Efficient AI for Media: VLM을 미디어 현장에 넣는다는 것

ㅁ **김태호** / NOTA AI

요 약

비전언어모델(VLM)은 이미지·영상을 언어로 이해하는 수준에서 빠르게 발전했지만, 성능 안정성과 배포 환경의 제약 때문에 산업 현장으로 넘어오는 길목에서 여전히 큰 간극을 남긴다. 이 글은 그 간극을 좁히는 축이 결국 효율화 기술에 있다는 관점에서, 노타가 최적화 플랫폼과 영상 관제 솔루션 두 갈래로 이 문제를 어떻게 다루는지를 정리하고, 방송·교통·산업 안전의 실제 사업 현장에서 어떻게 작동하는지를 보인다. 나아가 VLM의 근본적 한계와 다음 세대 흐름인 World Model을 조망하며, 모델 아키텍처가 어떻게 바뀌더라도 배포 조건은 그대로이므로 효율화 기술의 역할은 오히려 확장될 것임을 논한다.

요즘은 AI를 안 쓰는 사람이 오히려 눈에 띈다. Claude와 ChatGPT는 코딩, 작문, 자료 조사, 회의 정리까지 텍스트 기반 업무의 상당 부분을 이미 재편했다. 이 글의 초안 작업에도 나는 AI의 도움을 받았다. 이런 일이 아무렇지 않은 시대가 됐다. AI의 능력은 텍스트와 숫자에만 머물지 않는다. 지금의 프론티어 모델들은 이미지와 영상도 함께 이해한다. 이 흐름의 출발점 중 하나가 2023년 4월에 발표된 LLaVA다[1]. LLaVA를 계기로 언어 모델에 "이미지를 보는 눈"이 달렸고, 이 계열은 곧 비전언어모델(Vision-Language Model, VLM)이라는 이름으로 자리를 잡았다. VLM은 쉽게 말해 언어 모델에 이미지, 영상 입력을 붙여 확장한 모델이다. 사진이나 영상을 함께 넣으면 그 안의 상황을 문장으로 설명하거나, 특정 장면에 대한 질문에 답하거나, "이 화면이 지금 방송에 적합한가" 같은 판단까지 수행할 수 있다. 기존 컴퓨터 비전 모델이 "고양이가 있다" 정도의 라벨을 출력하는 데 그쳤다면, VLM은 "고양이가 창가에 앉아 밖을 보고 있다"처럼 상황을 언어로 풀어낸다. 여기서 언어 모델의 추론 능력이 그대로 결합되기 때문에, 규칙 기반으로 다루기 어려웠던 관제, 요약, 분류 같은 문제를 훨씬 유연하게 풀 수 있다.

초기 VLM은 이미지를 저해상도로밖에 볼 수 없었

고 이해 능력도 얕았다. 하지만 지난 2~3년 사이 GPT, Gemini, Qwen-VL, InternVL 같은 모델들이 잇달아 나오면서 영상 속 상황을 사람의 언어로 설명하고 질문에 답하는 수준이 빠르게 올라왔다. 미디어 산업 관점에서는 분명한 기회다. 방대한 영상 자산을 의미 기반으로 검색하고, 장면을 요약하고, 송출할 화면을 사람 대신 1차로 판단하는 일이 기술적으로 가능해졌다. 다만 데모 환경에서의 작동과 실제 현장에서의 안정적 운영은 여전히 다른 문제다.

Ⅰ. VLM은 여전히 "팔기 어려운 제품"이다

솔직히 말하면, 지금의 VLM은 판매하기에 여전히 문제가 많은 제품이다. 논문 벤치마크에서는 인상적이지만, 산업 현장의 요구를 그대로 통과시켜 보면 잘 안 된다.

• 성능이 아직 불안정하다

공개 벤치마크에서 좋은 점수를 받은 모델도, 실제 CCTV나 방송 영상처럼 학습 데이터 분포에서 벗어난 도메인에 넣으면 성능이 떨어질 수 있다. 야간 저조도 영상, 특수 각도의 관제 카메라, 재난 현장의 예외 상황, 프롬프트의 미묘한 표현 차이. 어느 하나의 조건 변화만으로도 결과가 흔들린다. 현장에서 진짜 중요한 건 벤치마크 점수가 아니라 "얼마나 실패하지 않는가"인데, VLM은 아직 그 지점을 온전히 넘지 못했다.

• 효율의 문턱이 높다

오늘날 대부분의 오픈 VLM은 NVIDIA의 고가 GPU를 전제로 설계돼 있다. Qwen2.5-VL 7B급을 안정적으로 구동하려면 최소 24GB VRAM급 GPU가 필요하고, 72B급으로 올라가면 최소 384GB VRAM급 서버가 있어야 한다[2]. 방송국 송출 시스템, 지자체 관제센터, 현장에 설치된 엣지 박스가 이런 자원을 갖추고 있을 리 없다. 학습된 모델의 능력과, 그 능력을 실제로 배포할 수 있는 환경 사이의 거리가 상당히 벌어져 있다.

