Smart Media Journal / Vol.15, No.4 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2026.15.4.84

B2B 플랫폼의 레거시 환경을 고려한 계층형 멀티모달 추천 모델의 마이크로서비스(MSA) API 설계 및 구현

(Design and Implementation of Microservice API for Hierarchical Multimodal Recommendation Model in Legacy B2B Platform Environment)

김재경*

(Jae Kyung Kim)

요 약

최근 B2B 플랫폼에서는 안정성과 호환성을 유지하면서도 최신 AI 기술을 접목하려는 시도가 늘고 있다. 이에 본 논문에서는 해외 바이어의 입력 이미지를 바탕으로 시각적·의미적 유사도가 높은 국내 스타트업 제품을 실시간 추천하는 마이크로서비스 기반 API 아키텍처를 설계하였다. 특히, 이미지 업로드나 URL 입력 시 계층형 CLIP 모델을 거쳐 최적의 제품군(Top-K)을 도출하는 과정을 실증하였다. 기존 Java/Spring 기반 모놀리식 구조에 GPU 연산 집약적인 멀티모달 모델을 직접 통합할 경우, 추론 지연에 따른 병목 현상과 시스템 간 높은 결합도 문제가 발생한다. 본 연구에서는 이러한 구조적 제약을 극복하고자 Python 기반의 독립적 추론 서버를 구축하고, RESTful API를 통해 레거시 환경과의 유연한 연동을 꾀하였다. 모델 사전 적재, 비동기 요청 처리, 장애 격리 구조를 적용하여 실시간 추천 서비스에 적합한 운영 환경을 구현하였다.

성능 평가 결과, 평균 응답 지연 시간은 210ms로 측정되었으며 초당 40건 이상의 트래픽에도 안정적인 처리량을 보였다. 이는 일반적인 웹 서비스 인터랙션 임계치인 300ms를 하회하는 수치로, 실무 환경에서도 사용자에게 끊김 없는 추천 경험을 제공할 수 있음을 시사한다. 본 연구는 고연산 멀티모달 추천 모델을 레거시 B2B 플랫폼 환경에 무중단으로 통합·운영할 수 있음을 실증적으로 제시한다.

■ 중심어 : 마이크로서비스 아키텍처 ; 계층형 멀티모달 추천 ; B2B 플랫폼 ; MLOps ; 레거시 시스템 통합

Abstract

Recent efforts in B2B platforms have focused on integrating cutting-edge AI technologies while maintaining system stability and compatibility. This study designs and implements a microservice-based API architecture that provides real-time multimodal recommendations of domestic startup products based on images provided by overseas buyers. Specifically, it demonstrates the process of deriving optimal product sets (Top-K) through a hierarchical CLIP model when a user uploads an image or enters a URL. Directly integrating GPU-intensive multimodal models into existing Java/Spring-based monolithic structures often leads to inference latency and high system coupling. To overcome these structural constraints, this research established a Python-based independent inference server and ensured flexible interworking with the legacy environment via RESTful APIs. An operational environment suitable for real-time services was realized by applying model preloading, asynchronous request processing, and fault isolation mechanisms.

Performance evaluation results show an average response latency of 210 ms and stable throughput of over 40 requests per second. This latency falls below the general web service interaction threshold of 300 ms, suggesting that a seamless recommendation experience can be provided in practical business environments. This study empirically demonstrates that high-computation multimodal recommendation models can be integrated and operated without service disruption in legacy B2B platform environments.

■ keywords : Microservice Architecture ; Hierarchical Multimodal Recommendation ; B2B Platform ; MLOps ; Legacy System Integration

Ⅰ. 서 론

최근 인공지능 기반 추천 기술은 이미지와 텍스트를 결합한 멀티모달(multimodal) 접근법으로 빠르

* 정회원, 한남대학교 경영정보학과
이 논문은 2025년도 한남대학교 교비학술연구비 지원에 의하여 연구되었음.

접수일자 : 2025년 12월 25일
수정일자 : 1차 2026년 01월 16일, 2차 2026년 02월 07일
게재확정일 : 2026년 03월 05일

게 진화하고 있다. 특히 시각 정보와 언어 정보를 동일한 표현 공간에 정렬하는 대규모 사전학습 모델의 등장은, 사용자가 복잡한 제품 정보를 직관적으로 이해하고 의사결정을 수행할 수 있도록 지원하는 핵심 기술로 주목받고 있다[1,2]. 정보 비대칭성이 높은 해외 바이어 - 국내 스타트업 매칭 플랫폼 환경에서 이러한 멀티모달 추천 기술은 기존의 키워드 중심 텍스트 검색이 갖는 한계를 보완할 수 있는 효과적인 대안으로 평가된다[3,4]. 특히 언어적 장벽이나 문화적 차이로 인해 텍스트만으로는 제품의 특성과 활용 맥락을 충분히 파악하기 어려운 글로벌 B2B 거래 상황에서, 이미지 기반의 직관적 추천은 매칭 성공률을 좌우하는 중요한 요인으로 작용한다.

본 연구에서 다루는 정보 비대칭성은 해외 바이어가 국내 스타트업 제품의 상세 사양, 기술적 맥락, 활용 가능성을 충분히 파악하기 어려운 상황을 의미하며, 제한적인 텍스트 정보에 의존한 기존 탐색 방식으로는 제품 간 비교와 의사결정에 한계가 존재하는 문제를 포함한다.

