VIBELENS

VINUNIVERSITY
College of Engineering and Computer Science

VIBELENS
Image-Based Music Recommendations System

TEAM MEMBERS

Nguyen Dai An - ML/NLP Specialist
Nguyen Tien Nhan - Data Engineer
Le Mai Thanh Son - Backend Developer
Nguyen Xuan Truong - DevOps Engineer
Nguyen Son Giang - Algorithm Specialist

INTRODUCTION

Background and context

- Social media platforms drive the need for creative, including **pairing music with images**.
- Manual music selection for images is **inconsistent and time-consuming**, leading to poor personalization.
- Most music recommendation work **relies on user behavior**; very little research connects visual input with music retrieval.

Problem statement and motivation

- Vibelens aims to **bridge the gap between visual content and musical aesthetics** by retrieve semantically relevant songs via vector similarity search.
- Our motivation is to create **a seamless and emotionally/ contextually personalized music experience** for social media users.

Objectives

- Enable users to upload an image and **receive 5–10 recommended songs.**
- Ensure that the **recommendations reflect the semantic content** and emotional tone of the image.
- **Build a scalable and responsive system** suitable for integration with social media or content creation platforms.

LITERATURE REVIEW – EXISTING SOLUTIONS

- Music recommendation systems typically **rely on collaborative filtering, audio-based features**, or metadata analysis (Schedl et al., 2018).
- Vision-Language Models (VLMs) can **convert images into semantically rich text** for downstream tasks (Ghosh et al., 2024)
- Vector databases (e.g., Pinecone) **enable fast semantic similarity search** in high-dimensional spaces (Jie et al., 2023).

Reference

[1] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions, 2024. URL https://arxiv.org/abs/2404.07214.

[2] James Jie, Jianguo Wang, Guoliang Li, and James Pan. Survey of Vector Database Management Systems. URL https: //arxiv.org/pdf/2310.14021.

[3] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. International Journal of Multimedia Information Retrieval, 7(2):95–116, Apr 2018. doi: https://doi.org/10.1007/s13735-018-0154-2.

RESEARCH GAP & VIBELENS CONTRIBUTION

Research Gap

- Few studies connect visual content with musical context.
- Lack of integrated systems combining CV + NLP + music semantics for real-time recommendations.

Vibelens Contribution

- Uses image captioning + semantic vector matching to recommend emotionally aligned songs.
- Enrich user experience on multimedia platforms through image-to-music retrieval.

VIBELENS | VINUNIVERSITY
College of Engineering and Computer Science

METHODOLOGY – APPROACH & TECHNOLOGIES

Overall Approach

- Analyze image content using **vision-language models** (**ViT-GPT2**) to generate captions.
- Match **image-derived captions to song lyrics** using semantic similarity.

