멀티모달 제조 파운데이션 모델 · 디지털 트윈 · 에이전트로 완성하는 Closed-loop 운영

2026.02.25

임학수 / 연구원, KAIST

목차

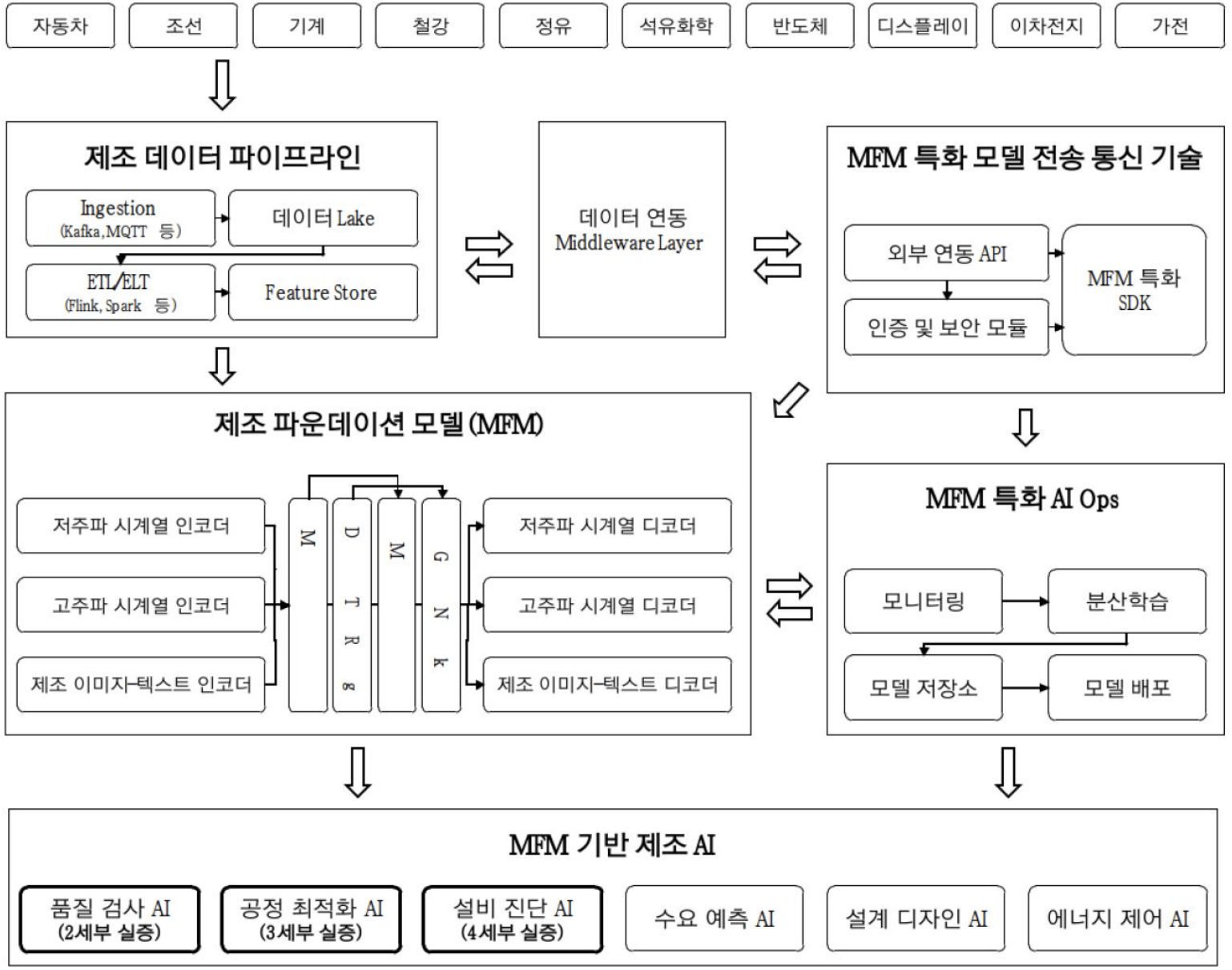
01.

MFM 소개

MFM

1. MFM이란?

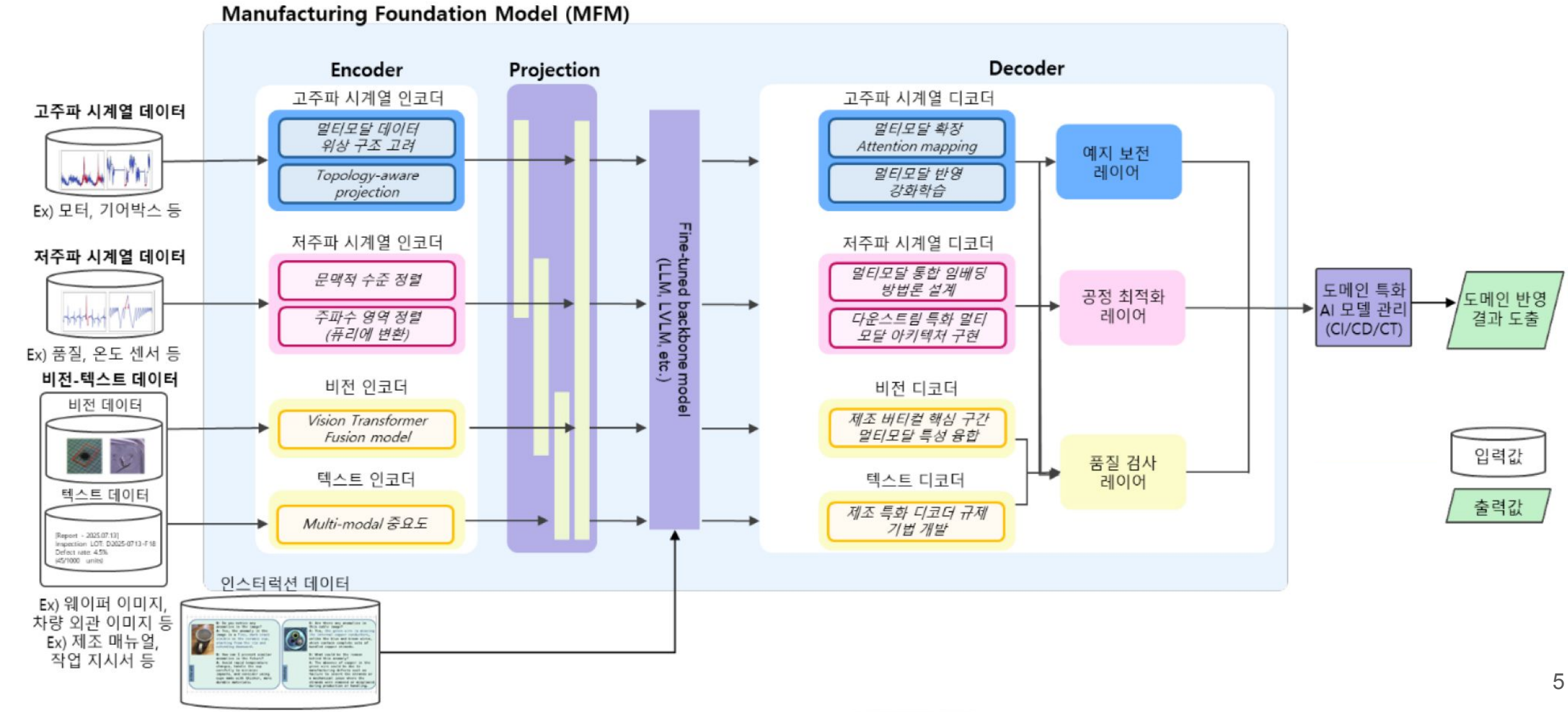
제조 특화 파운데이션 모델 (**Manufacturing Foundation Model, MFM**)은 제조 도메인에서 축적된 대규모 데이터를 통합 학습하여, 제조 데이터 전반을 분석 및 처리하는 기술.



