Dylumo: A Multimodal Agentic Image-to-Music Recommendation System

Muhammad Awab ur Rehman 22i-1068
Adeel Mehmood 22i-0979
Talha Azim 22i-1243

October 19, 2025

1 Introduction and Project Description

This project proposes the development of **Dylumo**, an innovative multimodal agentic system that bridges visual and auditory domains through intelligent image-to-music mapping. The core objective is to create a sophisticated recommendation system that analyzes the emotional and contextual content of images and generates personalized music recommendations that semantically align with the visual mood.

Dylumo represents a novel approach to cross-modal recommendation systems by leveraging a multi-agent architecture that enables dynamic interaction, adaptive learning, and conversational control over the recommendation process [6]. Unlike traditional static recommendation engines, Dylumo employs specialized agents that work collaboratively to understand image semantics, analyze music characteristics, perform cross-modal mapping, and generate personalized recommendations through natural language interaction [11].

The system addresses the complex challenge of translating visual emotional cues into musical expressions, requiring sophisticated understanding of both modalities and their semantic relationships [15]. This project contributes to the growing field of multimodal generative AI by exploring how agentic systems can enhance cross-modal understanding and recommendation quality.

2 Main Functions and Capabilities

2.1 Core Functionality

- **Image-to-Music Mapping**: Analyze uploaded images to extract mood, context, and emotional content, then recommend music that semantically matches these visual characteristics
- **Personalized Recommendations**: Generate music suggestions based on user preferences, listening history, and contextual factors
- **Playlist Integration**: Allow users to upload and search within their existing playlists from multiple music streaming services

2.2 Multi-Platform Integration

- **Music Service APIs**: Integration of multiple services like Spotify, YouTube Music and SoundCloud.

- **Playlist Management**: Retrieve user playlists and enable playlist-specific recommendations.

- **Cross-Platform Search**: Unified search across multiple music services for comprehensive recommendations

2.3 Advanced Features

- **Conversational Interface**: Natural language interaction with LLM-powered chatbot for refined recommendations [4]

- **Image-Music Alignment Evaluation**: Users can upload both image and song to evaluate compatibility [5]

- **Refinement Capabilities**: Users can provide additional input (e.g., ”give something upbeat,” ”something more melancholic”) to guide recommendations

- **Context Awareness**: Consider time of day, user history, and social context for enhanced personalization

3 Multi-Agent System Architecture

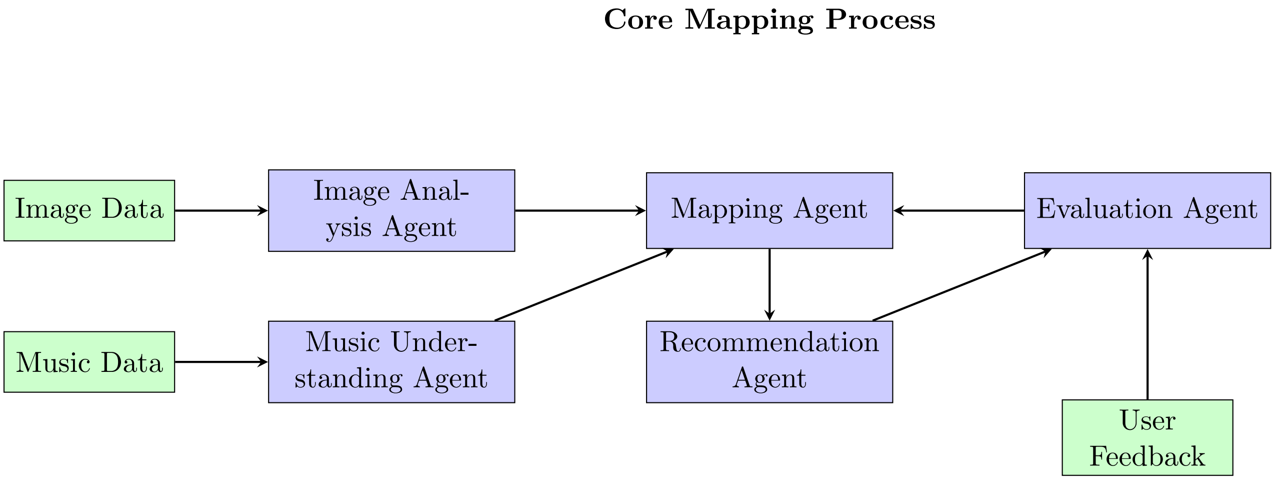

Figure 1: Dylumo Multi-Agent System Architecture

3.1 Agent 1: Image Analysis Agent

Role: Comprehensive analysis of input images to extract semantic understanding

- **Tools**: Computer vision models (ResNet, Vision Transformers) [10], mood classification models (VGG-Face, custom emotion classifiers), object detection frameworks (YOLO, Faster R-CNN)

- **Output**: Structured representation of image mood, context, and visual features

