Check for updates

Journal of the Korean Institute of Industrial Engineers
Vol. 52, No. 5, pp. 447-469, October 2026.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2026.52.5.447
© 2026 KIIE

사이버범죄 수사를 위한 도메인 특화 멀티모달 임베딩 기반 통합 검색 시스템

이승규[1] · 이수민[2] · 박현호[3,4] · 권은정[3] · 이민정[3,4†]

[1]서울과학기술대학교 국방인공지능응용학과 / [2]인천대학교 산업경영공학과
[3]한국전자통신연구원 과학치안ADX연구실 / [4]과학기술연합대학원대학교

A Domain-Specific Multimodal Embedding-Based Integrated Retrieval System for Cybercrime Investigation

Seunggyu Lee[1] · Su Min Lee[2] · Hyunho Park[3,4] · Eun Jung Kwon[3] · Minjung Lee[3,4]

[1]Department of Defense Applied Artificial Intelligence, Seoul National University of Science and Technology
[2]Department of Industrial and Management Engineering, Incheon National University
[3]Scientific Policing ADX Research Section, Electronics and Telecommunications Research Institute
[4]University of Science and Technology

This study proposes a domain-specific multimodal embedding model and an integrated retrieval system for cybercrime investigation. To address the limited domain adaptability of general-purpose embedding models, we fine-tune Jina Embeddings v4 via LoRA using synthetic triplets generated through an LLM-based pipeline. The proposed system indexes structured and unstructured evidence in a unified Qdrant vector database and orchestrates retrieval workflows using LangGraph. For image evidence, we propose an Image-Text Fusion strategy that combines original visual embeddings with VLM-generated descriptions and OCR text. Under optimized hyperparameters, the proposed Fusion strategy achieves a Hit Rate@1 of 0.9407 and MRR of 0.9641, improving upon the baseline Jina v4 Single Vector strategy by 48.1%p and 42.2%p, respectively. Furthermore, the proposed Single Vector strategy achieves an MRR of 0.9509, offering a resource-efficient alternative with an 8x smaller index size and over 6x faster retrieval speed.

Keywords: Cybercrime Investigation, Multimodal Embedding, Domain-Specific Fine-tuning, Synthetic Data Generation, LoRA

1. 서 론

사이버범죄는 디지털 기술의 발전과 함께 그 유형과 규모가 급속히 확대되고 있으며(Korean National Police Agency, 2023; Korean National Police Agency, 2025), 수사기관이 다루어야 할 디지털 증거의 양 또한 기하급수적으로 증가하고 있다(Casey, 2011; Lillis *et al.*, 2016). 사이버범죄 수사 과정에서 수사관은 금융거래 내역, 통화기록, 스크린샷, 채팅 로그 등 다양한 형식의 데이터를 분석해야 하며, 이러한 데이터에서 범죄 단서를 신속하게 검색하고 식별하는 것이 수사의 핵심 과제이다(Casey, 2011). 특히, 수사 현장에서는 수십만 건의 금융거래 기록이나 수천 장의 스크린샷을 수작업으로 검토해야 하는 경

이 논문은 2026년도 정부(경찰청)의 재원으로 과학치안진흥센터 사이버범죄 수사단서 통합분석 및 추론시스템 개발 사업의 지원을 받아 수행된 연구임(No. RS-2025-02218280).
† 연락저자 : 이민정 교수, 34113 대전광역시 유성구 가정로 217, 과학기술연합대학원대학교, Tel : 042-860-1365, E-mail : minjunglee@etri.re.kr
2026년 6월 9일 접수; 2026년 7월 20일, 2026년 7월 27일 수정본 접수; 2026년 7월 27일 게재 확정.

우가 빈번하며, 이는 수사관의 업무 부담을 가중시키고 수사의 적시성을 저해하는 주요 요인이 된다. 따라서 대규모 멀티모달 수사 데이터로부터 관련 증거를 자동으로 검색하고 식별할 수 있는 지능형 검색 시스템의 개발이 시급한 과제로 대두되고 있다.

임베딩 기반 의미론적 검색(semantic search)은 단순 키워드 일치를 넘어 문장의 의미와 맥락을 벡터로 변환하여 사용자의 쿼리와 가장 유사한 문서를 찾아내는 핵심 검색 기술이다(Reimers and Gurevych, 2019; Karpukhin *et al.*, 2020). 최근에는 대규모 언어 모델(LLM)의 발전에 힘입어 E5(Wang *et al.*, 2024a), Jina Embeddings(Günther *et al.*, 2025) 등 고성능 범용 임베딩 모델이 등장하였으며, 다양한 벤치마크에서 우수한 성능을 달성하고 있다. 그러나 범용 임베딩 모델을 사이버범죄 수사와 같은 특수 도메인에 그대로 적용할 경우, 도메인 고유의 의미가 왜곡되는 문제가 발생한다. Tang and Yang(2024)은 최신 LLM 기반 범용 임베딩 모델이 금융 등 전문 도메인에서 유의미한 성능 저하를 보임을 실증적으로 입증하였다. 예를 들어, '아이스'라는 단어는 범용 모델에서 얼음이나 냉음료를 의미하지만, 사이버범죄 수사 도메인에서는 마약의 은어로 사용된다. 마찬가지로, '작업'은 일반적 맥락에서는 업무를 뜻하지만, 사이버범죄 도메인에서는 보이스피싱 등의 범죄 행위를 지칭하는 용어로 통용된다. 이처럼 동일한 단어라도 도메인에 따라 전혀 다른 의미를 가질 수 있으므로, 도메인 특화 임베딩 모델의 개발이 필수적이다.

한편, 사이버범죄 수사 데이터는 텍스트 기반 정형 데이터(금융거래 내역, 통화기록 등)와 이미지 기반 비정형 데이터(스크린샷, 채팅 캡처 등)가 혼재된 멀티모달(multimodal) 특성을 가진다. 기존의 텍스트 전용 검색 시스템으로는 이미지에 포함된 텍스트 정보나 시각적 단서를 활용한 통합 검색이 불가능하며, 이는 수사 효율성의 저하로 이어진다. 따라서 텍스트와 이미지를 동일한 벡터 공간에서 통합적으로 검색할 수 있는 멀티모달 검색 시스템의 구축이 필요하다. 그러나 사이버범죄 수사 도메인에 특화된 멀티모달 임베딩 모델과 통합 검색 시스템에 관한 연구는 아직 미흡한 실정이다.

본 연구는 '사이버범죄 수사단서 통합분석 및 추론시스템 개발' 과제의 일환으로, 사이버범죄 수사 도메인에 특화된 멀티모달 임베딩 모델과 통합 검색 시스템을 개발하는 것을 목표로 한다. 구체적으로, 2025년 6월에 공개된 오픈소스 멀티모달 임베딩 모델인 Jina Embeddings v4(Günther *et al.*, 2025; Jina AI, 2025)를 기반 모델로 채택하고, LoRA(Low-Rank Adaptation)(Hu *et al.*, 2022) 기법을 활용하여 도메인 특화 파인튜닝을 수행한다. LoRA는 사전학습된 모델의 파라미터를 고정(freeze)한 상태에서 저랭크(low-rank) 행렬을 추가하고, 해당 행렬의 파라미터만 학습하는 방법이다. 파인튜닝을 위한 학습 데이터는 LLM을 활용한 합성 데이터 생성 파이프라인을 통해 확보하며, 쿼리-정답-오답(Triplet)으로 구성된 총 13,022개의 합성 데이터를 생성한다. 또한 Qdrant(Qdrant, 2023) 벡터 데이터베이스를 구축하고, 텍스트와 이미지가 혼재된 멀티모달 환경에서의 통합 검색 시스템을 구현한다. 검색 시스템은 흐름 관리를 위해 LangGraph(LangChain, 2024)를 활용하였다. 본 연구의 주요 기여는 다음과 같다.

- 사이버범죄 수사 도메인 특화 멀티모달 임베딩 모델 개발: Jina Embeddings v4를 기반으로 LoRA 파인튜닝을 수행하여, 사이버범죄 수사 환경에서 혼재되는 정형(통화·금융거래 등)과 비정형(스크린샷·채팅 캡처 등) 데이터를 단일 임베딩 공간에 정렬하는 멀티모달 임베딩 모델을 개발하였다.
- LLM 기반 3단계 합성 데이터 생성 파이프라인 설계: 도메인 전문가의 수동 라벨링에 의존하지 않고, LLM을 활용한 체계적인 3단계 파이프라인(데이터 전처리, 후보 Triplet 생성, LLM 기반 평가)을 통해 고품질 학습 데이터를 효율적으로 생성하는 방법론을 제시하였다.
- 이미지-텍스트 Multi Vector 통합 표현(Image-Text Fusion) 방식 제안: Vision-Language Model(VLM)이 산출한 설명문 및 OCR(Optical Character Recognition) 텍스트의 Multi Vector와 원본 이미지의 Multi Vector를 결합하는 통합 표현 방식을 설계하여, 텍스트 기반 검색과 시각적 유사성 기반 검색을 동시에 수행할 수 있는 검색 체계를 구현하였다.
- Qdrant 및 LangGraph 기반 멀티모달 통합 검색 시스템 구축: 정형 데이터와 비정형 데이터를 통합하는 벡터 데이터베이스와 검색 파이프라인을 구축하여, 사이버범죄 수사 환경에서 실용적으로 활용 가능한 end-to-end 검색 시스템을 구현하였다.

2. 관련 연구

2.1 임베딩 기반 의미론적 검색

문장 임베딩 기반 검색은 Sentence-BERT(Reimers and Gurevych, 2019)의 등장 이후 빠르게 발전하였다. Sentence-BERT는 Siamese 네트워크 구조를 활용하여 의미적으로 유사한 문장이 고차원 벡터 공간에서 인접하도록 학습함으로써, 대규모 코퍼스에 대한 효율적인 의미 검색을 가능하게 하였다. 이후 Dense Passage Retrieval(Karpukhin *et al.*, 2020)은 질문과 문서를 각각 Dense 벡터로 표현하여 기존 TF-IDF 기반 희소 검색의 한계를 보완하였으며, E5(Wang *et al.*, 2024a)는 약지도 대조 학습을 통해 다양한 검색 및 문장 유사도 태스크의 성능을 향상시켰다.

최근 범용 임베딩 모델은 인코더 기반 사전학습 모델에서 LLM 기반 임베딩으로 확장되고 있다. GTE(Li *et al.*, 2023a)는 범용 텍스트 임베딩 모델로서 다양한 자연어 처리 태스크에서 경쟁력 있는 성능을 보였으며, BGE(Xiao *et al.*, 2023)는 RetroMAE 기반 사전학습과 대조 학습을 결합하여 MTEB 벤

치마크(Muennighoff *et al.*, 2023)에서 우수한 성능을 달성하였다. 특히 다국어 및 멀티 태스크 지원을 강화한 BGE-M3(Chen *et al.*, 2024a)는 Dense·Sparse·Multi Vector 검색을 모두 지원하는 하이브리드 검색의 대표적 방법론으로 자리 잡았다. 더 나아가 E5-Mistral(Wang *et al.*, 2024b)은 Mistral 7B(Jiang *et al.*, 2023)를 백본으로 채택하고 LLM 기반 합성 데이터만으로 학습함으로써, 기존 BERT계열 임베딩 모델을 상회하는 성능을 입증하였으며, LLM2Vec(BehnamGhader *et al.*, 2024)은 디코더 전용 LLM을 양방향 인코더로 변환하는 방법론을 제시하여 LLM 기반 임베딩의 일반화 가능성을 확장하였다.

그러나 범용 임베딩 모델은 전문 도메인의 어휘와 문맥을 충분히 반영하지 못해 성능이 저하될 수 있다. Tang and Yang(2024)은 금융 도메인에서 범용 LLM 임베딩 모델의 성능 저하를 확인하였다. 의료 분야의 BioSentVec(Chen *et al.*, 2019)와 법률 분야의 Legal-BERT(Chalkidis *et al.*, 2020) 역시 도메인 코퍼스와 전문 어휘를 반영함으로써 해당 분야의 문장 표현 및 문서 처리 성능을 개선하였다. 이는 전문 도메인의 고유한 의미 체계를 반영한 특화 임베딩 모델이 필요함을 보여준다.

사이버범죄 수사 환경에는 계좌 정보, 거래 내역, 메신저 대화, 파일 경로, 은어뿐 아니라 이미지와 문서 등 다양한 형태의 증거가 혼재한다. 따라서 범용 텍스트 임베딩 모델만으로는 수사 맥락과 증거 간 의미적 연관성을 충분히 포착하기 어렵다. 그럼에도 사이버범죄 수사에 특화된 멀티모달 임베딩 연구는 제한적이다. 본 연구는 이러한 연구 공백을 바탕으로 사이버범죄 수사 도메인에 특화된 멀티모달 임베딩 모델을 개발한다.

2.2 멀티모달 임베딩 모델과 멀티모달 검색 시스템

멀티모달 임베딩 모델은 서로 다른 유형의 데이터를 공통 벡터 공간에 매핑하여 교차 모달 검색을 가능하게 한다. CLIP(Radford *et al.*, 2021)은 대조 학습을 통해 이미지와 텍스트를 정렬하는 대표적 방법을 제시하였으며, 이후 BLIP-2(Li *et al.*, 2023b)와 SigLIP(Zhai *et al.*, 2023) 등이 모달리티 간 표현 간극과 학습 효율을 개선하였다.

최근에는 VLM을 활용하여 이미지와 텍스트를 통합적으로 처리하는 연구가 활발하다. InternVL 2.5(Chen *et al.*, 2024b), Llama 3.2 Vision(Grattafiori *et al.*, 2024), PaliGemma 2(Steiner *et al.*, 2024), Qwen2.5-VL(Bai *et al.*, 2025a) 등은 각각 장문·비디오 이해, 배포 편의성, 전이 학습, 문서 분석과 객체 로컬라이제이션에서 우수한 성능을 보인다. 그러나 이들 모델은 이미지 추론과 자연어 생성에 최적화되어 있어 대규모 데이터를 색인하고 고속 유사도 검색을 수행하는 임베딩 태스크에 직접 활용하기에는 연산 및 구조적 제약이 있다.

이러한 한계를 보완하기 위해 VLM의 표현력을 검색용 벡터로 전이하는 멀티모달 임베딩 모델이 제안되었다. ColPali(Faysse *et al.*, 2025)는 문서 이미지를 토큰 수준에서 비교하는 late interaction 방식을 적용하였고, Nomic Embed Vision v1.5(Nussbaum *et al.*, 2024)는 Single Vector 기반의 멀티모달 검색을 지원한다. 다만 ColPali는 한국어 검색 성능에 대한 실험적 검증이 충분히 이뤄지지 않았으며, Nomic Embed Vision은 세밀한 토큰 수준 상호작용을 반영하기 어렵다.

본 연구는 이를 고려하여 Qwen2.5-VL-3B-Instruct(Bai *et al.*, 2025a; Qwen Team, 2025a)를 백본으로 사용하는 Jina Embeddings v4를 채택하였다. 이 모델은 Single Vector와 Multi Vector를 모두 지원하여 광범위한 탐색과 정밀 매칭을 병행할 수 있으며, ColBERT(Khattab and Zaharia, 2020)의 late interaction을 멀티모달 환경으로 확장하여 이미지 패치와 텍스트 토큰 간 상호작용을 계산한다. 또한 한국어를 포함한 30개 이상의 언어와 태스크별 LoRA 어댑터를 지원하여 제한된 연산 자원에서도 효율적인 도메인 적응이 가능하다. 특히 한국어 시각 문서 검색 벤치마크인 KoViDoRe에서도 우수한 검색 성능을 보여, 한국어 텍스트와 이미지가 혼재하는 국내 디지털 증거 검색 환경에 적합한 기반 모델로 판단하였다(Choi *et al.*, 2026).

멀티모달 검색 시스템은 서로 다른 형식의 데이터를 통합 색인·검색하는 구조로, RAG(Retrieval-Augmented Generation)와 함께 발전해 왔다. MuRAG(Chen *et al.*, 2022)와 UniIR(Wei *et al.*, 2023)은 멀티모달 지식 검색과 검색 태스크 통합 가능성을 제시했으나, 디지털 포렌식 실무에서 요구되는 정형·비정형 증거의 통합 처리와 세밀한 매칭에는 한계가 있다. 기존 법집행 분야 연구(Du *et al.*, 2020; Amato *et al.*, 2020; Scanlon *et al.*, 2023) 역시 주로 텍스트 중심으로 수행되었다.