[MFM 핵심 기술 개발 범위]

MFM

2. 목표 모델 구조


Manufacturing Foundation Model (MFM)
Encoder
Projection
Decoder
고주파 시계열 데이터
Ex) 모터, 기어박스 등
저주파 시계열 데이터
Ex) 품질, 온도 센서 등
비전-텍스트 데이터
비전 데이터
텍스트 데이터
Ex) 웨이퍼 이미지, 차량 외관 이미지 등
Ex) 제조 매뉴얼, 작업 지시서 등
인스터럭션 데이터
고주파 시계열 인코더
멀티모달 데이터 위상 구조 고려
Topology-aware projection
저주파 시계열 인코더
문맥적 수준 정렬
주파수 영역 정렬 (퓨리에 변환)
비전 인코더
Vision Transformer Fusion model
텍스트 인코더
Multi-modal 중요도
Fine-tuned backbone model (LLM, LVLM, etc.)
고주파 시계열 디코더
멀티모달 확장 Attention mapping
멀티모달 반영 강화학습
저주파 시계열 디코더
멀티모달 통합 임베딩 방법론 설계
다운스트림 특화 멀티 모달 아키텍처 구현
비전 디코더
제조 버티컬 핵심 구간 멀티모달 특성 융합
텍스트 디코더
제조 특화 디코더 규제 기법 개발
예지 보전 레이어
공정 최적화 레이어
품질 검사 레이어
도메인 특화 AI 모델 관리 (CI/CD/CT)
도메인 반영 결과 도출
입력값
출력값

02.

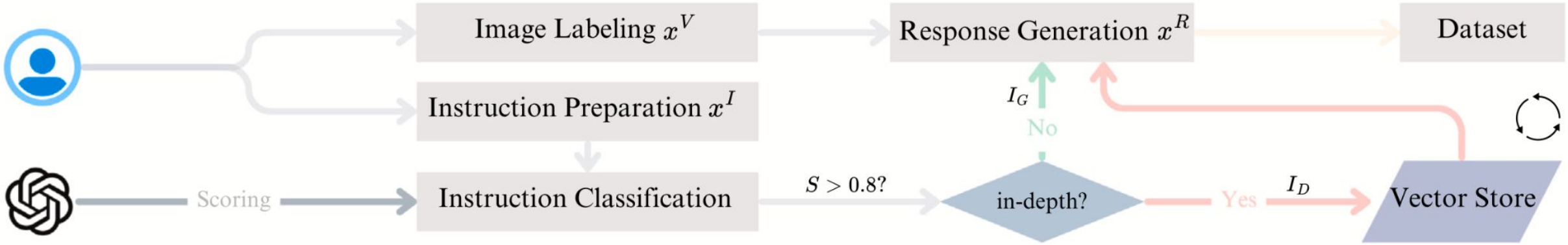
정부 기관 사례

정부 기관 사례

구분	MaViLa (🇺🇸 종료)	ELLIOT (🇪🇺 2025-2029)	DVPS (🇪🇺 2025-2029)
목표	제조 현장 특화 VLM로 공정 이해·진단·개선·제어	공간-시간 멀티모달 FM(MSTFM)로 스트리밍/시계열까지 범용화	언어+비전+센서 MMFM의 라벨효율·연산재사용·툴체인
데이터 전략	도메인 멀티모달 + 현장 이미지 (CAXTON 기반) + Q/A 생성	컨소시엄 실데이터 + European Data Spaces + 합성데이터	텍스트·이미지·영상·센서 (로보틱/공정) + 다중 소스
응용 (Agent)	에이전트형: Q/A→RAG·전문가 검색, 결함서술→탐지 2단계	로보틱스·모빌리티·워크플로 자동화로 전이학습	물리세계 이해 기반 제조/로봇 고신뢰 적용 공통 방법론
산출물	멀티모달 VLM + 에이전트 서비스 파이프라인(데모)	(계획) 오픈 멀티모달 FM	(계획) MMFM 프레임워크/툴킷