3.2 Agent 2: Music Understanding Agent

Role: Deep analysis of music characteristics and emotional content

- **Tools**: Audio analysis libraries (Librosa), music theory knowledge integration, genre classification models [1], Spotify audio features (valence, energy, tempo, mode, danceability)
- **Output**: Structured music mood and characteristic representation

3.3 Agent 3: Mapping Agent

Role: Establish optimal image-music semantic alignments

- **Tools**: Cross-modal models (CLIP-style architectures) [7], similarity metrics and embedding spaces, advanced reasoning modules
- **Output**: Ranked image-music alignments with confidence scores

3.4 Agent 4: Recommendation Agent

Role: Generate personalized music recommendations

- **Tools**: User preference learning algorithms, context awareness modules, ranking algorithms, music service API integrations
- **Output**: Ranked, personalized music recommendations with explanations

3.5 Agent 5: Evaluation Agent

Role: Continuous quality assessment and system improvement

- **Tools**: User feedback mechanisms, quality metrics (precision, recall, diversity), A/B testing frameworks
- **Output**: Performance assessments and improvement suggestions

4 Agentic Capabilities

4.1 Conversational Interface

- **Natural Language Interaction**: Users can describe images, express mood preferences, and refine recommendations through dialogue
- **Contextual Understanding**: System interprets user intent and coordinates appropriate agent responses
- **Dynamic Refinement**: Real-time adjustment of recommendations based on user feedback

4.2 Adaptive Learning

- **Feedback Integration**: Learn from user interactions, ratings, and implicit behavior
- **Preference Evolution**: Adapt to changing user tastes and listening patterns
- **Context Sensitivity**: Adjust recommendations based on situational factors

4.3 Multi-Modal Reasoning

- **Cross-Modal Understanding**: Sophisticated reasoning across visual, auditory, and textual modalities
- **Semantic Alignment**: Deep understanding of emotional and contextual relationships between images and music
- **Hierarchical Decision Making**: Coordinated decision-making across multiple agents

5 Innovation and Research Contributions

5.1 Novel Agent Architecture

- **Multi-Agent Coordination**: First application of multi-agent systems specifically for image-to-music recommendation [14]
- **Hierarchical Decision Making**: Advanced coordination strategies for cross-modal tasks
- **Adaptive Learning**: Integration of learning mechanisms within multimodal agentic frameworks

5.2 Conversational Control

- **Fine-Grained Control**: Users can exert precise control over recommendation processes through natural language [12]
- **Dynamic Interaction**: Real-time adaptation based on conversational context
- **Interpretable Recommendations**: Transparent reasoning and explanation generation

5.3 Cross-Modal Understanding

- **Semantic Bridge**: Novel approaches to bridging visual and auditory semantic spaces [13]
- **Emotional Alignment**: Advanced techniques for matching visual and musical emotional content [2]
- **Context-Aware Mapping**: Integration of multiple contextual factors in cross-modal alignment

6 Comparative Analysis Framework

6.1 System Architecture Comparison

- **Agent-based vs. Non-Agentic**: Compare Dylumo's multi-agent approach against traditional pipeline systems
- **Different Coordination Strategies**: Evaluate various inter-agent communication methods (hierarchical, peer-to-peer, blackboard)
- **Reasoning Approaches**: Compare neural embeddings vs. hybrid symbolic-neural reasoning

6.2 Feature and Model Comparison

- **Feature Set Analysis**: Compare different image and music feature combinations
- **Model Architectures**: Evaluate various cross-modal models and their effectiveness
- **Personalization Methods**: Compare different user preference learning approaches

6.3 Evaluation Metrics

- **Quantitative Metrics**: Precision, recall, F1-score, MAP, NDCG, diversity metrics
- **Qualitative Studies**: User satisfaction, perceived recommendation quality, usability
- **Adaptability Measures**: Learning effectiveness, preference evolution tracking

7 Evaluation Strategy

7.1 Quantitative Evaluation

- **Mapping Accuracy**: Precision/recall on curated image-music pairs
- **Recommendation Performance**: Standard recommendation metrics (MAP, NDCG)
- **Agent-Specific Metrics**: Individual agent performance evaluation
- **System Efficiency**: Response time, computational resource usage

7.2 Qualitative Evaluation

- **User Interaction Studies**: Usability and intuitiveness assessment
- **Perceived Quality**: User ratings of recommendation relevance and diversity
- **Adaptability Assessment**: Longitudinal studies of system learning
- **A/B Testing**: Comparative evaluation of different system configurations

7.3 Comparative Studies

- **Baseline Comparisons**: Against traditional recommendation systems
- **Ablation Studies**: Impact of individual agents and features
- **Cross-Domain Evaluation**: Performance across different image types and music genres

8 Dataset Sources

8.1 Music Data

- **Million Song Spotify Dataset**: Primary music dataset with Spotify audio features (valence, energy, tempo, mode, danceability, acousticness, instrumentalness, liveness, speechiness) [9]

- **User Playlists**: Integration with Spotify API for personalized playlist access
- **Music Service APIs**: Access to metadata and audio features from multiple platforms

