DMQA Open Seminar

Vision Language Model-based Anomaly Detection

2026. 02. 27

김성수

Data Mining and Quality Analytics Lab

고려대학교
KOREA UNIVERSITY

발표자 소개

❖ 김성수 (Sungsu Kim)

- 경희대학교 산업경영공학과 학부 졸업 (2022.02)
- 고려대학교 산업경영공학과 대학원 재학
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- 석박통합 과정 (2022.03 ~ Present)

❖ Research Interest

- Computer Vision
- Vision Language Model
- Foundation Model

❖ Contact

- 2022020650@korea.ac.kr

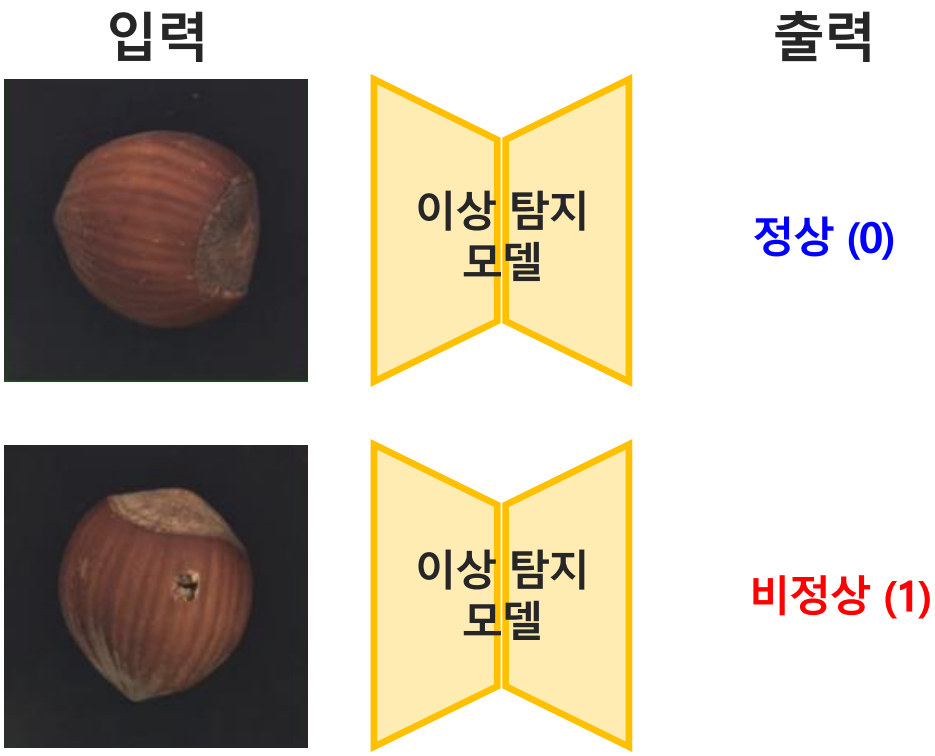
목차

Introduction

Introduction

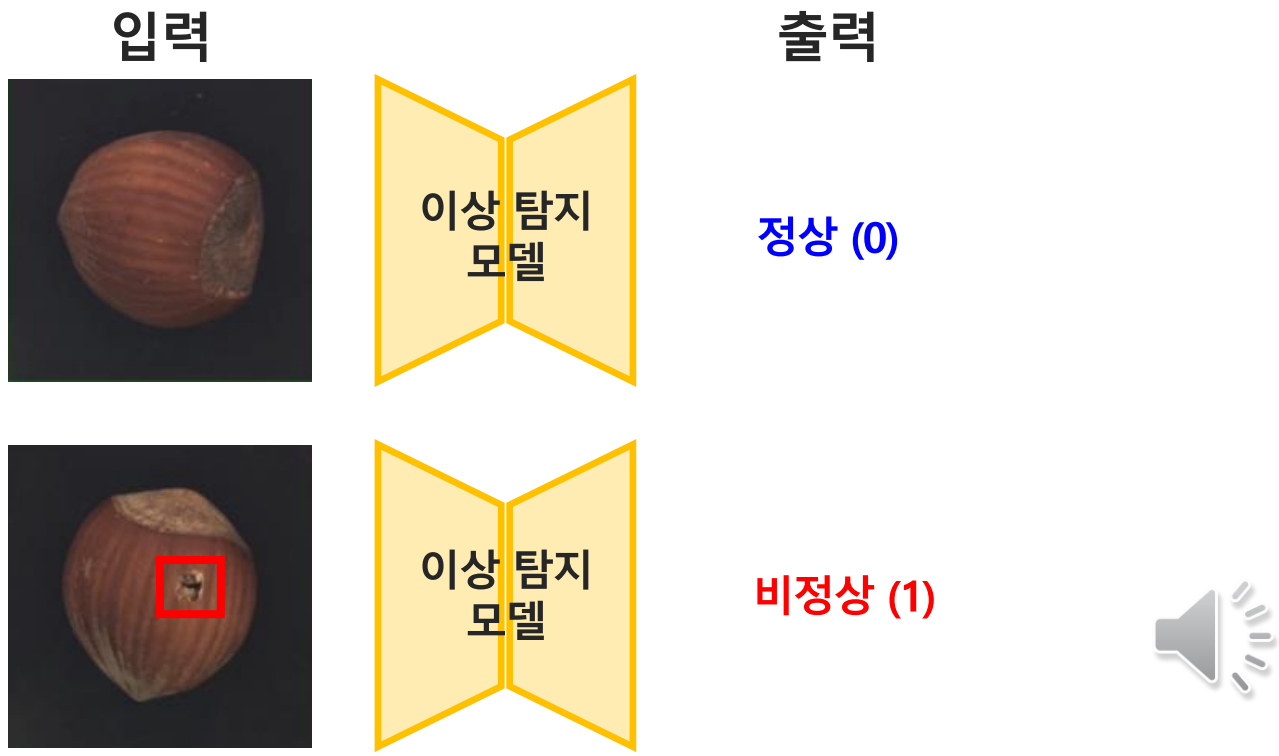
❖ Visual Anomaly Detection

- 이미지 또는 비디오 내 이상 여부를 탐지하는 Task
- 이미지 이상탐지 - 입력: 이미지 → 출력: 이상 여부 (0 or 1)



Introduction

❖ **Visual Anomaly Detection (이미지 이상 탐지) → 오직 시각적인 정보만 고려 가능**

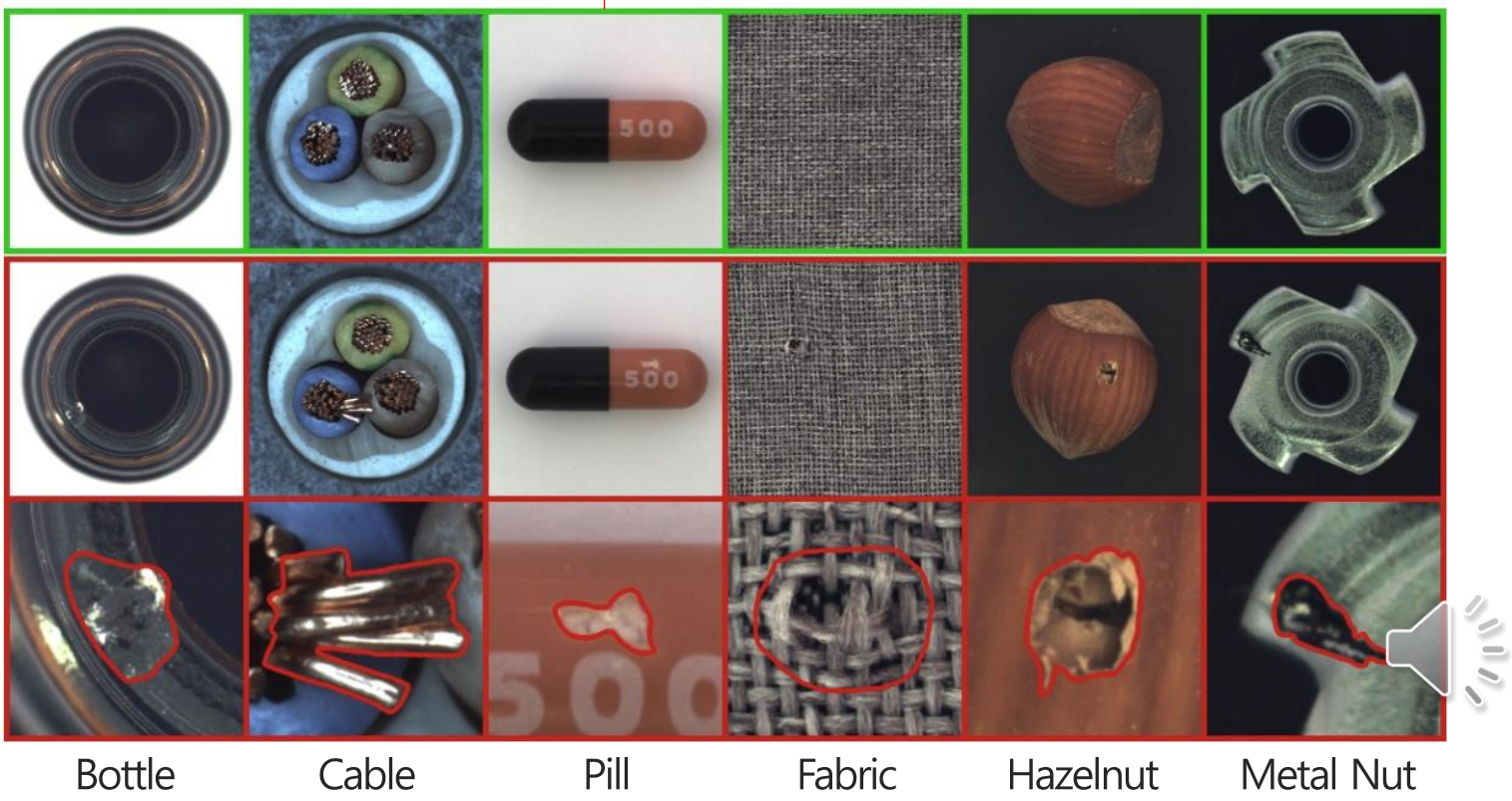


Introduction

❖ **Visual Anomaly Detection (이미지 이상 탐지)** → **오직 시각적인 정보만 고려 가능**

- **해결방향:** 각 객체 별 semantic 정보를 반영

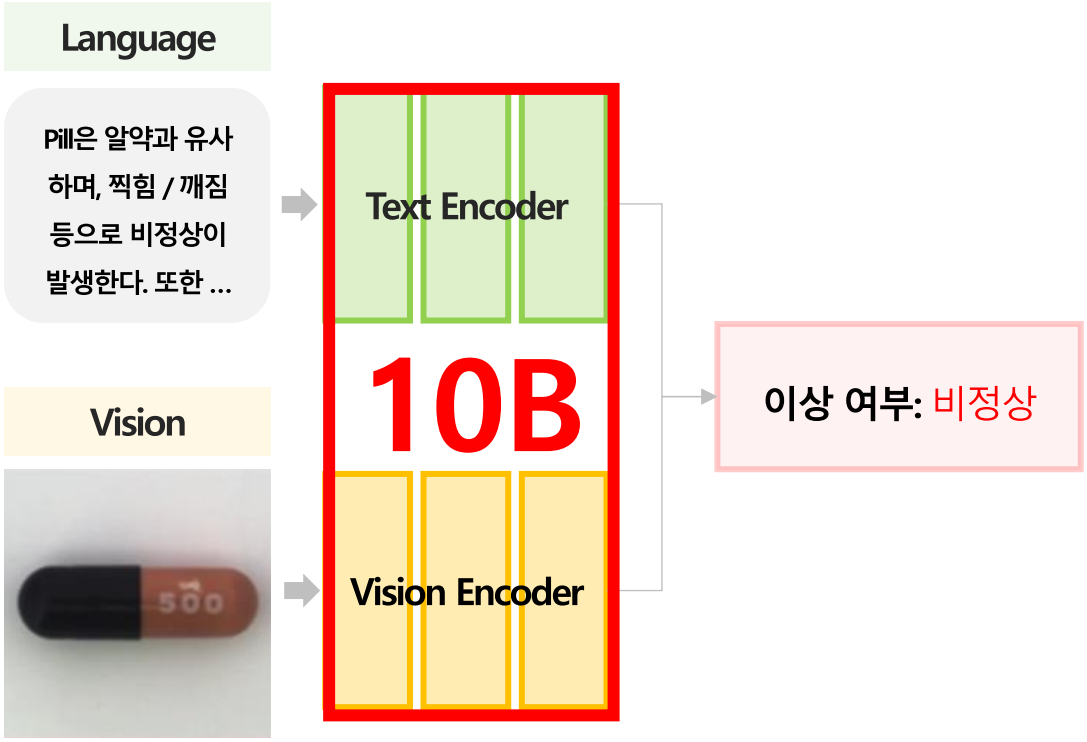
Pill의 불량 유형은 찍힘, 깨짐, 오염, 인쇄 불량 등이 있으며...