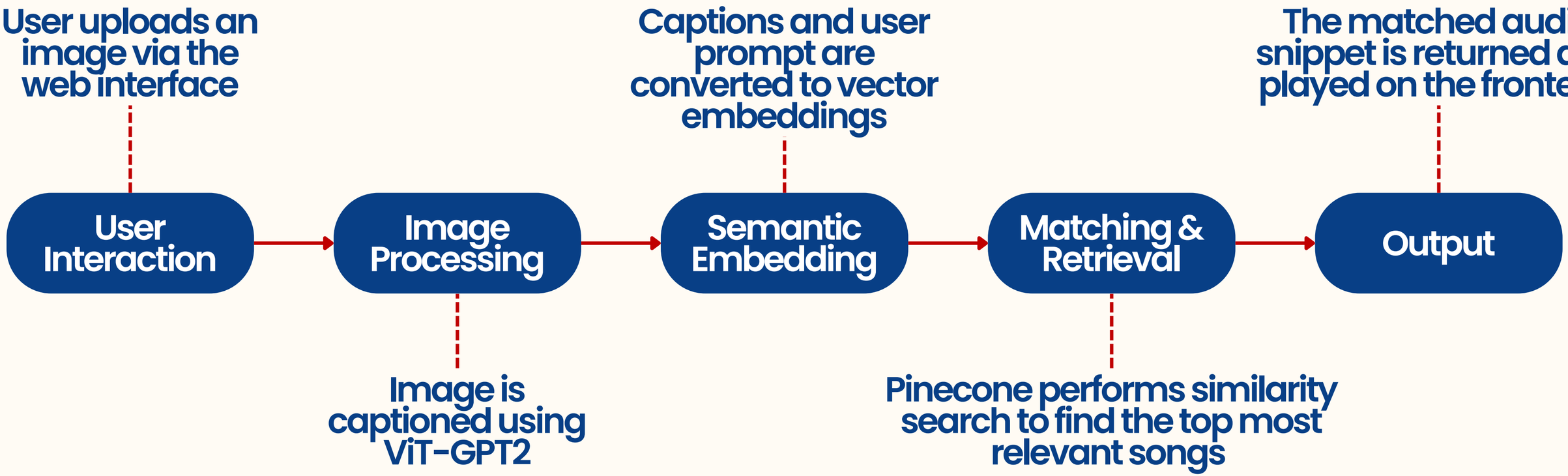
Core Technologies

- **Image Captioning:** ViT-GPT2
- **Text Embedding:** Sentence Transformers DistilUSE
- **Vector Search:** Pinecone vector database
- **Web Frameworks:** Flask (backend), Next.js (frontend)
- **Infrastructure:** Docker Swarm, Celery, Redis, PostgreSQL, MinIO, Kafka

VIBELENS

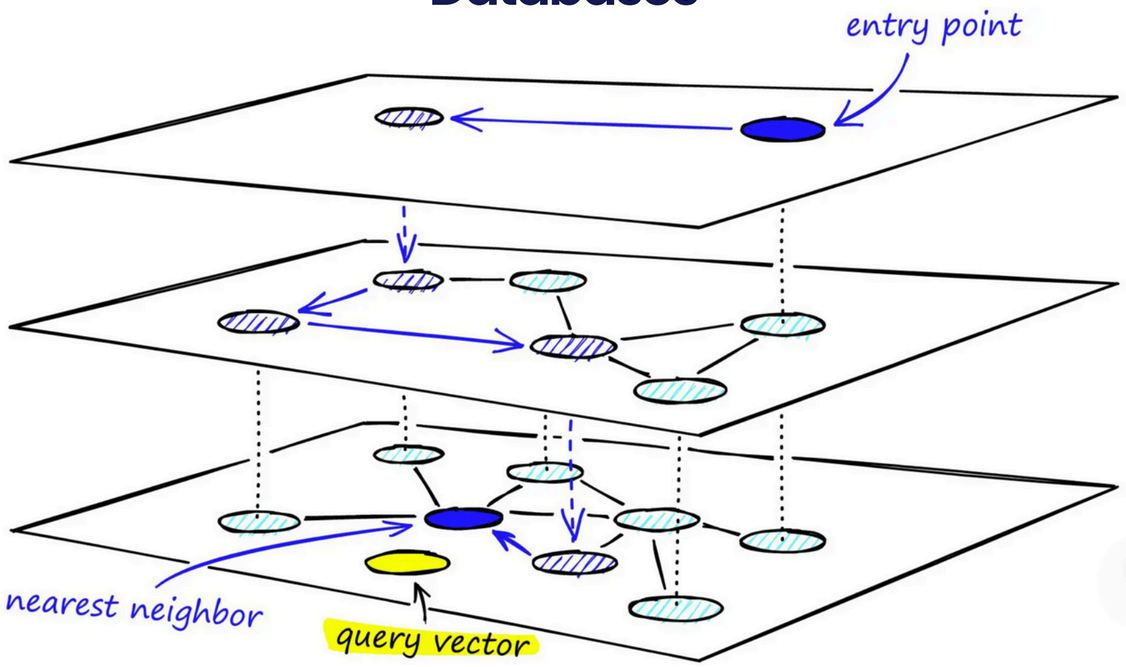
VINUNIVERSITY
College of Engineering and Computer Science