한편, 다수의 B2B 플랫폼은 거래의 안정성과 기존 시스템과의 호환성을 최우선으로 고려하여 Java/Spring 기반의 보수적인 레거시(monolithic) 아키텍처를 유지하고 있다. 반면 최신 딥러닝 기반 멀티모달 모델은 Python 실행 환경과 GPU 가속을 필수적으로 요구하므로, 이를 기존 시스템 내부에 직접 통합하는 것은 기술 스택 간 불일치와 구조적 제약을 수반한다. 이러한 이기종 환경 통합 문제는 시스템 복잡도를 증가시키고, 배포 및 운영 과정에서 장애 전파 가능성을 높이는 주요 원인으로 지적되어 왔다. 이에 따라 최근에는 모델 추론 기능을 독립적인 API 서버로 분리하고, 기존 플랫폼과는 느슨하게 결합(loose coupling)하는 방식이 현실적인 대안으로 논의되고 있다[5,6]. 이러한 접근은 고연산 AI 모델을 서비스 흐름에서 분리함으로써, 기존 B2B 플랫폼의 안정성을 유지하면서도 실시간 추천 기능을 확장할 수 있다는 점에서 주목받고 있다.

국내에서도 이미지 및 멀티모달 정보를 활용한 추천 기술이 영상 처리, 환경 정보 분석 등 다양한 응용 분야로 확장되고 있으며[7,8], 이는 멀티모달 추천 기법의 실무적 활용 가능성을 보여준다. 그러나 대부분의 연구는 모델 정확도나 알고리즘 성능 향상에 초점을 두고 있어, 실제 레거시 기반 플랫폼 환경에서의 실시간 운영 가능성이나 시스템 수준의 성능 검증은 충분히 다루어지지 않았다.

선행 연구[9]에서는 계층형 CLIP 기반 멀티모달 추천 모델을 제안하고, 단일 스테이지 모델 대비 추천 정확도가 유의미하게 향상됨을 Precision@10, MAP@10, NDCG@K10 등의 지표를 통해 정량적으로 검증하였다. 해당 연구는 복잡한 B2B 제품 도메인에서 계층적 추론 구조가 추천 성능 향상에 기여함을 실험적으로 입증하였다는 점에서 중요한 학술적 의의를 갖는다. 그러나 모델 구조와 알고리즘적 성능 검증에 초점을 둔 선행 연구의 특성상, 실제 B2B 플랫폼 환경에서의 실시간 서비스 적용 가능성, 시스템 지연 시간(latency), 처리량(throughput), 그리고 운영 안정성에 대한 분석은 포함되지 않았다.

본 연구는 이러한 한계를 보완하는 후속 연구로서, 해외 바이어의 이미지 기반 제품 탐색 시나리오를 전제로, 선행 연구[9]에서 검증된 계층형 CLIP 기반 멀티모달 추천 모델을 실시간 API 서버 형태로 구현하고, 이를 레거시 기반 B2B 플랫폼에 통합·운영하기 위한 시스템 아키텍처를 제안한다.

본 논문은 추천 알고리즘의 정확도 향상 자체보다는, 고성능 멀티모달 추천 모델을 실제 레거시 기반 B2B 플랫폼 환경에서 안정적으로 통합·운영할 수 있는 시스템 설계와 구현에 초점을 둔다.

본 연구에서는 추천 모델을 독립적인 마이크로서비스(MSA)로 분리하고, 비동기 추론 파이프라인과 단계적 배포 전략을 적용함으로써, 기존 서비스의 안정성을 저해하지 않으면서도 고성능 추천 모델을 운영할 수 있는 실무적 설계 방안을 제시한다. 이를 통해 알고리즘 성능을 넘어, 실제 서비스 환경에서의 적용 가능성과 확장성을 중심으로 한 공학적·운

영적 기여를 목표로 하여, 다음의 연구 질문에 답하고자 한다.

연구 질문 1. 고성능 계층형 멀티모달 추천 모델을 레거시 기반 B2B 플랫폼 환경에 실시간으로 통합·운영하는 것이 가능한가?

연구 질문 2. 계층형 추론 구조는 비계층형 방식 대비 시스템 성능(지연 시간, 처리량)과 운영 안정성 측면에서 실질적인 이점을 제공하는가?

Ⅱ. 관련 연구

본 장에서는 해외 바이어 - 국내 스타트업 매칭 플랫폼을 위한 계층형 멀티모달 추천 API 서버 구현과 관련하여, 멀티모달 추천 모델의 발전 흐름과 계층형 구조, 그리고 이를 지원하기 위한 마이크로서비스 기반의 시스템 통합 기술을 중점적으로 검토한다.

1. 멀티모달 추천 모델과 계층형 추론 구조

전자상거래 및 B2B 플랫폼 환경에서 이미지와 텍스트 정보를 결합한 멀티모달 추천은 텍스트 검색의 '희소성'과 이미지 검색의 '의미적 모호성'을 상호 보완하는 핵심 기술이다. 초기에는 시각적 특징과 텍스트를 단순 연결하는 방식이 주를 이루었으나, 최근 CLIP 계열 모델[1]과 같이 대규모 데이터를 대조 학습하여 두 모달리티를 동일한 잠재 공간에 투영하는 방식이 표준으로 자리 잡았다. 후속 연구들은 노이즈가 포함된 환경에서도 강건한 성능을 입증하였으며[2], FashionBERT[3]와 같이 도메인 특화 모델이나 단계적 융합(fusion) 방식[4]을 통해 추천 정밀도를 높이고 있다. 국내에서도 딥러닝 기반 영상처리[7]나 환경 인자 추천[10] 등 다양한 산업 분야로의 적용이 활발하다.

한편, 수백만 개 이상의 아이템을 다루는 대규모 B2B 환경에서는 실시간성을 위해 후보 생성(Retrieval)과 재순위화(Ranking)를 분리하는 계층형 구조가 필수적이다. YouTube[11]나 제품 검색[12] 분야에서 입증된 이 구조는 탐색 공간을 단계적으로 축소하여 연산 효율을 극대화한다. 또한 협업 딥러닝 모델[13]을 통해 사용자 상호작용과 콘텐츠 특징을 결합함으로써 추천의 표현력을 높일 수 있다. 본 연구는 이러한 계층형 접근 방식을 CLIP 기반 임베딩과 결합하여 레거시 파이프라인에 적용한다.