대규모 멀티모달 검색을 위해서는 벡터 데이터베이스도 필요하다. Faiss(Johnson *et al.*, 2019)와 Milvus(Wang *et al.*, 2021)는 대표적인 벡터 검색 도구이지만, 정형 데이터 필터링과 Multi Vector 기반 late interaction을 함께 처리하는 데 제약이 있다. 반면 Qdrant(Qdrant, 2023)는 HNSW(Hierarchical Navigable Small World) 인덱싱, 페이로드 필터링, Multi Vector 검색을 지원한다. 이에 본 연구는 Jina Embeddings v4의 이미지-텍스트 통합 표현과 Qdrant를 결합하여 정형 및 비정형 증거를 동일한 벡터 공간에서 검색하는 멀티모달 통합 검색 시스템을 제안한다.

2.3 LLM 기반 합성 데이터 생성

LLM을 활용한 합성 데이터 생성은 도메인 특화 모델 학습에서 라벨링 비용을 줄이는 효과적인 방법이다. InPars(Bonifacio *et al.*, 2022)는 GPT-3로 정보 검색용 쿼리를 생성하여 소량의 프롬프트만으로도 검색 성능을 개선할 수 있음을 보였다. 이후 InPars-v2(Jeronymo *et al.*, 2023)는 오픈소스 LLM과 자체 학습 루프(self-training loop)를 결합해 데이터 품질과 다양성을 높였으며, Promptagator(Dai *et al.*, 2023)는 few-shot 쿼리 생성과 일관성 필터링(consistency filtering)을 통해 데이터가 부족한 도메인에서도 고품질 학습 데이터를 확보할 수 있음을 보였다.

합성 데이터 기반 검색 모델 학습에서는 Hard Negative 마이

닝도 중요한 역할을 한다. Hard Negative는 정답과 의미적으로 유사하지만 실제로는 오답인 문서 또는 항목을 의미하며, 이를 학습에 포함하면 모델이 미세한 의미 차이를 구분하도록 유도할 수 있다. ANCE(Xiong *et al.*, 2021)는 학습 중 모델 자체를 활용하여 Hard Negative를 동적으로 생성하는 비동기 인덱스 갱신(asynchronous index update) 방식을 제안하였다. 또한 E5-Mistral(Wang *et al.*, 2024b)은 LLM 기반 합성 데이터를 활용하여 다양한 검색 태스크를 학습하였으며, 수동 라벨링 데이터가 제한적인 상황에서도 높은 성능을 달성할 수 있음을 보였다.

그러나 기존 연구는 주로 범용 도메인의 텍스트 검색을 대상으로 하여, 사이버범죄 수사와 같이 정형 데이터와 이미지 증거가 함께 존재하는 환경에서의 합성 데이터 생성 방법론은 충분히 탐구되지 않았다. 특히 사이버범죄 수사 도메인에서는 정답과 오답이 표면적으로 유사한 경우가 많고, 증거의 의미가 파일명, 행 단위 필드, 이미지 내 텍스트, 시각적 맥락 등에 분산되어 있기 때문에 일반적인 쿼리 생성 방식만으로는 충분한 학습 신호를 확보하기 어렵다. 본 연구는 이러한 한계를 보완하기 위해 LLM 기반 쿼리 생성, 벡터 데이터베이스 기반 Hard Negative 마이닝, LLM 기반 품질 평가를 결합한 3단계 합성 데이터 생성 파이프라인을 설계하였다. 이를 통해 사이버범죄 수사 도메인의 의미 구조를 반영한 Triplet 학습 데이터를 구축하고, 도메인 특화 멀티모달 임베딩 모델 학습에 활용한다.

3. 제안 방법론

본 장에서는 사이버범죄 수사 도메인에 특화된 멀티모달 임베딩 모델 학습과 통합 검색 시스템의 구축을 위한 전체 방법론을 기술한다. <Figure 1>은 본 연구의 전체 시스템 구조도를 보여준다. 제안 방법론은 (1) 합성 데이터 생성 파이프라인, (2) LoRA 기반 도메인 특화 멀티모달 임베딩 모델 학습, (3) Qdrant 및 LangGraph 기반 멀티모달 통합 검색 시스템 구축의 세 부분으로 구성된다.

3.1 합성 데이터 생성 파이프라인

사이버범죄 수사 도메인 특화 임베딩 모델의 파인튜닝을 위해서는 도메인 고유의 어휘와 의미 체계가 반영된 대규모 학습 데이터가 필요하다. 그러나 실제 수사 데이터는 기밀성과 개인정보 보호 문제로 직접 활용이 어려우며, 전문가의 수동 라벨링은 비용과 시간 측면에서 비현실적이다. 이에 본 연구에서는 과제 내 사이버범죄 수사 시뮬레이션 환경에서 수집된 이미지 데이터(사이버 범죄 사이트와 메신저앱·문자 범죄 대화 스크린샷 등)와 정형 데이터(통화 내역·금융 거래 내역 등)를 활용해 3단계로 구성된 LLM 기반 합성 데이터 생성 파이프라인을 설계하였다. <Figure 2>는 합성 데이터 생성 파이프라인을 나타내며, 사용 모델은 <Table 1>에 정리하였다. 활용된 원시 데이터의 구체적 구성은 4절의 <Table 2>에서 확인할 수 있다. 본 절에서 사용하는 주요 수식 기호는 <Appendix A>의 표에 정리하였다.

합성 데이터는 Query, Positive, Hard Negative로 구성된 Triplet 형태로 구축한다. Query는 수사관이 증거물 검색에 사용하는 자연어 질의이며, Positive는 이에 대응하는 정답 증거물이다. Hard Negative는 정답은 아니지만 의미적으로 유사하여 검색 모델이 혼동하기 쉬운 오답 증거물이다. 학습 목표는 Query와 Positive 간 거리는 최소화하고, Query와 Hard Negative 간 거리는 최대화하여 임베딩 공간의 검색 변별력을 높이는 것이다.

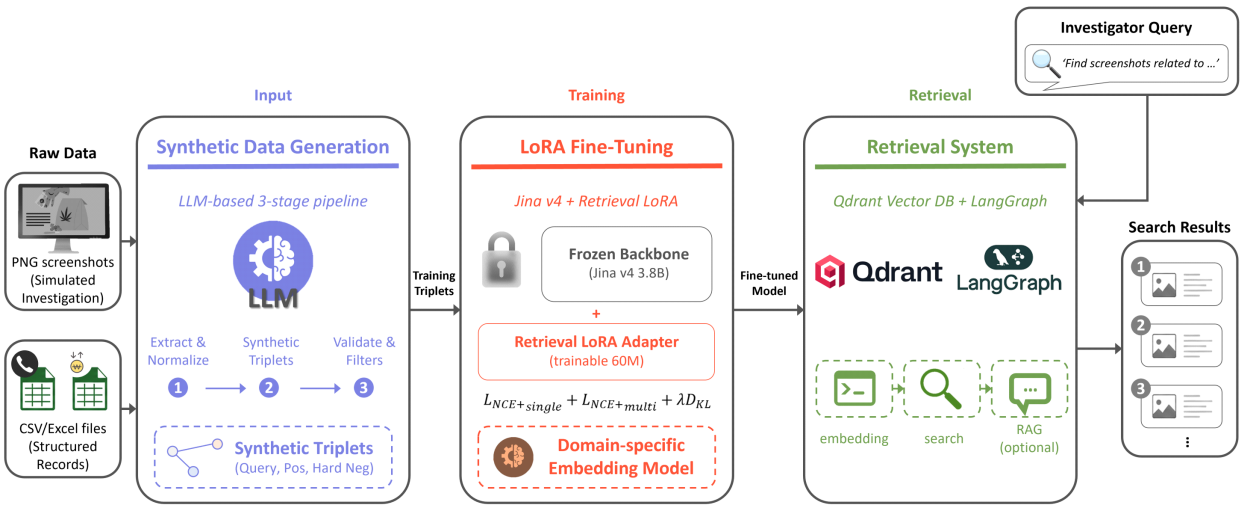


Figure 1. Overall architecture of the proposed domain-specific multimodal retrieval system for cybercrime investigation. The system consists of three main components: (1) a synthetic data generation pipeline, (2) LoRA-based domain-specific fine-tuning of the multimodal embedding model, and (3) an integrated retrieval system based on Qdrant and LangGraph.

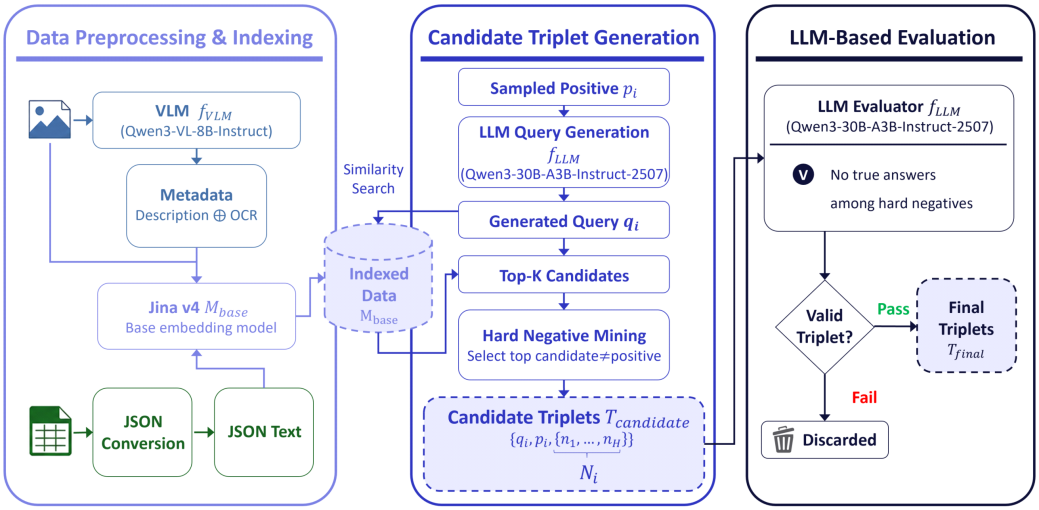


Figure 2. Three-stage synthetic data generation pipeline. Stage 1 preprocesses raw images via a VLM into metadata (Description + OCR) and converts tabular records into JSON, then indexes them into a vector database using the base embedding model. Stage 2 generates queries using an LLM and mines hard negatives from the vector database to construct candidate triplets. Stage 3 filters low-quality triplets via LLM-based evaluation, yielding the final training dataset.

Table 1. Roles of models in synthetic data generation

Model Category	Specific Model	Key Functional Role
Base Multimodal Embedding Model M_{base}	Jina Embeddings v4 (Günther *et al.*, 2025; Jina AI, 2025)	Serves as the core backbone architecture for domain-specific fine-tuning; maps structured tabular logs and unstructured evidence images into a unified high-dimensional dense vector space.
Metadata VLM f_{VLM}	Qwen3-VL-8B-Instruct (Bai *et al.*, 2025b; Qwen Team, 2025c)	Performs the evidence preprocessing stage, extracting textual metadata—visual descriptions and OCR text—from crime screenshot PNGs.
Query & Evaluation LLM f_{LLM}	Qwen3-30B-A3B-Instruct-2507 (Yang *et al.*, 2025; Qwen Team, 2025b)	Generates natural-language queries that emulate investigator search behavior from the given context, and performs post-generation quality filtering by removing false hard negatives.

첫 번째 단계는 데이터 전처리 및 인덱싱이다. 원시 데이터 $d_i \in D$는 이미지 데이터 또는 테이블형 데이터의 행으로 구성되며, 데이터의 유형에 따라 임베딩 가능한 형태로 변환된다. 원시 데이터가 이미지인 경우 VLM을 이용해 이미지별 메타데이터를 생성한다. 메타데이터는 이미지 내용에 대한 설명문(Description)과 OCR 결과 텍스트로 정의하였고, 이미지인 원시 데이터에 대한 메타데이터 $meta_i$는 식 (1)과 같이 표현할 수 있다.

$$meta_i = f_{VLM}(d_i, \text{prompt} = \text{Langfuse}[\text{'meta_generation'}]), \quad (1)$$

여기서 f_{VLM}은 이미지 전처리에 사용되는 VLM을 의미하며, 본 연구에서는 Qwen3-VL-8B-Instruct(Bai *et al.*, 2025b; Qwen Team, 2025c)를 사용하였다. 설명문과 OCR 텍스트를 생성하는 프롬프트를 포함해 제안 방법론에 사용된 모든 프롬프트는 LLM 엔지니어링 플랫폼인 Langfuse(Langfuse, 2023)를 통해 버전 관리하였다. 생성된 메타데이터는 합성 데이터 생성 파이프라인의 두 번째 단계에서 Query 생성용 LLM 컨텍스트로 활용된다.

테이블형 데이터의 행인 원시 데이터는 키-값 쌍으로 구성된 JSON 객체로 변환된다.

$$RowToJSON(d_i) = \{(col_k : val_{i,k})\}_{k=1}^{m}, \quad (2)$$

여기서 col_k는 k번째 컬럼명, $val_{i,k}$는 d_i의 k번째 컬럼값을 의미하며 m은 전체 컬럼 수이다. 이 표현 방식은 컬럼 간 의미적 관계를 유지하면서 데이터를 자연어 처리에 적합한 형태로 변환할 수 있다는 장점이 있다. 전처리된 모든 원시 데이터 $d_i \in D$는 기본 임베딩 모델 M_{base}인 Jina Embeddings v4로 벡터화된 후 벡터 데이터베이스에 인덱싱한다.

두 번째 단계는 후보 Triplet 생성이다. 먼저 전처리된 데이터에서 Positive p_i를 무작위로 샘플링하고, LLM f_{LLM}을 통해 Query q_i를 생성한 뒤 Hard Negative 마이닝을 수행하여 후보 Triplet을 구성한다. 본 연구에서는 Query 생성 모델로 Qwen3-30B-A3B-Instruct-2507(Yang *et al.*, 2025; Qwen Team, 2025b)를 사용하였다. LLM에 입력되는 컨텍스트 c_i는 데이터 유형에 따라 다르게 구성된다. 이미지 데이터의 경우 원본 이미지 대신 생성된 메타데이터 $meta_i$를 사용하고, 테이블형 데이터는 JSON 변환 결과인 $RowToJSON(d_i)$를 사용한다. f_{LLM}은 해당 컨텍스트 c_i를 바탕으로 수사관이 실제 수사 현장에서 사용할 법한 자연어 Query q_i를 생성한다. 이때 Positive 데이터를 지나치게 구체적으로 묘사하면 과적합이 발생할 수 있으므로, Query의 자연스러움을 우선하였다. 생성된 Triplet의 품질은 이후 LLM 기반 평가 단계에서 검증한다.

Hard Negative N_i는 Positive와 의미적으로 유사해 모델이 혼동하기 쉬운 오답 세트로, 대조 학습에서 미세한 의미 차이를 구분하도록 학습시키는 핵심 요소이다. 본 연구에서는 벡터 데이터베이스를 활용한 Hard Negative 마이닝 전략을 적용하였으며, 구체적인 절차는 Algorithm 2와 같다. 먼저 LLM이 생성한 Query q_i를 기본 임베딩 모델 M_{base}로 인코딩하고, 해당 벡터 v_{q_i}를 이용해 벡터 데이터베이스 V에서 코사인 유사도 기준 상위 후보군 $TopK$을 검색한다. 이후 검색 순위 r가 높은 순서대로 확인하여 Positive p_i와 일치하지 않는 후보 $cand_r$만 Hard Negative N_i에 선정하며, 이 과정은 하이퍼파라미터로 설정한 Hard Negative 세트 크기 H를 충족할 때까지 반복한다. 테이블형 데이터의 경우 각 행이 벡터 데이터베이스의 개별 데이터 포인트로 저장되므로, 동일한 테이블 내에서 Positive와 의미적으로 가장 유사한 행들이 Hard Negative로 선정된다. 이를 통해 모델은 유사한 행들 사이에서 정답 행을 정밀하게 식별하는 변별 능력을 학습한다.

