Journal of KIIT. Vol. 24, No. 3, pp. 11-21, Mar. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.3.11

2단계 지식 증류를 통한 설명 가능한 마약 증거 분석 시스템 구현

김태연*[1], 최주현*[2], 이준혁*[3], 김경종*[4]

Implementation of an Explainable Drug Evidence Analysis System via 2-Stage Knowledge Distillation

Taeyeon Kim*[1], Juhyun Choi*[2], Junhyeok Lee*[3], and Kyungjong Kim*[4]

본 연구는 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임.
(No. RS-2023-00225661, 디지털 증거의 증명력 제고를 위한 인과관계 추론 및 표현 기술 개발)

요 약

최근 디지털 플랫폼을 이용한 마약 범죄의 지능화로 인해 압수 이미지 증거가 급증하며 수사관의 분석 업무 부담이 가중되고 있다. 이에 사법 체계 내에서의 AI 기술의 도입이 논의되고 있으나, 결과의 신뢰성을 담보하기 위한 설명 가능성이 필수적이다. 본 연구는 XAI 원칙에 기반한 2단계 지시어 기반 미세 조정 방법론을 제안한다. 1단계에서 GPT-4o를 교사 모델로 활용하여 이미지 분류의 구체적인 근거가 포함된 데이터셋을 구축하고, 2단계에서 LLaVA 1.6 학생 모델을 미세 조정하여 분류와 근거 생성을 동시에 수행하도록 한다. 제안된 시스템은 분류 결과와 함께 자연어 설명을 제공함으로써 수사관의 신속한 의사결정을 보조하고 디지털 증거 분석 업무의 효율성을 개선하는 데 기여하고자 한다.

Abstract

As drug crimes leveraging digital platforms become increasingly sophisticated and covert, the volume of seized image evidence has exceeded the manual analysis capacity of investigators. While AI integration is essential to address this issue, explainability—providing the reasoning behind judgments—is critical for accountability in legal contexts. This study proposes a 2-stage instruction-tuning methodology based on XAI principles. Stage 1 utilizes GPT-4o as a teacher model to generate explanation datasets, and Stage 2 fine-tunes a LLaVA 1.6 student model to perform simultaneous classification and rationale generation. The proposed system provides natural language explanations to support investigators' decision-making processes and improve the efficiency of forensic evidence analysis.

Keywords

vision-language model, explainable AI, instruction tuning, image classification, drug crime investigation

* 경찰대학 치안데이터과학연구센터(*4 교신저자)
- ORCID[1]: https://orcid.org/0009-0002-1757-7845
- ORCID[2]: https://orcid.org/0009-0002-5119-2754
- ORCID[3]: https://orcid.org/0009-0005-3948-1803
- ORCID[4]: https://orcid.org/0009-0002-6682-1138

· Received: Dec. 02, 2025, Revised: Dec. 30, 2025, Accepted: Jan. 02, 2026
· Corresponding Author: Kyungjong Kim
Dept. of Police Science, Korean National Police University, Republic of Korea
Tel.: +82-41-968-2242, Email: leeyeongul@police.ac.kr

Ⅰ. 서 론

마약 범죄의 수법은 텔레그램과 다크웹을 비롯한 디지털 플랫폼을 매개로 지능화·은밀해지고 있으며, 거래 접촉·수거 지점 공유·자금 세탁에 이르기까지 전 과정이 온라인에서 신속하게 이루어지는 양상이 관찰된다[1]. 유엔마약범죄사무소(UNODC)는 다크웹과 소셜미디어의 결합이 마약 공급망의 '가속·분산·익명화'를 견인한다고 지적하고, 최근 세계 마약 보고서에서도 온라인 기반 유통의 확산과 증거 확보의 곤란을 반복적으로 경고하고 있다[1]. 이러한 환경 변화는 압수·수색으로 확보되는 디지털 기기 내부의 이미지 증거가 기하급수적으로 늘어나는 결과를 초래하며, 해당 이미지를 수사관이 일일이 선별·분석하는 기존 방식은 막대한 시간과 인력이 소요되어 구조적 병목을 낳는다. 이러한 데이터 폭증 문제를 해결하기 위해, 딥러닝 기반의 자동 분석 기술 방법론을 제안하고자 한다.

최근 VLM(Vision-Language Model)은 자연어 지시를 통해 이미지의 복잡한 맥락을 이해하는 능력에서 큰 발전을 이루었다[2]. 본 연구진은 선행 연구를 통해, LLaVA[2] 1.6 모델을 QLoRA(Quantized Low-Rank Adaptation)[3] 기법으로 효율적으로 파인튜닝하여 마약 관련 이미지 분류 작업에서 96.34%의 높은 정확도를 달성한 바 있다[4]. 이는 VLM이 특정 법 집행 도메인에서 기술적 실효성을 가질 수 있음을 입증한 성과였다. 하지만 이 모델은 왜 특정 이미지를 '마약 원본'으로 분류했는지에 대한 이유를 설명하지 못하는 블랙박스라는 한계를 지닌다. 법적 증거를 다루고 인간의 판단이 최종 책임을 지는 수사 환경에서, 근거 없는 AI의 예측 결과는 참고 자료 이상의 가치를 갖기 어려우며, 수사관의 온전한 신뢰를 얻을 수 없다[5].