2. API 기반 모델 서빙 및 마이크로서비스 아키텍처

딥러닝 모델을 실제 서비스에 배포할 때 애플리케이션과의 강한 결합은 유연성을 저해한다. 이를 해결하기 위해 Clipper[5]나 TensorFlow Serving[6]과 같은 API 기반 모델 서빙 패턴이 확산되고 있으며, 이는 트래픽 증가에 따른 유연한 확장과 시스템 안정성을 보장한다[14,15]. 특히 서로 다른 기술 스택을 사용하는 레거시 시스템과의 통합을 위해서는 서비스 간 느슨한 결합을 강조하는 마이크로서비스 아키텍처(MSA)가 가장 적합한 전략이다[16][17].

대규모 분산 환경에서는 Kubernetes와 같은 컨테이너 오케스트레이션[18]이 표준 운영 환경으로 자리 잡았으며, 꼬리 지연(Tail Latency)[19] 문제를 해결하기 위한 시스템 설계가 중요하다. 나아가 AI 도입 시 발생하는 기술 부채[20]를 관리하기 위해 MLOps 파이프라인[21]과 지속적인 재학습 및 성능 모니터링 체계[22]가 요구된다. 시스템의 신뢰성을 담보하기 위해서는 단순 모델 정확도 외에도, 해당 시스템이 실제 운영 환경에 적합한지를 판별하는 정량적 테스트와 관련 평가 지표[23]를 체계적으로 갖추어야 한다.

3. 선행 연구와의 차별점

기존의 딥러닝 추천 연구들은 주로 모델의 정확도 향상이나 범용 서빙 프레임워크 제안에 집중해 왔다. 반면, 본 연구는 고성능 멀티모달 모델을 안정성이 최우선인 레거시 B2B 플랫폼에 이식하는 실증적

과정에 초점을 둔다. 구체적으로 첫째, MSA 패턴을 통해 Java/Spring 레거시 환경과 Python/GPU AI 환경 간의 기술적 불일치를 해결하는 구체적인 아키텍처를 제시한다. 둘째, 선행 연구[9]의 계층형 알고리즘을 실제 서비스 가능한 엔드투엔드 시스템으로 구현한다. 셋째, 단순 추론 속도가 아닌 전체 서비스 관점에서 성능을 검증하여 실무 적용 가능한 참조 모델을 제공한다. 이는 알고리즘 연구와 시스템 연구 사이의 간극을 메우는 데 의의가 있다.

Ⅲ. 시스템 설계 및 구현

본 장에서는 레거시 B2B 플랫폼 환경을 유지하면서 대규모 멀티모달 추천 기능을 안정적으로 제공하기 위한 시스템 설계 원칙과 구현 구조를 설명한다. 특히 본 장에서는 제안하는 구조가 개념적 설계에 머무르지 않고, 실제 서비스 환경에서 실행 가능한 추천 API 서버로 구현되었음을 전제로 시스템 구성과 동작 흐름을 함께 제시한다.

1. 전체 시스템 아키텍처

그림 1은 본 연구에서 구현한 전체 시스템 아키텍처를 나타낸다. 본 시스템은 클라이언트, 레거시 플랫폼, 추천 API 서버, 데이터 저장 계층 간의 관심사 분리(Separation of Concerns)를 극대화한 아키텍처를 지향한다. 사용자 인증 및 비즈니스 로직을 처리하는 레거시 플랫폼과 고부하 GPU 추론을 전담하는 API 서버를 이기종 간 독립적 계층으로 구성하였다. 이러한 설계는 연산 집약적인 추론 프로세스가 핵심 트랜잭션 성능을 저해하지 않도록 물리적·논리적으로 격리하여 전체 시스템의 가용성을 보장한다.

본 시스템에서 사용한 제품 데이터베이스는 총 11,414개의 국내 스타트업 제품으로 구성되며, 238개의 중분류 카테고리 구조를 가진다. 계층형 추론 구조의 1단계에서는 전체 탐색 공간 중 평균 2~3% 수준의 후보군만이 2단계로 전달되도록 설계·구현되어, 대규모 제품 데이터 환경에서도 연산 비용을 효과적으로 감소시킨다. 이러한 구조는 이후 실험 결과에서 확인되듯이 실시간 추천 서비스를 가능하게 하는 핵심 요소로 작용한다.

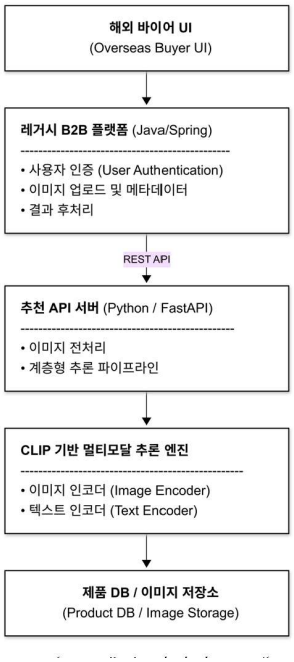


그림 1. 해외 바이어 - 국내 스타트업 매칭 플랫폼 전체 시스템 아키텍처

2. 계층형 CLIP 기반 추천 파이프라인

실시간 응답성을 확보하기 위해 그림 2와 같은 계층형 CLIP 기반 추론 파이프라인을 설계하고 이를 실제 추천 API 서버에 구현하였다.

1단계(Level-1): 입력 이미지와 상위 카테고리 텍스트 임베딩 간의 유사도를 계산하여 탐색 공간을 대폭 축소한다. 이는 전체 데이터베이스를 전수 검색(Full Scan)하는 방식에 비해 연산 비용을 크게 줄여, 대규모 B2B 데이터 환경에서도 실시간 추천이 가능하도록 하는 핵심 기제이다.

2단계(Level-2): 1단계에서 선별된 소수의 후보군을 대상으로 제품 레벨의 세부 임베딩 유사도를 계산하여 최종 Top-K 추천 결과를 도출한다. 이 단계에서는 제품의 세부적인 시각적 특징과 의미적 정보를 함께 고려함으로써 추천의 정밀도를 극대화한다.