• 비디오는 이미지보다 훨씬 무겁다

고해상도 이미지 한 장도 VLM 내부에서 수백에서 수천 개의 비주얼 토큰으로 변환된다. 영상은 여기에 시간 축이 붙는다. 초당 여러 프레임, 수 분에서 수십 분에 이르는 영상 하나를 그대로 넣으면 토큰 수가 기하급수적으로 늘어난다. VLM 추론 비용의 상당 부분이 이 비주얼 토큰 처리에서 발생한다. 정지 이미지에서는 잘 안 보이던 병목이, 영상 도메인에 오면 곧장 지연시간과 추론 비용 증가로 이어진다.

Ⅱ. 노타는 이 문제를 어떻게 다루는가

노타는 AI 최적화·경량화 회사다. 사업의 축은 두 갈래다. 하나는 AI 모델을 다양한 반도체에서 효율적으로 돌게 만드는 최적화 플랫폼 사업이고, 다른 하나는 교통, 산업안전, 미디어 CCTV 같은 버티컬 도메인에서의 솔루션 사업이다.

언뜻 보면 성격이 다른 사업 같지만, 실제로는 하나의 기술 위에 서 있다. 최적화 플랫폼 NetsPresso®는 외부 고객에게 제공되는 제품이면서, 동시에 우리 자체 솔루션이 현장에 배포되기 위한 기술적 근간이기도 하다. 내부에서 만든 최적화 도구를 우리 사업에 먼저 쓰고, 거기서 검증된 것을 다시 고객에게 제공한다. 이 글도 그래서 두 축을 자연스럽게 오가게 된다.

Ⅲ. NetsPresso, 모델을 현장에서 돌리는 기술

NetsPresso는 AI 모델의 크기와 연산량을 줄여 목표 하드웨어에서 최적의 성능으로 돌아가게 하는 통합 플랫폼

이다. 마법 같은 단일 기법은 없다. 서로 다른 비용 요소를 줄이는 기법들이 층층이 쌓인 스택에 가깝다.

• 양자화(Quantization)

양자화는 모델 가중치와 연산의 수치 정밀도를 낮춰 메모리, 연산량, 전력을 함께 줄이는 기법이다. 대표적으로 32비트 부동소수점(FP32)을 8비트 정수(INT8)나 4비트로 낮추는 식이다. 최근의 관건은 "얼마나 낮추느냐"가 아니라 "무엇을 지키면서 낮추느냐"다. 모델 전체를 일괄 압축하던 방식에서, 중요한 부분의 정밀도는 보존하고 덜 중요한 부분만 공격적으로 압축하는 선택적, 구조 인지 방식으로 발전해 왔다.

특히 요즘 대형 모델의 표준이 되고 있는 MoE(Mixture of Experts) 구조는 기존 양자화 기법이 그대로 통하지 않는다. 전문가(expert)들이 섞이는 과정에서 발생하는 표현 왜곡 때문에 INT4 양자화를 그대로 적용하면 성능이 무너진다. 우리는 이 문제를 풀기 위해 NotaMoEQuant(NMQ)이라는 MoE 특화 양자화 방법을 개발했다. 전문가별 캘리브레이션을 통해 라우팅 결과의 전체 일관성을 유지하면서 압축한다. 실제로 Upstage의 Solar-Open-100B에 NMQ를 적용했더니 메모리 사용량이 250GB에서 69GB로 약 72% 감소했고, 단일 H100 GPU(80GB) 한 장으로 100B급 모델을 돌릴 수 있게 됐다[3, 4]. 이 성과를 기반으로 우리 팀은 MoE 양자화 관련 논문 2편을 ICML 2026 워크샵(AdaptFM)에 올렸고[5], NVIDIA Nemotron Hackathon에서는 종합 대상을 받았다[6].

MoE 최적화 능력은 국내 대형 모델에도 적용된다. 얼마 전 우리는 퓨리오사AI의 데이터센터용 NPU 위에서 LG AI연구원의 K-엑사원(EXAONE) 236B 모델 최적화에 성공했다[7]. 2,360억 개 파라미터의 국내 대표적인 MoE 모델을 국산 NPU에서 돌게 만든 사례로, 거대 모델을 특정 하드웨어에 배포하기 위해 어떤 종류의 최적화 스택이 필요한지를 보여준다.

• 프루닝과 구조 최적화

프루닝은 기여도가 낮은 연결이나 채널을 잘라내 모델의 크기 자체를 줄인다. 양자화가 "정밀도"를 낮춘다면 프루닝은 "규모"를 줄인다. 둘은 배타적이지 않고 함께 적용할 때 이득이 누적된다. 여기에 그래프 최적화와 연산자 융합(operator fusion)을 적용해, 모델의 계산 그래프를 목표 하드웨어에서 효율적으로 실행되도록 재구성한다.