AI의 판단 과정을 인간이 이해 가능한 형태로 제시하는 설명 가능한 AI(XAI)[5]는 이러한 신뢰의 간극을 메우기 위한 기술이다. Grad-CAM과 같은 초기 XAI 연구들은 모델이 이미지의 어떤 영역에 집중했는지를 시각적으로 보여주었으나[6], 해당 영역을 왜 중요하게 판단했는지에 대한 논리적 설명은 제공하지 못했다. 이에 대해 VQA-X(Visual Question Answering with Explanations)[7]와 같은 연구는 정답과 더불어 자연어 설명을 함께 학습·평가하는 틀을 제안하여, 모델이 답변과 동시에 이유를 텍스트로 산출하도록 유도했으며, 이는 단순 주목 시각화에서 고수준 논리 서술로 XAI를 확장하는 방향성을 제시한다[7]. 나아가 Park 등의 연구는 텍스트 설명과 함께 모델이 참고한 이미지 내 특정 영역을 명시적으로 지목하는 멀티모달 설명 방식을 제안하며, 설명의 구체성과 신뢰도를 한층 더 높이는 연구 방향을 개척했다[7].

따라서 본 연구는 선행 연구[4]의 한계를 보완하고, 기존의 분류 모델을 설명 가능한 추론 모델로 진화시키는 2단계 파인튜닝 방법론을 제안한다. 이는 G. Hinton et al.[8]이 제안한 지식 증류(Knowledge distillation)의 교사-학생 모델 프레임워크를 멀티모달 환경에 맞게 변용한 것이다. 1단계에서는 GPT-4o를 교사 모델로 사용하여 각 이미지의 분류 근거를 자연어로 생성, 고품질의 설명 데이터셋을 구축한다. 2단계에서는 이 데이터셋을 LLaVA 1.6 학생 모델의 지시어 기반 미세 조정에 활용하여, 모델이 스스로 분류 결과와 함께 그 판단 근거를 서술형으로 생성하도록 학습시킨다.

궁극적으로 본 연구는 마약 범죄 수사관에게 단순히 분류 결과를 제시하는 것을 넘어, "AI가 어떤 시각적 증거에 기반하여 이러한 결론에 도달했는지"를 보여주는 의사결정 지원 시스템, XAI-LLaVA를 구축하는 것을 목표로 한다. 이는 AI의 투명성과 신뢰성을 제고하고, 생성된 설명 자체를 새로운 수사 단서로 전환함으로써 인간과 인공지능 간 상호보완적 의사결정 체계를 구축하고 실무적 활용성을 제고하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 Vision-Language Model의 발전 과정과 설명 가능한 AI의 주요 기법, 그리고 본 연구의 이론적 기반이 되는 교사-학생 모델에 대해 기술한다. 3장에서는 제안하는 2단계 지시어 기반 미세 조정 방법론의 상세한 과정과 실험 데이터셋 구축 과정을 설명한다. 4장에서는 구현된 XAI-LLaVA 모델의 정량적, 정성적 성능 평가 결과를 분석하고, 마지막으로 5장에서는 본 연구의 결론을 요약하고 한계점

및 향후 연구 방향을 제시한다.

II. 관련 연구

2.1 VLM와 LLaVA 1.6

VLM은 이미지와 텍스트를 통합적으로 이해하고 처리하는 멀티모달 인공지능으로, 시각적 질의응답(VQA), 이미지 캡셔닝 등 복합적인 과업에서 괄목할 만한 성능을 보이며 빠르게 발전해왔다. 본 연구에서 학생 모델로 채택한 LLaVA(Large Language and Vision Assistant)는 이러한 VLM 연구의 선두에 있는 대표적인 오픈소스 모델이다[2]. LLaVA 1.6 [9]은 그림 1과 같이 강력한 Vision Encoder와 언어 모델을 효율적으로 결합한 아키텍처를 기반으로 한다. 구체적으로, 이미지의 시각적 특징을 포착하는 눈의 역할로 CLIP ViT-L/14를 사용하며, 이를 통해 추출된 시각 정보는 간단한 MLP 프로젝터를 거쳐 뇌의 역할을 하는 Mistral-7B 백본으로 전달된다. CLIP ViT-L/14는 OpenAI의 CLIP(Contrastive Language–Image Pre-training) 모델 중 하나로, 이미지와 텍스트를 동시에 이해하도록 설계된 멀티모달 인코더 모델이다. Mistral-7B는 고성능 소형 언어 모델(SLM)로서, 입력된 시각 정보와 텍스트 지시를 종합적으로 이해하고 추론하여 최종적인 자연어 응답을 생성한다.

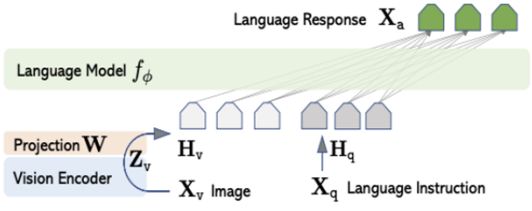


그림 1. LLaVA 네트워크 구조 [2]
Fig. 1. LLaVA network architecture [2]