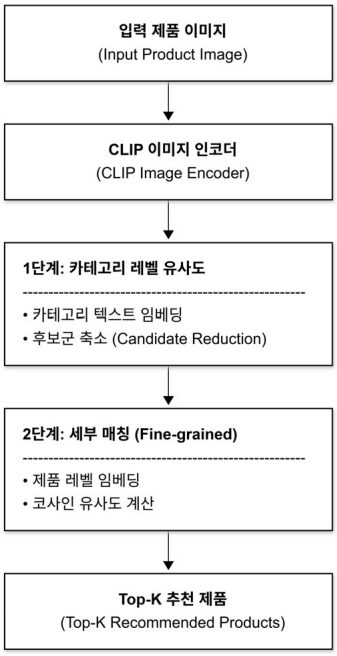


그림 2. 계층형 CLIP 기반 멀티모달 추천 파이프라인

3. 추천 API 서버 설계 및 호출 흐름

그림 3은 레거시 플랫폼과 추천 API 서버 간의 호출 시퀀스를 나타낸다. 추천 API 서버는 RESTful API 형태로 구현되어 레거시 플랫폼과 통신하며, 실제 서비스 환경에서 실행 가능한 형태로 구성되었다.

경량화 통신 및 데이터 흐름: 네트워크 대역폭 부하를 최소화하기 위해 이미지 파일 자체(Binary)를 직접 전송하는 대신, 공유 스토리지(S3 등)에 저장된 이미지 경로와 메타데이터만을 JSON 형태로 전달하는 방식을 채택하였다. 이는 대규모 실시간 추론 환경과 딥러닝 추천 시스템 서빙에서 권장되는 효율화 전략이다[14,15].

장애 격리 및 지연 관리: 레거시 플랫폼으로부터 요청을 수신한 추천 API 서버는 이미지를 비동기로 로드한 후 추론을 수행한다. 이 과정에서 Dean과 Barroso[19]가 지적한 꼬리 지연(Tail Latency) 문제를 완화하기 위해 엄격한 타임아웃 정책을 적용하였으며, Circuit Breaker 패턴을 도입하여 API 서버의 장애가 레거시 플랫폼 전체로 전파되는 것을 차단하였다.

이와 같은 분리 구조를 통해 추천 기능은 레거시 플랫폼의 트랜잭션 처리 흐름과 독립적으로 운영되도록 구현되었다. 추천 모델의 배포 및 업데이트 또한 플랫폼의 핵심 서비스에 영향을 주지 않으며, 이러한 효과는 이후 성능 평가에서 추천 API의 지연이나 오류가 플랫폼 서비스에 영향을 주지 않음을 통해 실증적으로 확인되었다.

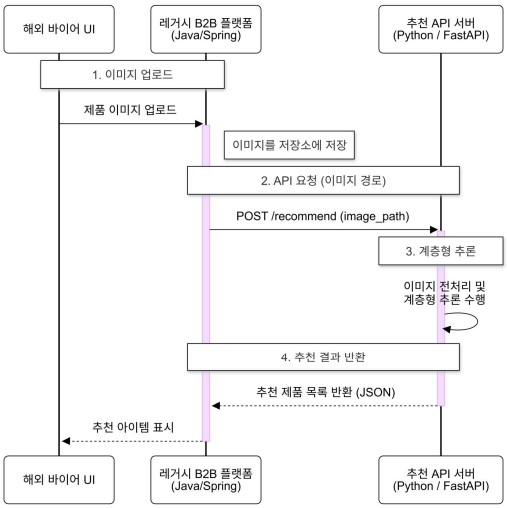


그림 3. 레거시 플랫폼 - 추천 API 서버 간 호출 시퀀스

4. 배포 및 운영 구조

그림 4는 추천 API 서버의 배포 및 운영 구조를 나타낸다. 추천 API 서버는 Docker 컨테이너로 패키징되어 Kubernetes 환경에서 오케스트레이션되며[18], 트래픽 변동에 따른 수평적 오토스케일링과 무중단 배포를 지원한다. 또한 보안 효율성을 고려하여 VPC 내부망을 구성하고, MLOps 파이프라인을 통해 추론 지연 시간, GPU 점유율, 처리량을 실시간으로 모니터링하도록 구현하였다.

다만 본 연구의 기여는 운영체제나 컨테이너 플랫폼이 제공하는 일반적인 배포·운영 기능 자체에 있지 않다. 본 연구는 이러한 인프라 환경 위에서 계층형 CLIP 기반 멀티모달 추천을 수행하는 실행 가능한 API 서비스를 설계·구현하고, 이를 실제 레거시 기반 B2B 플랫폼 환경에 통합하여 안정적으로 운영 가능함을 입증하였다는 점에서 차별성을 가진다.

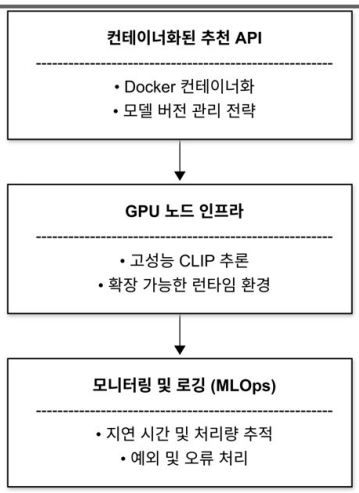
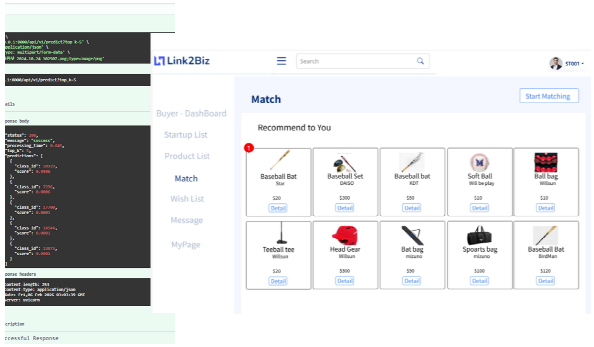


그림 4. 추천 API 서버의 배포 및 운영 구조

5. 추천 API 실행 흐름 및 인터페이스

본 연구에서 설계한 추천 API는 실제 서비스 환경에 즉시 투입 가능한 수준의 완성도로 구현되었다. 해외 바이어가 특정 이미지로 추천을 요청하면, 레거시 플랫폼은 RESTful API를 경유하여 해당 요청을 추천 서버로 라우팅한다. 이후 추천 서버 내에서는 이미지 전처리, 멀티모달 임베딩 생성, 계층형 후보군 필터링이 비동기적으로 수행되며, 최종 산출된 추천 리스트는 JSON 포맷으로 응답되어 레거시 시스템과의 데이터 정합성을 유지한다.