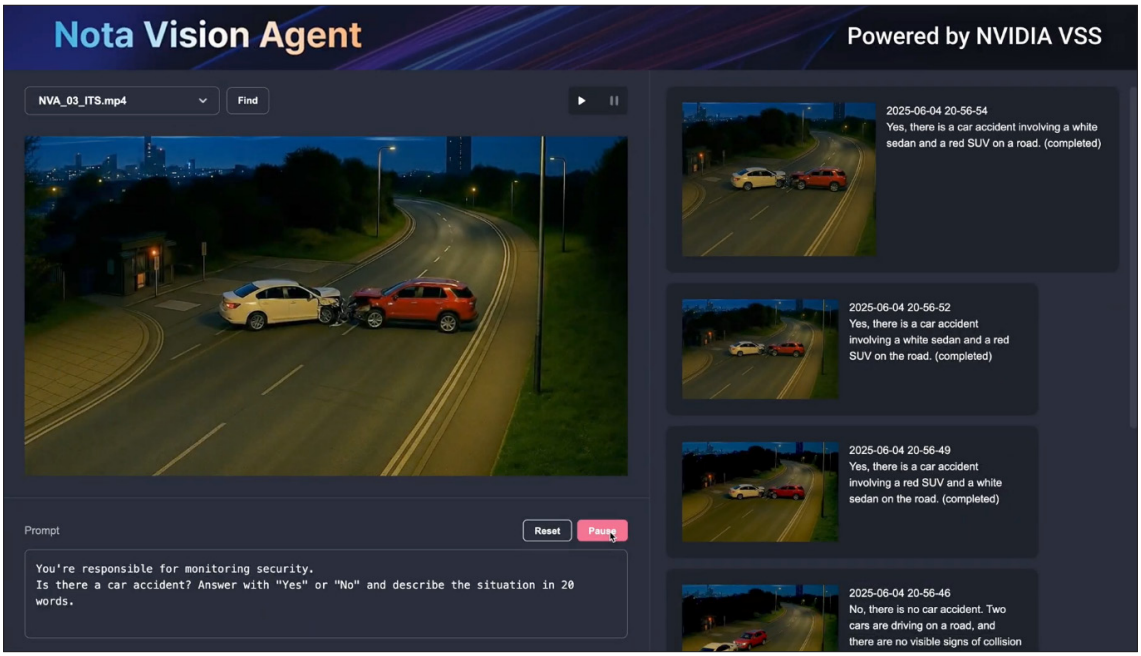
• 하드웨어 친화형 최적화

아무리 잘 압축한 모델이라도, 실행되는 반도체의 특성에 맞춰 매핑하지 않으면 이론상의 이득이 현장의 지연이나 처리량으로 옮겨지지 않는다. CPU, GPU, NPU는 각기 다른 병렬성 패턴과 메모리 계층을 가지고 있어, 같은 모델이라도 어느 칩에서 실행되느냐에 따라 최적의 형태가 달라진다.

글로벌 파트너십으로 확인된 대표적인 사례가 Arm과의 협력이다. 우리는 Arm이 주관하는 AI Virtual Tech Talk에서 NetsPresso 기반의 엣지 AI 최적화 사례를 발표했고[8], Arm의 AI 파트너 프로그램에 합류해 Arm Virtual Hardware 위에서 노타의 최적화 워크플로우를 개발자들이 바로 활용할 수 있도록 통합했다[9]. 이 협력은 2026년 3월 Arm과의 AI 소프트웨어 공급계약 체결로 이어졌다. 계약 기간은 2026년 3월부터 2028년 3월까지로, NetsPresso를 Arm 기반 반도체를 개발하는 글로벌 개발자 생태계에 통합하는 구조다[10]. Arm 아키텍처는 스마트폰, 자동차, 산업용 IoT까지 전 세계 엣지 디바이스의 사실상 표준이다. 여기에 노타의 최적화 스택이 결합됐다는 이야기는, 데이터센터에서 학습된 모델을 저전력 엣지에서 그대로 돌리는 경로가 표준 라인 위에 놓였다는 뜻이다.

• 그 외의 최적화 기법들

스택은 여기서 끝나지 않는다. 지식 증류(Knowledge Distillation)는 큰 교사 모델의 판단을 작고 빠른 학생 모델로 이전하는 기법으로, 태스크가 좁게 정의된 산업 현



<그림 1> 노타의 VLM 기반의 영상 관제 솔루션 NVA(Nota Vision Agent)

장에서 특화 소형 모델을 만들 때 특히 유용하다. 서빙 단계로 가면 continuous batching, PagedAttention 기반 KV 캐시 관리, prefix caching처럼 요청 단위의 자원을 효율화하는 기술이 표준이 됐고, 커널 수준에서는 FlashAttention 계열처럼 어텐션 연산을 하드웨어에 맞춰 재작성하는 방식이 지연을 크게 낮춘다. Speculative decoding은 작은 초안(draft) 모델이 다음 토큰을 미리 제안하고 큰 타깃 모델이 병렬로 검증하는 방식으로 지연을 줄인다. 대형 모델을 실제 서비스로 내보내는 단계에서는 이런 기법들의 조합이 결과를 크게 좌우한다.