본 연구가 LLaVA 1.6을 학생 모델로 채택한 이유는 VLM이 가진 고유의 고수준 의미론적 설명 능력과 검증된 성능, 그리고 실용적 효율성 때문이다. VLM의 등장은 AI의 판단 근거를 저수준의 픽셀이 아닌, 인간이 이해하는 고수준의 논리적 서술(자연어)로 제시할 가능성을 열었으며, LLaVA는 이러한 XAI를 구현하기에 적합한 SOTA(State-of-the-Art) 모델이다. 실제로 LLaVA-1.6 및 후속 NeXT 계열은 MMBench와 같은 주요 VLM 벤치마크에서 동급 모델 대비 최고 수준의 성능을 입증했다[10]. 특히 본 연구의 증거 분석에 필수적인 이미지 내 텍스트 인식(OCR, Optical Character Recognition), 논리적 추론, 배경지식 활용 영역에서 확인된 성능 개선은 복잡한 증거 이미지를 분석하기에 적합하다고 판단했다. 나아가 모델 가중치와 코드가 모두 공개된 오픈소스 특성은 연구의 투명성을 보장하며, QLoRA[3]와 같은 파라미터 효율적 파인튜닝(PEFT, Parameter-Efficient Fine-Tuning) 기법을 적용하기 용이하여, 제한된 컴퓨팅 자원 내에서도 특정 도메인에 맞게 효율적으로 미세 조정할 수 있다는 실용적 이점을 제공한다.

2.2 설명 가능한 AI와 VLM의 결합

설명 가능한 AI는 AI의 판단 과정을 인간이 이해 가능한 형태로 제시하여 모델의 신뢰성을 확보하고 책임 소재를 명확히 하기 위한 연구 분야다[5]. Grad-CAM[6], LIME[11] 등 초기의 대표적인 XAI 연구들은 모델 결정에 영향을 미친 이미지의 픽셀 영역을 히트맵으로 시각화하는 저수준 특징 기여도 분석에 중점을 두었다. 이는 AI가 이미지의 "어디를 보는가"에 대한 답은 주었지만, "그곳을 왜 중요하게 보는가"에 대한 맥락적 설명은 제공하지 못하는 한계를 가졌다.

VLM의 등장은 이러한 한계를 극복하고 XAI를 고수준 의미론적 설명의 단계로 발전시킬 새로운 가능성을 열었다. VLM은 단순히 픽셀의 중요도를 넘어, 이미지 속 객체, 관계, 상황 등 고차원적인 정보를 이해하고 이를 자연어로 표현할 수 있기 때문이다. VQA-X[7]와 같은 선행 연구들은 VLM이 질문에 대한 답과 함께 그 이유를 자유로운 형태의 텍스트로 설명하도록 학습시킬 수 있음을 실험적으로 증명했다. 이는 XAI가 전문가의 해석이 필요한 시각화 자료를 넘어, 최종 사용자가 직접 이해하고 활용할 수 있는 논리적 서술 형태로 발전할 수 있

음을 시사한다. 본 연구는 이러한 흐름을 실제 수사 도메인에 적용하여, VLM이 단순 분류기를 넘어 수사관의 추론 과정을 보조하는 설명 가능한 파트너가 될 수 있음을 보이고자 한다.

2.3 지식 증류와 교사-학생 모델

본 논문의 2단계 파인튜닝 방법론은 지식 증류[8]의 개념에 깊게 뿌리를 두고 있다. 지식 증류는 크고 복잡하며 성능이 뛰어난 교사 모델의 지식을 작고 효율적인 학생 모델에게 이전하는 학습 프레임워크이다. G. Hinton et al.[8]이 제안한 초기 방법론에서는 교사 모델이 예측한 정답 확률 분포, 즉 소프트 레이블을 학생 모델이 모방하도록 학습시켰다.

본 연구는 이러한 교사-학생 프레임워크를 멀티모달 환경으로 확장하고, 지식의 형태를 재정의한다. 최근 LLM 분야에서는 거대 언어 모델을 사용하여 고품질의 데이터를 생성하고, 이를 작은 모델의 지시어 기반 미세 조정에 활용하는 연구가 활발히 진행되고 있다[12]. 본 연구는 이 아이디어를 VLM에 적용하여, GPT-4o와 같은 초거대 상용 VLM을 교사 모델로 사용하여 단순히 '3: 마약 원본'과 같은 정답을 넘어 "왜 3번인지에 대한 서술형 근거"라는 풍부하고 구조화된 지식을 생성한다. 그리고 LLaVA 1.6 학생 모델은 이 고품질의 자연어 설명을 모범 답안으로 삼아 학습함으로써, 교사 모델의 시각적 추론 능력과 설명 능력을 효율적으로 습득하게 된다. 이는 자연어 설명을 매개로 한 멀티모달 추론 과정의 증류라는 새로운 형태로서, 학생 모델이 단순한 결과 모방을 넘어 교사의 사고 과정을 내재화하도록 유도할 수 있다.

III. 연구 방법

본 연구는 LLaVA 1.6 모델에 설명 가능성을 주입하기 위해, 교사-학생 모델 구조에 기반한 2단계 파인튜닝 전략을 설계했다. 전체 시스템 아키텍처는 그림 2와 같다.

모든 실험은 Ubuntu 20.04 운영체제 환경의 단일 NVIDIA A100 80GB GPU에서 수행되었다. 소프트웨어는 Python 3.12.11을 기반으로 CUDA 12.1 환경에서 PyTorch 2.5.1, transformers 4.57.1 라이브러리를 사용했다.

3.1 1단계: 실험 데이터셋 구축

첫째, 원본 데이터 수집 및 분류 단계에서는 실제 마약 범죄 수사 과정에서 법 집행기관이 압수한 스마트폰의 디지털 포렌식 분석 결과로부터 직접 확보한 이미지들을 중 파일 손상으로 인해 분류가 불가한 이미지를 제외하고 총 3,509개의 유효 이미지를 선별했다. 수집된 이미지는 내용 기반 분석을 통해 표 1과 같이 4가지 범주로 분류하였다.