METHODOLOGY – SYSTEM WORKFLOW

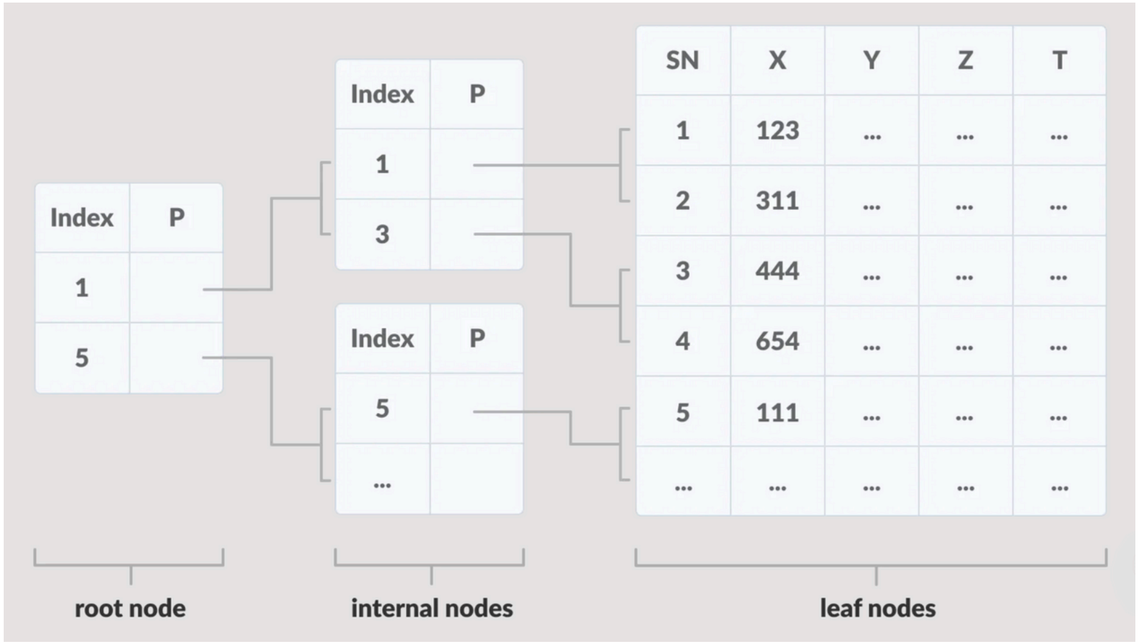


METHODOLOGY – ALGORITHMS & MODELS

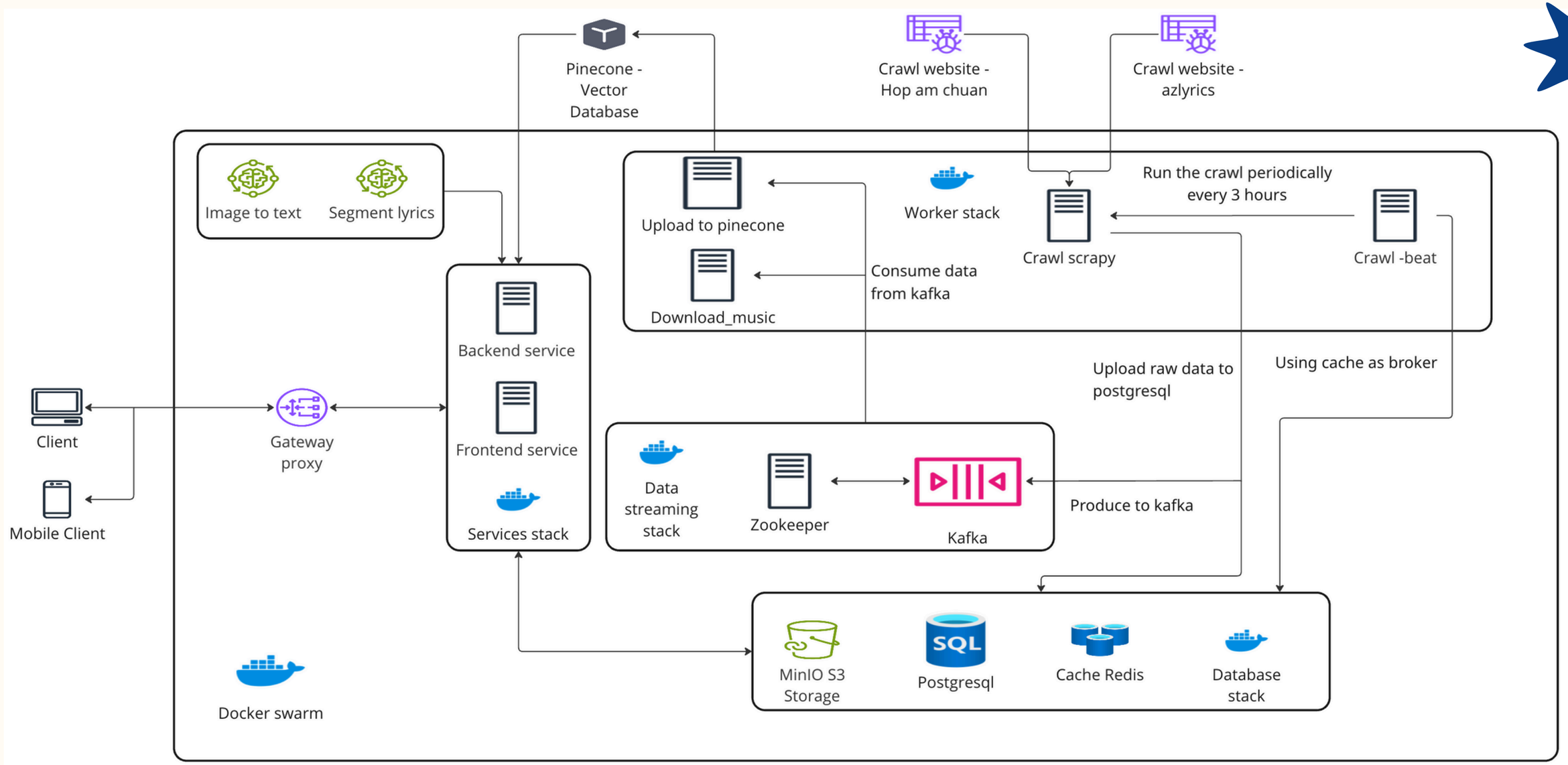
Hierarchical Navigable Small Worlds(HNSW) indexing in Vector Databases