Ⅳ. NVA - VLM을 산업 현장에 넣기 위한 기술

노타는 이런 최적화 기술 위에서 교통, 산업안전, 선별 관제 같은 버티컬 솔루션을 운영해 왔다. 여기에 VLM을 결합하면 좋은 결과가 나오리라 기대하기 쉽다. 그런데 실제 환경에 적용하면 그리 간단하지 않다. 몇 겹의 처방이 필요했고, 그 처방들을 모아서 제품으로 만든 것이 우리 팀의 VLM 기반 영상 관제 솔루션 NVA(Nota Vision Agent)다. NVA를 어떻게 만들었는지를 정리하는 것이 곧 VLM을 산업 현장에 넣기 위해 무엇이 필요한지를 정리하는 일이다.

먼저 프롬프트다. 같은 VLM이라도 프롬프트를 어떻게 설계하느냐에 따라 결과의 안정성이 크게 달라진다. 특히 관제, 재난, 산업안전처럼 오탐과 미탐의 비용이 큰 도메인에서는, 프롬프트 스코어링과 사용자 피드백을 결합한 튜닝이 성능의 상당 부분을 결정한다. NVA는 사용자가 원하는 유형의 장면을 프롬프트로 정의하고, 그 결과에 대한 사용자 피드백이 스코어링 로직과 프롬프트 개선에 다시 반영되는 구조로 설계됐다.

다음은 전통적인 컴퓨터 비전과의 결합이다. 사람 탐지나 객체 추적처럼 정확도와 지연이 함께 중요한 태스

크는 여전히 YOLO 계열의 특화 모델이 압도적으로 빠르고 안정적이다. VLM은 "이 장면이 방송에 적합한가", "이 상황이 위험한가" 같은 장면 판단에 강점이 있다. 서로 대체하려 하는 대신, 검출과 추적은 전통 CV에 맡기고 장면 판단은 VLM에 맡기는 하이브리드 파이프라인이 실제로는 훨씬 잘 돌아간다. NVA도 이 두 계열을 조합해 쓴다.

데이터 확보 역시 쉽지 않다. 관제, 산업안전 같은 도메인은 학습에 쓸 만한 실영상 데이터가 원천적으로 희소하다. 사건, 사고 영상은 자주 발생하지 않고, 발생하더라도 프라이버시와 규제 때문에 자유롭게 학습에 쓸 수 없다. 그래서 시뮬레이션과 생성 모델을 통한 데이터 확보가 사실상 필수다. 영상 생성 기술 자체는 급속히 발전했다. 손쉽게 영화의 한 장면 같은 영상을 만들 수 있는 도구들이 늘어났다. 그런데 산업 현장 특유의 환경, 예를 들어 지하 차도의 조명, 공장 내부의 시야, 재난 현장의 혼잡함은 일반 생성 모델로는 자연스러움이 크게 떨어진다. 우리는 이런 산업 환경에 맞춘 영상 생성 기술을 만들어 NVA의 현장 맞춤 파인튜닝에 쓰고 있다.

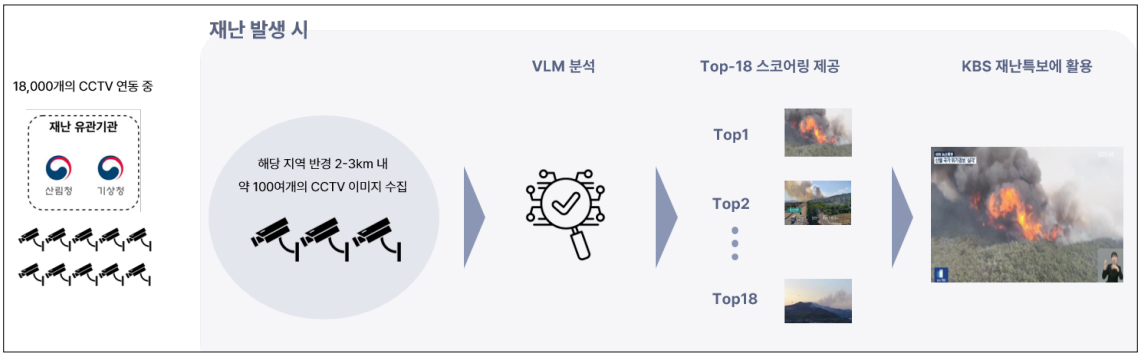
VLM 자체를 개선하는 일도 필요했다. VLM은 이미지를 균일하게, 대체로 저해상도로 본다. 그런데 관제 영상에서 정말로 중요한 정보는 프레임 전체가 아니라 일부 작은 영역에 있는 경우가 많다. 사람이 사진을 볼 때 자연스럽게 관심 영역을 골라 확대해서 보는 것과 다르게, 대부분의 VLM은 그런 능력이 없다. 우리 팀은 이 문제를 해결하기 위해 ERGO(Efficient Reasoning & Guided Observation)라는 방법을 제안했다. 먼저 저해상도 이미지로 전체 상황을 파악한 뒤, 시각적으로 불확실하거나 정밀 분석이 필요한 영역을 골라 원본 해상도로 다시 살펴본다. 어느 영역을 확대할지는 강화학습으로 학습한 보상 시스템이 결정한다. 그 결과 ERGO는 시각 토큰 사용량을 기존 대비 약 23% 수준으로 줄이면서 정확도를 유지했고, 고해상도 VLM 추론 속도는 약 3배 빨라졌다. 이 연구는 AI 학계 최상위 학회 중 하나인 ICLR 2026(채택률 약 28%)에 채택됐다[11]. ERGO의 아이디어는 NVA에도 적용돼 관제 영상의 관심 영역을 선택적으로 정밀 관측하는 축으로 쓰인다.