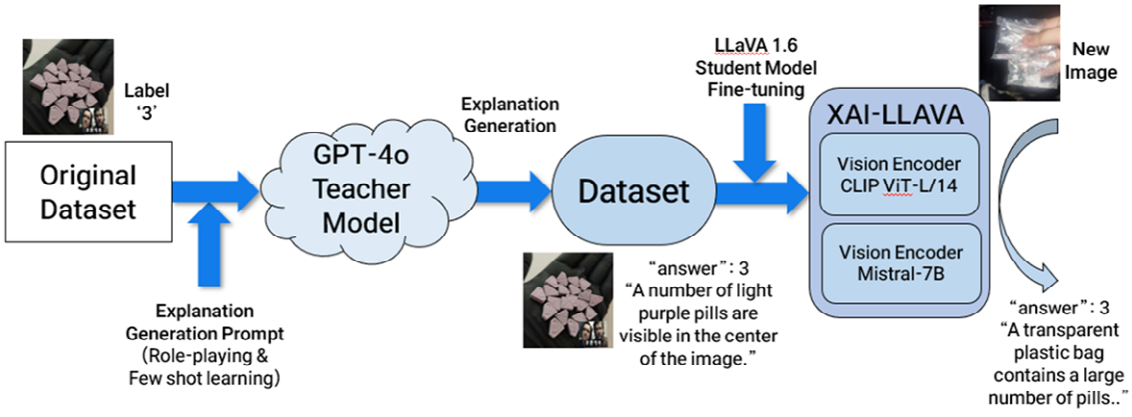


그림 2. 제안하는 설명 가능 VLM 파인튜닝 아키텍처
Fig. 2. Proposed 2-stage explainable VLM fine-tuning architecture

표 1. 마약 증거 이미지 범주 정의
Table 1. Definitions of evidence image categories

Category	Number of data samples	Description
0. Irrelevant	991	Not related to drug activity or context
1. Drug-related conversation	1144	Messenger screenshots (e.g., Telegram) containing drug-related slang
2. Drug deal location	145	Photos depicting buildings, streets, or meeting spots associated with drug exchanges
3. Confiscated drug image	1229	Images showing drugs directly photographed

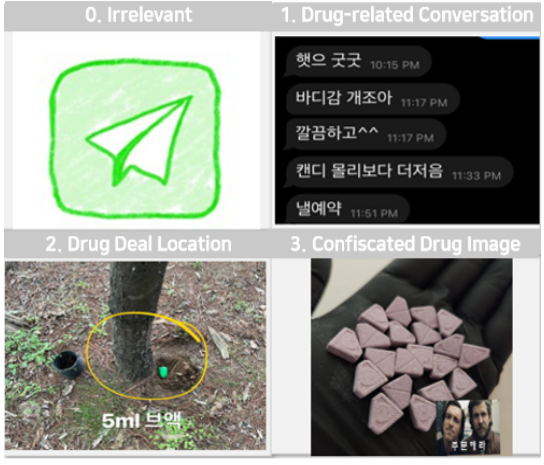


그림 3. 각 범주별 이미지 데이터 예시
Fig. 3. Examples of images for each category

둘째, 데이터 불균형 해소를 위해 데이터 증강 기법을 적용했다. '2: 거래 장소' 클래스는 원본 데이터가 29개로 다른 클래스에 비해 이미지 수가 적어 데이터 불균형이 존재했다. 이를 완화하고 학습의 안정성을 도모하고자 학습 데이터셋에 한하여 해당 클래스의 이미지에 대해 좌우 반전, 무작위 회전, 밝기/대비 조절 등을 적용하여 총 145개로 증강하였다.

셋째, 데이터셋 분할 단계에서는 최종적으로 완성된 3,509개의 사진 데이터를 계층적 샘플링을 통해 학습, 검증, 테스트용으로 각각 2,675개(약 76.2%), 417개(약 11.9%), 417개(약 11.9%)로 분할했다. 전체 데이터셋의 최종 구성은 표 2와 같다.

표 2. 전체 데이터셋의 실험용 데이터 분포
Table 2. Dataset distrubtion of model experiment

Dataset type	Number of data	Ratio (%)
Train	2,675	76.2
Validation	417	11.9
Test	417	11.9
Total	3,509	100

넷째, 설명 데이터 생성 및 검수 단계이다. 모델에 설명 가능성을 주입하기 위해 GPT-4o를 교사 모델로 정의하고, 각 이미지에 대한 분류 근거를 자연어로 생성하는 지식 증류 공정을 수행하였다. GPT-4o 모델을 사용한 이유는 최근 VLM 연구에서 GPT-4o는 텍스트, 오디오, 이미지 데이터를 End-to-End로 통합 처리할 수 있으며, MMBench 및 MME와 같은 최신 벤치마크에서 기존 모델들을 능가하는 성능을 보였기 때문이다[13]. 설명의 품질과 일관성을 확보하기 위해, 본 연구는 프롬프트 엔지니어링 기법[14]을 적용했다. 교사 모델에게는 단순히 이미지 분석을 요청하는 것을 넘어 표 3과 같이 구체적인 역할과 제약 조건을 부여했다.