8.2 Image Data

- **Emotion-Labeled Datasets**: AffectNet, Flickr-Creative Commons with emotion tags
- **General Image Datasets**: ImageNet, COCO for pre-training computer vision models
- **Custom Datasets**: Curated image collections with mood annotations [8]

8.3 Conversational Data

- **Dialogue Datasets**: Task-oriented conversation datasets for LLM fine-tuning
- **User Interaction Data**: Collected feedback and interaction patterns
- **Reinforcement Learning Data**: Human feedback for adaptive learning [3]

9 Implementation Scope and Feasibility

9.1 Project Scope

- **Leverage Pre-trained Models**: Utilize existing computer vision, music understanding, and LLM models
- **Focus on Integration**: Emphasis on agent coordination and system integration rather than training from scratch
- **Modular Development**: Parallel development of individual agents with clear integration points
- **Incremental Evaluation**: Continuous evaluation and refinement throughout development

9.2 Technical Feasibility

- **API Integration**: Well-documented APIs for music services and pre-trained models
- **Cloud Resources**: Utilize cloud computing for model inference and data processing
- **Open Source Tools**: Leverage existing frameworks for agent development and evaluation

9.3 Deliverable Timeline

- **Weeks 1-2**: Literature review and system design
- **Weeks 3-5**: Individual agent development and testing
- **Weeks 6-8**: System integration and comparative evaluation
- **Weeks 9-10**: Final testing, documentation, and paper preparation

10 Conclusion

Dylumo represents a significant advancement in multimodal recommendation systems by combining the power of multi-agent architectures with sophisticated cross-modal understanding. The project addresses real-world challenges in music discovery while contributing novel research insights to the fields of generative AI, multimodal learning, and agentic systems.

The proposed system is designed to be both innovative and practically valuable, offering users an intuitive and powerful way to discover music through visual cues while providing researchers with insights into effective cross-modal agent coordination. The comprehensive evaluation framework ensures rigorous assessment of the system's capabilities and provides clear metrics for success.

This project fulfills all course requirements by offering innovative work in agentic multimodal systems [14], comprehensive comparative analysis [15], and practical implementation within a semester timeframe. The combination of theoretical contributions and practical applications makes Dylumo an ideal project for demonstrating mastery of generative AI concepts and techniques.

References

[1] Xin Cai and Hongjuan Zhang. Music genre classification based on auditory image, spectral and acoustic features. *Multimedia Systems*, 28(3):779–791, 2022.

[2] Suwan Choi, Kyu Won Kim, and Myungjoo Kang. Mmva: Multimodal matching based on valence and arousal across images, music, and musical captions, 2025.

[3] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2023.

[4] Chris Donahue, Antoine Saillenfest, Weiyi Ji, Shun-Ching Vie Cheng, Zachary Novack, Ardavan Saeedi, Timo I. Denk, and Neil Zeghidour. Instruct-musicgen: Unlocking text-to-music editing for music language models via instruction tuning, 2024.

[5] Adwaita Janardhan Jadhav and Ishmeet Kaur. The melodies of an image: Exploring music recommendations based on an image's content and context. In *KaRS@RecSys 2023*, pages 65–68, 2023.

[6] Souraja Kundu, B. L. Kumar, and S. R. M. Prasanna. Emotion-guided image to music generation, 2024.

[7] Takayuki Nakatsuka, Masahiro Hamasaki, and Masataka Goto. Content-based music-image retrieval using self- and cross-modal feature embedding memory. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2173–2182, 2023.

[8] Jeongeun Park, Hyorim Shin, Changhoon Oh, and Ha Young Kim. "is text-based music search enough to satisfy your needs?" a new way to discover music with images. In *CHI 2024*, pages 504:1–504:21, 2024.

[9] Philippe Pasquier, Shlomo Dubnov, Jean-Pierre Briot, Francisco Gómez, Arne Eigenfeldt, Florian Colombo, Eduardo Coutinho, Stefan Lattner, Matthew E. P. Davies, Cumhur Erkut, Bob L. T. Sturm, and Alexander Lerch. Survey on the evaluation of generative models in music, 2024.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[11] Zelin Wang, Han Wu, Kang Zhang, Yu Wang, Zhiyong Zhang, and Le Sun. Continuous emotion-based image-to-music generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5776–5785, 2023.

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[13] Zeyu Xiong, Yan Zhang, and H. V. Zhao. Retaining semantics in image to music conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[14] Shunyu Yao, Jeffrey Zhao, David Yu, Naftali Shazeer, Michiel de Jong, Yonatan Bisk, Yoav Artzi, and Percy Liang. React: Synergizing reasoning and acting in language models, 2022.

[15] Xincheng Zhu, Bohan Nong, Yawen Wu, Yifeng Geng, Xu Tan, Chunfeng Wang, Yixuan Fu, Jiang Bian, Stephen Mutuvi, Sheng Zhao, Taylor Berg-Kirkpatrick, Roger Dannenberg, Gus Xia, and Rui Yan. Vision-to-music generation: A survey, 2024.