정부 기관 사례 - MaViLa

1. RAG 기반 데이터셋 파이프라인 구축

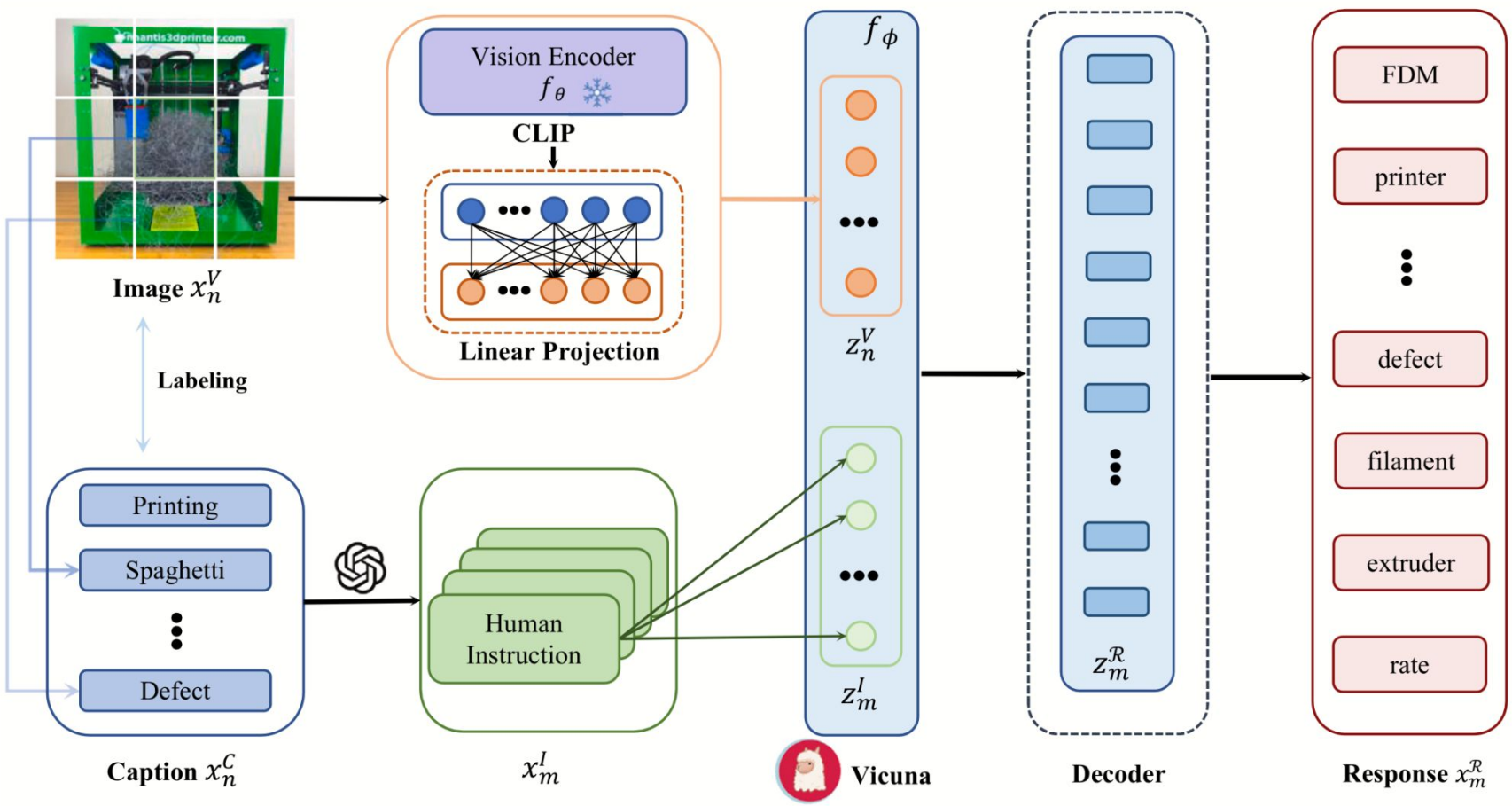


- Instruction을 Generic vs In-depth로 분류 (S>0.8 기준) → 제조 지식이 필요한 질문만 RAG로 보강.
- Vector Store(제조 문서/논문) 기반 근거 검색 → 고품질 Q/A 생성.
- 결과적으로 도메인 지식을 “모델 내부 학습 데이터”로 주입하여, 제조 특화 VLM을 위한 데이터 엔진을 설계.

* Fan, Haolin, et al. "MaViLa: Unlocking new potentials in smart manufacturing through vision language models." Journal of Manufacturing Systems 80 (2025): 258-271.

정부 기관 사례 - MaViLa

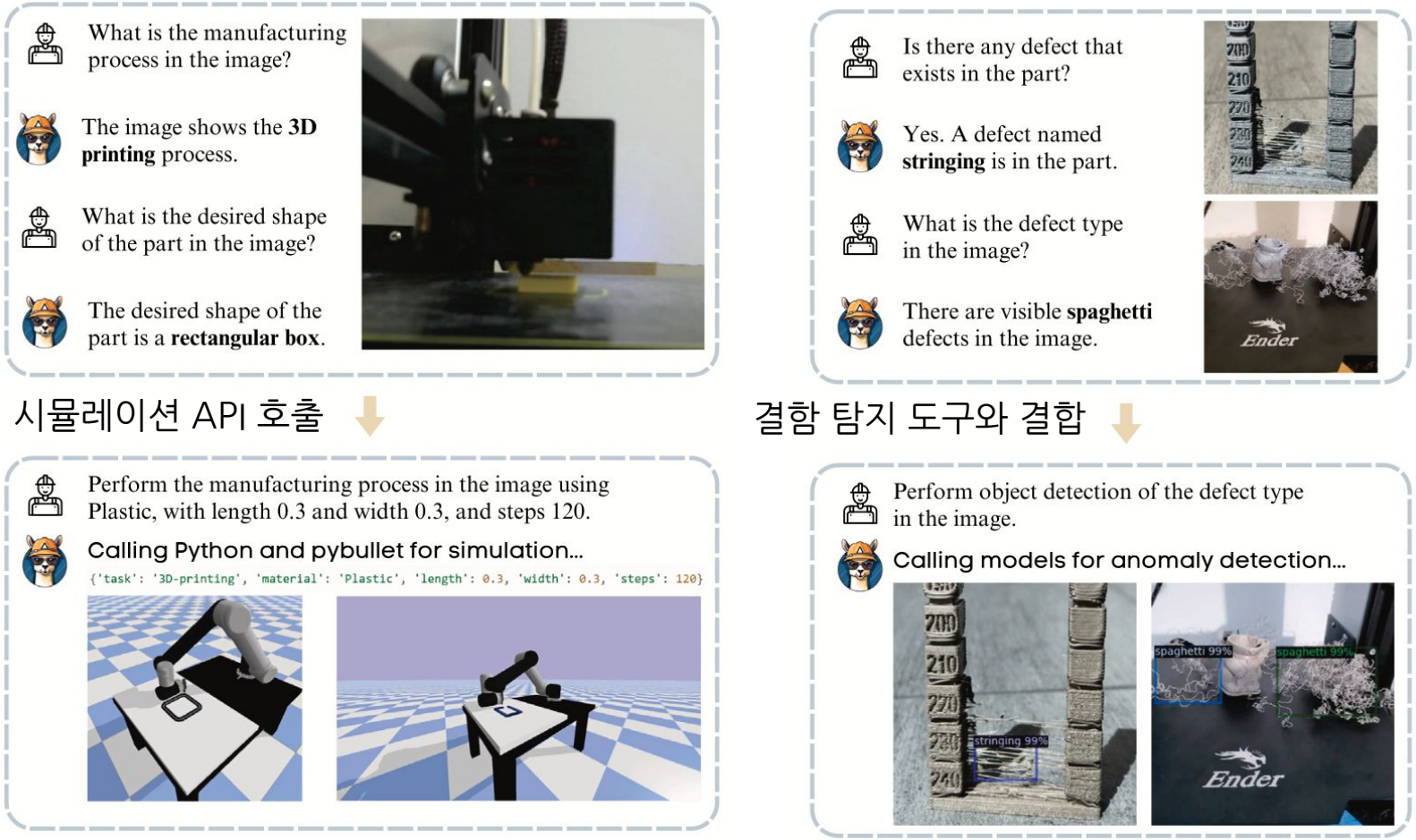
2. 모델 구조: CLIP + Vicuna 기반 VLM



* Fan, Haolin, et al. "MaViLa: Unlocking new potentials in smart manufacturing through vision language models." Journal of Manufacturing Systems 80 (2025): 258-271.

