Moderate	Limited	Very High (1.8B)
Granite3.2-Vision	Moderate–High	Very High	High	Limited	Very High (2B)

Figure 2: Comparison among different models

5. Model Development and Training

Architecture and Configuration

- The selected model, Llama 3.2 Vision (11B), follows a transformer-based architecture designed for multimodal reasoning tasks.
- **Transformer Architecture:** Combines a vision encoder with a language decoder, leveraging self-attention mechanisms to process and align visual and textual data effectively.
- **Attention Mechanisms:** Multi-head attention layers allow the model to focus on different aspects of visual and textual inputs simultaneously.
- **Number of Layers:** The model contains 11 billion parameters, striking a balance between performance and computational feasibility.
- **Pre-training vs Fine-tuning:** Llama 3.2 Vision was pre-trained on large-scale multimodal datasets and fine-tuned for specific tasks like VQA and OCR.

Training Process

- Since this project utilized a pre-trained LLVM for inference rather than training from scratch, no additional training or dataset splitting was performed.
- The pre-trained weights of Llama 3.2 Vision were directly used to analyze images and infer emotional contexts.
- No additional labeled datasets were required for fine-tuning.

Hyperparameter Tuning

- No hyperparameter tuning was performed in this project as the pre-trained configuration of Llama 3.2 Vision already provided optimal performance for the intended tasks.
- Default hyperparameters were used during inference.
- The reliance on pre-trained weights eliminated the need for grid search, random search, or other tuning methods.

For this project, the default configuration of Llama 3.2 Vision (11B) was used without modifications, as it already provided optimal performance for visual-emotional understanding tasks.

6. Evaluation and Comparison

Key Performance Metrics

To evaluate the performance of the models, the following metrics were used:

1. **Visual Question Answering (VQA) Accuracy:** Measures the model's ability to answer questions based on visual inputs, reflecting its multimodal reasoning capabilities.
2. **Visual Reasoning:** Assesses the model's ability to interpret complex visual inputs, such as emotional context, object relationships, and scene comprehension.
3. **OCR Accuracy:** Evaluates text extraction capabilities from images, which is critical for tasks involving document or infographic analysis.
4. **Consistency Across Prompts:** Determines how reliably the model provides structured and accurate responses across diverse prompts.
5. **Human Evaluation Metrics (ROUGE and Perplexity):**
    - **ROUGE Score:** Used to measure how well the generated responses match reference outputs in terms of recall.
    - **Perplexity:** Assesses the fluency and coherence of generated text by evaluating how well the model predicts sequences of words.

Types of Prompts Constructed

The evaluation involved constructing diverse prompts designed to test various aspects of multimodal reasoning:

1. **Emotion Extraction Prompts:**
    - Example: Analyze an image and provide its emotional tone, key elements, and suggested music traits.
    - Purpose: To evaluate how well models can infer emotional context from visuals.
2. **Scene Description Prompts:**
    - Example: Generate a detailed description of an image, including its themes, elements, and overall mood.
    - Purpose: To test the model's ability to identify objects, relationships, and abstract themes in images.
3. **Music Recommendation Prompts:**
    - Example: Suggest music characteristics (tempo, mood, genre) based on an image's visual mood and atmosphere.
    - Purpose: To assess how effectively models can align visual elements with audio traits.

Types of Images Compared

The evaluation involved analyzing a variety of image types to test the models' versatility:

1. **Natural Landscapes:**
    - Images depicting serene environments like forests, mountains, or lakes.
    - Purpose: To evaluate emotional tone extraction and nature-based music recommendations.
2. **Urban Scenes:**
    - Images featuring bustling streets or cityscapes with neon lights.
    - Purpose: To test the model's ability to capture energy and vibrancy in urban settings.
3. **Close-ups of Objects:**
    - Images like flowers in vases or picnic setups.
    - Purpose: To analyze how well models identify intricate details and suggest appropriate music traits.
4. **Animal Portraits:**

- Images featuring animals in natural or playful settings (e.g., a dog holding flowers).
- Purpose: To assess emotional inference and thematic understanding.

Why Llama 3.2 Vision (11B) is the Ideal Choice

After evaluating multiple models using these prompts and image types, Llama 3.2 Vision (11B) emerged as the most suitable choice for this project due to its superior performance across key metrics:

1. **Consistent Accuracy Across Prompts:**
    - Llama 3.2 Vision provided highly structured and accurate responses across all prompt types without requiring extensive post-processing.
2. **Superior Visual Reasoning Capabilities:**
    - It excelled in interpreting complex scenes, identifying abstract themes, and aligning them with appropriate music traits.
3. **High VQA Accuracy:**
    - The model demonstrated exceptional accuracy in answering questions about images, outperforming other open-source models like MiniCPM-V and BakLLaVA.
4. **Minimal Normalization Requirements:**
    - Unlike some other models that required output normalization for consistency (e.g., LLaVA), Llama 3.2 Vision delivered well-structured outputs directly.
5. **Human Evaluation Metrics:**
    - Llama 3.2 Vision achieved high ROUGE scores for recall-oriented tasks and low perplexity scores for fluency and coherence in text generation.

Llama 3.2 Vision (11B) consistently outperformed other models across diverse prompts and image types due to its strong multimodal reasoning capabilities, high accuracy in VQA tasks, minimal need for normalization, and structured outputs aligned with human evaluation metrics like ROUGE scores and perplexity. These strengths make it the ideal choice for this project's requirements in visual mood-based music recommendation tasks.