표 3. 지시어 설계를 위한 프롬프트 구성
Table 3. Prompt components for instruction design

Technique	Prompt instruction	Example instance
1. Persona assignment	Assign a persona such as “a skilled digital forensic investigator AI”	“You are an expert digital forensic investigator AI...”
2. Context provision	Provide the correct class label for each image in advance to establish context	“...This image is classified as ‘3: Drug Materials’. Your task is to explain why”
3. Task specification	Instruct the model to generate evidence-based key rationales in bullet points (2 - 4)	“Based only on the visual evidence... list 2 - 4 key rationales... Use bullet points”
4. Few-shot learning	Provide high-quality sample answers to help the model learn the desired output style	“Example 1 (Image: ...): \n*- A plastic bag containing white powder..”

이를 통해 획득한 서술형 데이터는 연구자의 수동 검수를 거쳐 단순 분류를 넘어 판단 논리를 내포한 지시어 기반 학습 데이터로 확정하였다, 최종적으로 구축된 학습 데이터는 표 4와 같이 분류 레이블과 서술형 설명이 결합된 answer 필드로 구성된다. 이러한 형식은 모델이 단순 분류를 넘어, 분류에 대한 근거를 함께 학습하도록 유도하는 지시어 데이터의 역할을 수행한다.

표 4. 최종 학습 데이터 예시
Table 4. Examples of the final training data

Label	Rationale
0. Irrelevant	- A selfie photo of a person. - The person is making a V-sign gesture with their hand. - No relevance to drugs.
1. Drug-related conversations	- An image capturing a mobile messenger chat screen. - Contains drug-related terms such as “rush”, “body convulsion”, and “tolerenace”. - The conversation discusses the effects and experiences of drug use.
2. Transaction locations	- The image shows the inside of an eletrical distribution box. - A specific spot is highlighted with a red arrow. - The text “Inside the upper right corner of the internet box” appears in the image, which is interpreted as indicating the location where drugs are hidden.
3. Drug materials	- A plastic bag containing white powder is visible on an electronic scale. - The scale display shows “1.30g”. - The image includes the text “Coca 1g” at the top.

3.2 2단계: 설명 가능성 주입을 위한 LLaVA 1.6 학생 모델 파인튜닝

본 연구는 Hugging Face TRL 라이브러리의 SFTTrainer와 같이 고도로 추상화된 도구 활용 시 발생할 수 있는 라이브러리 버전 간 상호 호환성 저해 요소를 배제하고, 데이터 전처리 공정의 정밀한 제어권을 확보하기 위해 transformers 라이브러리의 표준 Trainer와 사용자 정의 DataCollator를 결합한 학습 아키텍처를 설계하였다.

데이터 파이프라인은 Dataset 클래스가 저장소로부터 원본 샘플(PIL 이미지, 지시어 및 타깃 텍스트)을 호출하면, CustomDataCollator가 이를 배치 단위의 학습용 텐서로 가공하는 구조를 취한다. 해당 가공 공정은 그림 4와 같이 세 단계의 논리적 절차를 거쳐 수행된다.

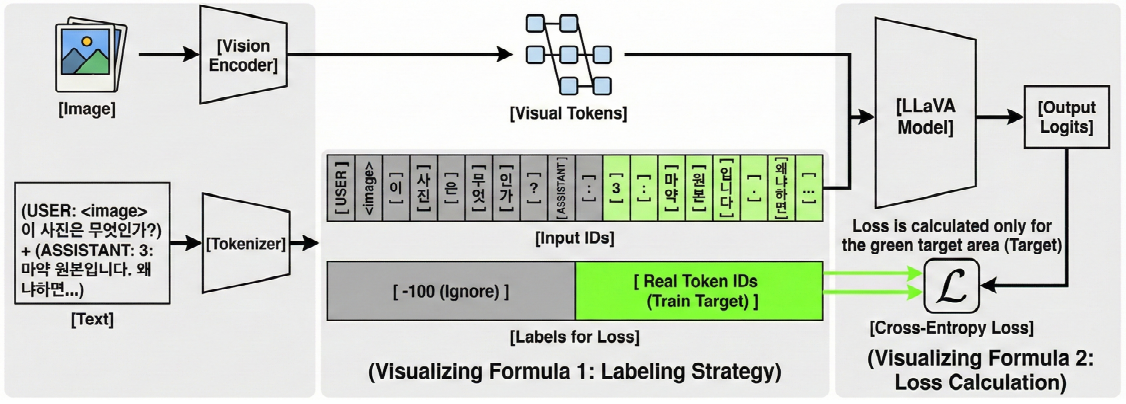
첫째, 통합 텍스트 생성 단계에서는 전달받은 지시어(T_{prompt})와 정답 시퀀스(T_{target})를 논리적으로 결합하여, 모델이 학습해야 할 전체 입출력 시퀀스 T_{full}(T_{full}= T_{prompt} $\oplus$ T_{target})을 구성한다. 이는 모델이 수사적 맥락을 이해한 뒤 적절한 추론 근거를 도출하도록 유도하는 기초 공정이다.

둘째, 배치 단위 처리 단계에서는 AutoProcessor를 사용하여 배치 내의 모든 이미지와 텍스트를 한 번에 처리한다. 이 과정에서 이미지는 Vision Encoder가 이해할 수 있는 고차원 벡터(pixel_values)로, 텍스트는 언어 모델이 이해할 수 있는 수치 시퀀스(input_ids)로 변환된다.