정부 기관 사례 - MaViLa

3. Agentic 확장: Simulation API / Detection Tool 연동



Perception(VLM) →

API/Tool(Action) →

Simulation/Detection으로

확장 가능한

제조 에이전트 형태 제시.

* Fan, Haolin, et al. "MaViLa: Unlocking new potentials in smart manufacturing through vision language models." Journal of Manufacturing Systems 80 (2025): 258-271.

03.

산업 기관 사례

산업 및 연구 기관 사례

기업/기관	적용영역	데이터/모달	핵심 접근	KPI·성과
NVIDIA 🇺🇸	로봇 작업 + 시뮬 훈련	인간 시연 + Omniverse 합성데이터	World Model(Cosmos) + 로봇 FM(GROOT N1)	Cosmos 출시 직후 200만+ 다운로드
General Electric 🇺🇸	X-ray 검사 + 터빈 예지보전 + 작업지시	X-ray 이미지 + 센서 + 수리기록	VLM(검사) + LLM Agent(작업지시 생성)	수치 언급 X (프로토타입/적용 확장)
IBM 🇺🇸	공정 전체 예측·최적화 (수율/설정값)	원자재·설정값·센서·품질의 다변량 시계열	Process-level FM + Digital Twin(공정 모사) + 실시간 최적화	예측정확도 7배↑, 배포시간 50%↓, 생산성 최대 5%↑, 에너지 4%↓
Landing AI 🇺🇸	부품 결함 탐지, 조립검사, OCR	대규모 사전학습 + 소량 라벨(수십~수백장)	제조 특화 LVM + Few-shot fine-tune + (클라우드/엣지)	50장으로 데모 정확도/재현율 100%, Pegatron 99.8%

산업 및 연구 기관 사례

기업/기관	적용영역	데이터/모달	핵심 접근	KPI·성과
Bosch 🇩🇪	생산라인 시각검사 성능 개선	230개 공장 데이터 + 공장당 평균 결함 합성 15,000장	Synthetic pretrain + Federated fine-tuning(공장별 최적화)	결함탐지율 ~100%, 적용기간 수개월→수주, 공장 사이클 타임 15%↓
Fraunhofer 🇩🇪	로봇/엔지니어 지원(문서 기반)	다기업 pooled data + 기술문서(RAG)	VLM+LLM 로봇 + RAG 엔지니어 어시스턴트	음성/텍스트 명령 반응 로봇 시연 성공
Siemens 🇩🇪	예지보전/이상탐지 + PLC/매뉴얼 자동화	PLC 코드 + IoT 시계열 + CAD + 시뮬	Multimodal IFM + Industrial Edge(플랫폼 통합 GenAI)	수치 언급 X (개발/운영 가속화 목표)
Huawei 🇨🇳	품질검사 + 공정최적화	정형 + 시계열 + 이미지 통합	Pangu 기반 멀티모달 적용 (검사/최적화)	결함 식별 ~40%↑, 수작업 25%↓, 30+ 산업·500+ 시나리오

04.

학술 연구 사례

학술 연구 사례

1. Vision-Language Foundation Model 연구 동향

[모델 자체 제안]

모델/프레임워크명	주요 내용 및 특징
AnomalyGPT	사전 학습된 시각 인코더와 대규모 언어 모델(LLM) 디코더를 사용하며, 시각적 특징을 텍스트 설명과 정렬함.
Multimodal LLM Framework	이미지, 시계열 센서 데이터, 생산 텍스트 기록을 통합하는 다중 모달 대형 모델 프레임워크를 도입함.
Large Manufacturing Decision Model	추론을 위한 LLM과 공장 현장 구성을 시각화하는 이미지 생성 모델(확산 모델)을 결합하여 고수준 의사결정을 제시함.

학술 연구 사례

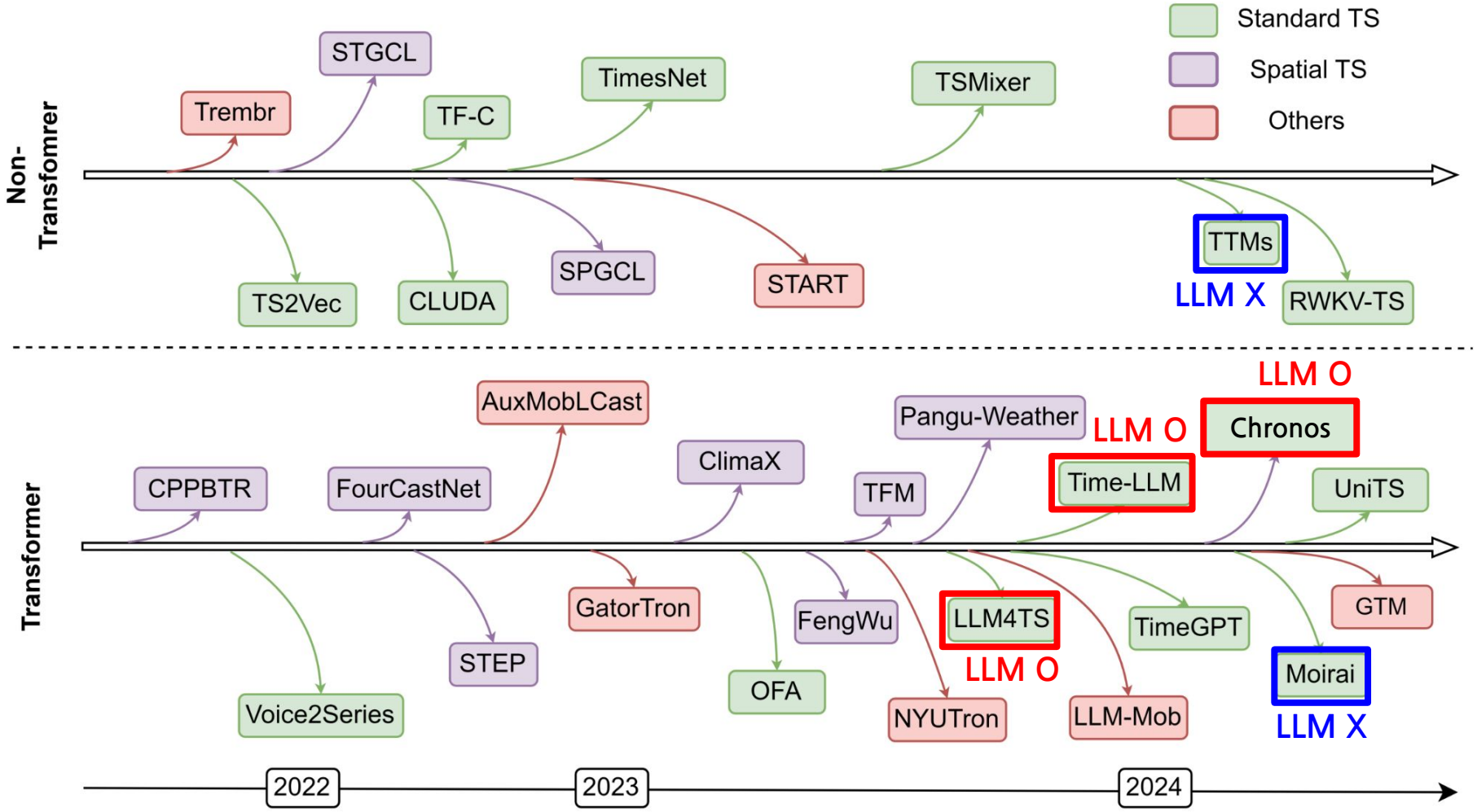
1. Vision-Language Foundation Model 연구 동향

