Seungwan, Hong

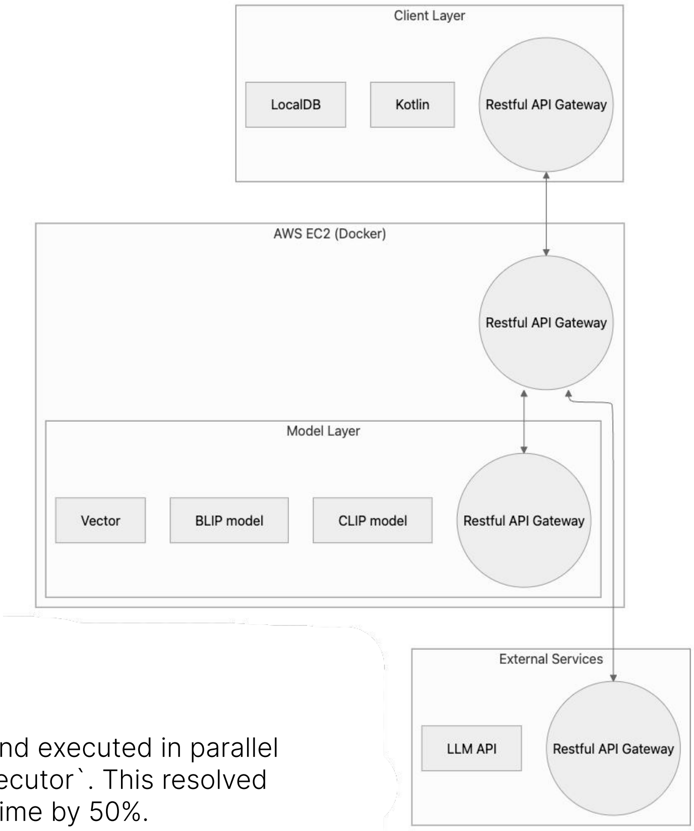
www.hosugator.com

Sodam Diary

Voice-Based Personalized Photo Narration Diary Service for the Visually Impaired
(2025 Korea Disabled People Hackathon Finalist Project)

System Architecture

Modulized core functionalities (Authentication, Map, Data Collection, NLP Processing, Recommendation) to ensure system flexibility and scalability.
Docker was used to build the local integration environment, and the final deployment was executed on an AWS EC2 Instance.



Core Technology

Multi-Model Pipeline for Image Analysis

- Object and Caption Generation: Utilized the BLIP model to generate specific and factual image descriptions.
- Atmosphere/Sentiment Tagging: Applied the CLIP model and 30+ pre-defined atmosphere texts to extract and assign the Top-3 image sentiments based on vector similarity.

RAG-based Personalized Narration

- STT-based Data Injection: Extracted key information from user voice input using STT.
- Dynamic Prompt Construction: Prompt Engineering to dynamically structure prompt using extracted user information.
- RAG Application: Integrated RAG to tag objects based on input data and return a personalized sentence.

Asynchronous Processing

The BLIP and CLIP model inference processes were separated and executed in parallel threads using the `concurrent.futures` module's `ThreadPoolExecutor`. This resolved synchronous blocking issues and reduced the overall response time by 50%.

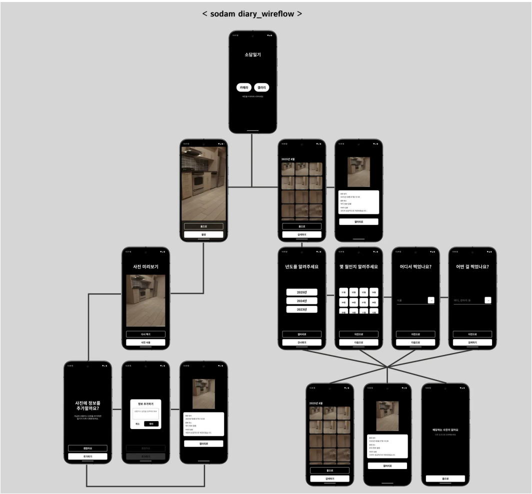
Problem Solving

Cost Reduction

To mitigate the high operational cost of exclusive LLM (GPT-4V) usage (approx. ₩1,300,000 per month for 50,000 responses), a multi-model structure utilizing more affordable open-source models was adopted. This achieved a total cost reduction of over 30%.

Performance Optimization

The average response time, which reached 30 seconds with exclusive LLM use, was reduced by 50% through the implementation of the multi-model parallel processing architecture.
4-bit NF4 Quantization was conditionally applied using `BitsAndBytesConfig` during BLIP/CLIP model loading (on non-Mac environments) to optimize memory usage and loading time.



Retrospective

Participating in a public service development project targeting the visually impaired user segment, I spearheaded the challenge to simultaneously achieve two goals: optimizing 'User Experience' and ensuring 'Business Viability (Cost Efficiency)', mirroring a real startup environment. The most significant experience was recognizing the high-cost problem and proposing and implementing the multi-model structure to solve the business challenge through technical means.
Future Enhancement Goals

- Personalization System Advancement: Implementing a personalized tagging recommendation system (Future work) by improving the DB structure to suggest names (e.g., of a dog, person) when similar objects are detected based on previous user input.
- Performance Optimization Completion: Maximizing model inference performance through OpenVINO integration to minimize voice guidance latency for visually impaired users.

Figma Jetpack Compose Hugging Face django Kotlin asyncio POSTMAN