그림 5는 개발된 추천 API의 실제 구동 양상을 시각화한 것이다. 그림 5(a)는 Postman 환경에서 추천 API 엔드포인트를 호출하여 요청 파라미터와 JSON 응답이 정상적으로 처리되는지를 확인한 실행 화면이며, 요청에는 이미지 URL(또는 업로드 이미지)과 Top-K 값 등의 입력이 포함된다. 추천 서버는 이미지 로딩 - 임베딩 추론 - 1단계 후보군 축소 - 2단계 재순위화 과정을 거쳐 Top-K 추천 결과를 반환하고, 응답에는 추천 제품의 제품 ID, 제품명, 카테고리, 유사도 점수, 대표 이미지 URL(및 처리 시간 등)이 포함되어 API 단독 수준에서도 기능과 결과 정합성을 검증할 수 있다. 그림 5(b)는 동일한 추천 결과가 레거시 기반 B2B 플랫폼 UI에 리스트/카드 형태로 렌더링된 화면을 보여주며, 해외 바이어가 입력한 이미지와 함께 추천 제품의 핵심 메타정보(제품명/카테고리/대표 이미지 등)를 제공함으로써 실제 탐색·비교 의사결정에 활용되는 사용자 인터페이스를 제시한다. 구체적으로는 이미지 기반 요청의 수신부터 계층형 CLIP 모델의 추론 프로세스, 그리고 최종 Top-K 제품 정보가 반환되기까지의 실행 로그와 인터페이스를 포함한다. 이러한 실증적 결과는 본 연구의 설계가 이론적 제안을 넘어, 실제 운영 환경의 제약 조건을 극복하고 안정적으로 작동하는 API 서비스임을 방증한다.



(a) (b)

그림 5. 계층형 CLIP 기반 멀티모달 추천 API의 실제 실행 화면: (a) Postman 기반 추천 API 호출(요청 파라미터 및 JSON 응답 예시), (b) 레거시 기반 B2B 플랫폼 UI에 통합된 이미지 기반 Top-K 추천 결과 화면

Ⅳ. 실험 및 성능 평가

본 장에서는 제안 시스템의 실무 적용 가능성을 검증하기 위해 응답 지연 시간, 처리량, 그리고 운영 안정성 측면에서 성능 평가를 수행하였다.

1. 실험 환경 및 구성

본 실험은 제안 시스템의 실효성을 다각도로 검증하기 위해 고사양 서버 자원(Intel i9-12900K, RTX 3090, 64GB RAM) 환경에서 오픈소스 부하 테스트 도구인 Locust를 활용하여 수행되었다. 여기서 정의하는 '운영 환경과의 유사성'은 단순한 하드웨어 수치의 일치를 넘어, Link2Biz 플랫폼 내 추천 API의 직무적 역할과 서비스 부하 특성에 초점을 맞춘 것이다. 실제 운영 환경과 동일하게 이미지 다운로드, 멀티모달 추론, 유사도 연산 등 고부하 작업이

메인 서버와 분리된 독립적 추론 노드에서 처리되도록 설계하여 실험의 외적 타당성을 확보하였다.

특히, 실제 운영망에서는 트래픽 증량에 따른 수평적 확장(Scale-out)이 가능함에도 불구하고, 본 실험에서는 의도적으로 단일 GPU 서버 환경에서의 성능을 측정하였다. 이는 인프라 최적화가 배제된 상태에서 추천 API가 보장할 수 있는 최소 성능 기준(Lower Bound)을 보수적으로 도출하기 위함이다. 이러한 가혹 조건 하에서의 검증은 역설적으로 본 시스템이 실무 환경의 변동성 속에서도 안정적인 서비스 하한선을 유지할 수 있음을 입증하는 강력한 논거가 된다.

시스템은 Java 기반 레거시 플랫폼과 Python 기반 추천 API 서버가 RESTful API를 통해 통신하는 구조로 구성되었다. 이러한 컨테이너 기반 분리 구조는 대규모 추론 환경에서 시스템 간 결합도를 낮추고[5,6], 장애 발생 시 신속한 복구를 가능하게 하는 장점을 가진다[16].

본 실험 환경은 추천 API가 담당하는 주요 연산인 이미지 다운로드, 멀티모달 임베딩 추론, 유사도 계산을 단일 GPU 서버에서 처리하는 실제 서비스 구조를 기준으로 설정되었다. 특히 파일럿 운영 단계의 B2B 플랫폼에서 추천 기능이 독립적인 추론 서버로 분리되어 운영되는 구조를 반영함으로써, 시스템 역할과 부하 특성 측면에서 실제 운영 환경과의 유사성을 확보하고자 하였다. 최대 동시 요청 수는 파일럿 운영 기간 동안 관찰된 해외 바이어의 평균 동시 활성 세션 규모가 10~20명 수준이었음을 고려하여, 해당 값의 약 2~3배에 해당하는 50으로 보수적으로 설정하였다.