[학습 방법론 제안]

연구명	Pain Point (문제점)	Solution (해결책)
Text-Align Anomaly Backbone	기존 산업 검사 모델은 ImageNet으로 사전 학습되어 특정 도메인의 이상 징후를 감지하는 데 한계가 있음.	CLIP을 사용한 정상 및 비정상 시각적 특징을 각각 해당하는 텍스트 설명과 정렬시키는 **'이상-텍스트 인식(Anomaly-Text-Aware)'** 사전 학습 전략을 도입함.
Dual-Interrelated Diffusion Model	실제 이상 데이터가 부족하면 생성된 샘플의 다양성이 제한되고 품질이 저하됨.	확산 모델(Diffusion Model)을 기반으로, 전체 이미지를 생성하는 **'글로벌 브랜치'**와 결함 부분만 생성하는 **'이상 브랜치'** 두 개로 구성.
Vision-Language-Diffusion Synergy	기존 이상 생성 방법들은 텍스트 설명으로 제어하기 어렵고, 별도의 학습 데이터 없이(Zero-shot) 사실적인 결함을 만드는 데 한계가 있음.	사용자가 텍스트로 결함의 유형, 위치 등을 설명하면 해당 이상 이미지를 생성할 수 있도록 구현함.
Unified Vision-Language Pre-Training	기존의 비전-언어 모델들은 단일 모달리티(이미지 또는 텍스트)와 다중 모달리티 작업을 별개의 구조로 처리하여 학습 효율이 떨어짐.	표준 트랜스포머의 FFN 계층을 비전, 언어, 비전-언어 전문가로 구성된 **MoME 모듈**로 대체하여 하나의 트랜스포머 프레임워크로 통합함.

학술 연구 사례

2. Time Series Foundation Model 연구 동향



* Liang, Yuxuan, et al. "Foundation models for time series analysis: A tutorial and survey.", KDD 2024.

학술 연구 사례

2. Time Series Foundation Model 연구 동향

모델	채널 처리	backbone	LLM 사용 방식	입력 표현 방식	pre-training 방식	fine-tuning 필요성
LLM4TS (2023)	univariate 채널 독립	Transformer + LLM	frozen LLM + alignment tuning	patch + temporal embedding	LLM pretrained → alignment 학습	필요 (alignment / LoRA 등)
TimeLLM (2024)	univariate 채널 독립	Transformer + LLM	frozen LLM + reprogramming	patch → text prototype 공간으로 projection	pretrained LLM 활용	projection head만 학습
Chronos (2024)	univariate 채널 독립	Transformer + LLM	LLM 자체를 TS 모델로 사용	quantized token (bin)	TS corpus로 LLM 직접 pretrain	fine-tuning 없이 가능
Moirai (2024)	multivariate 채널 상관	Transformer encoder	TS 전용 foundation model	patch + masked encoder	LOTSA dataset 기반 대규모 pretrain	fine-tuning 없이 가능
TTMs (2024)	multivariate 채널 상관 (파인 튜닝)	Mixer (non-Transformer)	LLM 사용 안함	patch + resolution embedding	channel-independent pretraining	fine-tuning 권장

* Chang, Ching, et al. "LLM4TS: Aligning Pre-Trained LLMs as Data-Efficient Time-Series Forecasters." *arXiv preprint arXiv:2308.08469* (2023).
* Jin, Ming, et al. "Time-LLM: Time Series Forecasting by Reprogramming Large Language Models.", ICLR 2024.
* Ansari, Abdul Fatir, et al. "Chronos: Learning the Language of Time Series.", *TMLR 2024.*
* Woo, Gerald, et al. "Unified Training of Universal Time Series Forecasting Transformers.", ICML 2024.
* Ekambaram, Vijay, et al. "Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series.", NeurIPS 2024.

05.

데이터셋

데이터셋

1. 이미지 오픈 데이터셋

MVTec AD (Anomaly Detection)

구성: 15개 카테고리(객체 10, 텍스처 5).
특징: 학습 데이터는 정상이미지만, 테스트 데이터는 80거자 아성 결함 유형 포함.
이상 현상: 긁힘, 오염, 변색, 구조적 결함 등.

VisA (Vision Anomaly)

구성: 12개 카테고리, 10,821개 이미지.
특징: PCB 기판, 캡슐 등 구조적으로 복잡한 객체 포함.
평가: 단일/다중 객체, 복잡한 구조 등 4가지 시나리오.
주 사용: 실제 공정형 결함 + 복잡한 외형.

MVTec LOCO AD (Logical Anomaly Detection)

구성: 5개 산업 시나리오 (아침 식사 상자, 주스 병, 나사 봉지 등).
특징: 논리적 이상 탐지 (부품 누락, 잘못된 배치 등).
요구 사항: 전역적 이해와 설명 가능성.

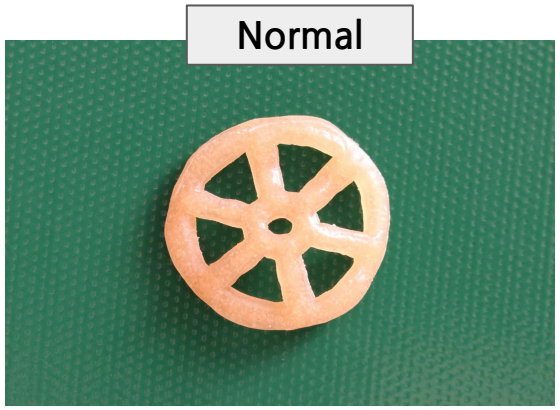
GoodsAD

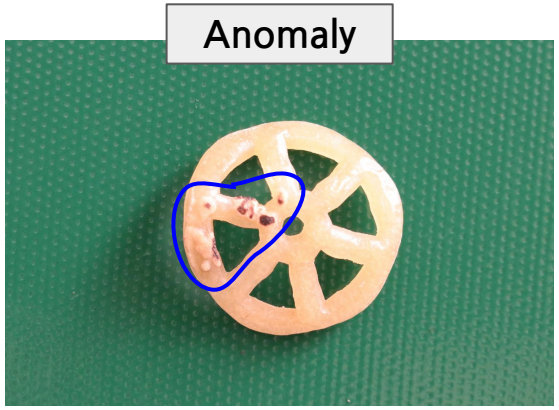
구성: 32개 제품 카테고리, 8개 브랜드, 16,871 이미지.
특징: 클래스 내 변동성↑, 현실 촬영 환경.
결함 유형: 긁힘, 찢어짐, 인쇄 오류, 내용물 누수 등.
주 사용: 제품·패키징 검사에 가까운 설정.