셋째, 프롬프트 마스킹 단계는 인과적 언어 모델(Causal language model)의 학습 방향을 정밀하게 통제하는 단계이다. 모델이 입력 프롬프트(T_{prompt})자체를 단순 복제하지 않고, 실제 추론 근거인 타깃 시퀀스(T_{target}) 생성에 설계하였다. 이를 위해, T_{prompt} 영역의 라벨 값을 PyTorch의 Cross-Entropy Loss에서 손실 계산을 무시하도록 하는 특수 값 −100으로 치환하는 마스킹 처리를 수행했다. 결과적으로 Trainer는 T_{target} 시퀀스에 대해서만 손실을 계산하며, 이는 식 (1)과 같이 표현된다.

$$L(\theta) = -\sum_{i=|T_{prompt}|+1}^{|T_{full}|} \log P_{\theta}(L_i | T_{full,} < i > \qquad (1)$$

식 (1)은 모델의 예측 확률 분포 $P(\theta)$가 오직 타깃 시퀀스 $L(\theta)$와 일치하도록 합산이 프롬프트가 끝난 $i=|T_{prompt}|+1$부터 시작하게 하여 모델 파라미터 Θ를 갱신함을 의미한다. 이러한 구조를 통해 데이터 전처리와 학습 루프를 정밀하게 통합함으로써 연구의 재현성을 확보하고 안정적인 파인튜닝 과정을 구현하였다.



(a) 데이터 입력 및 토큰화 (b) 마스킹 전략 (c) 손실 계산
(a) Data input & tokenization (b) Masking strategy (c) Loss calaulation

그림 4. 프롬프트 마스킹을 적용한 인스트럭션 튜닝 과정
Fig. 4. Instruction tuning process with prompt masking

최적화 알고리즘으로는 가중치 감쇠를 매개변수 업데이트와 분리함으로써 일반화 성능 향상을 보고한 AdamW 알고리즘[15]을 사용하였다. 학습률은 10^{-5}의 저율로 설정하였는데, 이는 미세 조정 과정에서 사전 학습된 모델의 시각적 표현이 과도하게 변화하는 것을 억제하여 재난적 망각(Catastrophic forgetting)을 완화하고, 학습 수렴의 안정성을 확보하기 위한 설정이다.

또한 QLoRA의 주요 파라미터인 어댑터 차원 r은 16, 스케일링 계수 α는 32로 정의하였다. 파라미터 효율적 미세 조정(PEFT) 가이드라인[16]에 따르면 α를 r의 2배로 설정하는 것이 학습 안정성에 유리하다. 본 연구에서 r값을 기본값 8보다 높은 16으로 설정한 이유는 마약 증거 이미지 내의 미세한 객체(가루, 도구 등)와 텍스트 특징을 정밀하게 학습할 수 있는 충분한 표현력을 확보하기 위해서이다.

Ⅳ. 파인튜닝 결과 분석

본 장에서는 2단계 파인튜닝을 통해 학습된 XAI-LLaVA 모델의 성능을 정량적, 정성적 측면에서 다각도로 분석한다. 평가는 최종적으로 분할된 417개의 테스트 데이터셋에 대해 수행하였다.

4.1 분류 결과의 정량적 평가

본 장에서는 제안 모델의 분류 성능을 정밀하게 검증하기 위해, 총 417개의 테스트 데이터셋을 활용하여 평가를 수행하였다. 성능 평가의 기준이 된 데이터셋은 '0. 관련 없음' 109개, '1. 마약 대화' 147개, '2. 거래 장소' 13개, '3. 마약 원본' 149개로 구성되었다. 모델이 생성한 추론 결과에서 숫자 레이블을 추출하여 정답과 비교한 정량적 지표는 표 5와 같다.

표 5. XAI-LLaVA 모델 분류 성능
Table 5. Classification performance of the XAI-LLaVA mode

Class	Precision	Recall	F1-score
0. Irrelevant	0.94	0.99	0.96
1. Drug-related conversation	1.00	0.97	0.99
2. Drug deal location	0.93	1.00	0.96
3. Confiscated drug image	0.99	0.96	0.97
Accuracy			0.98
Macro avg.	0.96	0.98	0.97
Weighted avg.	0.96	0.98	0.97

평가 결과, 제안 모델은 전체 정확도 97.6%, 가중 평균 F1-score 97%라는 높은 분류 성능을 달성했다. 이는 본 연구진의 선행 연구[4]에서 달성한 96.34%를 상회하는 수치이다. Macro Avg.는 모든

클래스에 동일한 가중치를 부여하여 산출한 평균값이다. 본 실험에서 Macro Avg.는 0.97의 F1-score를 기록하였으며, 이는 데이터의 불균형이 존재하는 환경에서도 모델이 특정 범주에 편향되지 않고 모든 클래스를 균등하게 학습하였음을 입증한다. 또한 각 클래스의 샘플 수에 비례하여 가중 평균을 낸 지표인 Weighted Avg. 역시 0.97의 높은 F1-score를 나타냈다. 이러한 결과는 실제 데이터셋의 분포를 고려한 모델의 종합적인 분류 성능이 안정적인 수준에서 유지되고 있음을 시사한다.

특히 주목할 점은 마약 관련 증거 클래스(1, 2, 3)에 대한 높은 정밀도와 재현율이다. '마약 대화'로 예측한 경우는 100% 정확했으며, 실제 '거래 장소' 이미지는 단 하나도 놓치지 않았다. 다만, '마약 대화' 클래스의 학습 및 테스트 데이터가 텔레그램 메신저 UI에 편중되어 있어, 이 100% 정밀도 수치는 텔레그램 환경에 과적합되었을 가능성이 있다.