Introduction

❖ **어떻게 semantic한 정보를 visual 정보와 함께 고려할 수 있을까?**

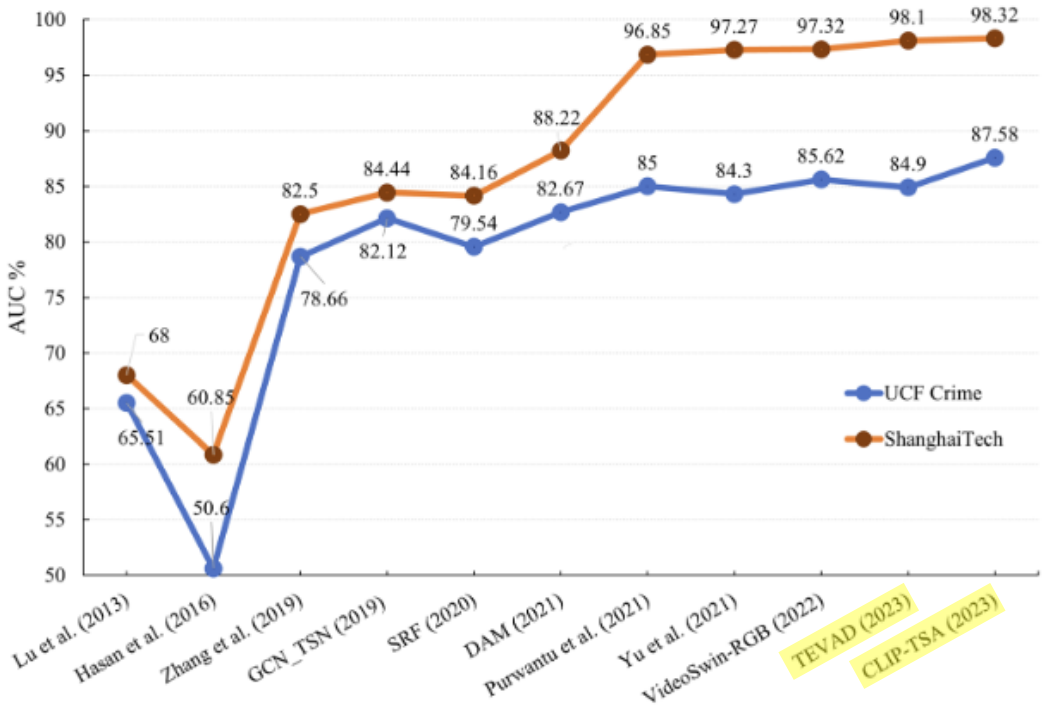
- **Vision Langauge Model (VLM):** 이미지와 텍스트를 한 번에 고려 할 수 있는 멀티모달 모델
- 최근에는 거대 모델 구조에 수 억개 이미지-텍스트 데이터로 학습한 Foundation VLM을 활용



Introduction

❖ VLM in Anomaly Detection

- 최근 Foundation VLM을 기반으로 visual anomaly detection에서 큰 성능 향상을 보임



[최근 video anomaly detection의 성능 변화 추이]

[1] Abdalla, M., Javed, S., Al Radi, M., Ulhaq, A., & Werghi, N. (2025). Video anomaly detection in 10 years: A survey and outlook. *Neural Computing and Applications*,

Algorithms

- Foundation VLM -

Algorithm

- Vision Language Model

CLIP (OpenAI, 2021/ICML)
→ 이미지와 텍스트 간 Align을 맞춤

LLaVA (Microsoft, 2023/NeurIPS)
→ 이미지와 텍스트를 활용하는 다양한 Instruction에 대응

Video-LLaVA (University of Wisconsin-Madison, 2024/EMNLP)
→ {이미지 및 비디오}와 텍스트를 활용하는 다양한 Instruction에 대응

Algorithm

- Vision Language Model (1/3)

❖ CLIP (2021/ICML – OpenAI)

- 가장 초창기 VLM
- 이미지와 텍스트 간 Align을 맞춘 최초의 연구

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

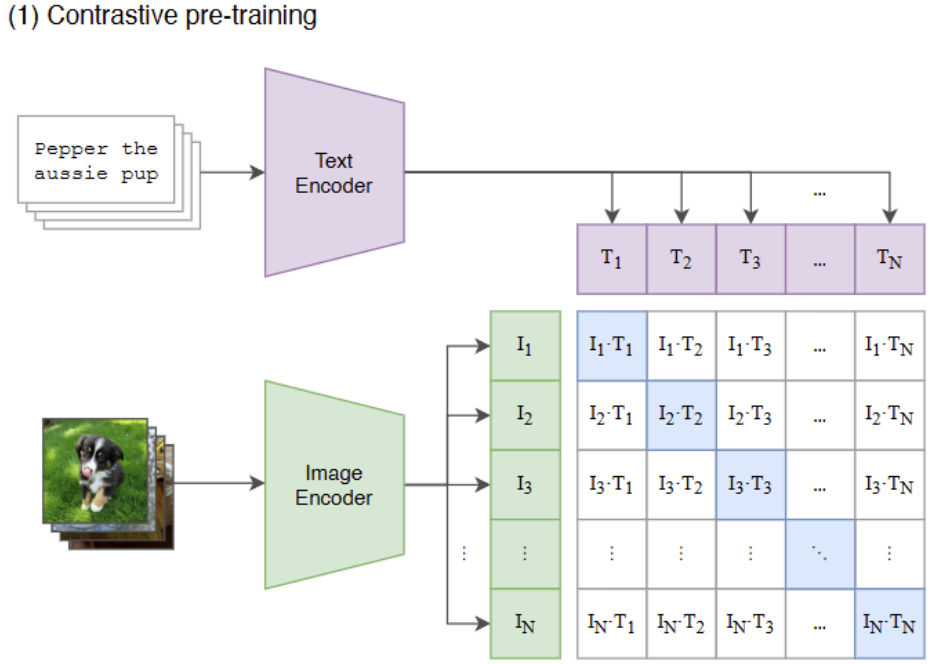
[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In ICML

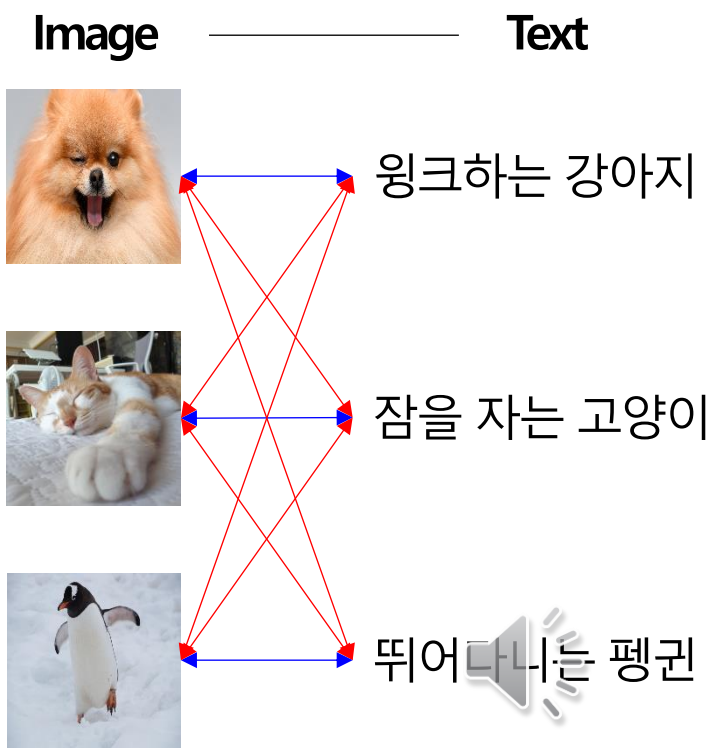
Algorithm

- Vision Language Model (1/3)

❖ CLIP (2021/ICML – OpenAI) - 학습

- 4억 개의 이미지–텍스트 쌍으로 400M 규모의 모델을 대조학습을 기반으로 충분히 학습
- 다양한 이미지와 텍스트에 학습 없이 활용가능한 똑똑한 Image & Text Encoder 확보



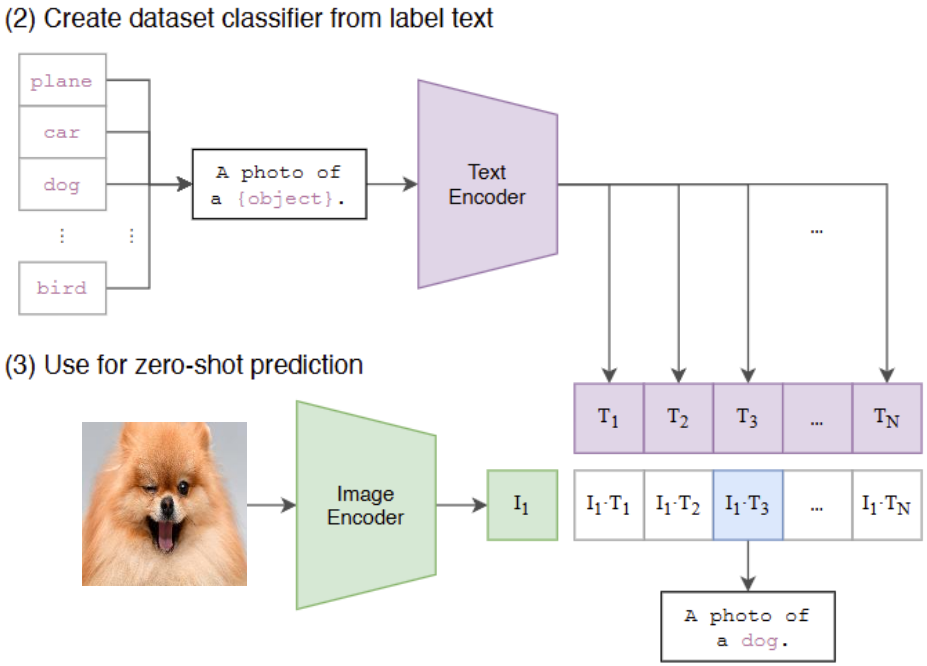


Algorithm

- Vision Language Model (1/3)

❖ CLIP (2021/ICML – OpenAI) - 추론

- **Image & Text Encoder:** 특성 Task에 적합하게 개발된 것은 아니며, Feature 추출 기능만 가능
- **기능:** 제한적 → 이미지와 텍스트 간 정렬 가능 (개 이미지와 개에 대한 텍스트 매칭 가능)
  - ➢ 추가적인 step을 도입하여 Classification 문제 등 확장 가능



	a photo of a {class}			
	plane	car	dog	bird
이미지	0.3	0.2	0.8	0.1

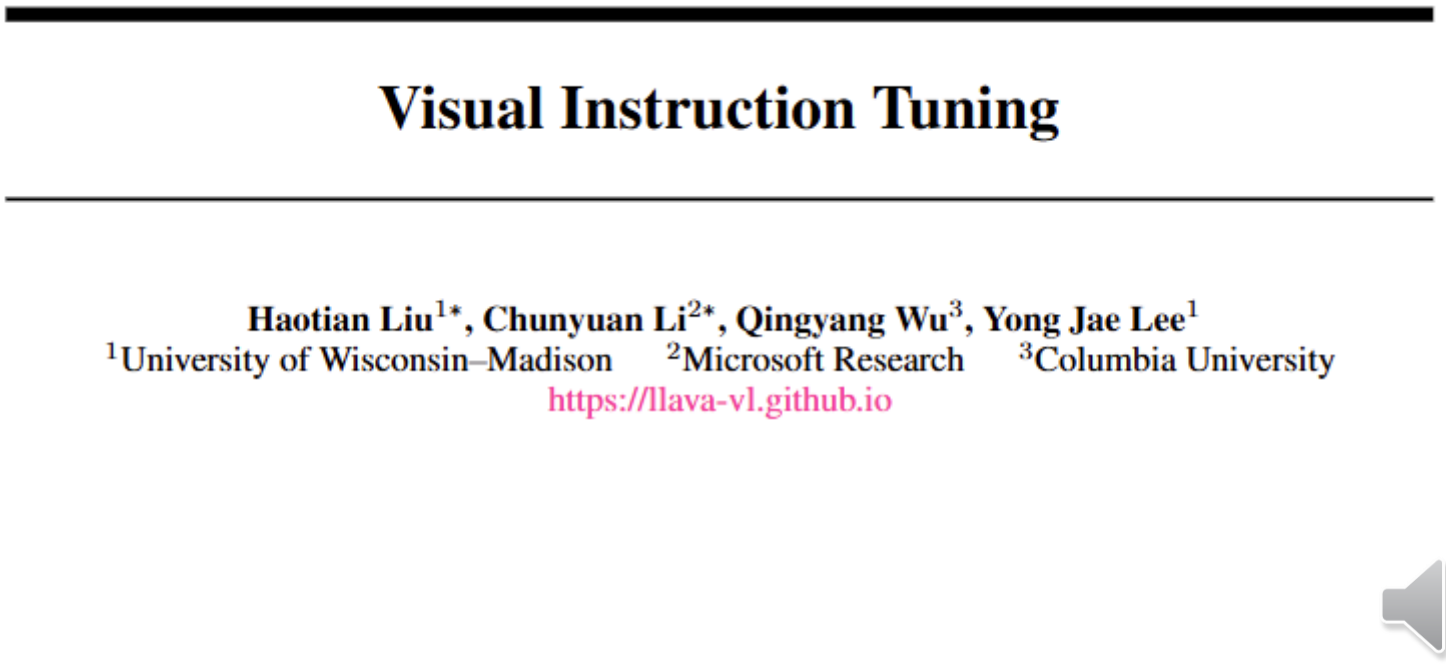
= dog

Algorithm

- Vision Language Model (2/3)