마지막 단계에서는 전체 후보 Triplet에 대해 LLM 기반 품질 평가를 수행하여 최종 학습 데이터셋 T_{final}을 구축한다. 여기서 $eval_i \in \{Pass, Fail\}$는 후보 Triplet에 대한 LLM 기반 품질 평가 결과를 의미하며, $eval_i$이 $Pass$인 경우에만 해당 후보 Triplet을 최종 학습 데이터셋 T_{final}에 포함한다. 평가 프롬프트는 생성된 Query에 대한 실제 정답이 Hard Negative로 잘못 분류된 경우를 식별하고 제거하여 학습 데이터의 품질을 높인다. 전체 합성 데이터 생성 절차는 Algorithm 1에 정리하였다.

Algorithm 1: Synthetic triplet data generation pipeline

Input: Raw dataset $D = \{d_1, ..., d_N\}$ consisting of images and tabular data rows,
VLM f_{VLM}, LLM f_{LLM},
Base embedding model M_{base}, Vector database V,
Size of the top-K candidate list K,
Number of hard negatives H
Output: Final filtered triplet dataset T_{final}

/* Stage 1: Preprocessing and indexing */
$P \leftarrow \varnothing$ // Initialize preprocessed tuple set P
for each d_i in D do
  if d_i is an image then
    $p_i \leftarrow d_i$
    $c_i \leftarrow meta_i$ // Eq. (1)
  else if d_i is a tabular data row then
    $p_i \leftarrow RowToJSON(d_i)$ // Eq. (2)
    $c_i \leftarrow RowToJSON(d_i)$ // Eq. (2)
  end if
  $P \leftarrow P \cup \{(p_i, c_i)\}$ // Accumulate preprocessed pair into set P
  Upload p_i into vector database V using M_{base}
end for
/* Stage 2: Candidate triplet generation */
$T_{candidate} \leftarrow \varnothing$ // Initialize candidate set
for each sampled tuple (p_i, c_i) in P do
  $q_i \leftarrow f_{LLM}$(context $= c_i$, prompt $=$ Langfuse['query_generation'])
  $N_i \leftarrow$ HardNegativeMining$(q_i, p_i, V, M_{base}, K, H)$ // See Algorithm 2
  if $|N_i| > 0$ then
    $T_{candidate} \leftarrow T_{candidate} \cup \{(q_i, p_i, N_i)\}$
  end if
end for
/* Stage 3: LLM-based evaluation and filtering */
$T_{final} \leftarrow \varnothing$ // Initialize final triplet set
for each (q_i, p_i, N_i) in $T_{candidate}$ do
  $eval_i \leftarrow f_{LLM}(q_i, p_i, N_i$, prompt $=$ Langfuse['synthetic_data_evaluation'])
  if $eval_i =$ Pass then
    $T_{final} \leftarrow T_{final} \cup \{(q_i, p_i, N_i)\}$
  end if
end for
return T_{final}

Algorithm 2: Hard Negative Mining

Input: Query q_i, Positive p_i,
Base embedding model M_{base}, Vector database V,
Size of the Top-K candidate list K,

Number of hard negatives H

Output: Hard negatives N_i $(|N_i| = H \leq K)$

/* Embed query and retrieve Top-K candidates */

$v_{q_i} \leftarrow M_{base}(q_i)$ // Encode query with base model

$TopK \leftarrow$ V.search(v_{q_i}, K) // Retrieve top-K by cosine similarity

/* Collect up to H hard negatives in rank order */

$N_i \leftarrow \varnothing$

for rank r = 1 to $|TopK|$ do

  $cand_r \leftarrow TopK[r]$ // r-th candidate

  if $cand_r = p_i$ then

    continue // Skip: same as positive

  else

    $N_i \leftarrow N_i \cup \{cand_r\}$ // Add to hard negative set

    if $|N_i| = H$ then

      break

  end if

end for

return N_i

3.2 LoRA 기반 도메인 특화 멀티모달 임베딩 모델 학습

본 연구에서는 LoRA 기반 경량 파인튜닝을 통해 사이버범죄 수사 도메인에 특화된 임베딩 모델을 학습한다. Jina Embeddings v4 모델은 Retrieval, Code Search, Text Matching 태스크에 대응하는 세 가지 LoRA 어댑터를 제공한다. 본 연구는 수사관의 자연어 질의를 기반으로 대규모 멀티모달 증거물에서 관련 증거를 검색하는 것을 목적으로 한다. 이에 따라 쿼리와 문서 간 의미적 유사도 검색에 특화된 Retrieval 어댑터를 파인튜닝 대상으로 선정하였다. 따라서 전체 약 3.8B 파라미터 중 Retrieval 어댑터에 해당하는 약 60M개 파라미터만 학습함으로써, 전체 모델을 파인튜닝하는 방식보다 학습에 필요한 연산량과 메모리 사용량을 줄이면서도 사이버범죄 수사 도메인에 대한 검색 성능을 향상하고자 하였다.

3.1절에 설명된 절차에 따라 생성된 합성 데이터는 학습, 검증, 평가 데이터로 각각 7:1.5:1.5의 비율로 분할하였다. 학습 과정에서는 Retrieval 어댑터가 적용된 백본 모델을 이용해 동일한 입력으로부터 Single Vector v^S와 Multi Vector V^M를 함께 생성한다. 이때 Retrieval 어댑터의 파라미터만 갱신하고, 나머지 모델 파라미터는 고정한다. 이하 수식에서 v_q, v_p, v_{hn_j}는 각각 Query, Positive, j번째 Hard Negative의 임베딩을 의미한다. 전체 학습 구조는 <Figure 3>에 제시하였다.

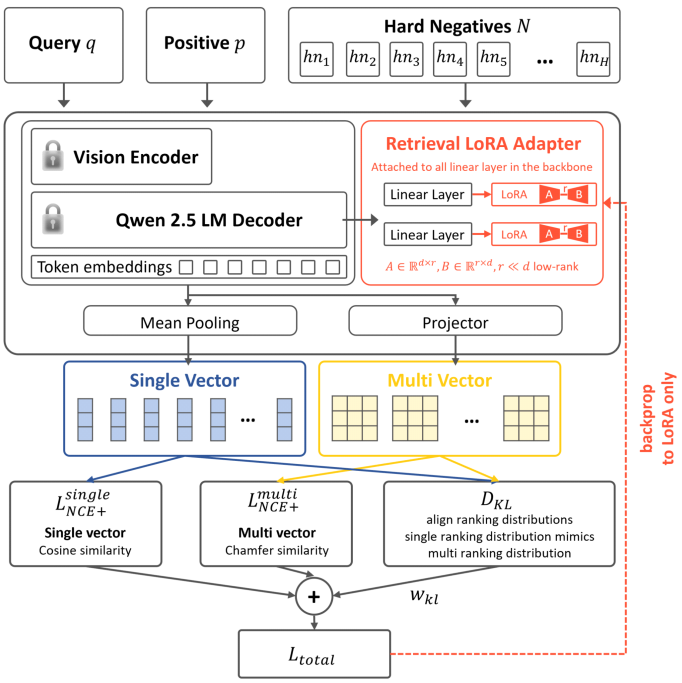


Figure 3. Training architecture of the LoRA-based domain-specific fine-tuning. The Jina Embeddings v4 backbone (~3.8B parameters) is frozen, while only the Retrieval LoRA adapter (~60M parameters, 1.58%) is updated. For each input triplet, the model produces both Single Vector and Multi Vector embeddings. The total loss combines Single Vector InfoNCE, Multi Vector InfoNCE, and KL Divergence regularization, where the KL term aligns the Single Vector ranking distribution with that of the Multi Vector representation.

식 (4) 손실함수에는 Extended InfoNCE Loss(Günther *et al.*, 2023)가 활용되었다. 이 함수는 쿼리가 문서를 찾는 과정(query-to-passage)과 문서가 쿼리를 찾는 과정(passage-to-query)을 동시에 고려하는 양방향 대조 학습 손실 함수로, 식 (3)과 같이 정의된다.

$$L_{NCE+}(B) := E_{r\sim B}\left[-\ln\frac{e^{s(v_q,\ v_p)/\tau}}{\sum_{i=1}^{k}\left[e^{s(v_q,\ v_{p_i})/\tau}+\sum_{j=1}^{H}e^{s(v_q,\ v_{hn_{j,i}})/\tau}\right]}\right] + E_{r\sim B}\left[-\ln\frac{e^{s(v_p,\ v_q)/\tau}}{\sum_{i=1}^{k}e^{s(v_p,\ v_{q_i})/\tau}}\right] \quad (3)$$
$$with\ r=(v_{q_i},\ v_{p_i},\ v_{hn_1},\ \cdots,\ v_{hn_H})$$

여기서 B는 배치 데이터를 의미하고, $s(v_q,\ v_p)$는 Query와 Positive 임베딩 벡터 간 유사도, τ는 Temperature 하이퍼파라미터로 InfoNCE 손실의 softmax 분포 예민도(sharpness)를 조절하는 스케일링 인자다.

전체 손실(Total Loss) 함수는 식 (4)와 같이 Single Vector 손실, Multi Vector 손실, KL Divergence의 가중합으로 구성하였다. 이는 Jina Embeddings v4의 기존 Retrieval 어댑터 학습에 사용했던 손실함수와 동일한 구조이다.

$$L_{total}=L_{NCE+}(B_{single})+L_{NCE+}(B_{multi}) + w_{kl}\cdot D_{KL}(P_{multi}\|P_{single}) \quad (4)$$

여기서 $L_{NCE+}(B_{single})$과 $L_{NCE+}(B_{multi})$는 각각 Single Vector 배치와 Multi Vector 배치에 대한 Extended InfoNCE Loss이며, $D_{KL}(P_{multi}\|P_{single})$은 두 벡터의 유사도 확률 분포 간 KL Divergence이다. w_{kl}은 KL Divergence의 가중치로, 두 손실 항 $L_{NCE+}(B_{single})$, $L_{NCE+}(B_{multi})$와의 균형을 조절한다.

Single Vector 간 유사도 $s(v_q^S,\ v_p^S)$는 코사인 유사도로 계산한다. Multi Vector 간 유사도 $s(V_q^M,\ V_p^M)$는 토큰 수준의 정밀한 상호작용을 반영하기 위해 챔퍼 유사도(Chamfer Similarity, Max-sim)를 사용하며, 식 (5)와 같이 정의된다.

$$s(V_q^M,\ V_p^M)=\frac{1}{T_q}\sum_{i=1}^{T_q}\max_{j\in[T_p]}\cos(v_{q_i}^M,\ v_{p_j}^M) \quad (5)$$

여기서 $v_{q_i}^M$는 Query Multi Vector의 i번째 유효 토큰 임베딩이고, $v_{p_j}^M$은 Positive Multi Vector의 j번째 유효 토큰 임베딩이다. T_q와 T_p는 각각 패딩 토큰을 제외한 Query와 문서의 유효 토큰 수를 의미한다. 문장마다 토큰 수가 달라 제로 패딩이 필요하므로, 패딩 토큰이 유사도 계산에 영향을 주지 않도록 어텐션 마스크를 적용하였다.

식 (4)의 $D_{KL}(P_{multi}\|P_{single})$은 식 (6)과 같이 정의된다.

$$D_{KL}(P_{multi}\|P_{single})= E_{r\sim B}\left[\sum_{i=1}^{|B|}P_{multi}(p_i|q)\ln\frac{P_{multi}(p_i|q)}{P_{single}(p_i|q)}\right] \quad (6)$$

여기서 $P_{single}(p_i|q)$와 $P_{multi}(p_i|q)$는 각각 쿼리에 대한 배치 내 문서의 유사도 점수를 각각 식 (7)과 식 (8)의 softmax 함수로 정규화한 확률분포이다. 두 확률분포는 Single Vector와 Multi Vector 표현 공간에서의 문서 순위 분포를 나타낸다. KL Divergence 항은 두 표현이 동일한 Query - 문서 쌍에 대해 유사한 순위 구조를 갖도록 유도하는 역할을 한다. 즉, Multi Vector가 포착한 정밀한 토큰 수준의 순위 정보를 Single Vector가 모방하도록 학습을 유도함으로써, Single Vector만 사용하더라도 높은 검색 성능을 확보하도록 한다.

$$P_{single}(p_i|q)=\frac{e^{s(v_q^S,\ v_{p_i}^S)/\tau}}{\sum_{m=1}^{|B|}e^{s(v_q^S,\ v_{p_m}^S)/\tau}} \quad (7)$$

$$P_{multi}(p_i\mid q)=\frac{e^{s(V_q^M,\ V_{p_i}^M)/\tau}}{\sum_{m=1}^{|B|}e^{s(V_q^M,\ V_{p_m}^M)/\tau}} \quad (8)$$

3.3 멀티모달 통합 검색 시스템 구축

본 연구에서는 멀티모달 수사 증거 데이터의 통합 저장과 검색을 위해 2.2절에서 검토한 Qdrant 벡터 데이터베이스와 LangGraph 기반 검색 파이프라인을 결합한 end-to-end 검색 시스템을 구축하였다. 사이버범죄 수사 데이터는 정형 데이터와 비정형 데이터가 혼재되어 있다. 두 유형의 데이터를 통합 검색하기 위해 Qdrant의 단일 컬렉션에 정형·비정형 데이터를 함께 인덱싱하였다. 정형 데이터는 행 단위로 식 (2)에 따라 JSON으로 변환 후 임베딩하여 행당 하나의 벡터 포인트를 생성하였으며, 이미지 데이터 역시 건당 하나의 벡터 포인트로 저장하였다.

통합 검색 시스템은 수사 목적과 연산 자원에 따라 Single Vector, Multi Vector, Image-Text Fusion의 세 가지 벡터화 및 검색 전략을 지원하도록 설계하였다. Single Vector 전략은 증거 문서 또는 행 데이터를 하나의 고밀도(Dense) 벡터로 표현하는 방식이다. 각 문서와 자연어 질의를 각각 2,048차원 Dense 벡터로 변환한 후, 코사인 유사도를 기준으로 검색한다. 이 방식은 인덱스 크기가 작고 검색 지연시간이 짧아 저장 공간 효율성과 검색 속도(latency)가 중요한 환경에 적합하다.

Multi Vector 전략은 ColBERT 방식의 late interaction 구조를 기반으로 문서를 토큰 단위의 벡터 집합으로 표현한다. 본 시스템에서는 각 토큰을 128차원 벡터로 변환하고, 질의와 문서의 토큰 벡터 간 챔퍼 유사도를 계산하여 검색한다. Single Vector보다 연산량은 증가하지만, 수사 은어와 세부 단서처럼 미세한 의미 정보를 정밀하게 정렬할 수 있다.

Image-Text Fusion 전략은 이미지 검색 성능을 높이기 위해 본 연구에서 제안한 이미지-텍스트 통합 표현 방식이다. 이 전략은 이미지 비언어적 시각 정보와 언어적 정보를 함께 활용하도록 원본 이미지와 텍스트 메타데이터의 Multi Vector를 결합한다. 구체적으로 3.1절의 데이터 전처리 단계에서 원본 이미지와 해당 이미지에서 생성된 텍스트 메타데이터를 각각 별도의 입력으로 처리하였다. 텍스트 메타데이터는 VLM이 생성한 이미지 설명과 이미지 내 문자 인식 결과로 구성되며, 두 정보를 하나의 문자열로 결합하여 Jina Embeddings v4의 텍스트 임베딩 입력으로 사용하였다. 이후 원본 이미지의 Multi Vector와 텍스트 메타데이터의 Multi Vector를 연결(concatenation)하여 최종 이미지-텍스트 통합 표현을 생성하였다.

검색 파이프라인은 LangGraph를 기반으로 구축하였다. LangGraph는 LLM 기반 애플리케이션을 상태(state) 기반 실행 그래프로 구성하는 프레임워크로, 선형적 체인(chain) 구조와 달리 노드 간 상태 전달, 조건부 분기 및 반복 실행을 지원한다. 이러한 특성은 질의 처리, 벡터 검색, 컨텍스트 구성 및 응답 생성이 결합된 멀티모달 검색 시스템에 적합하다.