V. 사업 현장

이런 접근을 바탕으로 우리는 최근 국내외 여러 프로젝트에서 NVA 기반의 실사업 성과를 만들어 왔다. 결이 다른 세 개의 사례를 짚어본다.

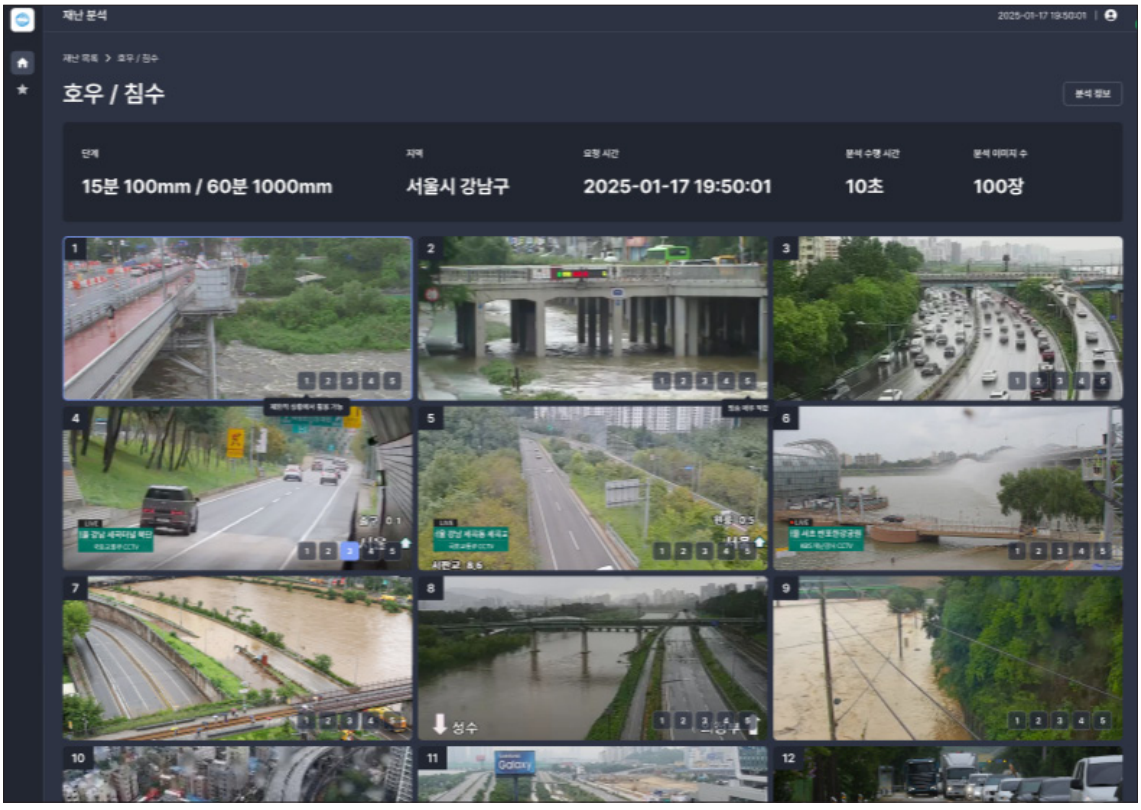
• KBS 재난특보

가장 가혹한 실시간 요구가 걸리는 사례는 방송 재난특보다. 우리는 KBS에 NVA를 적용해, 재난 발생 지역 인근 CCTV 100개 채널의 영상을 실시간으로 분석하고



<그림 2> KBS 재난특보 워크플로우: NVA가 100개 채널 영상을 실시간으로 분석해 방송 적합 18개 채널을 선별한다.

<그림 3> NVA를 적용해 방송에 적합한 KBS 재난특보 영상을 신속하게 확인할 수 있다.



<그림 4> NVA를 적용해 방송에 적합한 KBS 재난특보 영상을 신속하게 확인할 수 있다.

<그림 5> UAE 아부다비 교통청과 온디바이스 AI 기반 실시간 사고 관리 시스템 및 협력 지능형 교통 체계(C-ITS) 공동 연구개발을 위한 업무협약을 체결했다.