❖ LLaVA (2023/NeurIPS – Microsoft)

- 단순 이미지 & 텍스트 간 Align → 다양한 Task를 수행할 수 있는 VLM
- 입력되는 이미지와 텍스트에 대하여 다양한 Instruction에 대해 문장 형식으로 답변



Visual Instruction Tuning

Haotian Liu[1*], Chunyuan Li[2*], Qingyang Wu[3], Yong Jae Lee[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

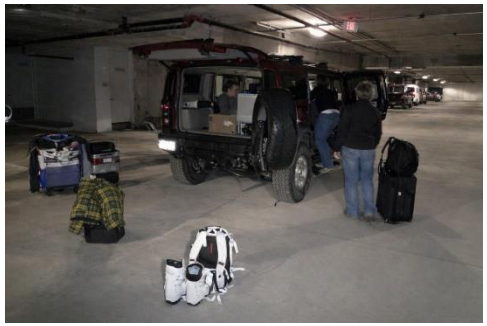
[3] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in neural information processing systems, 36, 34892-34916.

Algorithm

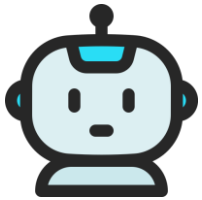
- Vision Language Model (2/3)

❖ LLaVA (2023/NeurIPS – Microsoft)

- "VLM을 활용하여 실제 다양한 Task를 수행할 수는 없을까?"
- **목표:** 다양한 Instruction에 대응할 수 있는 VLM을 만들어보자.



Q1. 자동차가 무슨 색이야?
Q2. 차 안에 사람은 몇 명이 있어?
Q3. 지금 사람들은 어떤 문제 상황에 있어?



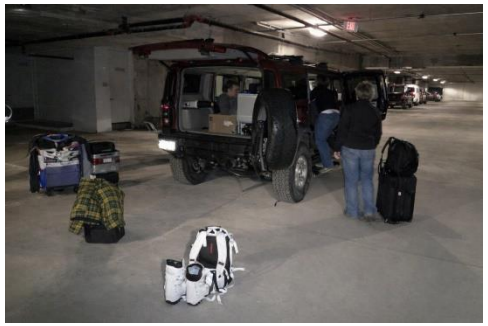
A1. 검정색.
A2. 1명.
A3. 짐이 많아 차에 적재가 어려움.

Algorithm

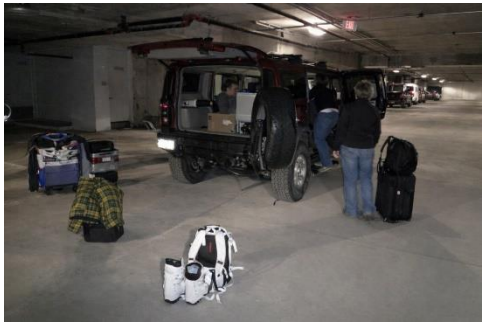
- Vision Language Model (2/3)

❖ LLaVA (2023/NeurIPS – Microsoft)

- 다양한 Instruction에 대해 대응하기 위해서는 (Instruction, Output) 형태 데이터 必
- 기존 데이터는 Instruction 없이, 이미지와 캡션만 존재 → Instruction 정보 전무
- {이미지, Text 캡션} 데이터를 {이미지, Instruction, Output}으로 확장할 수 있는 파이프라인 제안



자동차 주변에 다양한 짐이 있습니다.



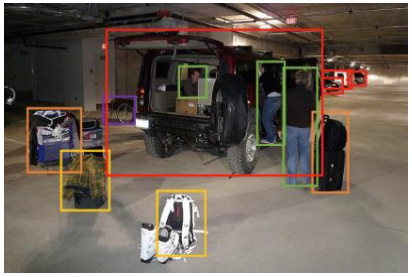
Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Algorithm

- Vision Language Model (2/3)

❖ LLaVA (2023/NeurIPS – Microsoft)

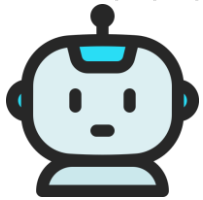
- 어떻게 데이터셋을 구축할까? → LLM을 활용하여 {이미지, Instruction, Output} 생성해보자.
- LLM은 이미지를 입력 받을 수 없음 → 이미지를 표현하는 캡션과 객체의 Box 정보 함께 활용
  - ➢ 이때 활용되는 캡션과 객체 Box 정보는 인간이 직접 레이블링 필요



캡션과 Box를 고려하여, 이미지 내 Content와 위치 정보에 대해 풍부한 Q&A 셋 확보 가능

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.
Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

위 Caption과 Box 정보를 고려하여 Q&A 세트를 만들어줘



Q1. 이미지에서 사람은 몇 명인가요?
A1. 이미지에는 총 2명의 사람이 있습니다.

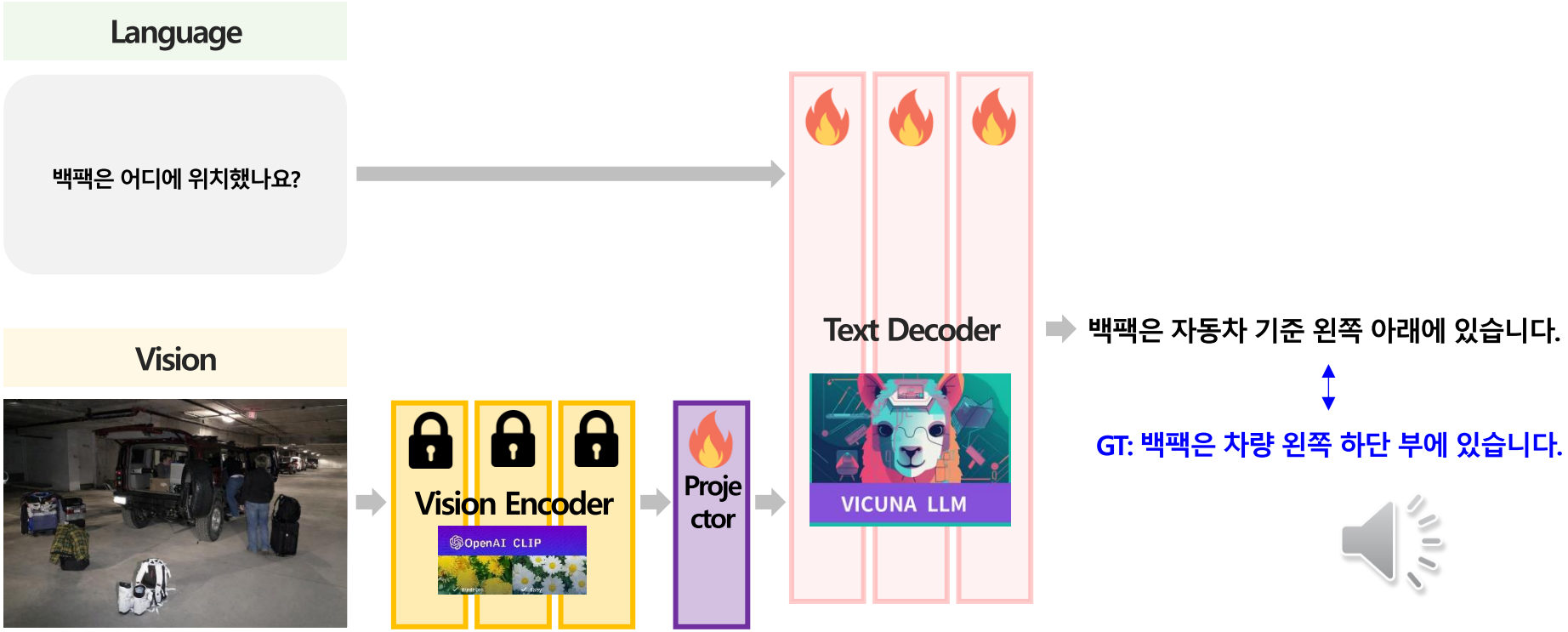
Q2. 백팩은 어디에 위치해있나요?
A2. 백팩은 차량 왼쪽 하단 부에 있습니다.

Algorithm

- Vision Language Model (2/3)

❖ LLaVA (2023/NeurIPS – Microsoft)

- 구축한 데이터셋 (158K)을 기반으로 VLM 미세조정
  - ➢ 이미 Visual 및 Text Decoding에는 충분한 학습이 되어있기에, 158K만으로도 충분한 학습 가능
- **VLM 구조:** CLIP Visual Encoder + Vicuna Text Decoder [4]



[4] Wei-Lin, C., Zhuohan, L., Lin, Z., Ying, S., Wu, Z., Hao, Z., ... & Ion, S. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. LMSYS.

Algorithm

- Vision Language Model (2/3)

❖ LLaVA (2023/NeurIPS – Microsoft)

- 이러한 Instruction에 대응 가능한 기존 Foundation VLM인 BLIP2보다 크게 우수한 성능을 보임

	Conversation	Detail description	Complex reasoning	All
OpenFlamingo [5]	19.3 ± 0.5	19.0 ± 0.5	19.1 ± 0.7	19.1 ± 0.4
BLIP-2 [28]	54.6 ± 1.4	29.1 ± 1.2	32.9 ± 0.7	38.1 ± 1.0
LLaVA	57.3 ± 1.9	52.5 ± 6.3	81.7 ± 1.8	67.3 ± 2.0
LLaVA†	58.8 ± 0.6	49.2 ± 0.8	81.4 ± 0.3	66.7 ± 0.3

Algorithm

- Vision Language Model (3/3)

❖ Video-LLaVA (2024/EMNLP – University of Wisconsin-Madison)

- {이미지 또는 비디오}와 텍스트를 입력 받아 다양한 Task를 수행할 수 있는 VLM
- 입력되는 {이미지 또는 비디오}에 대하여 다양한 Instruction에 대해 문장 형식으로 답변

Video-LLaVA: Learning United Visual Representation by Alignment Before Projection

Bin Lin[1], Yang Ye[1], Bin Zhu[1], Jiaxi Cui[4], Munang Ning[1,2,3], Peng Jin[1,2,3], Li Yuan[1,2,3]

[1]Peking University Shenzhen Graduate School, [2]Peng Cheng Laboratory,
[3]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School,
[4]PandaVilla Tech Limited