본 연구의 검색 파이프라인은 수사관의 자연어 질의를 입력으로 받아 세 단계로 수행된다. 첫째, 검색 노드(retrieve node)는 질의를 벡터화한 후 Qdrant 벡터 데이터베이스에서 유사도 검색을 수행하여 관련성이 높은 정형 데이터 행과 이미지 증거물을 검색한다. 둘째, 컨텍스트 생성 노드(generate context node)는 검색된 증거물의 텍스트와 메타데이터를 LLM 입력에 적합한 컨텍스트 형식으로 변환한다. 셋째, 응답 생성 노드(generate response node)는 구성된 검색 컨텍스트를 바탕으로 최종 응답을 생성한다. 다만 본 연구에서 구현한 LLM 응답을 생성하는 단계는 통합 검색 시스템의 동작 가능성을 확인하기 위한 구조로 실험에서는 LLM의 응답 품질보다 벡터화 전략별 검색 성능 평가에 초점을 두었다.

4. 실 험

4.1 실험 데이터

본 실험에서는 '사이버범죄 수사단서 통합분석 및 추론시스템 개발' 과제에서 구축한 사이버범죄 수사 시뮬레이션 데이터를 사용하였으며, 세부 구성은 <Table 2>와 같다. 해당 데이터는 정형 데이터 19.1%와 비정형 데이터 80.9%로 구성되며, 실제 수사 현장에서 활용되는 증거 유형과 속성 구조를 반영하였다. 또한 개인정보 및 민감정보를 보호하기 위해 실명, 전화번호, 계좌번호, IP 주소 등 개인 식별 가능 정보는 가상의 값으로 대체하거나 마스킹 처리하였다.

정형 데이터는 사기 의심 중고거래 게시글, 통화기록, 채팅 로그, 계좌 거래 내역 등으로 구성되며, 플랫폼 정보, 발·수신 정보, 대화 내용, 입출금 내역 및 IP 정보 등의 속성을 포함한다. 이미지 데이터는 사기 패턴, 유해 사이트, 악성 URL, 스팸 메시지 등의 스크린샷으로 구성된다. 이를 통해 실제 사이버범죄 수사에서 정형 및 비정형 증거가 복합적으로 활용되는 환경을 모사하였다.

본 연구는 위 데이터에 3.1절의 3단계 합성 데이터 생성 파이프라인을 적용하여 Query, Positive, Hard Negative로 구성된 후보 Triplet을 생성하였다. 이후 LLM 기반 품질 평가를 통해 부적절한 후보를 제거하고, <Table 3>과 같은 최종 Triplet 데이터셋을 구축하였다.

정형 데이터의 경우 모든 테이블의 모든 행을 합성 데이터 생성에 사용하지 않고 총 15,000개 행을 샘플링해 후보 Triplet을 생성하였다. 이 중 필터링 단계를 통과한 12,909개가 최종 데이터셋에 포함되어 86.1%의 유지율을 보였다. 이미지 데이터 기반 Triplet은 127개 중 113개가 선별되어 89.0%의 유지율을 보였다. 이에 따라 최종적으로 총 13,022개의 Triplet을 확보하였다. 확보된 데이터셋은 분할하여 학습 데이터 9,112개, 검증 데이터 1,953개, 평가 데이터 1,957개로 구성하였다. 각

Table 2. Raw dataset composition

Data	Modality	Description
Second-hand listings	Tabular	Second-hand marketplace listings flagged as suspected fraud, including platform, keyword, category, product name, post content, price, and posting date
Call records	Tabular	Call log records, including caller number, callee number, start/end time, call duration, origin location latitude/longitude, and call type
Chat logs	Tabular	Chat messages and related metadata, including datetime, sender, message content, and financial information mentioned within chats(e.g., account holder, account number, bank name)
Bank account transactions	Tabular	Transaction records containing deposits, withdrawals, balances, counterparty accounts, and IP information
Fraud/harmful-site screenshots	Image	Screenshots of scam patterns, harmful websites, and phishing pages originating from malicious URLs
Fraud-related messages screenshots	Image	Screenshots of spam and fraud-related messages

Triplet은 하나의 Query, 이에 대응하는 하나의 Positive, 그리고 7개의 Hard Negative로 구성된다.

Table 3. Summary of synthetic triplet dataset

Data Type	Total	Retained	Removed	Retention Rate
Excel-based	15,000	12,909	2,091	86.1%
Image-based	127	113	14	89.0%
Total	15,127	13,022	2,105	86.1%

4.2 하이퍼파라미터 민감도 분석

사전학습된 Jina Embeddings v4 모델을 사이버범죄 수사 도메인에 적응 시, 적절한 하이퍼파라미터 선택은 최종 검색 성능을 결정짓는 핵심 요인이다. 본 연구에서는 각 하이퍼파라미터의 개별 기여도를 정량적으로 분리하기 위해, 통제된 단일 인자 변경(one-at-a-time) 방식에 기반한 민감도 분석 실험을 설계하였다. 본 분석의 세부 목적은 세 가지이다. 첫째, Baseline 설정에서 하이퍼파라미터를 하나만 변경하여 해당 변수가 임베딩 표현 품질에 미치는 영향을 독립적으로 관측한다. 둘째, 동일 구성을 서로 다른 난수(seed)로 반복 실행하여 실행 간 분산 $\sigma_{baseline}$을 정량화한다. 이를 통해 후속 실험에서 관측된 변화량이 통계적 노이즈인지, 혹은 유의미한 개선인지를 판정하는 기준을 수립한다. 셋째, 민감도가 유의하게 확인된 변수만을 후속 Bayesian 최적화 탐색 대상으로 한정해 계산 자원의 불필요한 낭비를 방지한다.

본 연구는 Jina AI가 공개한 jinaai/jina-embeddings-v4를 기반 모델로 사용한다. 파인튜닝은 LoRA 기법을 채택하여, 전체 약 3.8B 파라미터 중 학습 대상 파라미터를 약 60M(전체의 1.58%)로 한정하고 나머지 파라미터는 값을 고정하였다. 손실 함수는 식 (4)의 Total Loss 구조를 따르며, Single Vector InfoNCE 손실, Multi Vector InfoNCE 손실, KL Divergence 정규화 항의 가중합으로 구성된다. 모든 실험은 NVIDIA A100 80GB PCIe 단일 GPU 환경에서 순차적으로 수행되었으며, 실험 추적 및 지표 시각화에는 Weights & Biases(Biewald, 2020)을 활용하였다. 학습 데이터는 사이버범죄 수사 시뮬레이션 환경에서 구축된 약 9,100쌍의 Triplet을 사용하였으며, 검증 데이터로는 1,953개의 Query – Positive 쌍을 활용하였다.

민감도 분석의 모든 ablation 실험은 동일한 공통 Baseline 구성에서 출발하며, 이 Baseline에서 정확히 하나의 하이퍼파라미터만 변경한다. Baseline 값은 Table 4와 같이 구성된다. 본 연구는 Jina Embeddings v4 파인튜닝에서 이론적·경험적으로 가장 큰 영향을 미칠 것으로 예상되는 세 가지 변수에 집중하여 ablation을 설계하였다. Ablation 대상은 학습률(η), contrastive 손실의 Temperature(τ), KL 정규화 가중치(w_{kl})의 세 하이퍼파라미터이며, 각 하이퍼파라미터에 대해 Baseline 대비 낮은 값과 높은 값 두 점을 선정하였다. 아울러 ablation에서 관측된 성능 변화가 하이퍼파라미터 변경의 실질적 효과인지 실험 실행 간 자연 발생하는 분산에 의한 것인지를 판별하기 위해, 동일 Baseline 구성을 세 개의 상이한 난수 seed(42, 123, 777)로 독립 실행하여 분산 $\sigma_{baseline}$을 추정하였다. 따라서 전체 실험은 총 9개의 실험으로 구성되며 구성은 <Table 5>와 같다. 해당 실험에서 민감도가 유의하게 확인된 변수만을 4.3절의 Bayesian 최적화 대상에 포함한다.

학습률(η)은 파인튜닝 과정에서 사전학습 표현의 보존과 도메인 적응 속도 사이의 균형을 결정하는 통제 변수이다. 과도하게 큰 학습률은 사전학습 가중치로부터의 급격한 이탈을 유발하여 catastrophic forgetting을 초래하며, 반대로 과도하게 작은 학습률은 도메인 특화 신호를 학습하지 못해 과소적합(underfitting)을 초래한다. 본 연구는 Baseline($3e-5$)을 중심으로 이 두 극단에 해당하는 $1e-5$와 $1e-4$를 선정하여 양측 경계에서의 학습 안정성을 함께 관찰하였다.

Table 4. Common baseline configuration

Category	Hyperparameter	Value
Batch	Micro batch size	32
	Gradient accumulation steps	4
	Effective batch size	128
Optimization	Learning rate(η)	3×10^{-5}
	Weight decay	0.01
	Warmup ratio	0.1
	Max grad norm	1.0
	Epochs	5
Loss	Temperature(τ)	0.040
	w_{single}/w_{multi}	1.0/1.0
	KL weight(w_{kl})	0.20
	Matryoshka dims	{128,256,512,1024,2048}
Reproducibility	Seed	42

Table 5. Hyperparameter sensitivity analysis experiment configurations

#	Run ID	Role	η	τ	w_{kl}	Seed
1	a01_baseline_seed42	Baseline	$3e-5$	0.040	0.2	42
2	a02_lr_low	$\eta\downarrow$	$1e-5$	0.040	0.2	42
3	a03_lr_high	$\eta\uparrow$	$1e-4$	0.040	0.2	42
4	a04_temp_low	$\tau\downarrow$	$3e-5$	0.020	0.2	42
5	a05_temp_high	$\tau\uparrow$	$3e-5$	0.065	0.2	42
6	a06_wkl_zero	no KL	$3e-5$	0.040	0.0	42
7	a07_wkl_high	$w_{kl}\uparrow$	$3e-5$	0.040	0.4	42
8	a08_baseline_seed123	Baseline rerun	$3e-5$	0.040	0.2	123
9	a09_baseline_seed777	Baseline rerun	$3e-5$	0.040	0.2	777

Temperature(τ)는 분포 sharpness를 조절하며 낮은 Temperature는 softmax 분포를 극단적으로 집중시켜 가장 유사한 오답 하나에 학습 신호를 편중시키며, 이는 임베딩 공간의 예리한 분리를 유도하지만 과적합과 일반화 약화의 위험을 동반한다. 반대로 높은 Temperature는 분포를 완만하게 하여 다수의 오답에 신호를 고르게 분산시킴으로써 넓은 margin과 유연한 의미 정렬을 형성한다. 탐색 범위는 Hard Negative 집중에 해당하는 0.02와 soft alignment에 해당하는 0.065로 설정하였다.

KL 정규화 가중치(w_{kl})는 식 (4)의 Total 손실에서 Single Vector 손실 및 Multi Vector 손실 항과 KL Divergence 항 사이의 균형을 조절하는 역할을 한다. KL Divergence 항은 동일 Query – Positive 쌍에 대해 Single Vector와 Multi Vector 표현 공간이 유사한 문서 순위 구조를 갖도록 유도하는 내부 정규화 기제로, 토큰 수준 정밀 매칭에 기반한 Multi Vector의 유사도 분포를 Single Vector가 모방하도록 학습을 유도함으로써 Single Vector만으로도 높은 검색 성능을 확보할 수 있도록 한다. $w_{kl}=0$ 조건은 이 정규화 항을 완전히 제거한 순수 대조(contrastive) 학습에 해당하므로, 단순한 탐색 변수로서의 역할을 넘어 본 연구에서 설계한 Single-Multi 표현 공간 정렬 방법론 자체의 기여를 정량화하는 방법론 ablation의 성격을 함께 가진다. 반대로 $w_{kl}=0.4$의 강한 정규화 조건은 KL 항의 과도한 반영이 contrastive 학습 신호를 약화시키는지 여부를 확인하기 위해 선정하였다.

주 평가지표로는 식 (9)의 Multi Vector Margin을 사용하였다.

$$Margin=\mu_{pos}-\mu_{neg}, \tag{9}$$

여기서 μ_{pos}는 Query와 Positive 간 평균 유사도, μ_{neg}는 Query와 Hard Negative 간의 평균 유사도이다. 여기서 μ_{neg}는 각 Query에 대한 7개 Hard Negative와의 유사도 평균값으로 계산한다. Margin이 클수록 모델이 정답 문서와 오답 문서를 임베딩 공간에서 명확히 분리함을 의미한다. Single Vector에는 코사인 유사도를, Multi Vector에는 챔퍼 유사도를 적용하여 각각 Single Vector Margin 및 Multi Vector Margin으로 구분하여 추적 가능하며, 본 분석의 주 평가지표는 Multi Vector Margin이다. 각 epoch 종료 시점마다 검증 데이터 전체에 대해 Margin을 산출하였으며, epoch 0(학습 전 초기 상태)에서도 동일한 지표를 측정하여 파인튜닝 전 기준선을 확보하였다.

보조지표로는 Single Vector Margin, Multi/Single Error Rate, 검증 손실을 epoch마다 함께 기록하여 학습 역학의 다각적 분석을 가능하게 하였다. 여기서 Error Rate는 1-Recall@1로 정의되며, 정답 항목이 검색 결과의 1순위에 위치하지 못한 비율을 의미한다. 따라서 Error Rate가 낮을수록 Top-1 검색 정확도가 높음을 나타낸다. Multi Error Rate와 Single Error Rate는 각각 Multi Vector 및 Single Vector 검색 방식에서 동일한 기준으로 산출하였다.

하나의 ablation 실험에서 관측된 Multi Margin의 변화량이 단순 실행 노이즈가 아닌 하이퍼파라미터 변경의 실제 효과임을 판정하기 위해 식 (10)의 기준을 채택한다.

$$\left|Margin_{ablation}-\overline{Margin}_{baseline}\right|\geq 2\sigma_{baseline} \tag{10}$$

여기서 $Margin_{ablation}$은 특정 하이퍼파라미터를 변경한 ablation 실험의 최종 epoch Multi Vector Margin 측정값이며, $\overline{Margin}_{baseline}$과 $\sigma_{baseline}$은 동일 구성을 서로 다른 세 개의 난수 seed(42, 123, 777)로 독립 실행한 Baseline 실험의 최종 epoch Multi Vector Margin 평균과 표준편차이다. $2\sigma_{baseline}$ 기준은 정규분포 가정하에서 약 95% 신뢰구간에 해당하며, 관측된 변화를 통계적으로 유의미한 효과로 판정하기 위한 최소 기준이다.

세 seed Baseline 실행(a01/a08/a09)의 Multi Vector Margin 평균은 0.4243, 표준편차는 0.0010으로 $2\sigma_{baseline}=0.0021$ 산출되었다. 실행 간 수치적 동요가 소수점 셋째 자리 수준으로 극히 작음을 확인하였으며, 이는 이후 ablation 판정의 엄격한 통계 기준이 된다. 각 ablation의 유의성 판정은 Table 6에 정리하였으며, 에폭별 Multi Vector Margin 학습 곡선은 Figure 4에 제시하였다. 전체 실험 결과와 보조 지표의 에폭별 학습 곡선은 각각 Appendix B, C에 정리하였다.

<Table 6>에서 세 가지 패턴이 명확히 드러난다. 첫째, Temperature는 모든 변수 중 가장 강한 영향력을 보였다. τ를

Table 6. Ablation results and statistical significance (Multi Vector Margin)

Run	Changed Variable	Multi Margin	△	Ratio to 2σ	Significant
Baseline (mean$\pm\sigma$)	Baseline	0.4243 ($\pm$0.0010)	-	-	-
a02_lr_low	$\eta\downarrow$	0.3789	-0.0454	21.6	Yes
a03_lr_high	$\eta\uparrow$	0.4606	+0.0363	17.3	Yes
a04_temp_low	$\tau\downarrow$	0.3129	-0.1114	53.0	Yes
a05_temp_high	$\tau\uparrow$	0.5192	+0.0949	45.2	Yes
a06_wkl_zero	no KL	0.4233	-0.0010	0.48	No
a07_wkl_high	$w_{kl}\uparrow$	0.4249	+0.0006	0.29	No

0.040에서 0.065로 상향한 a05는 $\triangle=+0.0949$로 Baseline 대비 22.4%의 상대적 향상을 기록하였으며, 반대로 0.020으로 하향한 a04는 $\triangle=-0.1114$로 26.3% 악화되었다. 두 방향 모두 2 $\sigma_{baseline}$의 45배 이상에 달하는 변화폭으로, 본 분석에서 관측된 효과 중 가장 강력한 인과 신호이다. 둘째, 학습률은 중간 수준의 유의한 영향을 보였다. η를 1e-4로 상향한 a03은 $\triangle=+0.0363$의 개선을, 1e-5로 하향한 a02는 $\triangle=-0.0454$의 악화를 나타내어 두 방향 모두 유의한 효과가 확인되었다. 이는 Baseline의 $\eta=3\times10^{-5}$가 본 데이터 규모에서 과도하게 보수적이었음을 시사한다. 셋째, KL Distillation 가중치는 완전 제거(a06)와 강화(a07) 모두에서 변화량이 $2\sigma_{baseline}$ 이하에 머물러 통계적 유의성을 확인할 수 없었다. $2\sigma_{baseline}$ 대비 배율이 각각 0.48×, 0.29×로 노이즈 수준임을 고려할 때, w_{kl}은 후속 탐색 대상에서 제외할 수 있는 충분한 근거를 확보하였다.