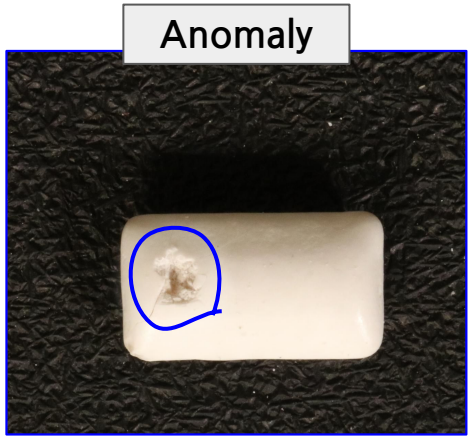
데이터셋

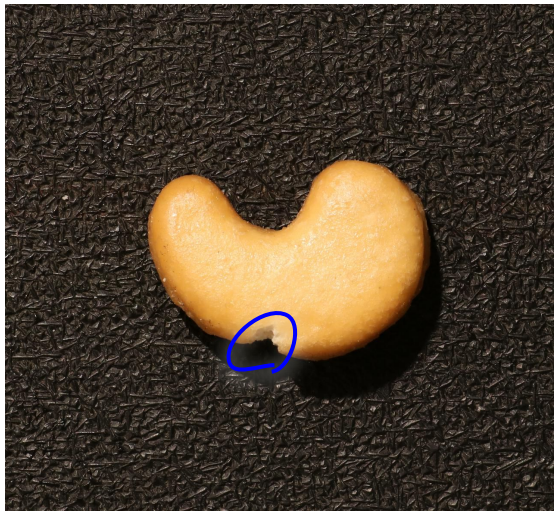
1. **이미지 오픈 데이터셋 -** VisA 예시

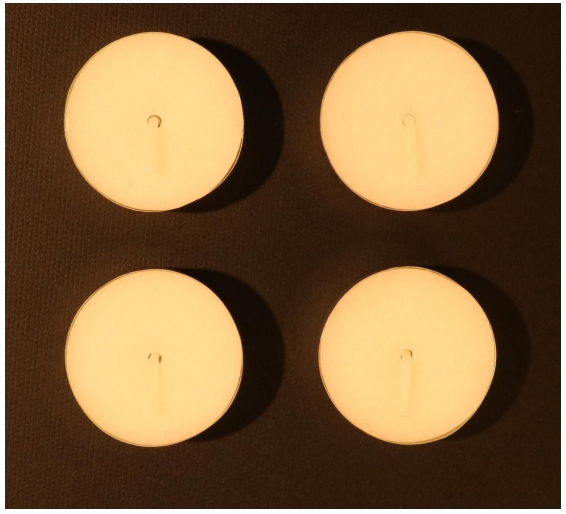
쉬운 문제

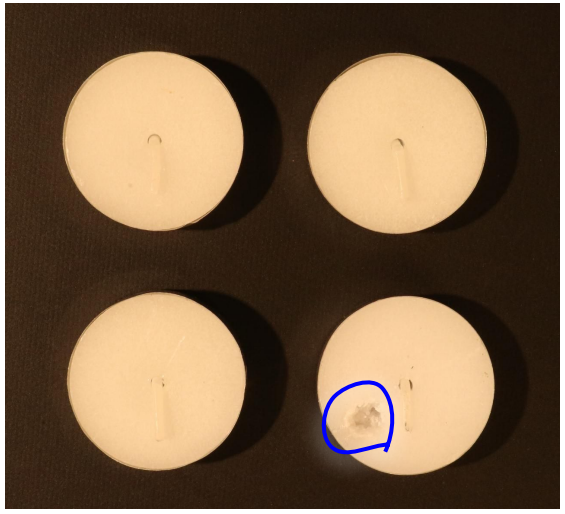








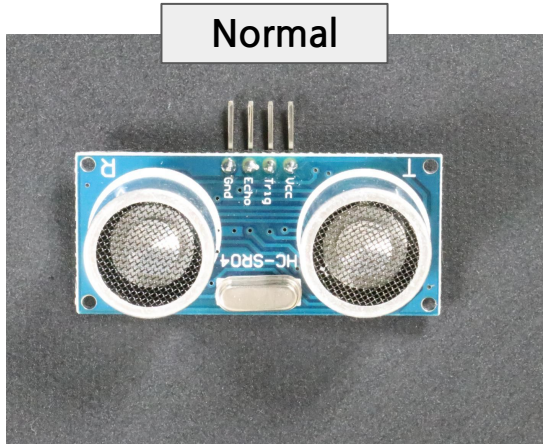


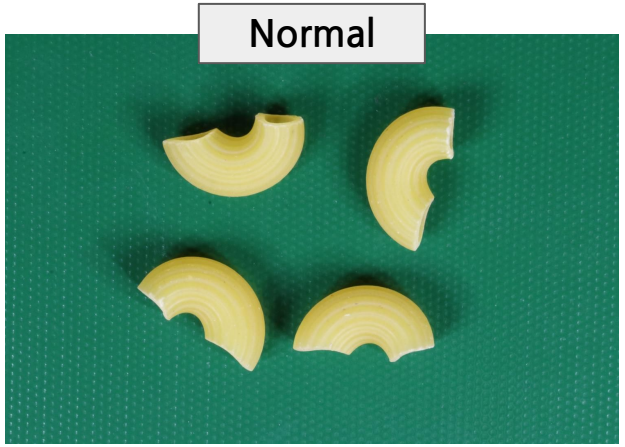


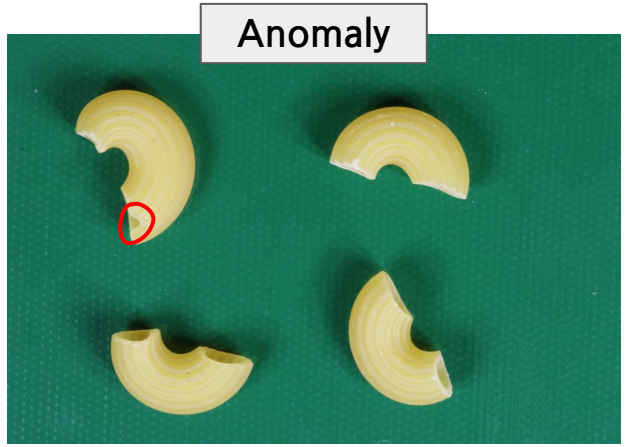
데이터셋

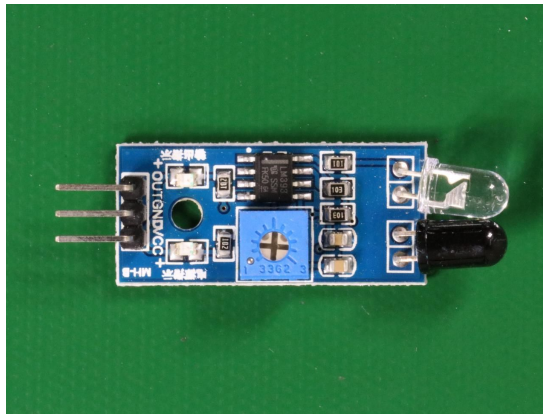
1. **이미지 오픈 데이터셋 -** VisA 예시

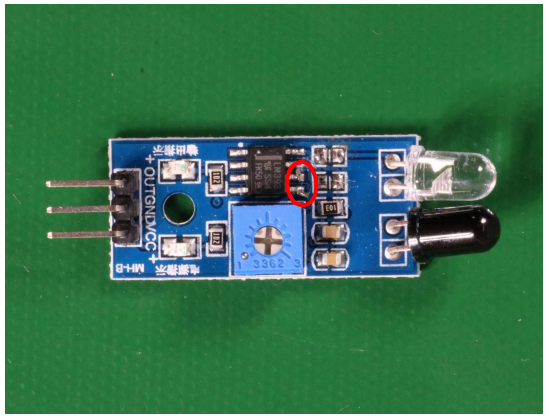
어려운 문제











데이터셋

2. 시계열 오픈 데이터셋

AI4I 2020 Predictive Maintenance (합성 데이터, 10,000 샘플)

- 입력 변수: Air temperature(K), Process temperature(K), Rotational speed(rpm), Torque(Nm)
- 라벨: Machine failure (binary)

Steel Energy Consumption (에너지/전력 사용 데이터)

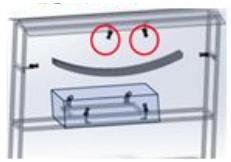
- 입력 변수: Usage_kWh, (Lagging/Leading) Reactive Power(kVarh), (Lagging/Leading) Power Factor