2. 성능 평가 결과

주요 성능 지표에 대한 실험 결과를 요약한 표 1을 통해 본 제안 시스템의 기술적 타당성을 검증하였다. 실험 결과, 기존의 전수 조사(Full Scan) 방식이 평균 620ms의 지연 시간을 기록한 데 반해, 계층형 구조를 적용한 제안 시스템은 약 210ms를 기록하며 지연 시간을 기존 대비 약 66% 수준으로 대폭 절감하는 성과를 거두었다. 이는 웹 기반 사용자 경험의 임계치로 통용되는 300ms 응답 기준을 안정적으로 충족하는 성능으로 , 실제 파일럿 운영 단계에서 요구되는 실시간 상호작용 요건을 충분히 상회함을 시사한다[24,25].

표 1. 계층형 CLIP 기반 추천 API 성능 평가 결과

평가 항목	측정 조건	결과
평균 응답 지연 시간	단일 요청	210 ms
95퍼센타일 지연 시간	단일 요청	280 ms
최대 동시 요청 수	이미지 기반 추천 요청	50 요청
처리량(Throughput)	안정 상태	약 40 req/s
오류 발생률	50명 동시 요청	0.5% 미만
서비스 영향도	레거시 플랫폼	영향 없음

응답 지연 시간 측면에서 보면, 이미지 다운로드부터 추론 완료까지의 전체 경로(End-to-End) 기준으로 단일 요청 시 평균 210ms의 응답 시간이 측정되었다. 또한 95퍼센타일 지연 시간은 280ms로 나타나, 간헐적인 네트워크 지연이나 연산 부하 상황에서도 꼬리 지연(Tail Latency)[19]이 효과적으로 억제되고 있음을 확인하였다. 이는 제안한 계층형 추론 구조가 실시간 서비스 환경에서 지연 시간 측면의 실질적인 개선 효과를 제공함을 보여준다.

처리량 및 동시성 측면에서는 50명의 동시 접속자가 지속적으로 요청을 보내는 부하 상황에서도 초당 약 40건(40 RPS)의 요청 처리가 가능하였다. 이는 파일럿 운영 단계에서 관찰된 실제 추천 요청 빈도를 기준으로 할 때 충분한 여유를 가지는 수준이며, 단일 서버 기준에서도 추천 기능이 병목 없이 제공될 수 있음을 보여준다[26]. 또한 B2B 서비스 특성상 추천 요청이 특정 시간대에 집중되는 경우에도 안정적인 서비스 제공이 가능함을 시사한다.

운영 안정성 측면에서는 테스트 과정에서 발생한 일부 예외 상황이 전체 요청의 0.5% 미만으로 관측되었으며, 해당 예외는 Circuit Breaker 메커니즘에 의해 즉시 차단되어 레거시 플랫폼으로 전파되지 않았다. 이는 마이크로서비스 아키텍처가 제공하는 장애 격리(Fault Isolation) 효과를 실증적으로 보여

주는 결과로, 추천 API 서버를 독립적으로 분리하여 운영하는 구조의 타당성을 뒷받침한다.

3. 추천 품질 평가

추천 시스템으로서의 실용성을 검증하기 위해 Precision@K, Recall@K, NDCG@K 지표를 사용하여 계층형 CLIP 모델과 단일 단계 CLIP 모델의 추천 성능을 비교하였다. 실험 결과, 계층형 구조는 모든 평가 지표에서 일관된 성능 향상을 보였으며, 이는 선행 연구에서 보고된 결과와 동일한 경향을 나타낸다.

표 2. 계층형 CLIP 기반 추천 모델과 단일 단계 CLIP 모델의 추천 품질 비교

평가 지표	단일 단계 CLIP	계층형 CLIP
Precision@10	.312	.348
Recall@10	.274	.309
NDCG@10	.356	.401

주: 본 평가는 선행 연구[9]와 동일한 데이터셋 및 실험 설정을 기반으로 재현한 결과이며, 본 논문에서는 해당 추천 품질을 전제로 시스템 및 서비스 성능을 평가하였다.

본 논문의 목적은 추천 알고리즘 자체의 성능을 새롭게 제안하거나 절대적인 정확도 향상을 검증하는 데 있지 않다. 선행 연구에서 이미 추천 품질의 유효성이 검증된 계층형 CLIP 기반 멀티모달 추천 모델을 전제로, 해당 모델이 실제 레거시 B2B 플랫폼 환경에서도 품질 저하 없이 안정적으로 실행·운영될 수 있는지를 시스템 및 운영 관점에서 검증하는 데 초점을 두었다. 따라서 본 절에서는 추천 품질 지표를 추가적인 비교 실험 대상으로 확장하기보다는, 실시간 응답성, 처리량, 장애 격리 요구 조건 하에서도 해당 추천 품질이 유지될 수 있음을 확인하는 것을 연구 범위로 설정하였다.

Ⅴ. 논의 및 결론

본 연구는 해외 바이어 - 국내 스타트업 매칭 플랫폼이라는 실제 B2B 서비스 환경을 대상으로, 해외 바이어가 업로드하거나 선택한 제품 이미지와 시각적으로 유사한 국내 스타트업 제품을 추천하는 멀티모달 추천 시스템을 설계·구현하고, 이를 레거시 기반 플랫폼에 통합하는 방안을 제시하였다. 본 논문은 추천 알고리즘의 단순한 정확도 비교를 넘어, 이미지 기반 제품 추천 기능을 실제 플랫폼 서비스로 안정적으로 통합·운영하기 위한 구조적·공학적 설계와 구현 과정에 초점을 두고 논의를 전개하였다.

1. 연구 결과 논의

그림 5에서 확인되듯이, 본 연구의 핵심 결과는 해외 바이어가 입력한 제품 이미지에 기반하여 시각적으로 유사한 국내 스타트업 제품을 실시간으로 추천하는 기능을 기존 레거시 플랫폼 구조에 무리 없이 통합할 수 있음을 실증적으로 제시한 것이다. 기존 연구들이 주로 멀티모달 표현 학습이나 추천 정확도 향상에 초점을 맞추어 온 반면[1][3], 본 논문은 해당 추천 모델이 실제 서비스 환경에서 어떻게 호출되고, 어떤 흐름으로 실행되며, 어떠한 형태로 결과를 반환하는지를 실행 화면과 함께 구체적으로 제시하였다.