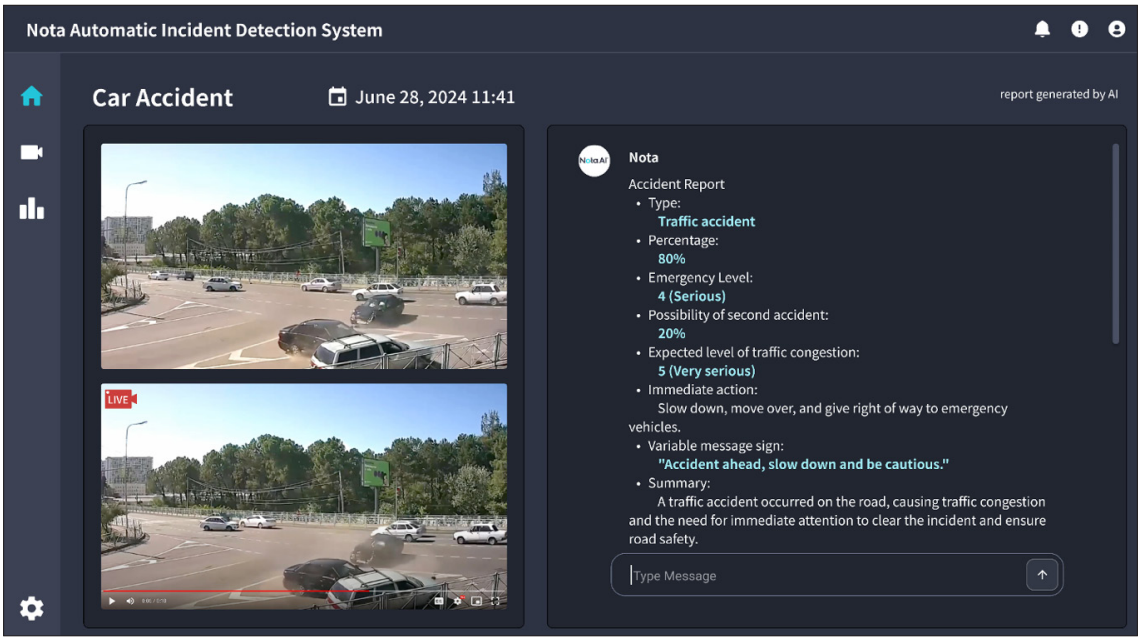
이 가운데 방송에 적합한 18개 채널을 선별하는 역할을 맡았다[12]. 먼저 프롬프트로 원하는 장면 유형을 정의한다. 프롬프트 스코어링으로 보도에 적합한 영상과 부적합한 영상을 가른다. 그리고 선별 결과에 대한 사용자 피드백을 다시 스코어링 로직과 프롬프트 개선에 반영한다. KBS는 이를 통해 재난 상황에 적합한 영상을 신속히 확인해 송출하는 재난 뉴스특보 워크플로우를 국내 최초로 구현했다. 100개 채널을 실시간으로 분석해 18개로 좁히는 작업을, 송출 지연을 늘리지 않으면서 해내야 한다.

• UAE 아부다비 교통청

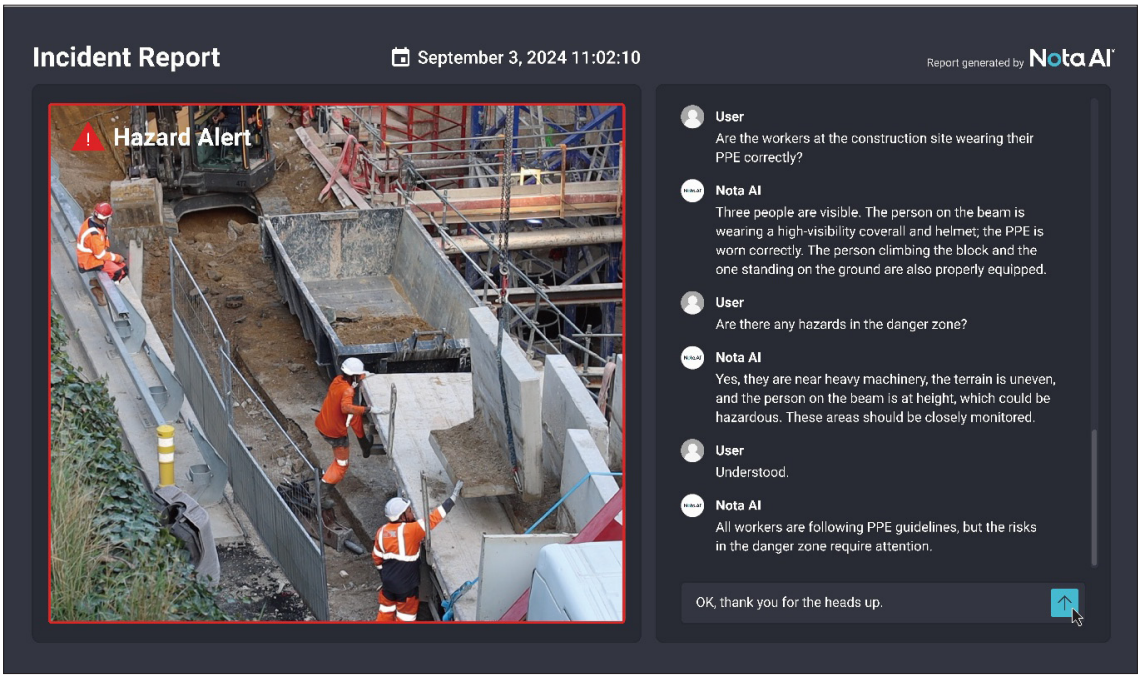
두 번째는 해외다. 우리는 UAE 아부다비 교통청과 온디바이스 AI 기반 실시간 사고 관리 시스템 및 협력 지능형 교통 체계(C-ITS) 공동 연구개발을 위한 업무협약을 체결했다[13].