Industrial IoT Fault Detection (1,000 샘플)

- 입력 변수: Vibration(mm/s), Temperature(°C), Pressure(bar)
- 라벨: Fault class {0,1,2}

데이터셋

3. 이미지 수집 데이터셋

불량 유형	용접 불량	스크레치	덴트	버	홀 유무
특징	용접 타수: 30타 검사 항목: 용접 피놀, 용접 누락, 비드 길이 불량	스크레치 발생	형상이 변형	외부에 돌출부 발생	홀의 유무 확인
촬영 방식	전용 지그 2대 	스테이션 운반 중 2대 → 전용 지그 2대 	스테이션 운반 중 2대 → 전용 지그 4대 	전용 지그 2대 	전용 지그 2대 
촬영 단면	-T면	+-H면	+-h면, +-T면	+-L면	+T면

데이터셋

4. 시계열 수집 데이터셋

실제 현업 공정 데이터 (삼양)

- 3-stage 연속 공정: Melter → Saturator 1/2(병렬) → Saturator 3
- 변수 구성 및 수집 주기
  - 원료 변수: 1일 1회(07:00) 성분 분석, 다음 측정 전까지 동일 값 유지.
  - 공정 변수: 15분 간격 센서 기반 모니터링과 제어.
  - 설비 상태 변수: 1일 1회(09:00) 설비 상태 측정, 다음 측정 전까지 동일 값 유지.
  - 타겟: 센서 기반 제품 생산량(yield)

06.

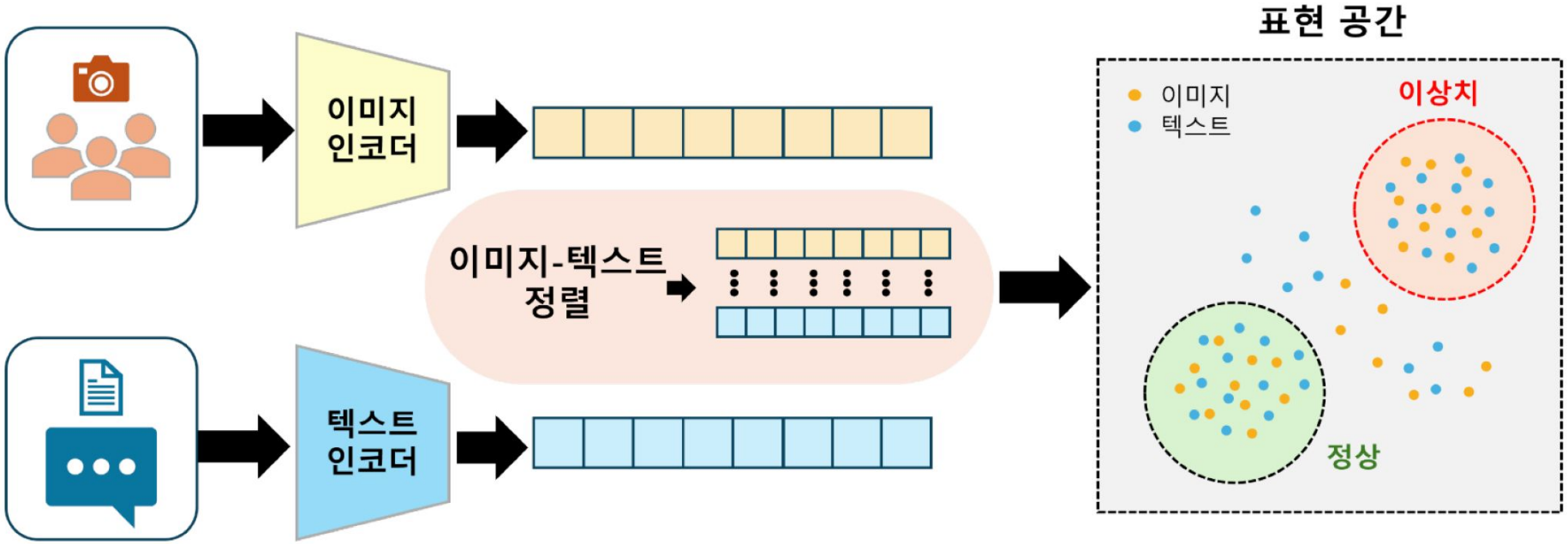
연구 진행 사항

연구 진행 사항

1. 모델 아키텍처

- Vision Encoder: CLIP (사전학습 가중치 사용, freeze)
- LLM: LLaVA v1.6-7B (사전학습 가중치 사용, freeze)
- Adapter: LoRA + **자체 설계 소형 Projection Layer**