특히 계층형 CLIP 기반 추론 구조는 대규모 제품 집합을 다루는 B2B 플랫폼 환경에서 실시간성을 확보하는 데 핵심적인 역할을 수행하였다. 이는 대규모 추천 시스템에서 후보 생성과 재순위화를 분리하는 계층형 구조가 효과적이라는 기존 연구 결과와 일관된 흐름을 보이며[9][11][12], 본 연구는 이러한 접근을 이미지 - 텍스트 멀티모달 추천 시나리오에 적용하여 실시간 API 서버 환경에서도 안정적인 응답 성능과 운영 가능성을 동시에 확보할 수 있음을 보여주었다.

2. 플랫폼 및 스마트미디어 관점의 시사점

스마트미디어 환경에서 추천 시스템은 단순한 정보 제공 기능을 넘어 사용자와 콘텐츠 간 상호작용 방식을 재구성하는 핵심 요소로 작용한다[27,28]. 특

히 해외 바이어 - 국내 스타트업 매칭 플랫폼과 같이 언어 및 문화적 배경이 상이한 사용자 집단이 공존하는 환경에서는, 텍스트 설명 이전에 이미지 기반 탐색을 통해 제품 후보를 직관적으로 좁혀 나갈 수 있는 추천 방식이 사용자 경험 개선에 중요한 역할을 한다.

본 연구에서 제안한 시스템은 해외 바이어가 관심 있는 제품 이미지를 기준으로 유사한 국내 스타트업 제품들을 추천함으로써, 텍스트 정보의 한계나 언어적 장벽으로 인한 정보 비대칭을 완화하는 데 기여한다. 이러한 접근은 데이터 기반 플랫폼 서비스가 사용자의 인지 부담을 줄이고, 보다 효율적인 탐색과 의사결정을 지원하는 방향으로 진화하고 있다는 스마트미디어 연구의 흐름과도 부합한다[29].

3. 시스템·공학적 기여

공학적 관점에서 볼 때, 본 연구의 기여는 운영체제, 컨테이너 플랫폼, 또는 오케스트레이션 도구가 제공하는 일반적인 기능 자체에 있지 않다. 본 연구는 이러한 환경 위에서, 해외 바이어의 이미지 입력을 기반으로 국내 스타트업 제품을 추천하는 계층형 멀티모달 추천 API를 실제로 설계·구현하고, 이를 레거시 기반 B2B 플랫폼 서비스 흐름에 통합하여 검증하였다는 점에서 차별성을 가진다.

구체적으로, 멀티모달 추천 연산을 레거시 플랫폼 내부에 포함시키지 않고 독립적인 API 서버로 분리하여 구현함으로써 시스템 간 결합도를 낮추고 확장성을 확보하였다. 이는 대규모 실시간 추론 환경에서 권장되는 모델 서빙 아키텍처와도 부합하는 접근이다[5,6]. 또한 마이크로서비스 아키텍처 기반의 분리 구조를 통해 추천 모델의 업데이트와 운영을 기존 플랫폼 서비스와 독립적으로 관리할 수 있는 기반을 마련하였다. 이러한 접근은 분산 시스템 운영 사례에서 강조되어 온 설계 원칙과 일관성을 가지며[14,15], 머신러닝 시스템에서 기술 부채가 누적되는 문제를 완화하는 데에도 기여할 수 있다[19-21].

아울러 본 연구는 알고리즘 성능 분석은 선행 연구에 위임하고[9], 본 논문에서는 선행 연구에서 검증된 추천 품질을 전제로 해당 모델이 실제 B2B 서비스 환경에서 실행 가능하고 안정적으로 운영될 수 있는지를 중심으로 분석하였다. 이를 통해 알고리즘 연구와 서비스·운영 연구의 역할을 명확히 구분하고, 멀티모달 추천 연구의 적용 범위를 실제 플랫폼 서비스 맥락으로 확장하는 공학적 연구 방향을 제시하였다.

4. 연구의 한계 및 향후 연구

본 연구는 실제 서비스 환경에 적용 가능한 시스템 구현과 운영 성능 검증을 중심으로 분석을 수행하였으나, 추천 결과에 대한 사용자 행동 기반의 정량적 평가는 포함하지 못하였다. 향후 연구에서는 해외 바이어의 클릭 로그, 추천 후 문의·매칭 전환 데이터를 활용하여 이미지 기반 멀티모달 추천이 실제 비즈니스 성과에 미치는 영향을 보다 정밀하게 분석할 필요가 있다.

또한 본 연구의 시스템 구조는 특정 B2B 매칭 플랫폼 환경을 기반으로 설계되었기 때문에, 다른 도메인이나 데이터 분포에 적용하기 위해서는 카테고리 구조나 임베딩 전략에 대한 추가적인 조정이 필요할 수 있다. 그럼에도 불구하고, 이미지 기반 추천을 API 서버 형태로 분리하여 운영하는 설계 원칙과 계층형 추론 구조는 다양한 스마트미디어 서비스 환경으로 확장 가능하다는 점에서 일정 수준의 일반화 가능성을 가진다.

5. 결어

본 논문은 해외 바이어 - 국내 스타트업 매칭 플랫폼을 대상으로, 해외 바이어의 제품 이미지 입력을 기반으로 유사한 국내 스타트업 제품을 추천하는 계층형 CLIP 기반 멀티모달 추천 시스템을 실시간 API 서버 형태로 구현하고 이를 레거시 플랫폼 환경에 통합하는 방안을 제시하였다. 제안된 시스템

은 실제 서비스 환경에서 안정적인 응답 성능과 운영 가능성을 확보하였으며, 멀티모달 추천 기술을 플랫폼 서비스로 구현하는 데 필요한 설계 원칙과 실무적 고려 사항을 구체적으로 제시하였다. 본 연구의 결과는 스마트미디어 환경에서 인공지능 기반 추천 시스템을 설계·운영하고자 하는 연구자와 실무자 모두에게 의미 있는 시사점을 제공한다.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning (ICML),* pp. 8748 - 8763, 2021.