이 협약은 앞서 아부다비 현지에서 진행한 지능형 교통 체계 실증에서 도로 돌발상황 검지 정확도 95% 이상을 기록한 결과에 이어[14], 2025년 11월 이재명 대통령의 UAE 순방을 계기로 열린 한-UAE 비즈니스 라운드테이블에서 아부다비 전역을 대상으로 한 ITS 솔루션 도입 및 확산 방안이 논의된 이후의 후속 조치다.

여기서 핵심은 VLM을 온디바이스에서 돌리는 일이다. NetsPresso®로 VLM을 경량화해 클라우드를 거치지 않고 교통 CCTV와 관제 시스템의 현장 기기에서 사고나 교통 약자를 인식하게 만들었다. 통신 지연이 없어 골든타임을 잡을 수 있고, 영상을 외부로 안 내보내기에 개인정보 규제가 민감한 중동 지역의 장벽도 통과된다. 두바이에 이어 아부다비까지 이어진 이 흐름은, 규제 조건이 다른 시장에서는 최적화가 곧 시장 진입 조건이 된다는 얘



<그림 6> NVA를 교통 시스템에 활용한 화면: NetsPresso로 VLM을 경량화해 클라우드를 거치지 않고 교통 CCTV와 관제 시스템의 현장 기기에서 사고나 교통 약자를 인식한다.



<그림 7> 산업현장에서 활용되는 NVA

기이기도 하다.

• 코오롱인더스트리 김천 공장

산업 현장 사례로는 코오롱인더스트리 김천 2공장이 있다. 2025년 8월 코오롱베니트와의 협력을 통해 NVA를 김천 2공장에 도입했다[15]. 작업자의 안전 상태, 위험 지역 접근, 안전 수칙 위반을 실시간으로 감지해 중대 재해를 사전에 방지하는 것이 목표다. 공장 환경은 관제 카메라 각도, 조명, 작업자 복장, 장비 배치 등 조건이 매우 특수하다. 이런 조건에서 VLM이 "저 사람은 지금 안전 수칙을 지키지 않고 있다"를 신뢰성 있게 판단하려면, 앞서 얘기한 프롬프트 튜닝, 산업 환경에 맞춘 파인튜닝 데이터, ERGO 계열의 확대 관찰이 모두 필요하다.

VI. 다음 세대의 모델, World Model

VLM 자체에는 근본적인 한계가 남아 있다. VLM은 영상을 프레임 단위 이미지로 이해한다. 사람의 시각 인지가 시간의 흐름 속에서 세계를 연속적으로 파악하는 방식과는 다르다. 장기기억이라 부를 만한 구조도 사실상 없어서 긴 영상, 긴 상황을 안정적으로 유지하며 판단하기 어렵다. 무엇보다 물리 세계의 인과에 대한 감각이 얕다. 사물이 어떻게 움직이고 상호작용하며 다음 순간에 어떻게 될지에 대한 이해가 제한적이다.

이 한계를 정면으로 겨냥하는 흐름이 World Model이다. 영상 생성과 이해를 통합해 세계 자체를 시뮬레이션하려는 접근으로, Physical AI라는 더 큰 흐름의 기반이 되고 있다. 대표적인 갈래는 Meta의 Yann LeCun이 2022년에 제안한 JEPA(Joint Embedding Predictive Architecture) 계열이다. JEPA는 픽셀 공간에서 미래를 예측하는 대신 잠재 표현(latent) 공간에서 예측하도록 학습해, 모델이 세계의 의미 구조 자체를 배우도록 유도한다. Meta가 2025년에 공개한 V-JEPA 2는 62시간 분량의 로봇 데이터로 파인튜닝한 뒤 zero-shot 로봇 제어에서 80% 성공률을 기록했다. 같은 태스크에서 Octo가 15%였던 것과 비교하면 격차가 크다[16]. NVIDIA 역시 Cosmos 시리즈를 통해 피지컬 AI를 위한 파운데이션 모델을 확대하고 있다. 2026년 5월 COMPUTEX 기간 중 열린 NVIDIA GTC Taipei에서 발표된 Cosmos 3는 로봇, 자율주행차, 비전 AI 에이전트를 위한 오픈 파운데이션 모델이다. Agile Robots, Doosan Robotics, LG전자, 삼성전자, Skild AI, Li Auto 등 여러 기업이 이를 도입해 활용하고 있다[17].