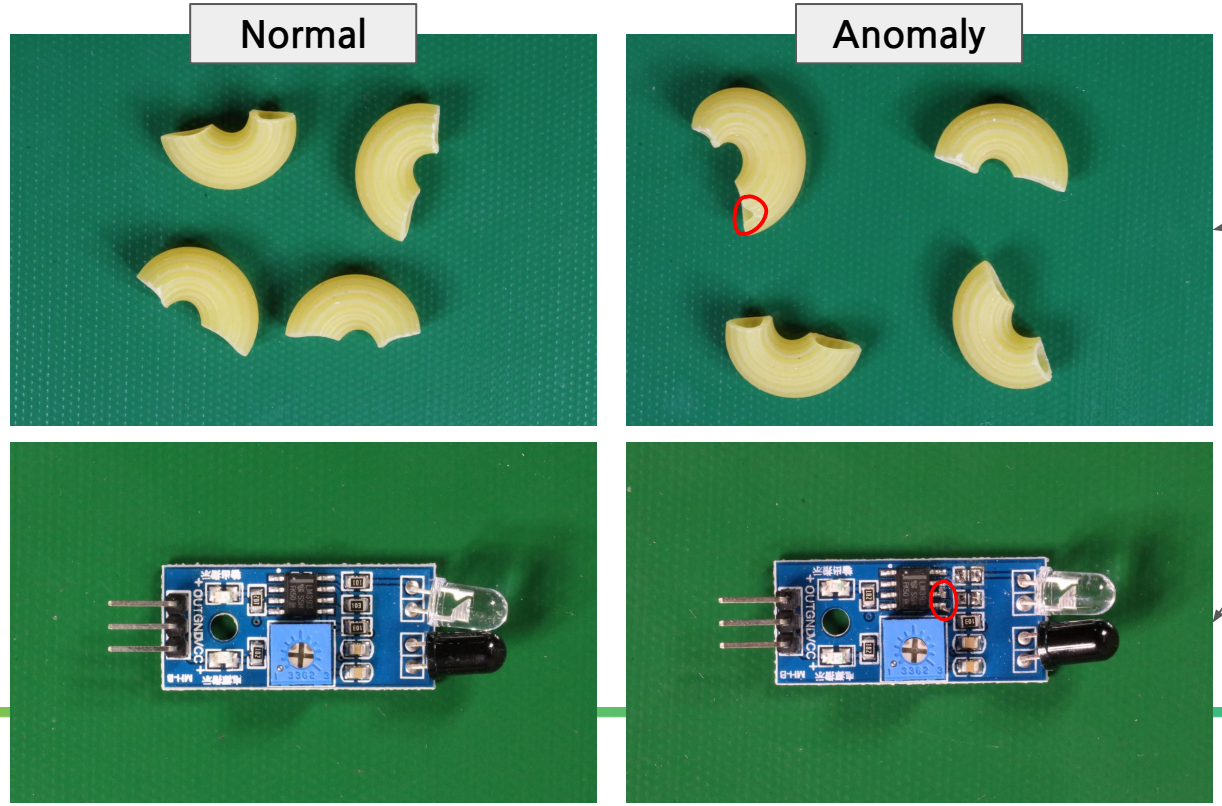
→ 대부분 가중치 재사용 + 소수 파라미터만 학습해 빠르게 성능 달성 목표.



연구 진행 사항

2. KPI 달성

- 카테고리별 점수 분포가 달라 Global threshold 대신 클래스별 최적 임계값을 개별 적용.
- **mAP 0.92로 KPI 초과 달성.**

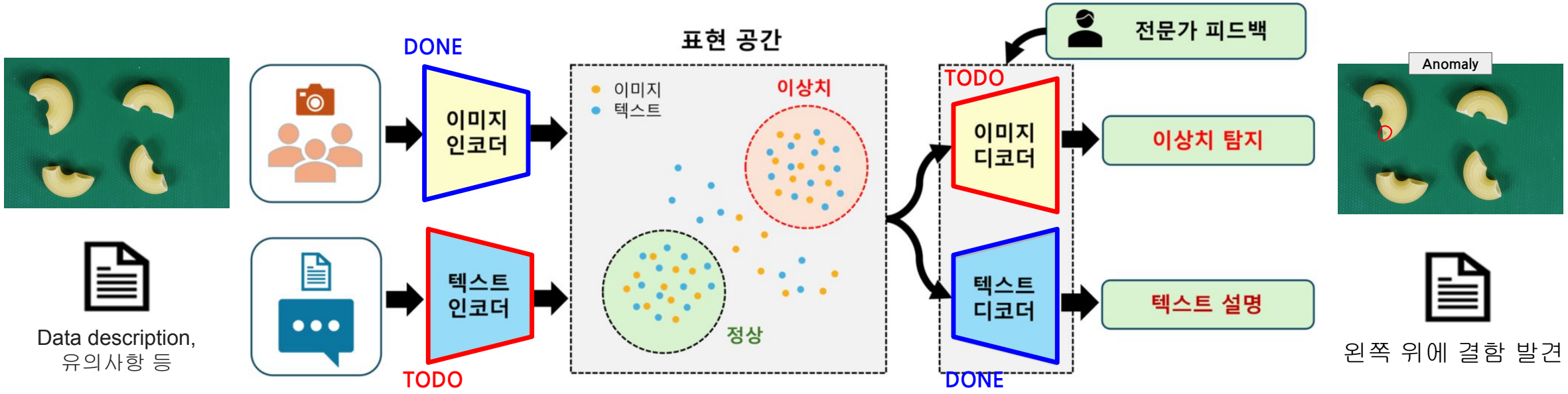


Category	AP	AUROC
candle	0.9835	0.9981
capsules	0.9129	0.9596
cashew	0.9397	0.9642
chewinggum	0.9585	0.9629
fryum	0.9512	0.9729
macaroni1	0.8719	0.9578
macaroni2	0.8058	0.9338
pcb1	0.8845	0.9794
pcb2	0.8772	0.9354
pcb3	0.8544	0.9498
pcb4	0.9400	0.9851
pipe_fryum	0.9750	0.9920

mAP (Mean Average Precision): 0.9229
mAUROC (Mean AUROC) : 0.9659

연구 진행 사항

3. 추후 계획



Q & A

Thank You!

References

- Fan, Haolin, et al. "MaViLa: Unlocking new potentials in smart manufacturing through vision language models." *Journal of Manufacturing Systems* 80 (2025): 258-271.