<Figure 4>는 변수별 에폭에 따른 Multi Vector Margin 학습 곡선을 나타낸다. 회색 밴드는 3-seed Baseline의 평균$\pm 2\sigma$ 범위를 나타내며, 각 ablation 곡선이 이 밴드를 벗어나는 정도가 해당 변수의 실질적 영향력을 시각적으로 보여준다.

Temperature (중앙)에서 $\tau=0.065$의 곡선이 Epoch 1부터 Baseline 밴드를 명확히 상회하며 수렴하는 반면, $\tau=0.020$은 Baseline 밴드 아래로 크게 이탈한다. Learning Rate (좌)에서는 $\eta=1\times10^{-4}$가 Baseline을 안정적으로 상회하나 Temperature 만큼의 격차는 아니다. KL Distillation Weight (우)에서는 $w_{kl}=0.00$과 $w_{kl}=0.40$ 두 곡선이 Baseline 밴드와 구분되지 않아 영향이 없음을 시각적으로 확인할 수 있다.

본 분석 결과는 후속 Bayesian 최적화 실험의 탐색 공간을

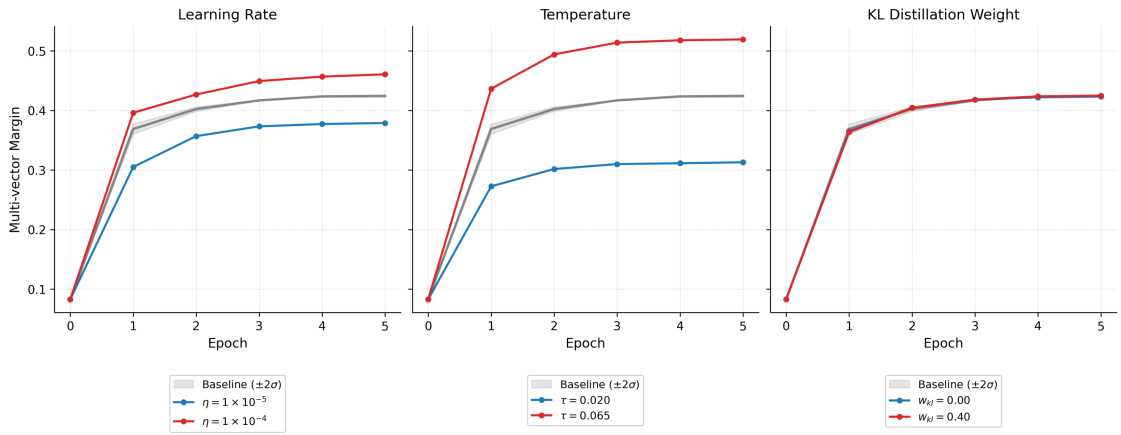


Figure 4. Multi Vector Margin learning curves across epochs. The three panels show ablation results for Learning Rate (left), Temperature (center), and KL Distillation Weight (right). The gray band denotes the 3-seed Baseline mean ± 2σ range.

Table 7. Revised hyperparameter search space for subsequent Bayesian optimization

Hyperparameter	Previous Range	Revised Range	Rationale
η	[1e-5, 1e-4]	[3e-5, 2e-4] Log-uniform	Improvement continues at 1e-4; upper bound unexplored
τ	[0.02, 0.065]	[0.04, 0.10] Uniform	Monotonic improvement up to 0.065; saturation point unconfirmed
w_{kl}	[0.00, 0.40]	Fixed at 0.20	No statistically significant in-domain effect; excluded from search

<Table 7>과 같이 재조정하는 근거를 제공한다. τ와 η는 모두 현재 Baseline보다 상향 조정이 유리함이 확인되어 탐색 범위를 상향 확장하는 반면, w_{kl}은 성능에 유의미한 영향이 없는 것으로 판명되어 고정값으로 전환함으로써 탐색 차원을 3차원에서 2차원으로 축소하였다.

4.3 Bayesian 최적화 기반 하이퍼파라미터 탐색

4.2절의 통제된 단일 변수 ablation 분석을 통해 학습률(η)과 Temperature(τ)가 Multi Vector Margin에 유의한 영향을 미치는 변수임을 확인하였으며, 두 변수 모두 Baseline 대비 상향 조정이 유리하되 포화점이 미확인 상태임을 규명하였다. 그러나 해당 실험은 본질적으로 변수 간 상호작용 효과(interaction effects)를 포착하지 못한다는 구조적 한계를 가지며, 따라서 단일 변수 최적화의 합집합이 다변량 최적해와 일치한다는 보장이 없다. 이 한계를 해소하고 다변량 공간에서의 진정한 최적 구성을 식별하기 위해, 본 연구는 Bayesian 최적화 기반 다변량 탐색을 수행하였다.

<Table 7>에 정리된 바와 같이 각 변수의 탐색 범위는 학습률은 자릿수에 따른 효과가 비대칭이므로 Log-uniform 분포로 sampling하여 자릿수별 균등 탐색을 보장하였으며, Temperature는 관심 범위 내에서 효과가 선형에 가깝다고 관찰되어 uniform 분포를 채택하였다. η 하한은 4.2절 Baseline, 상한은 catastrophic forgetting 위험을 고려하여 2e-4로 제한하였다. τ 하한은 4.2절 Baseline, 상한은 0.065에서 단조 개선이 확인되었으나 포화점이 미확인 상태임을 고려하여 0.10까지 확장하였다.

본 연구는 Weights & Biases Sweep이 제공하는 Gaussian Process(GP) 기반 Bayesian 최적화를 채택하였다. GP 기반 Bayesian 최적화는 지금까지 관측된 (하이퍼파라미터, 성능 지표) 쌍으로부터 목적 함수의 사후 분포(posterior distribution)를 추정한 뒤, 획득 함수(acquisition function)를 최대화하는 지점을 다음 sampling 후보로 선정한다(Snoek *et al.*, 2012).

Bayesian 최적화의 최대화 대상은 검증 데이터 1,953개 샘플에 대해 산출한 Multi Vector Margin으로 설정하였다. 해당 지표는 4.2절에서 다른 어떤 지표보다 하이퍼파라미터 변화에 민감하고, Error Rate와 달리 포화 현상이 나타나지 않음이 확인된 바 있다. 본 실험은 4.2절과 동일한 학습 엔진, 데이터 구성 및 하드웨어 환경에서 수행되었다.

전체 탐색의 trial 상한은 20회로 설정하였으며, 계산 효율성을 제고하기 위해 Hyperband 기반 조기 종료 정책을 함께 적용하였다. Hyperband는 각 trial의 중간 학습 결과를 평가하여 성능이 낮은 trial은 조기에 중단하고, 성능이 상대적으로 우수한 trial만 끝까지 학습시키는 방식이다. 실제 sweep 수행 결과, 계획된 20회 trial 중 9회가 5 epoch 전체 학습을 완료하였다. 나머지 11회는 중간 평가 단계에서 성능이 낮은 trial로 판정되어 조기 종료되었다.

<Table 8>은 5 epoch 학습을 완료한 9개 trial의 검증 결과를 Multi Margin 기준 내림차순으로 정리한 것이다. 상위 4개 trial은 모두 학습률 $\eta = 2 \times 10^{-4}$와 $\tau \in [0.078, 0.094]$ 영역에 집중되어 있으며, 이는 4.2절 단변량 하이퍼파라미터 분석이 시사한 방향과 정확히 일치한다.

Multi Vector Margin만을 기준으로 할 경우 apricot-sweep-10이 0.6134로 가장 높은 값을 보였다. 그러나 해당 trial은 Best Validation Loss가 0.4276으로 상대적으로 높게 나타나, margin 개선이 검증 손실의 감소와 반드시 일치하지 않음을 보여준다. 반면 eternal-sweep-7은 Multi Vector Margin이 0.6004로 상위권에 해당하면서도 Best Validation Loss가 0.2361로 낮게 나타났다. 이는 단순히 margin을 최대화하는 조합보다, margin과 검증 손실 간의 균형을 고려한 하이퍼파라미터 선정이 필요함을 시사한다.

이에 본 연구는 완료 trial의 결과를 단일 목적 함수만으로 해석하지 않고, Multi Vector Margin과 Best Validation Loss의 동시 최적화 관점에서 추가적으로 Pareto front를 분석하였다. Figure 5는 Multi Vector Margin은 최대화하고 Best Validation Loss는 최소화하는 방향으로 산출한 Pareto front를 나타낸다. Pareto front에 포함된 trial은 다른 trial에 의해 두 지표 모두에서 동시에 열등하다고 판단되지 않는 비지배 해(non-dominated solution)로 해석된다. 즉, 해당 trial들은 margin 향상과 검증 손실 감소 사이에서 서로 다른 trade-off를 형성하는 후보군이다.

Table 8. Validation results for Bayesian optimization completed trials (n = 9, sorted by Multi Vector Margin)

Rank	Run Name	η	τ	Multi Margin	Best Val Loss
1	apricot-sweep-10	1.99e-4	0.0933	0.6134	0.4276
2	winter-sweep-6	1.97e-4	0.0875	0.6064	0.3030
3	noble-sweep-8	1.95e-4	0.0838	0.6057	0.2696
4	eternal-sweep-7	1.99e-4	0.0784	0.6004	0.2361
5	cool-sweep-4	9.75e-5	0.0979	0.5845	0.5498
6	lively-sweep-5	1.03e-4	0.0907	0.5835	0.4516
7	rosy-sweep-3	1.28e-4	0.0984	0.5782	0.6143
8	clear-sweep-3	5.21e-5	0.0964	0.5594	0.5899
9	summer-sweep-2	1.11e-4	0.0519	0.5114	0.1525

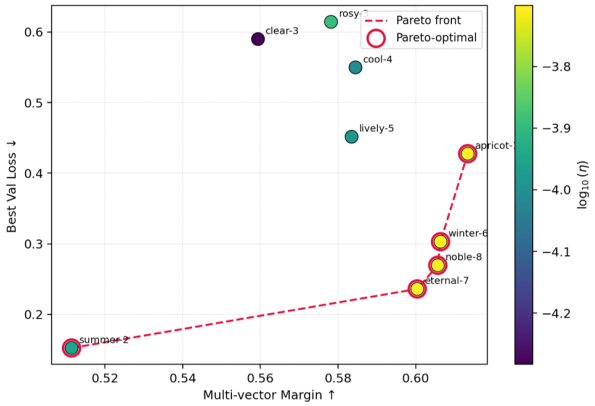


Figure 5. Pareto Front of Completed Bayesian Optimization Trials

<Figure 5>에서 summer-sweep-2, eternal-sweep-7, noble-sweep-8, winter-sweep-6, apricot-sweep-10이 Pareto-optimal 후보로 확인되었다. 이 가운데 summer-sweep-2는 가장 낮은 Validation Loss를 보였으나 Multi Vector Margin이 0.5114로 낮아 최종 후보로 보기 어렵다. 반대로 apricot-sweep-10은 가장 높은 Multi Vector Margin을 달성하였으나 Validation Loss가 상대적으로 크게 증가하였다. winter-sweep-6과 noble-sweep-8은 높은 margin을 유지하면서 Validation Loss를 낮춘 후보이나, eternal-sweep-7은 이들보다 낮은 Validation Loss를 보이면서도 Multi Vector Margin이 0.6004로 0.60 수준을 유지하였다.

따라서 본 연구에서는 Multi Vector Margin 단일 지표상 최고값을 보인 apricot-sweep-10과 Pareto front상에서 margin과 검증 손실 간의 절충이 가장 안정적으로 나타난 eternal-sweep-7을 최종 최적 하이퍼파라미터 조합 후보로 선정하였다. 두 해당 조합의 재현 가능성을 검증하기 위해 후속 실험에서는 동일 하이퍼파라미터 조건에 대해 복수 seed 기반 반복 실험을 수행하였다.

<Table 9>는 <Table 6>의 Baseline과 Bayesian 최적화에서 도출된 두 Pareto-optimal 후보의 값을 3개의 random seed에 대해 동일한 학습 조건으로 반복 실험한 결과의 평균과 표준편차로 제시하였다. 비교 지표는 본 연구의 주요 목적 함수인 Multi Vector Margin, top-1 검색 오류율을 나타내는 Multi Vector Error Rate, 그리고 학습 안정성을 반영하는 Best Validation Loss로 구성하였다.

두 Bayesian 최적화 후보는 모두 Baseline 대비 Multi Vector Margin을 크게 개선하였다. apricot-sweep-10은 0.6041 ± 0.0042를 기록하여 Baseline의 0.4243 ± 0.0010 기준 0.1798(상대 비율 42.4% 향상) 개선되었으며, eternal-sweep-7 역시 0.6019 ± 0.0045로 절댓값 기준 0.1776(상대 비율 41.9% 향상)의 향상을 보였다. 하지만, 두 후보 간 Multi Vector Margin 차이는 0.0022에 불과하여, margin 개선 효과는 실질적으로 유사한 수준으로 해석된다.

그러나 Multi Vector Error Rate와 Best Validation Loss에서는 두 후보 간 차이가 뚜렷하게 나타났다. apricot-sweep-10은 가장 높은 Margin을 유지하였으나, Error Rate가 0.0415 ± 0.0029로 증가하였고 Best Validation Loss 또한 0.4494 ± 0.0112로 크게 상승하였다. 반면 eternal-sweep-7은 Error Rate가 0.0288 ± 0.0016으로 Baseline의 0.0273 ± 0.0016과 유사한 수준을 유지하였으며, Best Validation Loss도 0.2333 ± 0.0036으로 apricot-sweep-10 대비 낮게 나타났다.

따라서 eternal-sweep-7은 apricot-sweep-10과 거의 동일한 수준의 Margin 개선을 달성하면서도, Error Rate의 악화를 최소화하고 검증 손실의 증가를 상대적으로 억제한 구성으로 판단된다. 이에 본 연구에서는 단일 목적 함수상 최고 Margin을 보인 apricot-sweep-10이 아니라, Margin 개선과 검색 오류율 유지, 검증 손실 안정성을 종합적으로 고려하여 eternal-sweep-7 ($\eta = 1.99e-4$, $\tau = 0.0784$)을 최종 하이퍼파라미터 조합으로 채택하였다.

4.4 멀티모달 통합 검색 시스템 평가

(1) 검색 시스템 평가 실험 설정

앞선 4.2절과 4.3절의 하이퍼파라미터 최적화 프로세스는 검증 데이터셋의 Multi Vector Margin을 대리 지표(proxy metric)로 채택하여 진행되었다. Multi Vector Margin이 임베딩 공간에서 쿼리-정답 쌍과 쿼리-오답 쌍 간의 분리 정도를 효과적으로 반영하는 지표일지라도, 해당 지표만으로는 실제 검색 환경에서의 다운스트림 성능, 즉 Hit Rate@K, NDCG@K, Mean Rank 등으로 대변되는 절대적 검색 품질을 직접 확인하기 어렵다. 또한 4.3절에서 관측된 성능 향상이 특정 벡터화 방식에만 국한된 결과가 아닌, 3.3절에서 제안한 멀티모달 통합 검색 시스템의 세 가지 벡터화 전략 전반에서 일관되게 재현되는지 규명할 필요가 있다.