B-trees & Hash Indexes



IMPLEMENTATION – SYSTEM ARCHITECTURE



IMPLEMENTATION

KEY MILESTONES

Week	Milestones
1 - 2	Project setup, team roles, initial research
3 - 4	Web crawler (Scrapy + Celery), MP3 & lyrics collection
5 - 6	Image captioning (ViT-GPT2), vector embedding & Pinecone integration
7 - 8	Frontend development (Next.js), backend integration (Flask)
9 - 10	Testing, bug fixes, and final presentation

CHALLENGES & SOLUTIONS

- **Infrastructure Limitations**

Running large models required more compute than available → Used local machines and optimized workflows with Docker Swarm.

- **Deployment Complexity**

Coordinating services (Flask, Redis, Kafka, Celery) caused integration bugs → Applied CI/CD automation and service orchestration.

- **Data Quality Issues**

Scraped lyrics often had formatting issues and mismatched audio → Implemented data cleaning and alignment scripts.

- **Model Adaptation**

No existing model was trained to link image content with music → Adapted general VLMs and bridged gap using semantic similarity.

RESULTS – VIDEO DEMO

VIBELENS

VINUNIVERSITY
College of Engineering and Computer Science

RESULTS – ANALYSIS

Final Outputs & Key Findings

- A **functional web-based system** that recommends emotionally relevant song segments from user-uploaded images.
- **Successfully integrated image captioning (ViT-GPT2), semantic matching via Pinecone**, and an end-to-end music retrieval pipeline.
- Developed a scalable backend using Flask, Kafka, Redis, and Docker Swarm; frontend built with Next.js.
- **Collected and processed a curated dataset of lyrics and MP3 files**, segmented and cleaned for precise recommendation.

Comparison with Expectations

Expectations

- Accurate, fast, personalized music recommendations.
- Full-stack deployment with modular components.
- Use pretrained model specific to music-visual matching.