VLM이 "장면을 이해하는 모델"이었다면, World Model은 "세계 자체를 이해하고 예측하는 모델"에 가깝다. 그래서 이 계열이 앞으로 VLM 및 VLA(Vision-Language-Action) 모델의 자리를 상당 부분 대체하리라 보는 연구자들이 늘고 있다.

우리 입장은 단순하다. 모델이 VLM에서 World Model로 바뀌어도, 지연, 메모리, 전력, 규제 같은 배포 조건은 바뀌지 않는다. 오히려 모델이 무거워질수록 최적화가 있어야 할 자리는 넓어진다. 지금 VLM에서 하고 있는 일을 다음 세대에서도 그대로 이어갈 것이다.

참 고 문 헌

[1] LLaVA: Visual Instruction Tuning, Haotian Liu et al., arXiv:2304.08485 (2023.4). https://arxiv.org/abs/2304.08485
[2] Qwen2.5-VL 72B VRAM Needs, Novita AI Blog, https://blogs.novita.ai/qwen2-5-vl-72b-vram/
[3] NotaMoEQuantization: An MoE-Specific Quantization Method for Solar-Open-100B, Nota AI Tech Blog, https://www.nota.ai/community/notamoequantization-an-moe-specific-quantization-method-for-solar-open-100b

참 고 문 헌

[4] Nota AI Reduces Memory Usage of Upstage's Solar LLM by 72%, PR Newswire. https://www.prnewswire.com/news-releases/nota-ai-reduces-memory-usage-of-upstages-solar-llm-by-72-demonstrating-proprietary-quantization-technology-302704670.html
[5] Nota AI Has Two MoE Quantization Papers Accepted at ICML 2026 Workshop, PR Newswire. https://www.prnewswire.com/news-releases/nota-ai-has-two-moe-quantization-papers-accepted-at-icml-2026-workshop-demonstrating-global-competitiveness-in-large-scale-ai-optimization-302796634.html
[6] Nota AI Wins Grand Prize at NVIDIA Nemotron Hackathon, PR Newswire. https://www.prnewswire.com/news-releases/nota-ai-wins-grand-prize-at-nvidia-nemotron-hackathon-proving-moe-quantization-prowess-with-synthetic-data-technology-302752573.html
[7] 노타, 퓨리오사AI NPU에서 K-엑사원 236B 최적화 성공, 이데일리(2026.6.30). https://edaily.co.kr/News/Read?mediaCodeNo=257&newsId=03896646645486968
[8] 노타, 'Arm AI Tech Talk' 발표…AI 인공지능 '넷츠프레소' 베타서비스 출시, 데일리시큐. https://www.dailysecu.com/news/articleView.html?idxno=125722
[9] Arm, 국내 대표 엣지 AI 스타트업 노타와 파트너십, 인공지능신문(2022.11). https://www.aitimes.kr/news/articleView.html?idxno=26553
[10] 노타, ARM과 AI 소프트웨어 공급 계약 체결(2026.3.24), Daum 뉴스. https://v.daum.net/v/20260324162201816
[11] 노타, AI 경량화 기술 'ERGO'로 ICLR 2026 채택, 인공지능신문(2026.1). https://www.aitimes.kr/news/articleView.html?idxno=38370
[12] 노타-KBS, AI 기반 재난 특보 영상 분석 시스템 구축, 벤처스퀘어(2026.2). https://www.venturesquare.net/1039524
[13] 노타, UAE 아부다비 교통청과 온디바이스 AI 사고관리 MOU, 이코노믹데일리(2026.1). https://www.economidaily.com/view/20260123144419108
[14] 중동 교통판 흔든다…노타, 아부다비와 온디바이스 AI 동맹 체결(아부다비 ITS 실증 돌발상황 검지 95%+ 명기), 이지경제(2026.1). https://www.ezyeconomy.com/news/articleView.html?idxno=230894
[15] 코오롱베니트, 노타 NVA 기반 프리패키지 출시(코오롱인더 김천 2공장 적용), 데일리시큐(2025.8). https://www.dailysecu.com/news/articleView.html?idxno=168867
[16] Introducing the V-JEPA 2 World Model and New Benchmarks for Physical Reasoning, Meta AI(2025.6). https://ai.meta.com/blog/v-jepa-2-world-model-benchmarks/
[17] NVIDIA Launches Cosmos 3, the Open Frontier Foundation Model for Physical AI, NVIDIA Newsroom(2026.5.31). https://nvidianews.nvidia.com/news/nvidia-launches-cosmos-3-the-open-frontier-foundation-model-for-physical-ai

저 자 소 개

김 태 호

- 2008년 ~ 2013년 : 한국과학기술원 바이오및뇌공학 학사
- 2013년 ~ 2015년 : 한국과학기술원 전기및전자공학 석사
- 2015년 ~ 2017년 : Nota AI CEO & Co-Founder
- 2017년 ~ 2020년 : 한국과학기술원 선임연구원
- 2020년 ~ 현재 : Nota AI CTO & Co-Founder