이에 본 절에서는 4.3절에서 최종 채택한 하이퍼파라미터 조합인 eternal-sweep-7($\eta = 1.99 \times 10^{-4}$, $\tau = 0.0784$, $w_{kl} = 0.20$)을 적용한 모델을 실제 Qdrant 기반 검색 환경에 배치하고, 다운스트림 검색 성능을 평가하였다. 본 평가는 파인튜닝을 거치지 않은 Retrieval 어댑터를 사용한 Jina Embeddings v4 모델(Base Jina v4)과 생성한 합성 Triplet 데이터를 사용해 LoRA 파인튜닝한 제안

Table 9. Performance comparison of the previous baseline and two pareto-optimal Bayesian optimization candidates

Model	Multi Vector Margin	Multi Vector Error Rate	Best Validation Loss
Baseline	0.4243 ± 0.0010	0.0273 ± 0.0016	0.1539 ± 0.0023
apricot-sweep-10	0.6041 ± 0.0042	0.0415 ± 0.0029	0.4494 ± 0.0112
eternal-sweep-7	0.6019 ± 0.0045	0.0288 ± 0.0016	0.2333 ± 0.0036

모델을 대상으로, Single Vector, Multi Vector, Image-Text Fusion의 세 가지 벡터화 전략을 각각 적용하여 총 6가지 조건에서 수행되었다. 이러한 2×3 비교 설계는 모델 요인과 벡터화 전략 요인을 분리하여, 도메인 특화 파인튜닝이 각 검색 전략에 미치는 순효과를 독립적으로 분석하기 위한 것이다.

특히 Base Jina v4 모델과 제안 모델 간의 공정한 비교 실험을 위해, 파인튜닝 적용 여부를 제외한 모든 실험 통제 변수를 동일하게 설정하였다. 구체적으로 Base Jina v4 모델 평가에도 제안 모델과 완전히 동일한 Retrieval 어댑터(파인튜닝 전 상태)를 사용하였으며, 입력 프롬프트 템플릿 및 데이터 포맷 역시 동일하게 적용하였다. 아울러 Qdrant는 내부적으로 HNSW(Hierarchical Navigable Small World) 그래프 기반 근사 최근접 이웃 인덱스를 사용하며, 본 평가에서는 검색 정확도와 속도의 균형점으로 일반적으로 채택되는 hnsw_ef=128 설정을 모든 조건에 동일하게 적용하였다. 그 외 평가 코퍼스 및 쿼리셋, 하드웨어 인프라 환경까지 통제하여 오직 도메인 적응 학습 여부에 따른 순효과만 명확하게 비교되도록 아키텍처를 구성하였다.

평가 코퍼스는 사이버범죄 수사 시뮬레이션 환경에서 수집된 다출처 데이터로 구성하였다. 각 벡터화 전략별로 정형 데이터는 Excel 및 Parquet 파일의 행 단위로, 비정형 이미지 데이터는 PNG 파일 단위로 단일 Qdrant 컬렉션에 30,180개의 데이터 포인트를 인덱싱하였다. 평가 쿼리는 4.1절에서 분할한 평가 데이터셋 1,957개를 그대로 사용하였으며, 각 쿼리에 대해 컬렉션에서 Top-100 검색을 수행한 뒤 정답 매칭 규칙에 따라 첫 번째 정답의 순위를 기록하였다. 이미지 데이터는 파일명을 기준으로, 정형 데이터는 파일명, 시트명, 행 번호의 조합을 기준으로 정답 여부를 판정하였다.

성능 지표는 표준 검색 지표와 운영 관점의 확장 지표로 구성하였다. 본 평가는 각 쿼리에 대해 정답이 단일하게 정의되는 환경이므로, 표준 검색 지표로는 정답의 순위 정보에 기반한 식 (11)-(13)의 세 가지 지표를 사용한다. 쿼리에 대한 첫 번째 정답의 순위를 r_i, 전체 쿼리 집합을 Q라 할 때, MRR(Mean Reciprocal Rank)(Voorhees, 1999)은 정답 순위의 역수의 평균으로 정의되며 식 (11)과 같다.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{r_i} \tag{11}$$

Hit Rate@K는 정답이 상위 K순위 안에 회수된 쿼리의 비율로 정의되며 식 (12)와 같고, $K \in \{1, 5, 10\}$에 대해 산출한다. 1[·]은 조건이 참이면 1, 그렇지 않으면 0을 반환하는 지시함수이다.

$$HR@K = \frac{1}{|Q|}\sum_{i=1}^{|Q|}1[r_i \le K] \tag{12}$$

NDCG@K(Normalized Discounted Cumulative Gain)(Järvelin and Kekäläinen, 2002)는 정답의 순위에 따른 누적 이득을 이상적인 순위(정답이 1위인 경우)로 정규화한 지표로, 단일 정답 환경에서 식 (13)과 같이 단순화된다. 본 연구에서는 K=10에 대해 산출한다.

$$NDCG@K = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1[r_i \le K]}{\log_2(r_i+1)} \tag{13}$$

MRR과 NDCG@K는 정답의 순위 자체를 고려하므로 상위 결과의 품질에 민감한 반면, Hit Rate@K는 임계 순위 K 안에 정답이 회수되었는지 여부만 판단한다.

Long-tail 검색 성능을 확인하기 위해 Hit Rate@50, Mean Rank, Fail Rate@100을 추가로 산출하였다. Mean Rank는 정답이 회수된 쿼리에 한해 첫 번째 정답의 평균 순위를 의미하며, Fail Rate@100은 top-100 검색 결과 내에 정답이 포함되지 않은 쿼리의 비율을 의미한다.

(2) 검색 시스템 평가 실험 결과

<Table 10>은 총 6가지 실험 조건별 최종 다운스트림 검색 성능을 요약한 표이다. 첫째, 도메인 특화 파인튜닝에 따른 검색 성능 향상 효과는 Single Vector 전략에서 가장 두드러지게 나타났다. Base Jina v4의 Single Vector 조건은 MRR 0.5417, HR@1 0.4599로 동일 모델의 Multi Vector 조건에 비해 현저히 낮은 성능을 보였다. 그러나 제안 모델의 Single Vector 조건은 MRR 0.9509, HR@1 0.9208을 기록하여 Baseline 대비 각각 40.92%p, 46.09%p 검색 성능 향상을 달성하였다. 또한 Fail Rate@100은 0.2090에서 0.0015로 감소하여, Top-100 내 정답 회수 실패 비율이 거의 제거되었다.

해당 결과는 운영 환경에서 중요한 의미를 갖는다. Single

Table 10. Retrieval performance comparison across six conditions

Model	Strategy	MRR	HR@1	HR@5	HR@10	HR@50	NDCG@10	Mean Rank	Fail Rate@100
Base Jina v4	Single	0.5417	0.4599	0.6423	0.6919	0.7690	0.5746	5.88	0.2090
Base Jina v4	Multi	0.9231	0.8769	0.9806	0.9882	0.9959	0.9391	1.47	0.0031
Base Jina v4	Fusion	0.9222	0.8753	0.9806	0.9882	0.9954	0.9385	1.47	0.0031
Ours	Single	0.9509	0.9208	0.9842	0.9918	0.9974	0.9609	1.35	0.0015
Ours	Multi	0.9622	0.9382	**0.9918**	**0.9954**	**0.9990**	0.9704	1.22	0
Ours	Fusion	**0.9641**	**0.9407**	**0.9918**	**0.9954**	**0.9990**	**0.9718**	1.21	0

Vector 방식은 Multi Vector 방식에 비해 저장 공간과 검색 latency 측면에서 유리하지만, Baseline 모델에서는 정확도 손실이 커 실용적 적용에 한계가 있었다. 반면 본 연구의 파인튜닝 모델은 Single Vector 조건에서도 Baseline Multi Vector 조건의 MRR 0.9231을 상회하는 성능을 보였다. 이는 제안한 학습 방식이 Single Vector 표현 공간에서도 도메인 내 정답과 오답의 분리 구조를 효과적으로 학습하였음을 의미한다. 다만 4.2절과 Appendix B 및 C에서 확인할 수 있듯이, KL Divergence 가중치 변화에 따른 Single Vector Margin의 차이는 통계적으로 유의하지 않았다. 따라서 이러한 성능 향상을 KL 정규화 항의 독립적인 효과로 해석하기는 어려우며, 전체 손실 함수를 적용한 파인튜닝 과정의 종합적인 효과로 보는 것이 타당하다.

둘째, 새롭게 제안된 Image-Text Fusion 전략의 유효성이 실제 검색 환경에서 확인되었다. 제안 모델에서 Fusion 전략은 모든 지표에서 6개 조건 중 가장 높은 성능을 보였다. 이는 동일 제안 모델의 Multi Vector 조건과 비교해도 MRR은 0.19%p, HR@1은 0.25%p 높은 값이다. 비록 개선 폭은 크지 않지만, 모든 주요 지표에서 일관된 우위를 보였다는 점에서 의미가 있다. 또한, Base Jina v4에서는 Fusion 전략이 Multi Vector 전략보다 미세하게 낮은 MRR을 보였으나, 제안 모델에서는 Fusion 전략이 Multi Vector 전략을 상회하였다. 이는 Base Jina v4 임베딩 공간에서는 이미지와 텍스트 정보의 결합이 반드시 성능 향상으로 이어지지 않을 수 있으나, 도메인 특화 파인튜닝 이후에는 두 모달리티의 정보 결합이 검색 성능 개선에 기여할 수 있음을 보여준다. 따라서 본 연구에서 제안한 Image-Text Fusion 방식은 단순한 정보 결합을 넘어, 파인튜닝된 임베딩 공간에서 멀티모달 단서의 상호보완성을 활용하는 전략으로 해석된다.

셋째, 제안 모델은 Multi Vector 및 Fusion 전략에서 Fail Rate@100 = 0.0000을 달성하였다. 이는 1,957개 평가 쿼리 전체에 대해 정답이 Top-100 내에 모두 회수되었음을 의미한다. 실제 수사 지원 검색 시스템에서 사용자가 관련 증거물을 검색할 때, 정답이 코퍼스 내에 존재함에도 검색 후보군 안에 포함되지 않는 상황은 사용성 측면에서 치명적이다. 따라서 제안 모델이 Multi Vector 및 Fusion 조건에서 완전한 Top-100 recall coverage를 달성한 것은 운영적 안정성 측면에서 중요한 결과이다.

또한 Hit Rate@K가 K = 1, 5, 10, 50 전 구간에서 일관되게 개선되었다는 점은 성능 향상이 특정 상위 순위에만 국한되지 않음을 보여준다. 즉, 제안 모델은 쉬운 쿼리의 정답 순위를 일부 끌어올린 것에 그치지 않고, 검색 난이도 분포 전반에서 정답 회수 능력을 향상시킨 것으로 해석할 수 있다.

마지막으로, 본 평가 결과는 4.2절과 4.3절에서 사용한 Multi Vector Margin의 대리 지표로서의 타당성을 사후적으로 뒷받침한다. 4.3절에서 최종 채택된 eternal-sweep-7 구성은 검증셋에서 Multi Vector Margin을 Baseline 0.4243에서 0.6019로 절댓값 기준 0.1776(상대 비율 41.9% 향상) 개선하였으며, 본 절의 다운스트림 검색 평가에서도 Multi Vector 조건의 MRR을 0.9231에서 0.9622로 총 3.91%p(상대 비율 4.23% 향상) 개선하였다. Margin의 상대적 개선 폭이 MRR의 개선 폭보다 크게 나타나는 것은 두 지표의 측정 단위와 포화 구간이 다르기 때문이다. MRR은 Baseline 단계에서 이미 0.92 이상의 높은 수준에 위치해 추가 개선 가능 범위가 제한적인 반면, Multi Vector Margin은 임베딩 공간의 분리도를 연속적으로 반영하므로 상대적 변화가 더 크게 관찰될 수 있다. 그럼에도 두 지표가 모두 Baseline 대비 일관된 향상 방향을 보였다는 점은, 본 연구의 대리 지표 기반 하이퍼파라미터 선택이 실제 다운스트림 검색 성능으로도 연결되었음을 확인시켜준다.

(3) 이미지 증거 검색 성능 및 신뢰성 평가

<Table 10>에서 제시한 6개 조건의 평가 결과는 전체 1,957개 평가 쿼리에 대한 평균 성능이다. 그러나 평가 쿼리의 구성은 정형 데이터(Excel·Parquet) 행을 정답으로 갖는 쿼리가 1,939개(99.08%)로 대부분을 차지하며, 이미지를 정답으로 갖는 쿼리는 18개(0.92%)에 불과하다. 따라서 전체 평균 성능은 주로 정형 데이터 검색 성능을 반영한다고 볼 수 있다. 반면 3.3절에서 제안한 Image-Text Fusion 전략은 본질적으로 이미지 데이터 검색의 정밀도 향상을 목적으로 설계되었으므로, 그 효과가 전체 쿼리 평균에서는 정형 데이터 쿼리에 의해 희석될 가능성이 있다. 이러한 한계를 보완하기 위해 본 연구는 이미지를 정답으로 갖는 18개 쿼리만을 별도로 추출하고, 해당 부분집합에 대해 제안 모델이 이미지만 사용한 Multi Vector 검색 조건에 비해 Image-Text Fusion 전략이 어떤 효과를 나타내는지 추가로 평가하였다. 부분집합의 표본 크기가 작은 점을 고려하여 비율형 지표에 대해서는 95% Wilson 신뢰구간을 함께 산출하여 결과 해석의 통계적 신뢰성을 보강하였다. 이미지 정답 쿼리 부분집합에 대한 제안 모델의 Multi Vector 전략과 Image-Text Fusion 전략의 검색 성능은 <Table 11>과 같이 나타났다.

이미지 정답 쿼리 부분집합 분석의 결과는 다음과 같이 해석된다. 첫째, Image-Text Fusion 전략은 동일한 학습 가중치에서 출발하였음에도 Multi Vector 전략 대비 현저한 검색 성능 개선을 보였다. Fusion 전략의 MRR은 0.9491·Hit Rate@1은 0.9444인 반면, Multi Vector 전략은 MRR 0.7503·Hit Rate@1 0.6667에 그쳐, 두 전략 간 MRR 19.88%p, Hit Rate@1 27.77%p

Table 11. Retrieval performance of the proposed model on the image-only query subset

Model	Strategy	MRR	HR@1	HR@5
Ours	Multi (Image only)	0.7503	0.6667	0.9444
Ours	Multi (Image-Text Fusion)	**0.9491**	**0.9444**	**0.9444**

의 격차가 관측되었다. Table 10에서 전체 쿼리 기준으로 두 전략의 MRR 차이가 0.19%p에 머물렀던 것과 비교하면, 이미지 정답 쿼리에 한정할 경우 Fusion 전략의 효과가 정형 데이터 쿼리에 의한 희석 없이 본질적 수준으로 드러남을 확인할 수 있다. 이는 3.3절에서 설계 의도로 제시한 '텍스트 기반 의미 검색과 시각적 유사성 기반 검색의 동시 수행' 이 이미지 검색 시나리오에서 결정적 성능 이득으로 직접 연결됨을 정량적으로 입증한다.

둘째, 이러한 격차의 메커니즘은 학습 데이터의 모달리티 분포 편향에서 찾을 수 있다. 4.1절에서 보고한 바와 같이 본 연구의 데이터셋은 Excel 기반 12,909개와 이미지 기반 113개로 구성되어 이미지 데이터 비율이 약 0.87%에 불과하다. 이러한 분포에서 LoRA 기반 대조 학습은 텍스트 도메인 적응에 압도적으로 치우치며, 그 결과 시각 정보에만 의존하는 Multi Vector 전략은 이미지 쿼리에서 도메인 특화 학습의 혜택을 충분히 받지 못한다. 반면 Fusion 전략은 VLM이 생성한 텍스트 메타데이터(설명문 및 OCR 결과)를 시각 임베딩과 결합하므로, 도메인 특화 학습으로 강화된 텍스트 처리 능력을 이미지 검색에 동원할 수 있다. 즉, Fusion은 텍스트와 시각 정보의 상호 보완을 통해 학습 분포 편향의 영향을 완화하는 구조적 이점을 가진다.

다만, 이미지 관련 부분집합의 표본 수가 18개로 제한적이므로, 본 연구의 멀티모달 및 Fusion 전략에 대한 결과는 예비적 근거로 해석될 필요가 있다. Hit Rate@1을 기준으로 95% Wilson 신뢰구간을 산출한 결과, 제안 모델의 Multi Vector 전략은 [0.437, 0.837], Fusion 전략은 [0.742, 0.990]으로 나타났으며, 두 신뢰구간 사이에 일부 중첩이 확인되었다. 이는 소표본 평가 환경에서 두 전략 간 차이를 통계적으로 강하게 단정하기보다는, 성능 경향성을 중심으로 해석하는 것이 적절함을 시사한다.