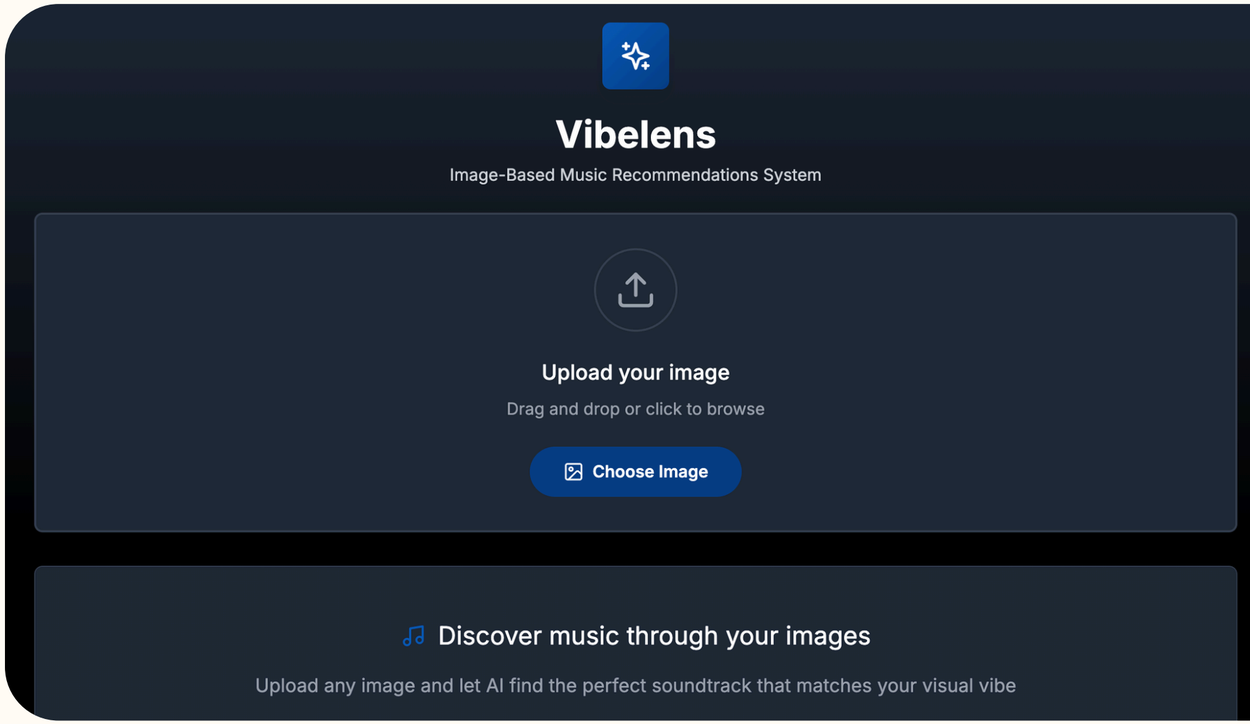
Achievements

- Good semantic relevance and <3s latency in most cases; some limitations due to compute and data noise.
- Engineered a workaround using general-purpose VLMs + custom similarity logic.

DISCUSSION

Project Limitations

- **Infrastructure Constraints:** Limited budget restricted access to high-performance GPU/cloud servers, resulting in slower inference and limited testing capacity.
- **Data Quality Issues:** Scraped lyrics often contained noise or inconsistent formatting. Some MP3 files were outdated or did not align properly with lyrics.
- **Model Generalization:** Used general-purpose vision-language models (e.g., ViT-GPT2) that were not specifically trained for music-related tasks, reducing alignment accuracy.
- **Limited Evaluation Scope:** Reliance on cosine similarity and qualitative feedback for evaluation, without large-scale user testing or A/B benchmarking.
- **Language and Cultural Bias:** Most dataset entries were in Vietnamese, limiting diversity in music genres, languages, and emotional expression.



CONCLUSION & FUTURE WORK

Recap of Achievements

- Built a **complete image-to-music recommendation system** using vision-language models, semantic vector search, and a modular full-stack architecture.
- **Successfully connected visual semantics with music content**, delivering emotionally resonant song suggestions.
- **Developed and deployed core components**: image captioning, data crawling, audio segmentation, vector embedding, and web interface.

Broader Implications

- Vibelens demonstrates a novel way to **enhance multimedia experiences through AI-driven content personalization**.
- **Has potential for integration into social media**, photo apps, or creative tools to enrich user-generated content.
- **Showcases how cross-modal AI** (vision + language + audio) can unlock deeper contextual understanding.

Future Development

- **Optimize latency and scalability** for real-world deployment (e.g., GPU inference, cloud scaling).
- **Expand the dataset** with global music diversity and higher-quality metadata.
- **Enable multi-language support** and mood-based filters for more personalized user experience.

VIBELENS | VINUNIVERSITY
College of Engineering and Computer Science

Acknowledgments

Our Progress

We would like to express our sincere gratitude to the individuals who supported and guided us throughout the development of Vibelens:

Supervisors & Faculty:

- Prof. Kok-Seng Wong – Project Supervisor
- Prof. Pham Huy Hieu – Course Instructor (COMP3080)

Our Team:

- Nguyen Dai An – ML/NLP Specialist
- Nguyen Tien Nhan – Data Engineer
- Le Mai Thanh Son – Backend Developer
- Nguyen Xuan Truong – DevOps Engineer
- Nguyen Son Giang – Algorithm Specialist

Thank you for your guidance, dedication, and teamwork!

Thank You So Much!

Any question?