Correspondence: yuanli-ece@pku.edu.cn

GitHub: https://github.com/PKU-YuanGroup/Video-LLaVA

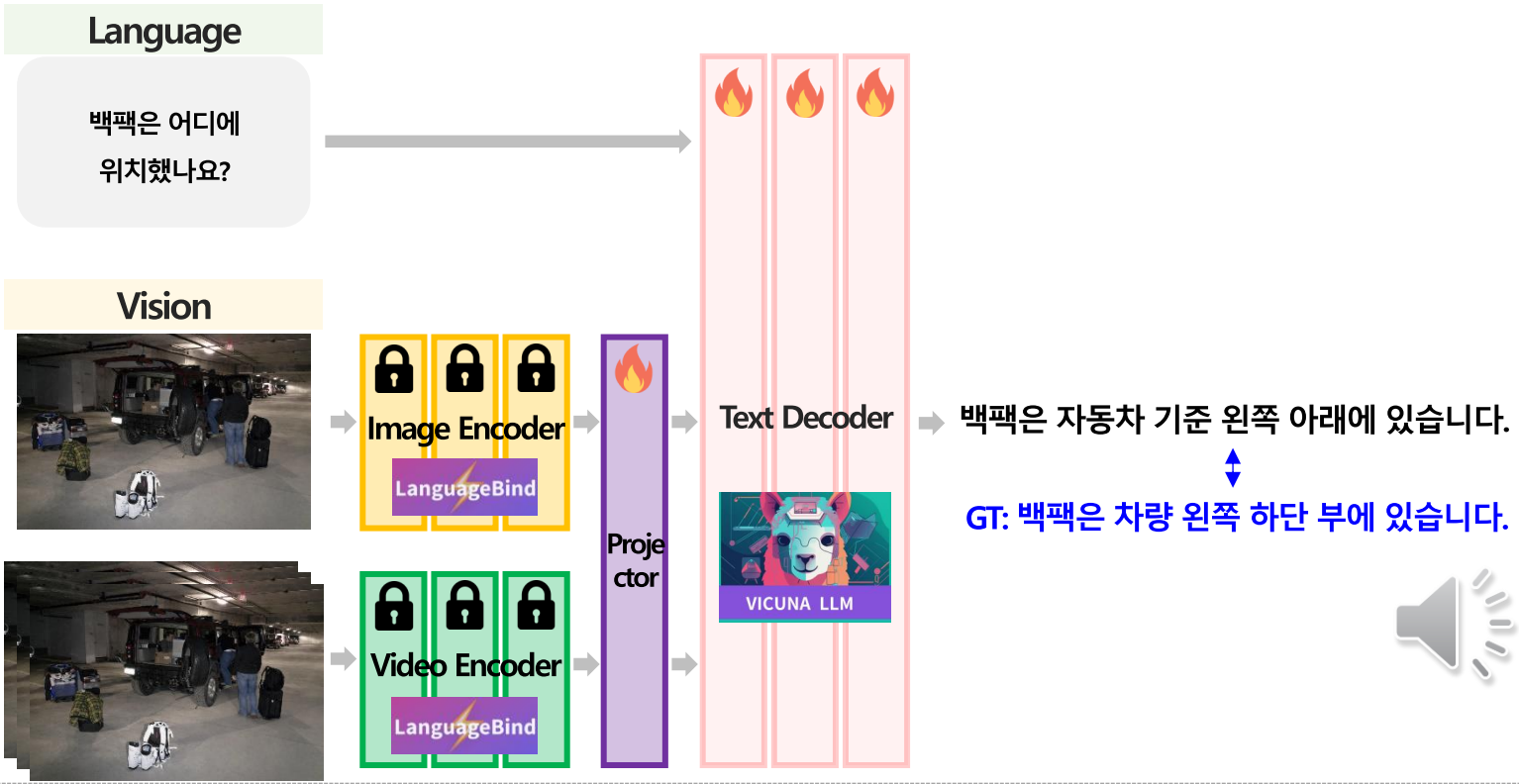
[5] Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., & Yuan, L. (2024, November). Video-llava: Learning united visual representation by alignment before projection. In EMNLP.

Algorithm

- Vision Language Model (3/3)

❖ Video-LLaVA (2024/EMNLP – University of Wisconsin-Madison)

- 이미지 뿐만 아니라, 비디오에 대해서 호환되도록 모델 구성
- 이미지 & 비디오 Instruction을 활용하여 Projector와 Text Decoder 학습
  - ➢ LLaVA 1.5 instruction dataset (665K) & Video-ChatGPT instruction dataset (100K)

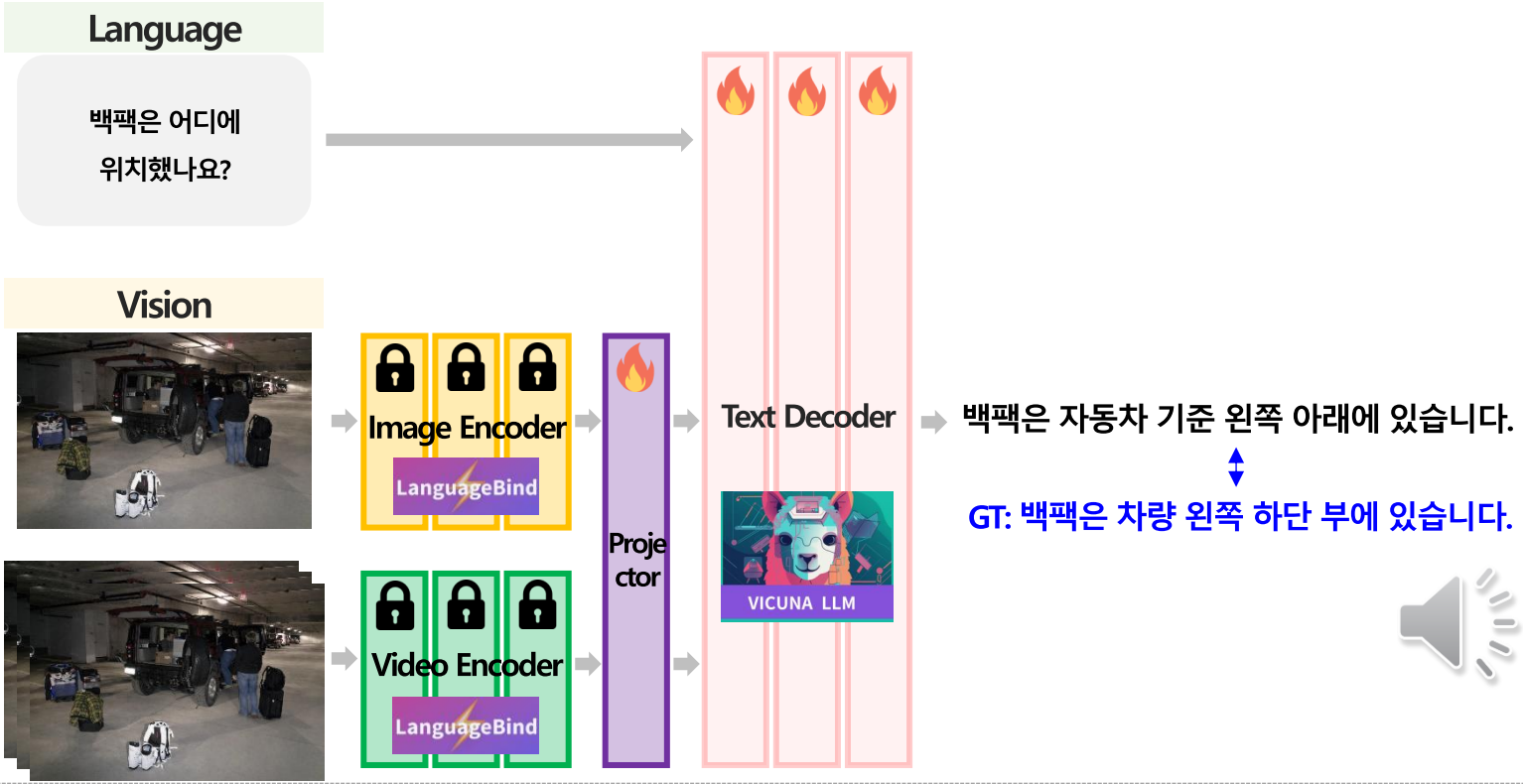


Algorithm

- Vision Language Model (3/3)

❖ Video-LLaVA (2024/EMNLP – University of Wisconsin-Madison)

- 비슷한 문맥의 이미지와 비디오에 대해서는, 두 Feature가 유사한 결과를 출력해야 함.
- Image와 Video에 대해 잘 정렬되도록 학습된 Visual Encoder인 LanguageBind 활용



Algorithm

- Vision Language Model (3/3)

❖ Video-LLaVA (2024/EMNLP – University of Wisconsin-Madison)

- 기존 VLM과 비교했을 때, 다양한 Task에서 큰 폭으로 성능 개선

Table 2: **Comparison between different LVLMs on video reasoning benchmarks.** We employ ChatGPT-Assistant to evaluate the performance following Video-ChatGPT (Maaz et al., 2023). The version of ChatGPT is "gpt-3.5-turbo".

Methods	LLM size	MSVD-QA		MSRVTT-QA		TGIF-QA		ActivityNet-QA	
		Accuracy	Score	Accuracy	Score	Accuracy	Score	Accuracy	Score
FrozenBiLM	1B	32.2	-	16.8	-	41.0	-	24.7	-
VideoChat	7B	56.3	2.8	45.0	2.5	34.4	2.3	-	2.2
LLaMA-Adapter	7B	54.9	3.1	43.8	2.7	-	-	34.2	2.7
Video-LLaMA	7B	51.6	2.5	29.6	1.8	-	-	12.4	1.1
Video-ChatGPT	7B	64.9	3.3	49.3	2.8	51.4	3.0	35.2	2.7
Chat-UniVi	7B	65.0	3.6	54.6	3.1	60.3	3.4	**45.8**	3.2
Video-LLaVA	**7B**	**70.7**	**3.9**	**59.2**	**3.5**	**70.0**	**4.0**	45.3	**3.3**

Algorithms

- VLM-based Anomaly Detection -

Algorithm

- VLM-based Anomaly Detection (1/4)

❖ MMAD (2025/ICLR – Tencent)

- 다양한 Foundation VLM들의 Anomaly Detection에 대한 가능성 평가

MMAD: A COMPREHENSIVE BENCHMARK FOR MULTIMODAL LARGE LANGUAGE MODELS IN INDUSTRIAL ANOMALY DETECTION

Xi Jiang[1] **Jian Li**[2] **Hanqiu Deng**[3] **Yong Liu**[2] **Bin-Bin Gao**[2] **Yifeng Zhou**[2]
Jialin Li[2] **Chengjie Wang**[2,4] **Feng Zheng**[1*]
[1]Southern University of Science and Technology [2]Tencent YouTu Lab
[3]University of Alberta [4]Shanghai Jiao Tong University
jiangx2020@mail.sustech.edu.cn, hanqiu1@ualberta.ca,
{swordli, choasliu, danylgao, joefzhou, jarenli, jasoncjwang}
@tencent.com, f.zheng@ieee.org

[6] Jiang, X., Li, J., Deng, H., Liu, Y., Gao, B. B., Zhou, Y., ... & Zheng, F. (2025) MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection. In ICLR.

Algorithm

- VLM-based Anomaly Detection (1/4)

❖ MMAD (2025/ICLR – Tencent)

- 전체적으로 GPT-4o가 가장 좋은 성능을 보임 (코드가 공개되지 않음)
- 오픈소스들 중에서는 InternVL2 계열이 동일 파라미터 개수 대비 우수한 성능을 보임
- InternVL2-8B 기준 인간과 약 20% 격차 존재 → Foundation VLM은 여전한 한계를 가짐

Model	Scale	Anomaly	Defect				Object		Average
		Discrimination	Classification	Localization	Description	Analysis	Classification	Analysis	
Random Chance	-	50.00	25.00	25.00	25.00	25.00	25.00	25.00	28.57
Human (expert)	-	95.24	75.00	92.31	83.33	94.20	86.11	80.37	86.65
Human (ordinary)	-	86.90	66.25	85.58	71.25	81.52	89.58	69.72	78.69
Claude-3.5-sonnet	-	60.14	60.14	48.81	67.13	79.11	85.19	79.83	68.36
Gemini-1.5-flash	-	58.58	54.70	49.10	66.53	82.24	91.47	79.71	68.90
Gemini-1.5-pro	-	**68.63**	60.12	**58.56**	70.38	82.46	89.20	82.25	73.09
GPT-4o-mini	-	64.33	48.58	38.75	63.68	80.40	88.56	79.74	66.29
GPT-4o	-	**68.63**	**65.80**	55.62	**73.21**	**83.41**	**94.98**	**82.80**	**74.92**
AnomalyGPT	7B	65.57	27.49	27.97	36.86	32.11	29.84	35.82	36.52
Qwen-VL-Chat	7B	53.65	31.33	28.62	41.66	63.99	74.46	67.94	51.66
LLaVA-1.5	7B	51.33	37.04	36.62	50.60	69.79	68.29	69.53	54.74
Cambrian-1*	8B	55.60	32.53	35.39	43.46	49.14	78.15	67.22	51.64
SPHINX*	7B	53.13	33.93	52.27	50.96	71.23	85.07	73.10	59.96
LLaVA-NEXT-Interleave	7B	57.64	33.79	47.72	51.84	67.93	81.39	74.91	59.32
InternLM-XComposer2-VL	7B	55.85	41.80	48.27	57.52	76.60	74.34	77.75	61.73
LLaVA-OneVision	7B	51.77	46.13	41.85	62.19	69.73	90.31	80.93	63.27
MiniCPM-V2.6	8B	57.31	49.22	43.28	65.86	75.24	**92.02**	80.80	66.25
InternVL2	8B	59.97	43.85	47.91	57.60	78.10	74.18	80.37	63.14
LLaVA-1.5	13B	49.96	38.78	46.17	58.17	73.09	73.62	70.98	58.68
LLaVA-NeXT	34B	57.92	48.79	52.87	71.34	80.28	81.12	77.80	67.16
InternVL2	76B	**68.25**	**54.22**	**56.66**	**66.30**	**80.47**	86.40	**82.92**	**70.75**

Algorithm

- VLM-based Anomaly Detection (2/4)

❖ Triad (2025/ICCV)

- Foundation VLM을 Anomaly Detection Task를 잘 수행할 수 있도록 개선
- 입력 이미지 & 텍스트 개선 + Foundation VLM 미세조정 수행

Triad: Empowering LMM-based Anomaly Detection with Expert-guided Region-of-Interest Tokenizer and Manufacturing Process

Yuanze Li[1†] Shihao Yuan[1,2†] Haolin Wang[1] Qizhang Li[1,2]
Ming Liu[1(✉)] Chen Xu[2] Guangming Shi[2] Wangmeng Zuo[1,2]
sqlyz@hit.edu.cn, csshihao@outlook.com, why_cs@outlook.com, csqizhang@gmail.com
csmliu@outlook.com, xc.xc@qq.com, gmshi@xidian.edu.cn, wmzuo@hit.edu.cn
[1]Harbin Institute of Technology, [2]Pengcheng Lab, Guangzhou

[7] Li, Y., Yuan, S., Wang, H., Li, Q., Liu, M., Xu, C., ... & Zuo, W. (2025). Triad: Empowering LMM-based Anomaly Detection with Expert-guided Region-of-Interest Tokenizer and Manufacturing Process. In ICCV.