4.2 설명 품질의 자동화된 정량 평가

모델이 생성한 설명의 품질을 객관적으로 평가하기 위해 본 연구는 심사위원 모델 기반 평가 방법론인 "LLM-as-a-Judge"[17]를 도입했다. 이는 인간 평가자를 모방하도록 설계된 LLM(GPT-4o)을 심사위원으로 사용하여, 학생 모델이 생성한 설명의 품질을 다각도로 채점하는 방식이다. 평가 과정은 심사위원 모델이 평가 프롬프트($P_{eval,}$), 원본 질문(Q), 학생 모델의 답변(A_{model}), 그리고 평가 기준(C_c)을 종합적으로 분석하여 점수를 도출하는 방식으로 진행된다. 식 (2)는 이러한 평가 프로세스를 정의한 것이다.

$$Score_c = Judge(P_{eval,}Q,A_{\mathrm{model},}C_{c)} \quad (2)$$

구체적인 점수 산출 방식은 다음과 같다. 심사위원 모델인 GPT-4o는 각 평가 기준(충실성, 타당성, 유용성)에 대해 사전에 정의된 5점 척도의 정성적 상세 채점 기준표를 입력받는다. 심사위원 모델은 학생 모델이 생성한 설명을 이미지 내 실제 증거와 대조하여 논리적 일관성을 추론한 뒤, 상세 채점 기준표의 기준에 가장 부합하는 점수를 제로샷 방식으로 할당한다[17]. 이러한 다각도 평가 구조를 통해 단순 정답 여부를 넘어 생성된 설명이 수사관의 의사결정에 기여할 수 있는 질적 수준을 수치화하였다.

평가는 이미지 내 증거에 기반하는 정도를 의미하는 충실성, 분류 결과를 논리적으로 뒷받침하는 정도인 타당성, 수사관에게 실질적인 단서를 제공하는 정도인 유용성의 세 가지 기준에 따라 5점 척도(1점: 매우 나쁨 ~ 5점: 매우 좋음)로 이루어졌다. 그 결과는 표 6과 같다.

표 6. "LLM-as-a-Judge"를 통한 설명 품질 평가
Table 6. Explanation quality evaluation via "LLM-as-a-Judge"

Evaluation criterion	Scale meaning	Score
Faithfulness	The degree to which the explanation is grounded in visual evidence from the image.	3.58
Relevance	The degree to which the explanation logically supports the classification result.	3.27
Helpfulness	The degree to which the explanation provides practical clues for investigators.	2.77

가장 높은 점수를 받은 충실성(3.58점)은 제안 모델이 생성한 설명이 대부분 이미지에 실제로 존재하는 시각적 증거에 기반하고 있으며, 허위 정보를 생성하는 환각 현상이 현상이 유의미하게 억제되었음을 확인하였다. 타당성(3.27점)은 모델이 제시하는 근거가 대부분 논리적으로 타당함을 의미하지만, 일부 사례에서는 분류에 결정적인 단서보다는 부차적인 특징을 제시하는 경향이 있음을 시사한다. 한편 유용성 점수는 2.77점으로 '보통' 이하 수준을 기록했다. 이 점수는 현재 모델의 설명이 사실에 기반하고 논리적이지만, 실제 수사관에게 새로운 통찰이나 단서를 제공하기보다는 시각적 특징을 서술하는 수준에 머무르는 경우가 많다는 한계를 나타낸다. 예를 들어, "흰 가루가 보인다"는 설명은 사실이지만 수사관에게는 당연한 정보일 수 있다.

Ⅴ. 결론 및 향후 과제

본 논문은 마약 범죄 수사 과정에서 발생하는 디지털 이미지 증거 분석의 한계를 극복하고자, 설명 가능한 AI 기술을 Vision-Language Model에 접목하는 새로운 방법론을 제안하고 그 실효성을 검증했다. 제안하는 2단계 지시어 기반 미세 조정 방법론은 GPT-4o 교사 모델을 통해 설명 데이터셋을 구축하고, 이를 LLaVA 1.6 학생 모델의 학습에 활용하여 분류와 설명 생성을 동시에 수행하는 XAI-LLaVA를 구축했다.

실험 결과, 제안 모델은 97.6%라는 높은 분류 정확도를 달성했을 뿐만 아니라, "LLM-as-a-Judge" 방법론을 통한 자동 평가에서 평균 3.58점의 충실도를 기록하며 논리적인 자연어 설명을 생성하는 데 성공했다. 이는 AI를 단순 자동화 도구에서 신뢰할 수 있는 의사결정 지원 시스템으로 격상시켰다는 점에서 학술적·실용적 의의를 가진다. 본 연구는 인공지능 모델의 불투명성으로 인한 신뢰성 문제를 개선함으로써 인간과 기술이 협력할 수 있는 의사결정 기반을 마련하고, 마약 수사 현장에서의 실질적인 활용 가치를 입증하였다. 나아가 본 연구의 방법론은 향후 마약 외에도 사이버범죄, 아동성착취물 탐지 등 다양한 수사 분야로 확장될 수 있다.