[2] J. Jia, Y. Tang, B. L. Zoph, et al., "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," *Proceedings of the 38th International Conference on Machine Learning (ICML),* pp. 4904 - 4916, 2021.

[3] S. Gao, X. Xiong, J. Xu, et al., "FashionBERT: Text and Image Matching with BERT for Fashion Item Retrieval," *Proceedings of the 43rd International ACM SIGIR Conference,* pp. 2193 - 2196, Xi'an, China, 2020.

[4] X. Li, C. Zhang, Y. Li, et al., "Align Before Fuse: Vision and Language Representation Learning With Momentum Distillation," *Advances in Neural Information Processing Systems,* pp. 9694 - 9705, 2021.

[5] D. Crankshaw, X. Wang, G. Zhou, et al., "Clipper: A Low-Latency Online Prediction Serving System," *Proceedings of the 14th USENIX Symposium on Networked Systems Design and Implementation (NSDI),* pp. 613 - 627, Boston, USA, 2017.

[6] M. Olston, N. Fiedel, D. Gorajek, et al., "TensorFlow-Serving: Flexible, High-Performance ML Serving," arXiv preprint, arXiv:1712.06139, 2017.

[7] 이준환, "지능형 관제시스템을 위한 딥러닝 기반의 다중 객체 분류 및 추적에 관한 연구," *스마트미디어저널*, 제12권, 제5호, 65 - 72쪽, 2023년

[8] 장재영, 정수용, 김현일, 서창호, "메타버스 환경에서의 효율적인 사용자 인증을 위한 다중 서명 기법 연구," *스마트미디어저널*, 제12권, 제2호, pp. 27 - 35쪽, 2023년

[9] 오소진, 김재경, "계층형 이미지-텍스트 멀티모달 모델을 활용한 국내 스타트업-해외 바이어간 B2B 추천 성능 향상 연구", *Journal of Information Technology Applications and Management,* 제32권, 제6호, 107-117쪽, 2025년

[10] 조한진, "디지털 농업을 위한 딥러닝 기반의 환경인자 추천 기술 연구," *스마트미디어저널*, 제12권, 제5호, 58 - 64쪽, 2023년

[11] P. Covington, J. Adams, and E. Sargin, "Deep Neural Networks for YouTube Recommendations," *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 191 - 198, Boston, USA, 2016.

[12] Q. Ai, V. Azizi, X. Chen, and Y. Zhang, "Learning Hierarchical Representations for Product Search," *Proceedings of the 42nd International ACM SIGIR Conference*, pp. 885 - 894, Paris, France, 2019.

[13] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative Deep Learning for Recommender Systems," *Proceedings of the 21st ACM SIGKDD Conference,* pp. 1235 - 1244, Sydney, Australia, 2015.

[14] H. Fang, D. Zhang, Y. Shu, and G. Guo, "Deep Learning-based Recommender Systems: A Survey," *IEEE Transactions on Knowledge and Data Engineering,* vol. 35, no. 1, pp. 1 - 20, 2023.

[15] S. Gupta and M. Cohen, "Model Serving for Real-Time Inference at Scale," *Proceedings of the USENIX Workshop on Hot Topics in Cloud Computing*, 2020.

[16] S. Newman, Building Microservices, O'Reilly Media, 2015.

[17] Fowler, M. (2014). Microservices. https://martinfowler.com/articles/microservices.html (accessed May 10, 2024).

[18] B. Burns, B. Grant, D. Oppenheimer, et al., "Borg, Omega, and Kubernetes," *Communications of the ACM*, vol. 59, no. 5, pp. 50 - 57, 2016.

[19] J. Dean and L. A. Barroso, "The Tail at Scale," *Communications of the ACM,* vol. 56, no. 2, pp. 74 - 80, 2013.

[20] D. Sculley, G. Holt, D. Golovin, et al., "Hidden Technical Debt in Machine Learning Systems," *Advances in Neural Information Processing Systems,* pp. 2503 - 2511, 2015.

[21] M. Zaharia, A. Chen, A. Davidson, et al., "Accelerating the Machine Learning Lifecycle with MLflow," *IEEE Data Engineering Bulletin,* vol. 41, no. 4, pp. 39 - 45, 2018.

[22] E. Breck, S. Cai, E. Nielsen, et al., "What's your ML test score? A rubric for ML production systems,"

in Proceedings of the MLSys Conference, pp. 1-12, 2019.
[23] E. Breck, S. Cai, E. Nielsen, et al., "The ML test score: A rubric for ML production readiness," *in Proceedings of the IEEE International Conference on Big Data,* pp. 1123-1132, 2017.
[24] J. Johnson, *Designing with the Mind in Mind: Simple Guide to Understanding User Interface Design Rules* (2nd ed.). Morgan Kaufmann, 2014.
[25] A. Dix, Finlay, J., Abowd, G. D., & Beale, R. *Human-Computer Interaction* (3rd ed.). Pearson, 2004.
[26] M. Kleppmann, *Designing Data-Intensive Applications.* O'Reilly Media, 2017.
[27] L. Manovich, *The Language of New Media,* MIT Press, 2001.
[28] H. Jenkins, *Convergence Culture: Where Old and New Media Collide,* New York University Press, 2006.
[29] R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, SAGE Publications, 2014.

저 자 소 개

김재경(정회원)
1991년 아주대학교 경영학과 학사 졸업.
2000년 Miami University MBA 졸업.
2004년 University of Nebaraska-Lincoln Ph.D. in Managemen 졸업.
<주관심분야 : 지식공유, 딥러닝, 정보교육, 인공지능>