한편 운영 관점에서는 두 전략의 차이가 평가 시나리오에 따라 달라질 수 있다. Hit Rate@1과 MRR에서는 Fusion 전략이 더 우수한 경향을 보였으나, 사용자가 상위 여러 개의 후보를 함께 검토하는 상황에서는 두 전략의 실용적 차이가 상대적으로 줄어들 수 있다. 특히 Hit Rate@5 기준에서는 Multi Vector 전략과 Fusion 전략 간의 차이가 크게 나타나지 않았다. 종합하면, 이미지 데이터의 비중이 높거나 최상위 검색 결과의 정확도가 중요한 시나리오에서는 Image-Text Fusion 전략이 보다 적합한 선택이 될 수 있다. 다만 향후에는 이미지 중심 평가 쿼리를 확대하여, 멀티모달 및 Fusion 전략의 효과를 보다 안정적으로 검증할 필요가 있다.

(4) 시스템 자원 요구량 및 검색 지연시간 분석

본 연구에서 제안하는 사이버범죄 수사 지원 멀티모달 통합 검색 시스템의 현업 배치 가능성과 자원 효율성을 검토하기 위해, 실험 조건별 모델 파일 저장 크기, 벡터 인덱스 디스크 점유 크기, 쿼리당 평균 검색 latency, 그리고 처리량(QPS, Queries Per Second)을 측정하였다. 모든 실험은 NVIDIA A100 80GB PCIe 단일 GPU와 단일 Qdrant 서버 환경에서 수행하였다. 전체 검색 대상은 30,180개의 데이터 포인트로 구성하였으며, 모든 실험 조건에서 동일하게 사용하였다.

초기 실행으로 인한 지연시간 왜곡을 줄이기 위해 각 조건마다 20회의 사전 조회(Warmup Query)를 수행한 후, 평가 데이터셋(n=1,957)의 전체 쿼리를 대상으로 배치 크기 1(batch=1) 조건하에 쿼리당 평균 지연시간과 처리량을 산출하였다. 전체 측정 결과는 Table 12에 제시하였다. Index Size는 Qdrant 벡터 데이터베이스 내에 구축된 단일/다중 벡터 인덱스가 디스크 상에서 차지하는 물리적 저장 용량을 의미한다. Model Storage는 추론 장비에 모델을 배포할 때 실제 하드디스크에 적재해야 하는 딥러닝 모델 가중치 자체의 물리적 용량이다. 본 연구의 제안 모델은 3.8B 기본 가중치(7.0 GB)를 그대로 공유하면서 파인튜닝된 LoRA 어댑터 가중치(+687 MB)만 추가로 탑재한다. Embed는 수사관의 자연어 질의가 입력되었을 때, 이를 임베딩 모델을 통해 검색용 벡터 표현으로 변환하는 데 소요된 평균 지연시간이다. Search는 생성된 질의 벡터를 기반으로 Qdrant 벡터 데이터베이스 내에서 유사도 연산을 수행하여 상위 K개의 후보 문서를 최종 검색해내는 데 소요된 평균 지연시간이다. Total은 쿼리당 소요되는 총 지연시간으로, 질의의 벡터화(Embed)와 실제 DB 탐색(Search) 시간을 합산한 값이다. p95는 전체 테스트 질의 중 가장 느린 극단치 5%를 제외한, 상위 95%의 질의가 해당 시간 이내에 안정적으로 처리되었음을 보장하는 꼬리 지연시간(Tail Latency) 지표이다. QPS는 초당 시스템이 처리할 수 있는 평균 쿼리 처리 건수를 의미하며, 시스템의 전반적인 처리량(Throughput) 성능을 대변한다.

<Table 12>의 실험 결과 분석을 통해 도출된 핵심 관찰 사항은 다음과 같다. 백본 모델의 3.8B 기본 가중치는 디스크 상에서 약 7.0 GB를 점유한다. 제안 모델은 전체 파라미터를 완전 복제하여 독립된 모델로 배포하는 방식 대신 LoRA 파인튜닝 기법을 적용하였다. 이에 따라 백본 모델의 3.8B 기본 가중치(7.0 GB)를 완전히 공유하면서, 도메인 적응을 위해 학습된 소량의 LoRA 어댑터 가중치(+687 MB)만 추가로 적재하도록 설계되어 물리적 저장 인프라 소모를 극적으로 절감하였다.

반면 벡터 인덱스가 디스크에서 차지하는 용량(Index Size)은 모델 매개변수 크기와 무관하며, 전적으로 3.3절에서 정의한 벡터화 전략에 종속된다. Single Vector 전략의 인덱스 크기는 331MB로 가장 조밀한 반면, Multi Vector와 Fusion 전략은 문서당 다중 토큰 벡터를 개별 저장하는 구조적 특성으로 인해 약 2.6GB 수준의 디스크 공간을 소모하였다.

제안 모델의 쿼리당 총 지연시간을 Base Jina v4와 대조한 결과, 모든 벡터화 전략에서 그 편차가 단 3 ms 이하(측정 노이즈 범위 내)로 수렴하였다. 이는 수사 도메인 적응을 위해 LoRA 기반 학습 가중치를 아키텍처에 추가했음에도, 실제 추론 과정에서는 연산적 병목이나 실질적인 속도 저하를 전혀 유발하

Table 12. Storage footprint, retrieval latency, and throughput

Model	Strategy	Index Size (MB)	Model Storage	Embed (ms)	Search (ms)	Total (ms)	p95 (ms)	QPS
Base Jina v4	Single	331	7.0GB	78.3	8.7	87	95.2	11.5
Base Jina v4	Multi	2,619	7.0GB	78.4	477	555.4	710.7	1.8
Base Jina v4	Fusion	2,634	7.0GB	77	478.4	555.4	714.3	1.8
Ours	Single	331	+687MB	80.2	9.8	89.9	98.9	11.1
Ours	Multi	2,620	+687MB	83	475.5	558.5	713.2	1.8
Ours	Fusion	2,662	+687MB	78.4	474.2	552.6	711.3	1.8

지 않았음을 실증하는 지표이다.

본 검색 시스템의 물리적 연산 병목은 임베딩 생성(Embed, 전 조건 약 77 – 83 ms 수준으로 균일 유지) 단계가 아닌 Qdrant 상에서의 탐색(Search) 단계에서 발생하는 것으로 규명되었다. Single Vector 전략은 Search 연산에 약 8.7 – 9.8 ms(QPS 11.1 – 11.5)만을 소모하였으나, 다중 벡터 간 챔퍼 유사도 계산 및 전수 검색 매칭 연산이 요구되는 Multi Vector와 Fusion 전략은 Search 단계에서만 약 474.2 – 478.4 ms를 소모하여 QPS가 약 1.8 수준으로 하락하였다.

결론적으로, Multi Vector 및 Fusion 전략은 고성능의 하드웨어 자원이 확보되고 검색 정확도의 극대화가 최우선인 특수 수사 시나리오에 적합하다. 반면 제한된 연산 자원이나 온프레미스 단말 환경에서 실시간 응답성이 핵심인 현장에서는, 앞선 4.4절 (2)에서 입증된 바와 같이 베이스라인 다중 벡터 모델의 성능을 상회하는 정확도(MRR 0.9509)를 달성하면서도, 저장공간은 8배 작고 처리 속도는 6배 이상 빠른 'Ours × Single Vector' 전략이 성능과 자원 효율성을 모두 충족하는 실무 최적의 타협점이 될 수 있다.

5. 결론 및 향후 연구 방향

본 연구는 정형 로그 데이터와 비정형 이미지 증거가 혼재하는 사이버범죄 수사 환경에서 이종 증거를 유기적으로 통합 검색하기 위한 도메인 특화 멀티모달 임베딩 모델과 통합 검색 시스템을 제안하였다. 범용 임베딩 모델이 사이버범죄 수사라는 특수 도메인의 고유한 의미론과 은어 맥락을 포착하지 못하는 한계를 극복하고자, Qwen3 계열의 대규모 언어 및 비전 모델을 활용한 3단계 합성 데이터 생성 파이프라인을 설계하여 Triplet 학습 데이터를 확보하였다. 이를 바탕으로 약 3.8B 파라미터 규모의 Jina Embeddings v4 백본 모델에서 Retrieval 어댑터만을 효율적으로 파인튜닝하고 최적화하였다. 또한, 오픈소스 벡터 데이터베이스 엔진인 Qdrant와 상태 기반 실행 그래프 프레임워크인 LangGraph를 결합하여 Single Vector, Multi Vector 및 Image-Text Fusion 검색 전략을 지원하는 통합 검색 아키텍처를 구현하였다.

실험 결과, 본 연구에서 제안한 최적 하이퍼파라미터 조합($\eta = 1.99 \times 10^{-4}$, $\tau = 0.0784$, $w_{kl} = 0.20$)하에서 Image-Text Fusion 전략은 Hit Rate@1 0.9407, MRR 0.9641을 달성하며 전체 실험 조건 중 최적의 검색 성능을 입증하였다. 이는 파인튜닝을 거치지 않은 Baseline Jina Embeddings v4의 Single Vector 검색 대비 Hit Rate@1에서 48.1%p, MRR에서 42.2%p 향상된 수치로, 도메인 특화 파인튜닝 및 멀티모달 표현 융합 기법의 유효성을 실증한다. 특히, 이미지를 정답으로 갖는 소표본 부분집합(n=18) 평가에서 Image-Text Fusion 전략은 이미지 단독 Multi Vector 방식 대비 Hit Rate@1을 0.6667에서 0.9444로, MRR을 0.7503에서 0.9491로 개선하여, VLM이 추출한 설명문과 OCR 정보가 이미지 임베딩의 의미론적 공백을 보완함을 실험적으로 보여주었다. 한편, 제안 모델의 Single Vector 전략은 MRR 0.9509를 기록하여 Baseline Multi Vector 조건의 성능 MRR 0.9231을 상회하면서도, 인덱스 크기를 약 8배 절감하고 검색 처리 속도를 6배 이상 단축시켰다. 이는 제한된 연산 자원과 실시간성이 요구되는 환경에서 실용적인 대안이 될 수 있다.

본 연구는 수동 라벨링 데이터 확보가 극히 제한적인 치안·사이버범죄 수사 분야에서 LLM 기반 합성 데이터 생성 기법을 활용한 멀티모달 임베딩 모델의 도메인 적응 가능성을 검증하고, 시스템의 운영 목적에 따른 벡터화 전략의 명확한 정량적 선택 기준을 제시했다는 점에서 학술적·실무적 의의를 지닌다. 그러나 본 연구는 이미지 평가 쿼리의 수가 제한적이고, 학습 및 평가가 합성 쿼리와 시뮬레이션 데이터에 기반한다는 한계가 있다. 또한 제안하는 Image-Text Fusion 전략의 성능이 VLM이 도출하는 메타데이터 품질에 구조적으로 종속되어 cascade 오류 전파 가능성이 존재한다. 마지막으로 LangGraph 기반 검색 증강 생성 시스템까지 구축되어 있음에도 불구하고 검색 결과 이후 LLM이 생성하는 최종 수사 지원 응답의 정밀도 평가는 본 논문의 범위에서 제외되었다.

상기 한계점을 보완하고 제안 시스템의 실무 완성도를 제고하기 위해, 향후 연구는 다음과 같은 세부 방향으로 추진하고자 한다. 첫째, 시각적 다양성을 확보한 이미지 중심 평가 데이터셋의 확대이다. 현재 평가 아키텍처가 정형 텍스트 로그에 편중되어 이미지 검색 성능 검증의 통계적 신뢰성이 제한적이므로, 향후 다양한 범죄 유형과 레이아웃 특성을 반영한 이미지 전용 벤치마크셋을 추가 구축하여 Multi Vector와 Image-Text Fusion 전략 간

의 성능 편차를 보다 엄밀하게 검증할 계획이다.

둘째, 실제 수사관 작성 질의 기반의 현장 일반화 가능성 검증이다. 본 연구의 합성 쿼리는 실무 수사관의 검색 행위를 모사하도록 생성되었으나, 실제 수사 현장에서 발생하는 축약 표현, 통사적 불완전성, 수사 은어 및 맥락 의존적 구문을 완벽히 모사하는 데 한계가 존재한다. 따라서 실제 수사관의 질의 로그를 확보하여 합성 쿼리와의 어휘적·통사적 분포 차이를 정량적으로 분석하고, 도메인 적응 모델의 현장 타당성을 사후 검증하고자 한다.

셋째, RAG 파이프라인의 종단 간 평가 및 설명 가능한 근거 제시 설계이다. 향후 연구에서는 질의 처리 이후 검색된 증거를 바탕으로 생성되는 최종 LLM 응답의 사실적 정확성(Factual Accuracy), 검색 맥락과의 일치성(Faithfulness), 설명 가능성(Explainability) 및 환각 발생 비율을 정량적으로 평가할 필요가 있다. 아울러 수사관이 검색 결과와 최종 답변의 생성 근거를 직접 역추적할 수 있도록 검색 증거의 출처(파일명, 행 번호, 타임스탬프 등)를 명시적으로 매핑하는 구조를 인터페이스에 통합 설계함으로써, 치안 지원 시스템으로서의 신뢰성과 실무적 유용성을 극대화하고자 한다.

참고문헌

Amato, F., Castiglione, A., Cozzolino, G., and Narducci, F. (2020), A Semantic-Based Methodology for Digital Forensics Analysis, *Journal of Parallel and Distributed Computing*, **138**, 172-177.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., … *et al*. (2025a), Qwen2.5-VL Technical Report, arXiv preprint arXiv:2502.13923.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., *et al*. (2025b), Qwen3-VL Technical Report, arXiv preprint arXiv:2511.21631.

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., and Reddy, S. (2024), LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders, arXiv preprint arXiv:2404.05961.

Biewald, L. (2020), Experiment Tracking with Weights and Biases, Available at https://www.wandb.com.

Bonifacio, L., Abonizio, H., Fadaee, M., and Nogueira, R. (2022), InPars: Data Augmentation for Information Retrieval Using Large Language Models, arXiv preprint arXiv:2202.05144.

Casey, E. (2011), *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*, 3rd ed., Academic Press, San Diego, USA.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020), LEGAL-BERT: The Muppets straight out of Law School, Findings of the Association for Computational Linguistics: EMNLP 2020.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., and Liu, Z. (2024a), BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, arXiv preprint arXiv:2402.03216.

Chen, Q., Peng, Y., and Lu, Z. (2019), BioSentVec: Creating Sentence Embeddings for Biomedical Texts, *Proc. 7th IEEE Int. Conf. on Healthcare Informatics (ICHI)*.

Chen, W., Hu, H., Chen, X., Verga, P., and Cohen, W. (2022), MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text, *Proc. 2022 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 5558-5570.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., … Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Zhang, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., and Wang, W. (2024b), Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling, arXiv preprint arXiv:2412.05271.

Choi, Y., Song, Y., and Sung, M. (2026), KoViDoRe: A Benchmark for Korean Visual Document Retrieval, *Proc. 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2026)*, 54-80.

Dai, Z., Zhao, V. Y., Ma, J., Luan, Y., Ni, J., Lu, J., Bakalov, A., Guu, K., Hall, K. B., and Chang, M.-W. (2023), Promptagator: Few-shot Dense Retrieval From 8 Examples, *Proc. Int. Conf. on Learning Representations (ICLR)*.

Du, X., Le, N. H., and Bhatt, S. (2020), Leveraging Semantic Search for Digital Forensics, *Proc. IEEE Int. Conf. on Big Data*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., … *et al.* (2024), The Llama 3 Herd of Models, arXiv preprint arXiv:2407.21783.

Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., and Colombo, P. (2025), ColPali: Efficient Document Retrieval with Vision Language Models, *Proceedings of the International Conference on Learning Representations (ICLR)*, 61424-61449.

Günther, M., Ong, J., Mohr, I., Abdessalem, A., Abel, T., Akram, M. K., Guzman, S., Mastrapas, G., Sturua, S., Wang, B., Werk, M., Wang, N., and Xiao, H. (2023), Jina Embeddings 2: 8192-Token General-Purpose Text Embeddings for Long Documents, arXiv preprint arXiv:2310.19923.