본 연구는 의미 있는 성과를 거두었으나 다음과 같은 한계점을 가지며, 이는 향후 연구의 방향을 제시한다. 학생 모델의 설명 능력은 자동 평가 결과에서 나타났듯이 교사 모델 데이터의 품질에 의존하며, 특히 유용성 점수가 낮아 사실 나열 수준에 머무르는 경향을 보였다. 또한, '마약 대화' 클래스의 학습 데이터가 텔레그램 메신저의 사용자 인터페이스에 국한되어, 카카오톡 등 다른 인터페이스 환경에서는 분류 성능이 저하될 수 있는 일반화의 한계가 존재한다. 또한, 현재 모델은 텍스트 설명만 제공하여 근거가 되는 이미지 내 특정 위치를 지목하지 못하며 실험실 환경에서의 검증에 머물러 다양성을 띄는 실환경 적용성은 아직 확인되지 않았다. 따라서 향후 연구에서는 수사 맥락상 중요한 의미를 추론하도록 프롬프트 엔지니어링을 고도화하고, D. H. Park et al.[7]의 연구처럼 설명과 함께 관련 이미지 영역을 시각화하는 멀티모달 설명 시스템을 개발할 필요가 있다. 또한 실제 포렌식 도구와의 통합 등을 통해 현재 수사관의 피드백을 받고 이를 반영 하는 등의 방식으로 모델의 완성도를 계속 높여나갈 필요가 있다.

Acknowledgement

2025년도 한국정보기술학회 하계종합학술대회에서 발표한 논문(마약 범죄 수사 지원을 위한 VLM (LLaVA 1.6) 기반 디지털 포렌식 이미지 자동 분류 시스템 연구)[4]을 확장한 것임.

References

[1] United Nations Office on Drugs and Crime (UNODC) – "World Drug Report 2023: Executive Summary", UNODC, pp. 29, Jun. 2023. https://www.unodc.org/res/WDR-2023/WDR23_Exsum_fin_DP.pdf. [accessed: Oct. 08, 2025]

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning", Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, Louisiana, USA, Vol. 36, pp. 34656-34670, Dec. 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[3] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs", Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, Louisiana, USA, Vol. 36, pp. 10088-10115, Dec. 2023. https://doi.org/10.48550/arXiv.2305.14314.

[4] T. Kim, J. Choi, and K. Kim, "A Study on an Automated Digital Forensic Image Classification System to Support Drug Crime Investigations Using a Vision-Language Model (LLaVA 1.6)", Proc. KIIT Conference, Jeju, Korea, pp. 390-394, Jun. 2025.

[5] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)", IEEE Access, Vol. 6, pp.

52138-52160, Sep. 2018. https://doi.org/10.1109/ACCESS.2018.2870052.

[6] R. R. Selvaraju, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, pp. 618-626, Oct. 2017. https://doi.org/10.1109/ICCV.2017.74.

[7] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence", Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, USA, pp. 8779-8788, Jun. 2018. https://doi.org/10.1109/CVPR.2018.00915.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network", arXiv preprint, arXiv:1503.02531, pp. 2-4, Mar. 2015. https://doi.org/10.48550/arXiv.1503.02531.

[9] llava-hf - llava-v1.6-mistral-7b-hf: Model Card, https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf. [accessed: Oct. 08, 2025]

[10] LLaVA Project - LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, https://llava-vl.github.io/blog/2024-01-30-llava-next. [accessed: Oct. 08, 2025]

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier", Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., San Francisco, California, USA, pp. 1135-1144, Aug. 2016. https://doi.org/10.1145/2939672.2939778.

[12] Y. Wang, et al., "Self-Instruct: Aligning Language Model with Self-Generated Instructions", Proc. 61st Annu. Meet. Assoc. Comput. Linguist. (ACL), Toronto, Canada, pp. 13489-13512, Jul. 2023. https://doi.org/10.18653/v1/2023.acl-long.754.

[13] Y. J. Fan, et al., "A Comprehensive Survey on Evaluation of Multimodal LLMs", arXiv preprint, arXiv:2411.15296, pp. 12-13, Nov. 2024. https://doi.org/10.48550/arXiv.2411.15296.

[14] T. B. Brown, et al., "Language Models are Few-shot Learners", Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Oline, pp. 1877-1901, Dec. 2020. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[15] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization", Proc. 7th Int. Conf. Learn. Representations (ICLR), New Orleans, LA, USA, pp. 1-18, May 2019. https://doi.org/10.48550/arXiv.1711.05101.

[16] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, and C. Paul, "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods", GitHub Repository, 2022. https://github.com/huggingface/peft. [accessed: Dec. 26, 2025]

[17] L. Zheng, et al., "Judging LLM-as-a-judge with MT-bench and Chatbot Arena", Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, LA, USA, Vol. 36, pp. 46655-46673, Dec. 2023. https://doi.org/10.48550/arXiv.2306.05685.

저자소개

김 태 연 (Taeyeon Kim)

2026년 3월 : 경찰대학 법학과(학사)
2024년 9월 ~ 2026년 3월 : 경찰대학 안데이터과학연구센터 학생연구원
관심분야 : 멀티모달 딥러닝, 디지털 포렌식

최 주 현 (Juhyun Choi)

2026년 3월 : 국민대학교 소프트웨어융합대학원(석사)
2024년 2월 ~ 2026년 3월 : 경찰대학 치안데이터과학연구센터 연구원
관심분야 : 컴퓨터 비전

이 준 혁 (Junhyeok Lee)

2023년 3월 ~ 현재 : 경찰대학
행정학과 학사과정
2025년 9월 ~ 현재 : 경찰대학
치안데이터과학연구센터
학생연구원
관심분야 : 멀티모달 딥러닝,
디지털 포렌식

김 경 종 (Kyungjong Kim)

2006년 2월 : 경찰대학 행정학과
(학사)
2018년 2월 : 고려대학교
경찰법학과(석사)
2025년 2월 : 카이스트
데이터사이언스 대학원(박사수료)
2022년 2월 ~ 현재 : 경찰대학
치안데이터과학연구센터 지도교수
관심분야 : LLM 응용, ML, SNA