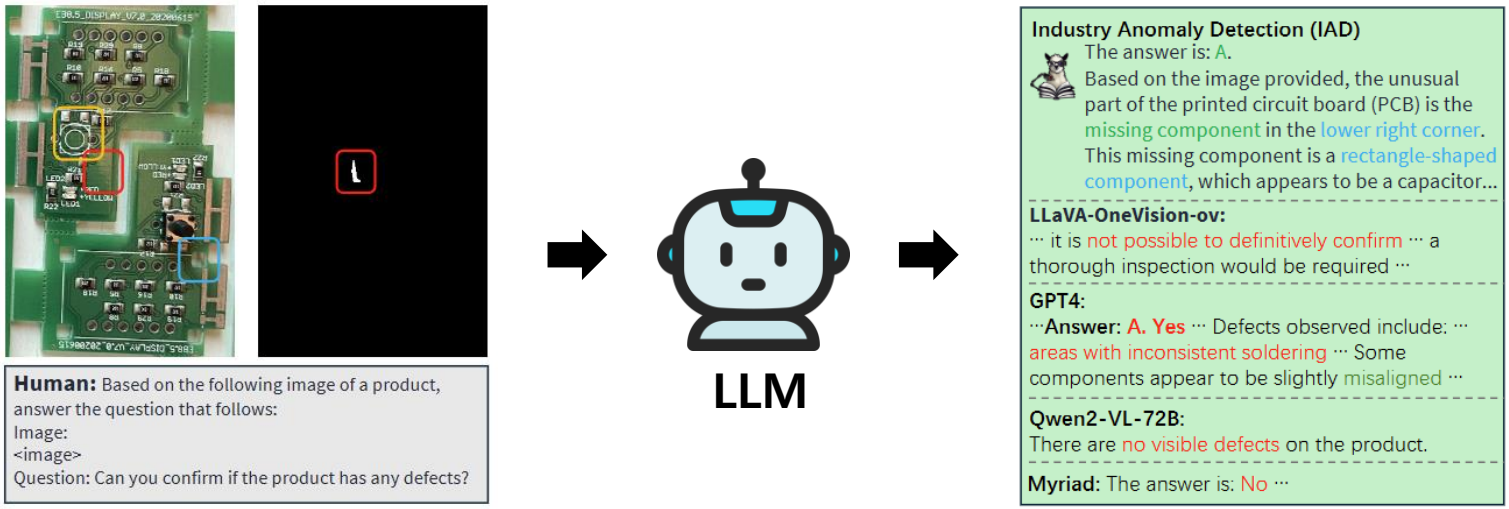
Algorithm

- VLM-based Anomaly Detection (2/4)

❖ Triad (2025/ICCV)

① 이미지에 대한 이해 부족 → 비정상은 작은 영역에서 발생하나, 미세 영역을 잘 고려하지 못함

② 텍스트에 대한 이해 부족 → 특수한 도메인인 제조 공정에 대한 깊은 이해 부족

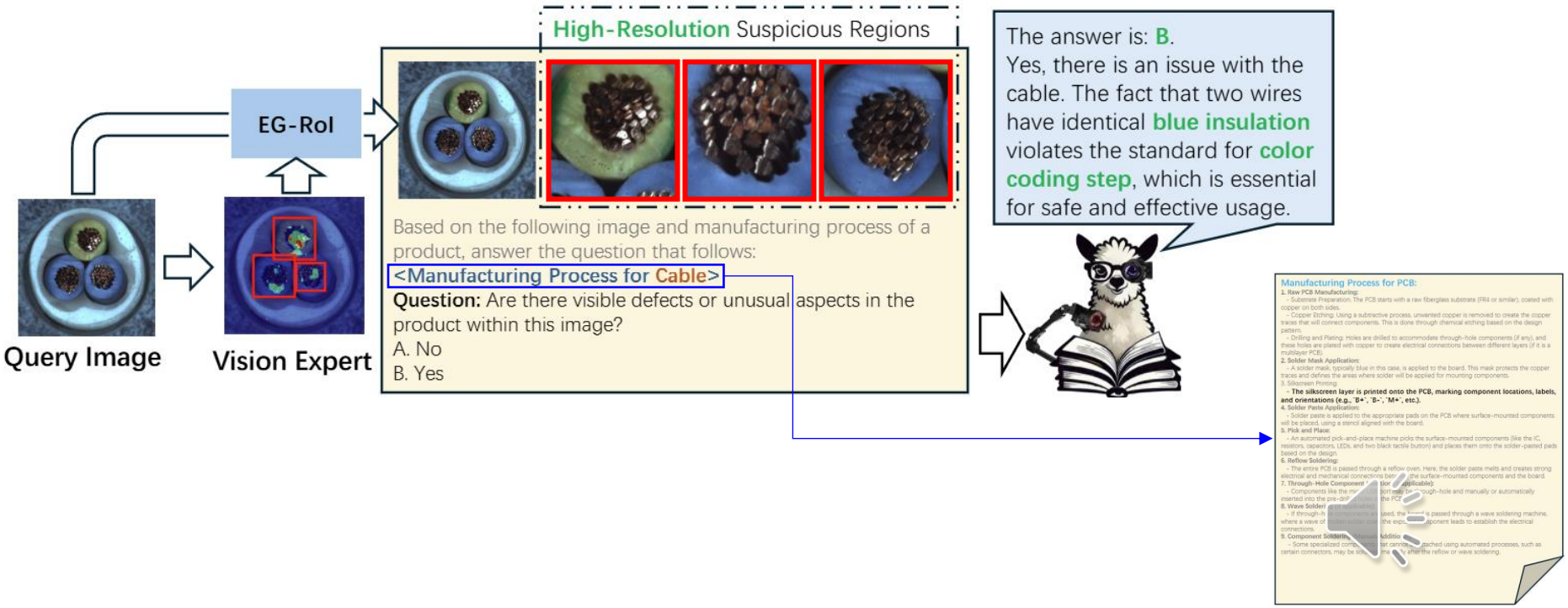


Algorithm

- VLM-based Anomaly Detection (2/4)

❖ Triad (2025/ICCV)

- 이미지 이해력 향상 → 원본 이미지와 비정상이 의심되는 영역을 Crop하여 함께 고려
- 텍스트 이해력 향상 → 제조 공정에 대한 정보를 텍스트로 추가 제공



Algorithm

- VLM-based Anomaly Detection (2/4)

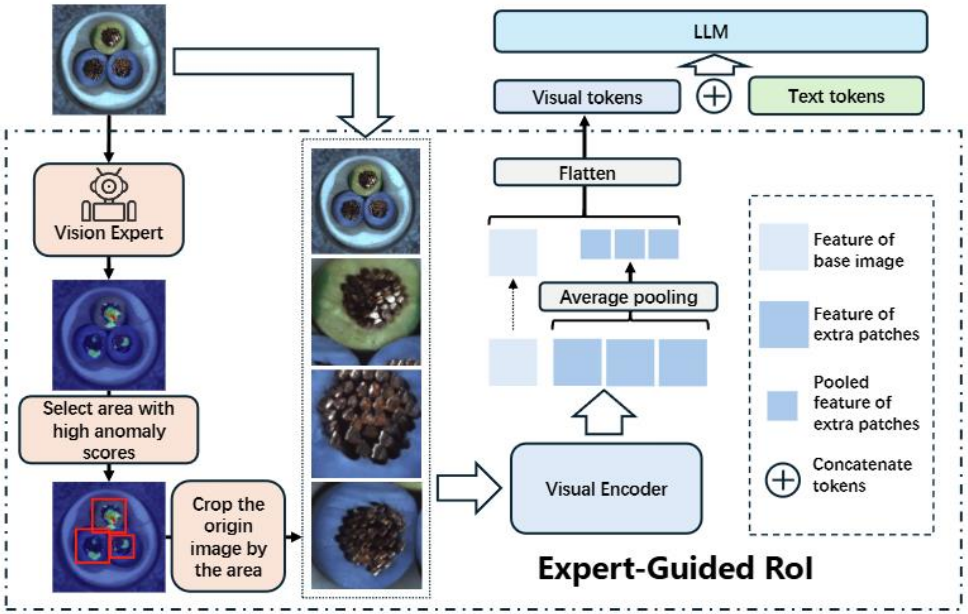
❖ **Triad (2025/ICCV) – Anomaly 미세 영역에 대한 정보 부족 → 이미지에 대한 이해 개선**

① Anomaly Map을 알 수 있는 Vision Expert를 기반으로 Anomaly Region 식별

② Anomaly Score가 높은 영역들을 포함하는 픽셀에 대해 3개 Bounding Box로 Crop

③ 원본 이미지를 포함하여 4개 이미지를 Visual Encoder에 입력

➢ 원본 이미지 Feature와 Crop된 이미지 Feature 모두 합쳐서 활용



Algorithm

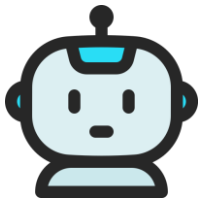
- VLM-based Anomaly Detection (2/4)

❖ **Triad (2025/ICCV) – 제조 도메인에 정보 및 사고력 부족 → 텍스트에 대한 이해 개선**

- **텍스트 프롬프트에 제조 공정 정보 추가:** 질문 + 제조공정 프로세스(MPG) 설명 추가
- **이상 상황 레이블 구체화:** Chain of Thought (CoT) 형태로 이상 상황에 대해 구체화

다음 제품의 제조 공정을 설명해줘

제품명: Cable
캡션 (사람 입력): 세 개의 전선이 각각 파란색/갈색/초록색 피복을 가지며, 회색 외피로 감싸져 있음



LLM

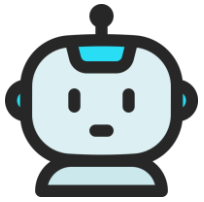
(1) 전선 제작: 구리선을 꼬아서 휘어짐 없이 전선 제작
(2) 색상 코딩: 세 전선을 각각 다른 색(파란색/갈색/초록색)으로 피복
(3) 외부 피복: 세 전선을 회색 외피로 감쌈

[MPG]

해당 이미지는 비정상이야. 아래 정보를 바탕으로 CoT 형태 답변을 만들어줘.

캡션 (사람 입력): 파란색 피복을 가진 전선이 두 개 존재함. 왼쪽 하단 전선의 구리 가닥이 중심축에서 30도 휘어져 있음. 회색 외피로 감싸져 있음.