Günther, M., Sturua, S., Akram, M. K., Mohr, I., Ungureanu, A., Wang, B., ... , and Xiao, H. (2025), jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval, *Proc. 5th Workshop on Multilingual Representation Learning (MRL 2025)*, 531-550.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022), LoRA: Low-Rank Adaptation of Large Language Models, *Proc. Int. Conf. on Learning Representations (ICLR)*.

Järvelin, K. and Kekäläinen, J. (2002), Cumulated Gain-based Evaluation of IR Techniques, ACM Transactions on Information Systems, **20**(4), 422-446.

Jeronymo, V., Bonifacio, L., Abonizio, H., Fadaee, M., Lotufo, R., Zavrel, J., and Nogueira, R. (2023), InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval, arXiv preprint arXiv:2301.01820.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023), Mistral 7B, arXiv preprint arXiv:2310.06825.

Jina AI (2025), Jina Embeddings v4 [Embedding model], Hugging Face, Available at https://huggingface.co/jinaai/jina-embeddings-v4.

Johnson, J., Douze, M., and Jégou, H. (2019), Billion-Scale Similarity Search with GPUs, IEEE Transactions on Big Data, **7**(3), 535-547.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen,

D., and Yih, W.-T. (2020), Dense Passage Retrieval for Open-Domain Question Answering, **Proc. 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)**, 6769-6781.

Khattab, O. and Zaharia, M. (2020), ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT, *Proc. 43rd Int. ACM SIGIR Conf.*

Korean National Police Agency (2023), *Cybercrime Trend Report 2023*, Cyber Investigation Division, Seoul, Korea.

Korean National Police Agency (2025), Cybercrime Statistics, e-National Indicators, Available at https://www.index.go.kr/unity/potal/main/EachDtlPageDetail.do?idx_cd=1608.

LangChain, Inc. (2024), LangGraph: Build Resilient Language Agents as Graphs, Available at https://github.com/langchain-ai/langgraph.

Langfuse (2023), Langfuse: Open-Source LLM Engineering Platform, Available at https://langfuse.com.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023b), BLIP-2: Bootstrapping Language-Image Pre-Training With Frozen Image Encoders And Large Language Models, *Proc. 40th Int. Conf. on Machine Learning (ICML)*, 19730-19742.

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. (2023a), Towards General Text Embeddings with Multi-Stage Contrastive Learning, arXiv preprint arXiv:2308.03281.

Lillis, D., Becker, B., O'Sullivan, T., and Scanlon, M. (2016), Current Challenges and Future Research Areas for Digital Forensic Investigation, *Proc. 11th ADFSL Conf. on Digital Forensics, Security and Law*, 9-20.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023), MTEB: Massive Text Embedding Benchmark, *Proc. 17th Conf. of the European Chapter of the Association for Computational Linguistics (EACL)*, 2014-2037.

Nussbaum, Z., Duderstadt, B., and Mulyar, A. (2024), Nomic Embed Vision: Expanding The Latent Space. arXiv preprint arXiv:2406.18587.

Qdrant (2023), Qdrant: High-Performance Vector Search Engine and Database, Available at https://github.com/qdrant/qdrant.

Qwen Team (2025a), Qwen2.5-VL-3B-Instruct [Vision language model], Hugging Face, Available at https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct.

Qwen Team (2025b), Qwen3-30B-A3B-Instruct-2507 [Large language model], Hugging Face, https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507.

Qwen Team (2025c), Qwen3-VL-8B-Instruct [Vision language model], Hugging Face, https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021), Learning Transferable Visual Models From Natural Language Supervision, *Proc. 38th Int. Conf. on Machine Learning (ICML)*, 8748-8763.

Reimers, N. and Gurevych, I. (2019), Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks, *Proc. 2019 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 3982-3992.

Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J. N., and Sheppard, J. (2023), ChatGPT for Digital Forensic Investigation: The Good, The Bad, and The Unknown, *Forensic Science International: Digital Investigation*, **46**, 301609.

Snoek, J., Larochelle, H., and Adams, R. P. (2012), Practical Bayesian Optimization of Machine Learning Algorithms, *Advances in Neural Information Processing Systems (NeurIPS)*, 25.

Steiner, A., Pinto, A. S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., Qin, S., Ingle, R., Bugliarello, E., Kazemzadeh, S., Mesnard, T., Alabdulmohsin, I., Beyer, L., and Zhai, X. (2024), PaliGemma 2: A Family of Versatile VLMs for Transfer, arXiv preprint arXiv:2412.03555.

Tang, Y., and Yang, Y. (2024), Do We Need Domain-Specific Embedding Models? An Empirical Investigation, arXiv preprint arXiv:2409.18511.

Voorhees, E. M. (1999), The TREC-8 Question Answering Track Report, *Proc. 8th Text Retrieval Conf. (TREC-8)*, 77-82.

Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., Yu, K., Yuan, Y., Zou, Y., Long, J., Cai, Y., Li, Z., Zhang, Z., Mo, Y., Gu, J., Jiang, R., Wei, Y., and Xie, C. (2021), Milvus: A Purpose-Built Vector Data Management System, *Proc. 2021 ACM SIGMOD Int. Conf.*

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2024a), Text Embeddings by Weakly-Supervised Contrastive Pre-training, arXiv preprint arXiv:2212.03533.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. (2024b), Improving Text Embeddings with Large Language Models, arXiv preprint arXiv:2401.00368.

Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A., and Chen, W. (2023), UniIR: Training and Benchmarking Universal Multimodal Information Retrievers, arXiv preprint arXiv:2311.17136.

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. (2023), C-Pack: Packaged Resources To Advance General Chinese Embedding, arXiv preprint arXiv:2309.07597.

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J., and Overwijk, A. (2021), Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval, *Proc. Int. Conf. on Learning Representations (ICLR)*.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. (2025), Qwen3 Technical Report, arXiv preprint arXiv:2505.09388.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023), Sigmoid Loss for Language Image Pre-Training, *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*.

Appendix

Appendix A. 합성 데이터 생성 파이프라인 주요 기호 정의

Table A1. Notation used in the synthetic data generation pipeline

Symbol	Definition
D	Raw dataset
d_i	The i-th raw data instance, either a PNG image or a CSV row
$meta_i$	Textual metadata extracted from image data d_i using a VLM(f_{VLM})
$RowToJSON(d_i)$	JSON text obtained by converting a CSV row d_i into key-value pairs
c_i	Context input to the LLM(f_{LLM}) for query generation
q_i	Synthetic natural-language query generated by the LLM(f_{LLM})
p_i	Positive evidence corresponding to the query q_i
N_i	Set of mined hard negative evidence corresponding to the query q_i
M_{base}	Base multimodal embedding model
f_{VLM}	Vision-language model for image preprocessing
f_{LLM}	Large language model for query and training data synthesis
T_{final}	Final filtered triplet training dataset constructed after checking validity
K	The size of the top-K candidate list retrieved from the vector database (Hyperparameter)
H	The target number of hard negatives mined per query for contrastive learning (Hyperparameter)
P	The comprehensive set of preprocessed tuples (p_i, c_i) generated in Stage 1
v_{q_i}	The high-dimensional dense vector encoded from query q_i using the base model M_{base}
$TopK$	The list of top-K nearest neighbors retrieved from vector database V based on cosine similarity
$cand_r$	The r-th candidate item positioned within the retrieved $TopK$ results list
r	The rank index variable utilized to iterate through the retrieved candidates $(1 \le r \le K)$

Appendix B. 하이퍼파라미터 민감도 분석을 위한 실험 전체 결과

Table B1. Full experimental results (epoch 5)

Run ID	η	τ	w_{kl}	Seed	Multi Margin	Single Margin	Multi ER	Single ER	Val Loss
a01_baseline_seed42	$3e-5$	0.040	0.2	42	0.4231	0.5848	0.0287	0.0261	0.1537
a02_lr_low	$1e-5$	0.040	0.2	42	0.3789	0.5144	0.0256	0.0246	0.1807
a03_lr_high	$1e-4$	0.040	0.2	42	0.4606	0.6288	0.0251	0.0225	0.1357
a04_temp_low	$3e-5$	0.020	0.2	42	0.3129	0.4513	0.0271	0.0292	0.1277
a05_temp_high	$3e-5$	0.065	0.2	42	0.5192	0.6785	0.0271	0.0312	0.2236
a06_wkl_zero	$3e-5$	0.040	0.0	42	0.4233	0.5873	0.0271	0.0261	0.1532
a07_wkl_high	$3e-5$	0.040	0.4	42	0.4249	0.5825	0.0246	0.0261	0.1529
a08_baseline_seed123	$3e-5$	0.040	0.2	123	0.4248	0.5797	0.0256	0.0266	0.1563
a09_baseline_seed777	$3e-5$	0.040	0.2	777	0.4249	0.5843	0.0276	0.0266	0.1517

Appendix C. 하이퍼파라미터 민감도 분석을 위한 실험 보조 지표 학습 곡선

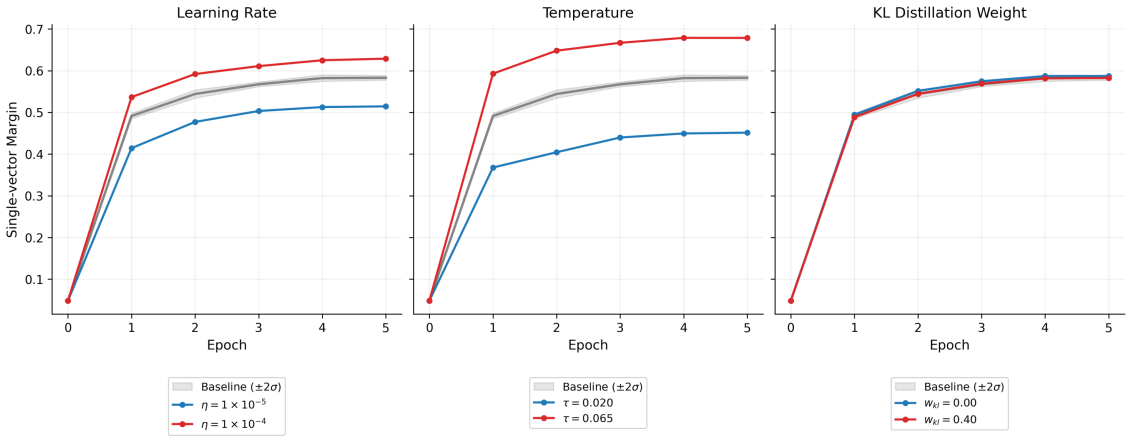


Figure C1. Single Vector Margin learning curves across epochs. Three panels correspond to Learning Rate (left), Temperature (center), and KL Distillation Weight (right). The gray band denotes the 3-seed Baseline mean ± 2σ range.

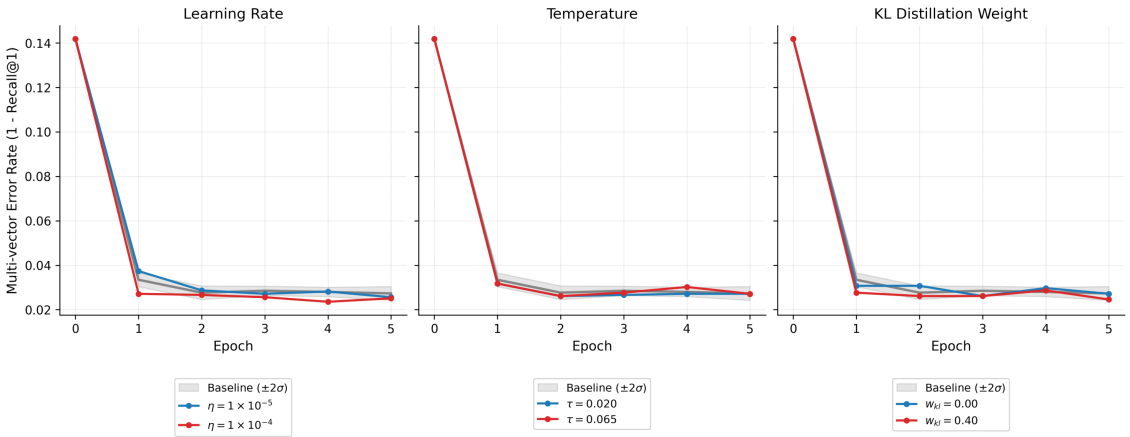


Figure C2. Multi Vector Error Rate (1 − Recall@1) learning curves across epochs. Three panels correspond to Learning Rate (left), Temperature (center), and KL Distillation Weight (right). The gray band denotes the 3-seed Baseline mean ± 2σ range.

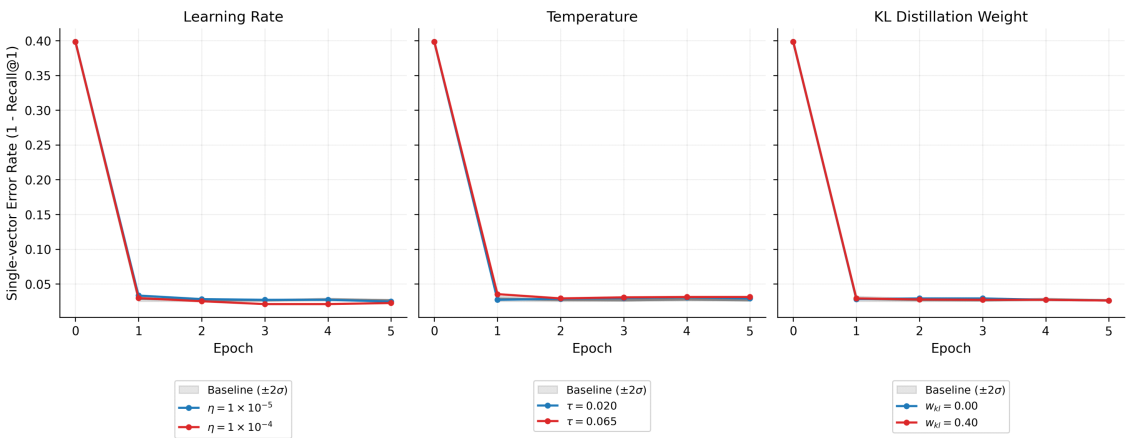


Figure C3. Single Vector Error Rate (1 − Recall@1) learning curves across epochs. Three panels correspond to Learning Rate (left), Temperature (center), and KL Distillation Weight (right). The gray band denotes the 3-seed Baseline mean ± 2σ range.

저자소개

이승규: 서울과학기술대학교 산업정보시스템전공에서 2026년 학사학위를 취득하고 서울과학기술대학교에서 국방인공지능응용학과 석사과정에 재학 중이다. 연구분야는 자기지도 학습, 멀티모달 학습, 자연어 처리이다.

이수민: 인천대학교 산업경영공학과에서 2026년 학사학위를 취득하고 인천대학교에서 산업경영공학과 석사과정에 재학 중이다. 연구분야는 딥러닝을 활용한 이상 탐지 및 잔여 수명 예측이다.

박현호: 2005년 경북대학교 전자전기컴퓨터학부에서 학사, 2007년 광주과학기술원 정보기전공학부 석사학위를 취득하고 2014년 과학기술연합대학원대학교 광대역네트워크공학 박사학위를 취득하였다. 2014년부터 현재 한국전자통신연구원 책임연구원으로 재직하고 있다. 관심분야는 딥러닝기반 공공안전 기술개발, 자연어 처리이다.

권은정: 1997년 동덕여자대학교 전산정보학과에서 학사, 2000년 이화여자대학교 컴퓨터공학과에서 석사학위를 취득하고 2024년 충남대학교 컴퓨터공학과에서 박사학위를 취득하였다. 2000년부터 현재 한국전자통신연구원 책임연구원으로 재직하고 있고 현재 과학치안공공ICT연구센터장을 맡고 있다. 관심분야는 재난안전분야 의사결정지원시스템, 기계학습, 빅데이터 분석, 자연어 처리이다.

이민정: 2017년 홍익대학교 산업공학과에서 학사학위를, 2023년 고려대학교 산업경영공학과에서 석·박사 통합과정을 통해 박사학위를 취득하였다. 2023년부터 현재 한국전자통신연구원 선임연구원으로 재직하고 있으며, 2026년 3월부터 과학기술연합대학원대학교 조교수를 겸임하고 있다. 주요 연구 관심 분야는 인공지능 기반 공공안전, 치안 기술개발 및 의사결정지원시스템 구축이다.