MPG:
(1) 전선 제작: 구리선을 꼬아서 휘어짐 없이 전선 제작
(2) 색상 코딩: 세 전선을 각각 다른 색(파란색/갈색/초록색)으로 피복
(3) 외부 피복: 세 전선을 회색 외피로 감쌈



LLM

예측 결과: 비정상

(1) 전선 제작 확인: 왼쪽 하단 전선의 30도 휘어짐 → 비정상
(2) 색상 코딩 확인: 파란색 피복 전선이 두 개 → 비정상
(3) 외부 피복 확인: 회색 외피로 감싸져 있음 → 정상

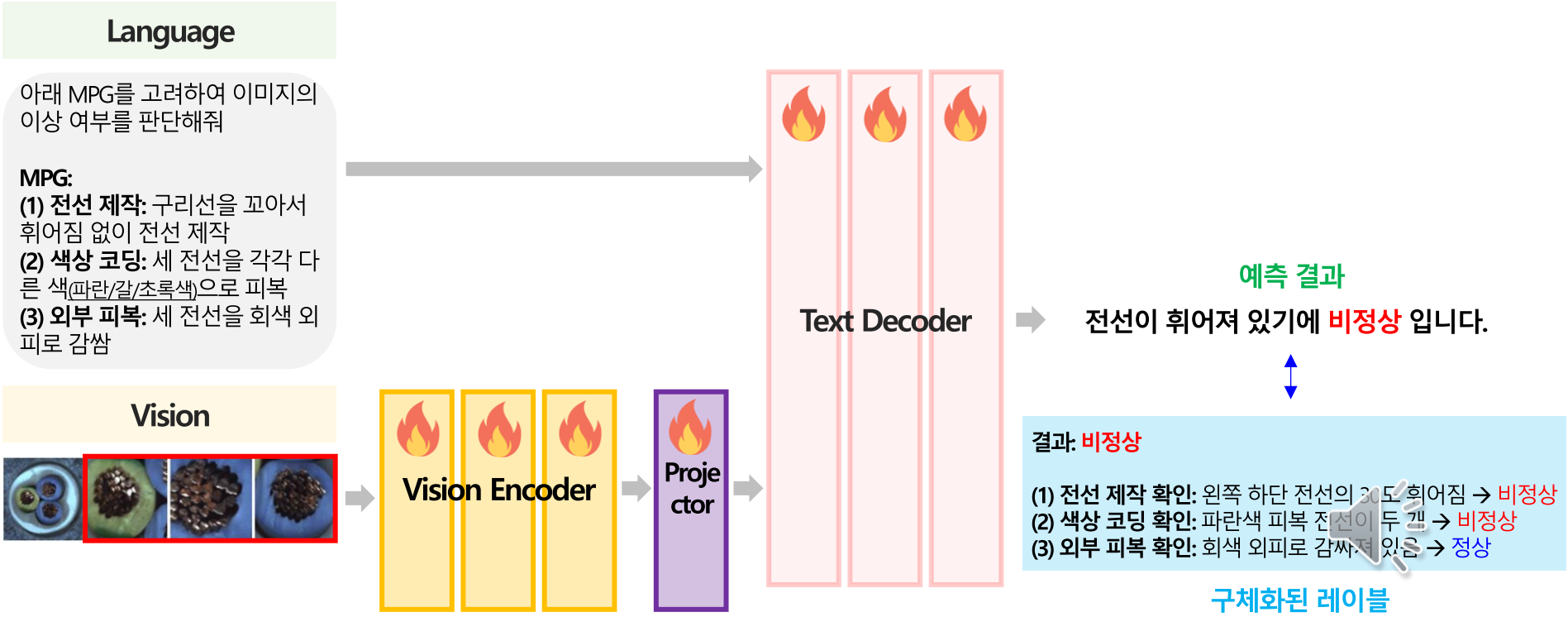
[구체화된 이상 레이블]

Algorithm

- VLM-based Anomaly Detection (2/4)

❖ Triad (2025/ICCV) – 모델 미세조정

- 고도화된 이미지와 텍스트 정보를 기반으로 모델 전체 미세조정



Algorithm

- VLM-based Anomaly Detection (2/4)

❖ Triad (2025/ICCV) – 실험결과

- **비교분석:** Vanilla Foundation VLM 대비 성능 크게 개선
- **추론 시 MFG 활용효과:** 미세조정 시에는 효과적이나, 그렇지 않으면 효과적이지 않음
  - ➢ 미세조정 없이 제조공정 정보를 주는 것은 오히려 모델에 혼란을 부여할 수 있음

Model	Params	MVTec-AD		WFDD	
		0-shot	+ MFG Proc.	0-shot	+ MFG Proc.
GPT-4o [11]	-	82.2%	67.9% (14.3%↓)	78.5%	77.3% (1.2%↓)
Qwen2-VL [31]	2B	77.0%	46.7%(30.3%↓)	70.6%	45.2%(25.4%↓)
LLava-1.6 [22]	7B	76.9%	75.9% (1.0%↓)	63.8%	64.0% (0.2%↓)
MiniCPM-V [34]	8B	62.3%	51.6%(10.7%↓)	70.3%	52.1%(18.2%↓)
LLaVA-OneVision-si [16]	7B	77.7%	60.6%(17.1%↓)	65.2%	61.4% (3.8%↓)
LLaVA-OneVision-ov [16]	7B	91.0%	80.8%(10.2%↓)	79.8%	80.3% (0.5%↑)
Qwen2-VL [31]	7B	84.4%	61.1%(23.3%↓)	74.4%	61.4%(13.0%↓)
Qwen2-VL [31]	72B	87.1%	79.5% (7.6%↓)	**81.1%**	74.2% (6.9%↓)
LLaVA-OneVision-ov [16]	72B	87.3%	75.5%(11.8%↓)	75.0%	74.6% (0.4%↓)
Myriad [19]	7B	79.3%	81.5% (2.5%↑)	60.5%	61.7% (1.2%↑)
Triad-llava-1.6	7B	85.0%	87.5% (2.5%↑)	67.3%	69.9% (2.6%↑)
Triad-ov	7B	**91.2%**	**92.6%** (1.4%↑)	80.2%	**81.1%** (0.9%↑)

Algorithm

- VLM-based Anomaly Detection (2/4)

❖ Triad (2025/ICCV) – 실험결과

- 미세조정 → 1.9% 향상
- + CoT 기반 레이블 활용 → 1% 추가 향상
- **+ 이미지를 Crop하여 활용 → 5.2% 향상**
  - ➢ 이미지 Crop 영역을 탐지 모델 선정 중요 → 잘못 선정 시 Random보다 저하 (April-GAN)

InstructIAD	CoT-M	EG-RoI	0-shot
✗	✗	✗	76.9%
✓	✗	✗	78.8%
✓	✓	✗	79.8%
✓	✗	✓	85.4%
✓	✓	✓	85.0%

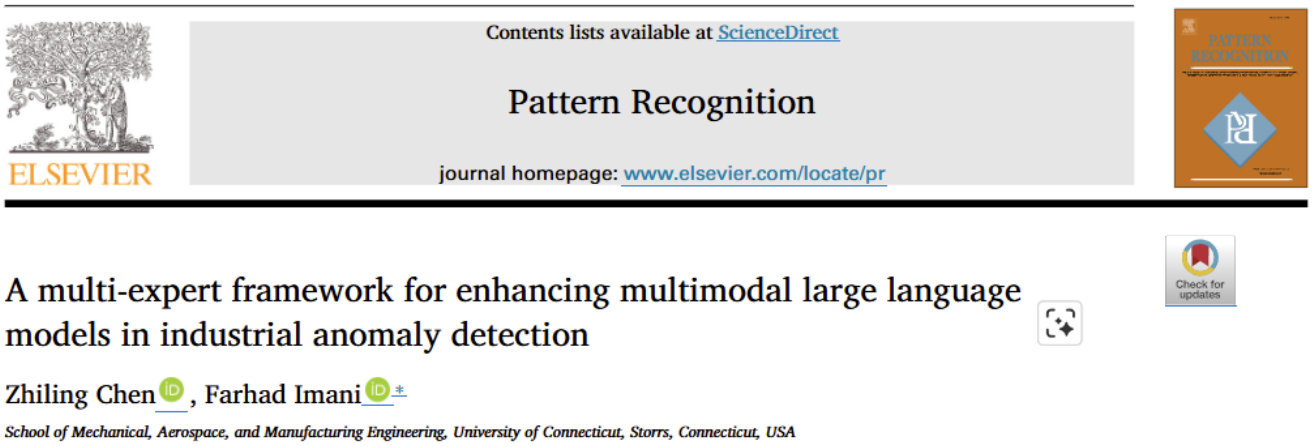
Vision Expert	Expert P-AUROC	base
Null	-	83.3%
AnyRes	-	84.0%
April-GAN [7]	87.6%	83.0%
AnomalyClip [40]	91.1%	84.6%
MuSc [18]	97.3%	85.0%

Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR)

- 다양한 전문가 모델들을 활용하여 이상 탐지에 대한 VLM 입력정보 개선



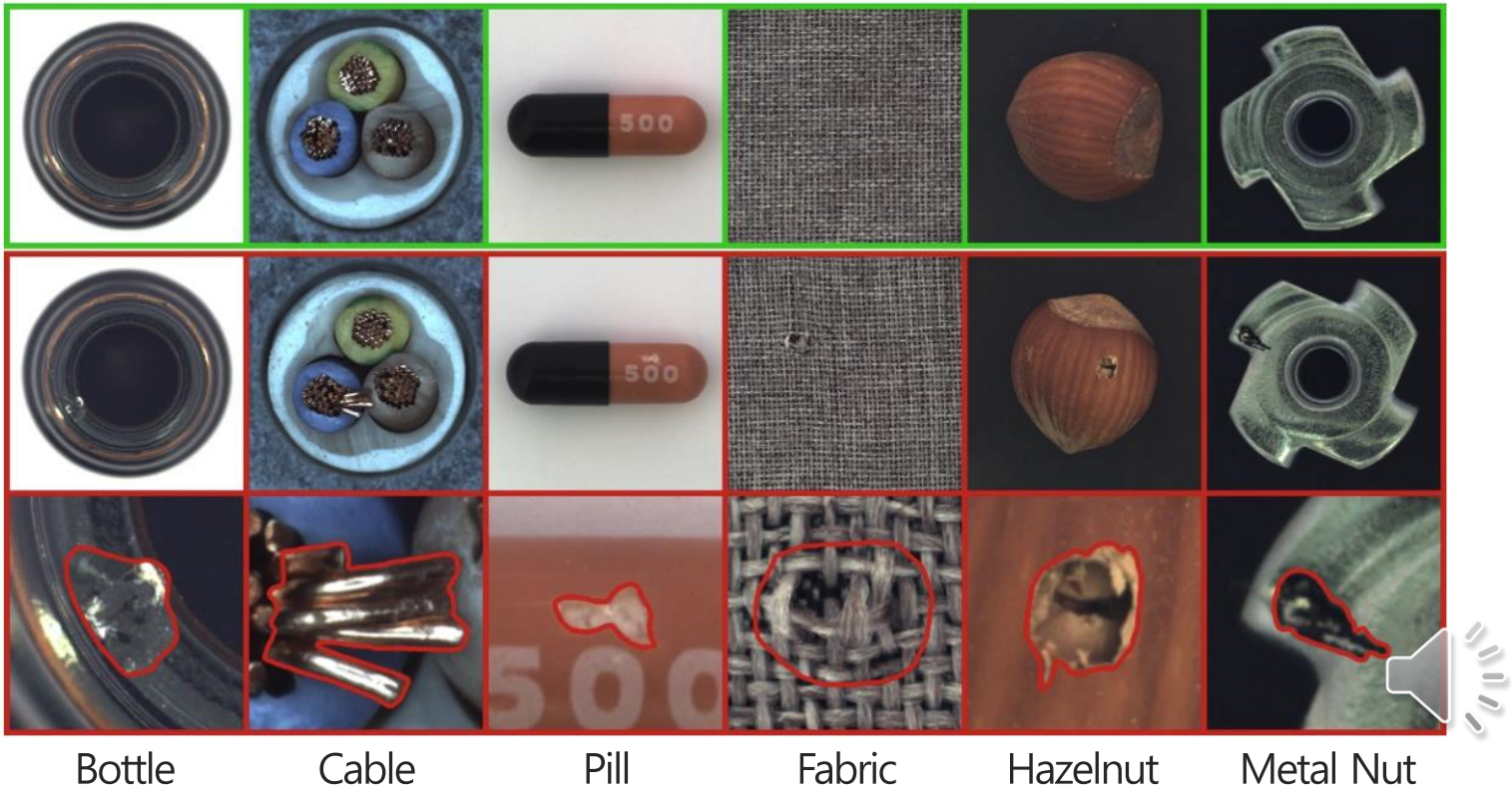
[8] Chen, Z., & Imani, F. (2026). A multi-expert framework for enhancing multimodal large language models in industrial anomaly detection. Pattern Recognition, 112752.

Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR)

- VLM은 이상 탐지에서 큰 성능 향상을 이룸 → 그러나, 여전히 만족할만한 수준은 아님
- 산업 현장에 대한 이미지와 텍스트 수준 지식이 충분하지 않음 → 주입 필요

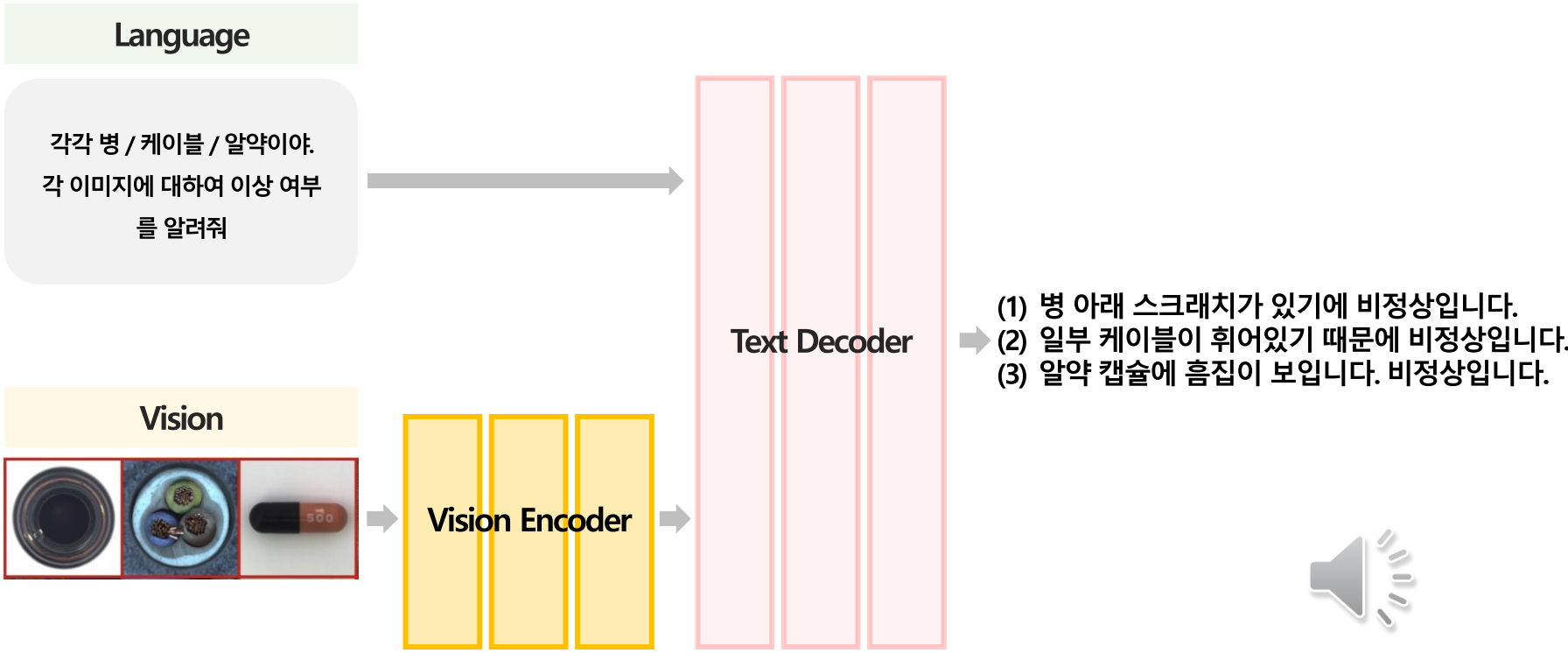


Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR)

- Training-free로 접근 → 외부 가이드 없이 프롬프팅만으로 개선하는 것은 한계를 가짐

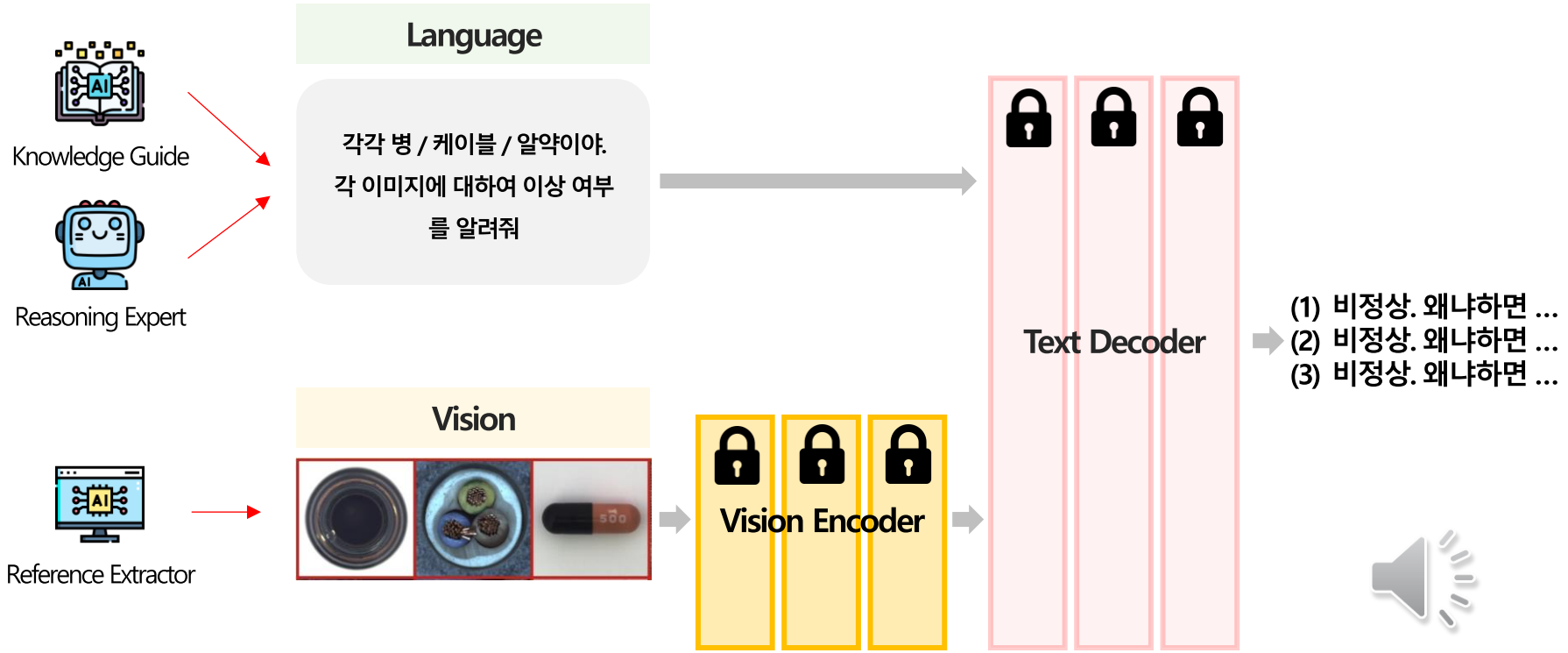


Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR)

- **목표:** VLM에 외부 지식을 부여하여, Training-free로 접근해보자.
- 이때, 외부 지식을 주입하기 위해, 3개의 전문가를 함께 활용

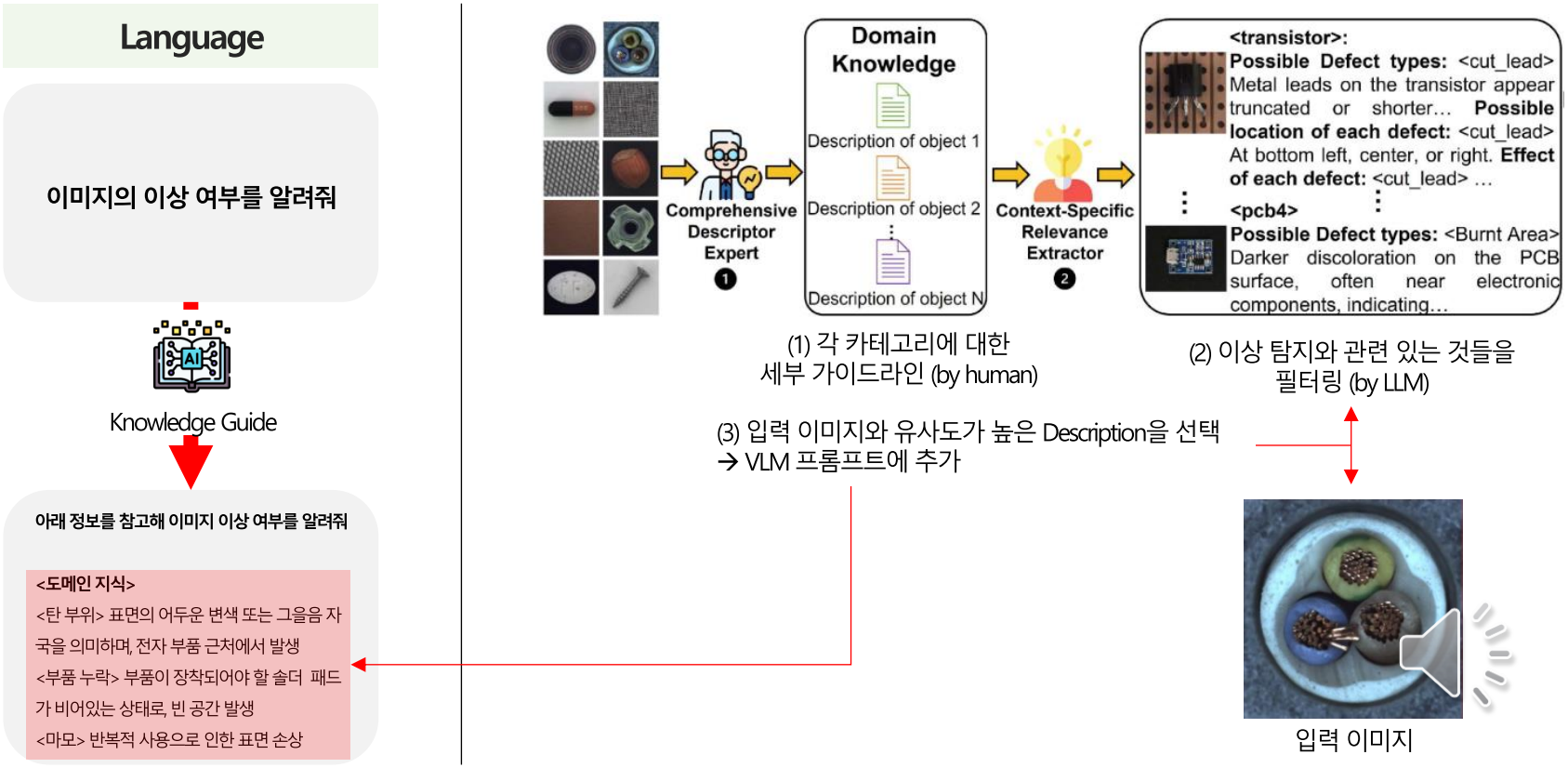


Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR) – (1) Knowledge Guide

- **역할:** 텍스트 수준의 외부 지식을 주입할 수 있는 전문가
- 객체에 대한 외부 지식 정보를 저장 → LLM으로 필터링 → 이미지와 유사한 정보만 Retrieval

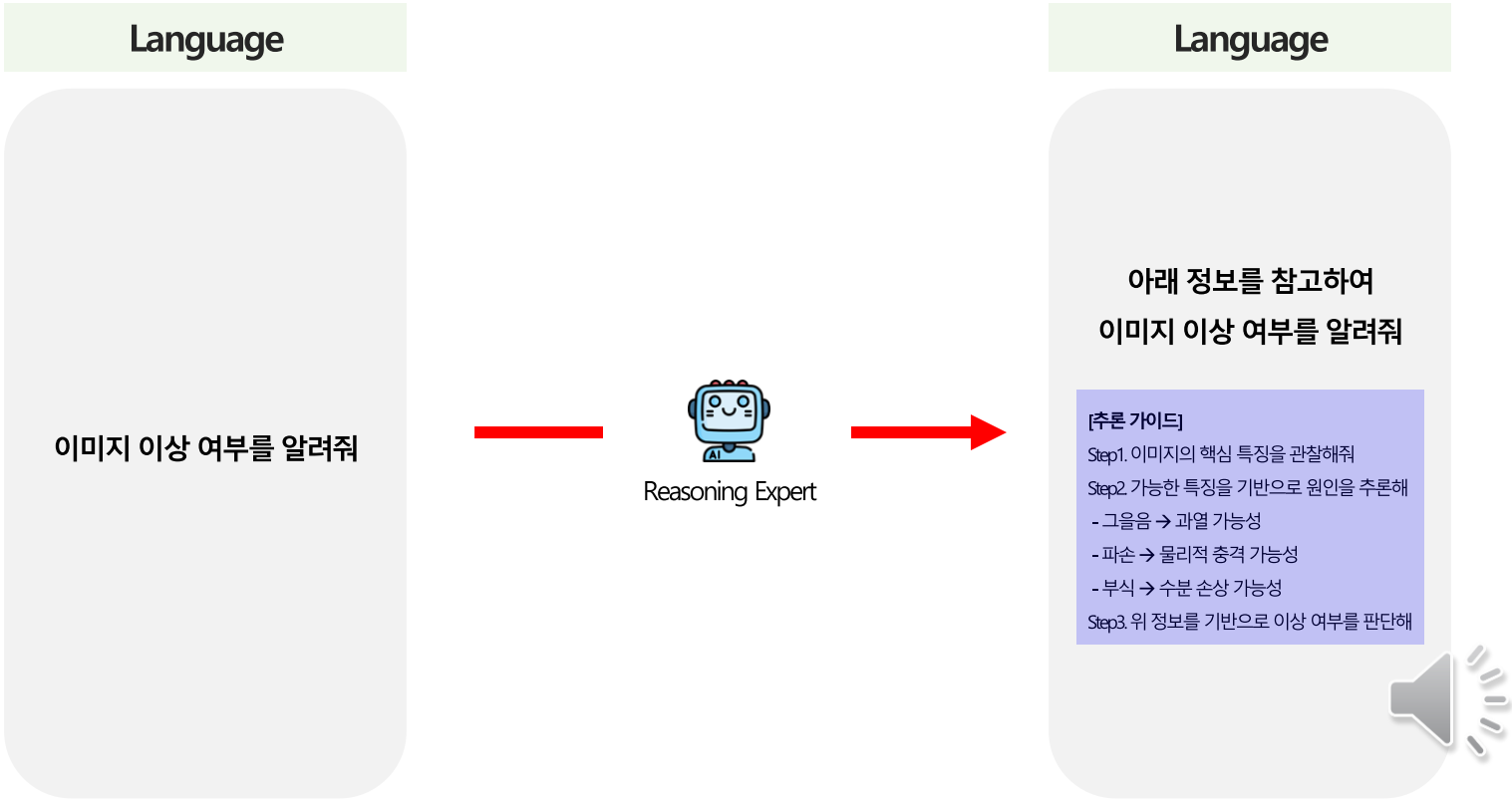


Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR) – (2) Reasoning Expert

- **목표:** 한 번에 추론하는 것이 아닌, Step-by-Step 추론 (CoT)을 유도하는 전문가
- 인간처럼 문제를 순차적으로 해결하는 방향을 알려주어, 문제를 쉽게 해결할 수 있도록 유도

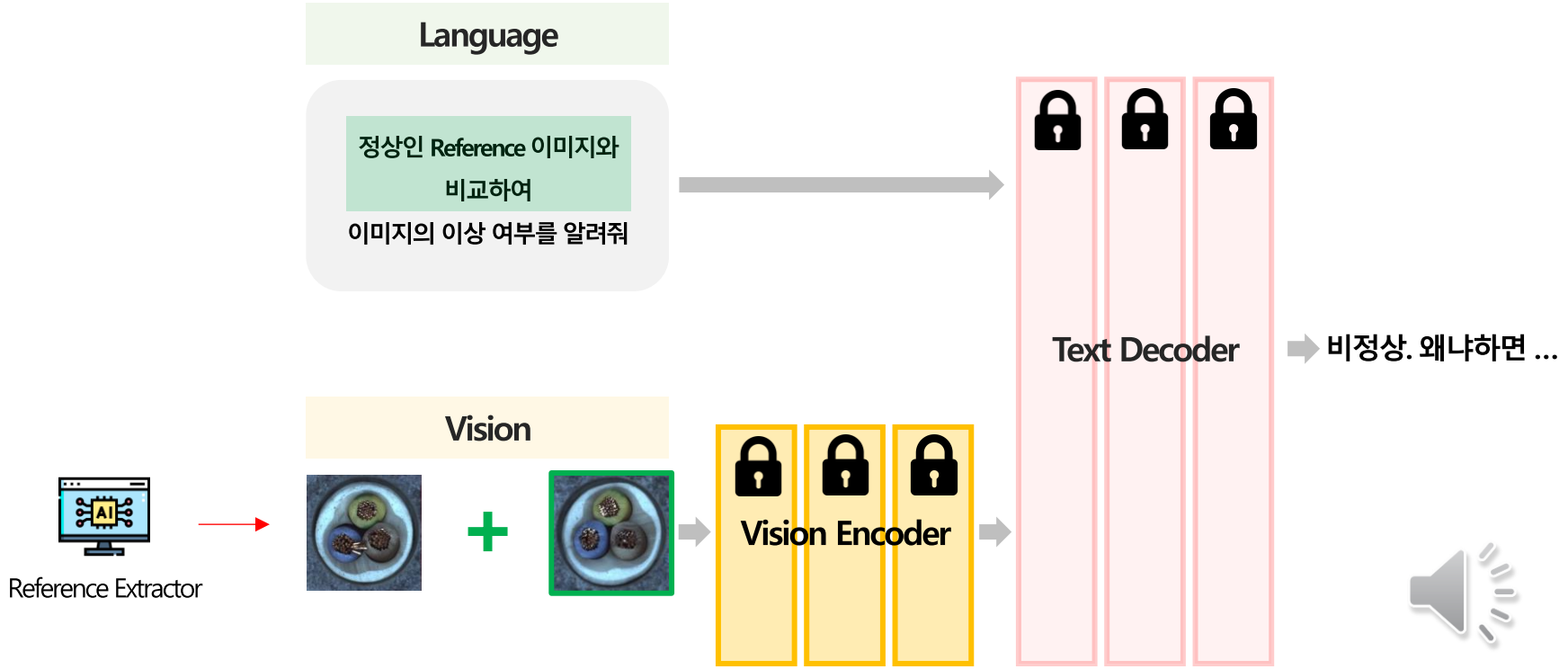


Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR) – (3) Reference Extractor

- **목표:** 입력 이미지 외에 힌트를 얻을 수 있는 Reference 이미지를 함께 활용하는 전문가
- 정상 샘플에 대한 이미지 함께 활용 → VLM이 특정 샘플과 대조하여 이상여부 판단 가능

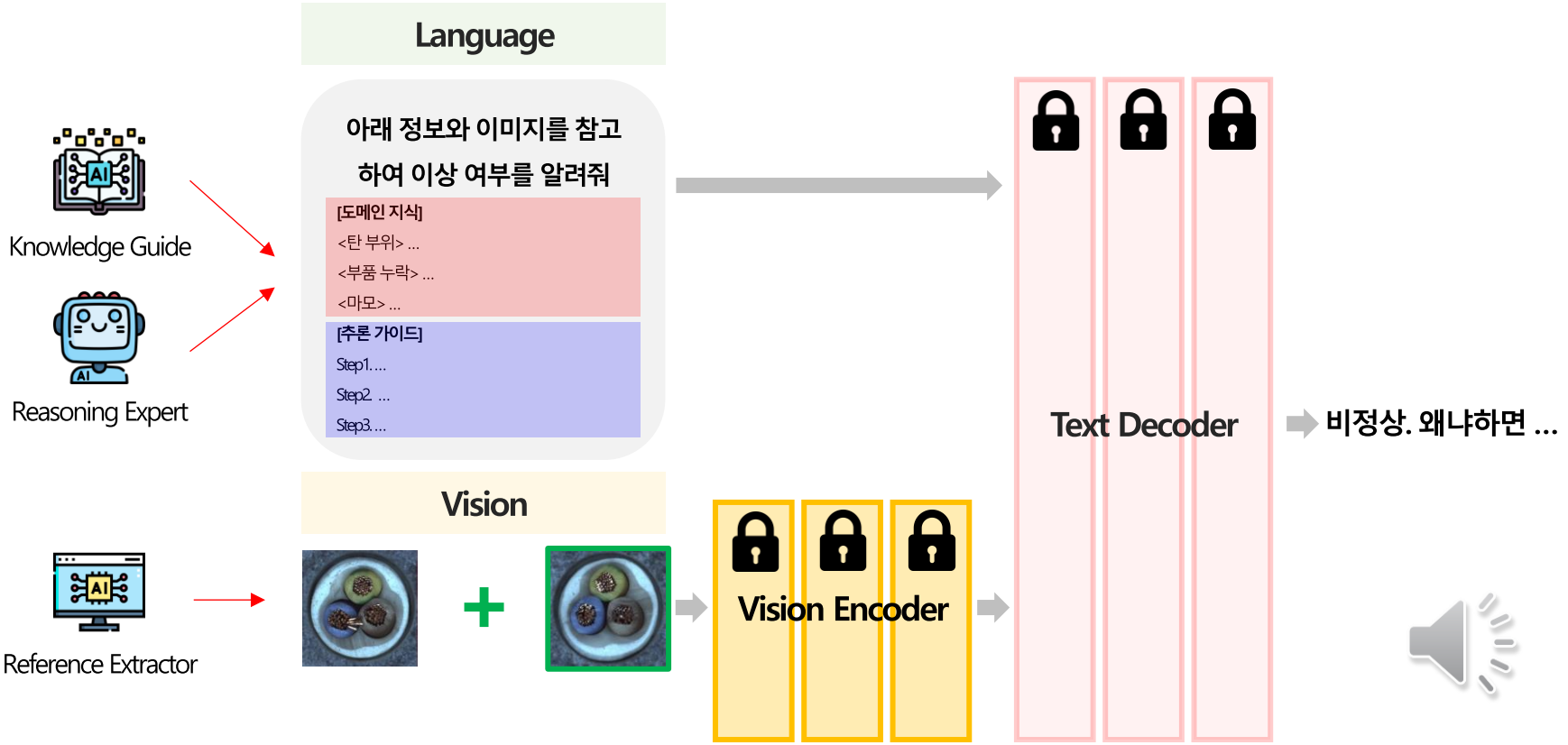


Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR) – (4) 활용

- 전문가들의 지식을 합쳐서 VLM 성능 개선

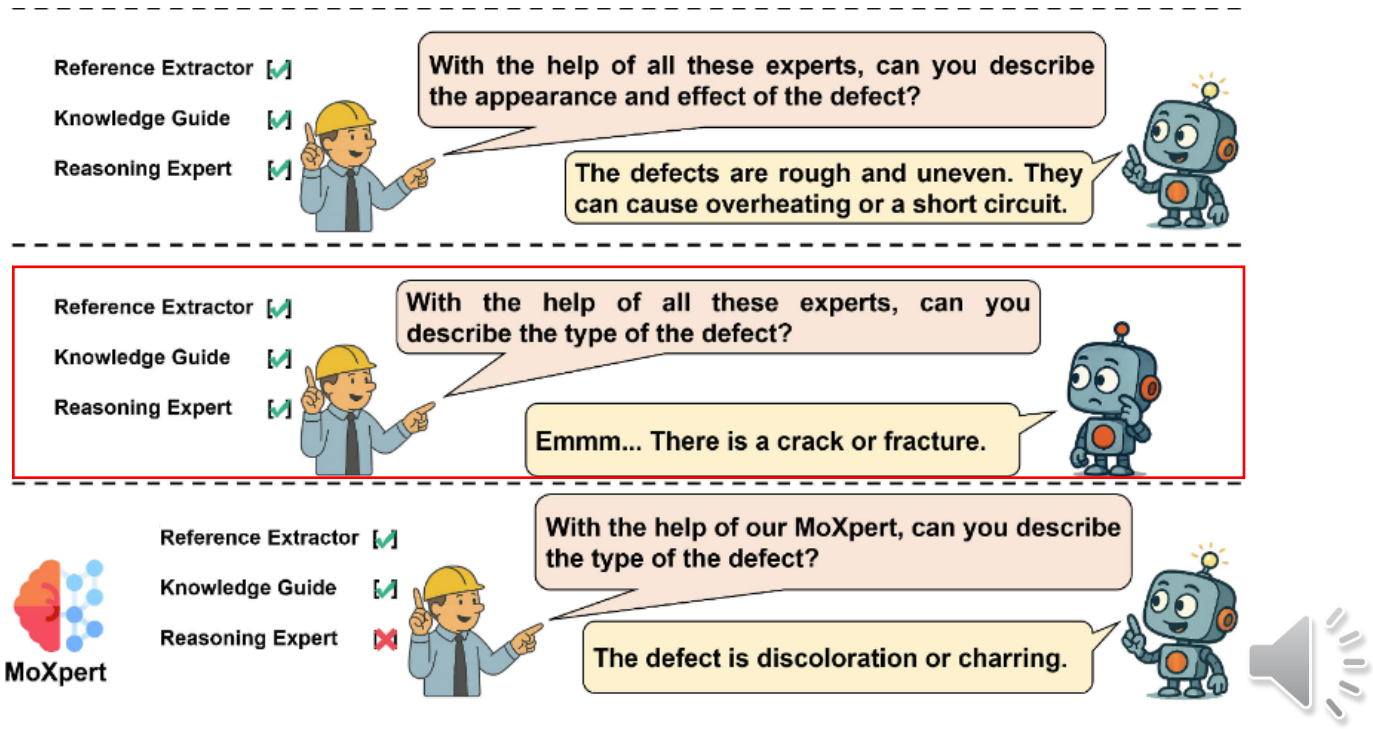


Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR) – (5) 전문가 선정

- **문제상황:** 정보량이 과도하면 VLM이 혼란을 겪을 수 있음 → 핵심 정보에 집중이 어려움
- 문제상황에 특정 전문가를 선택하는 Router 네트워크를 추가 학습

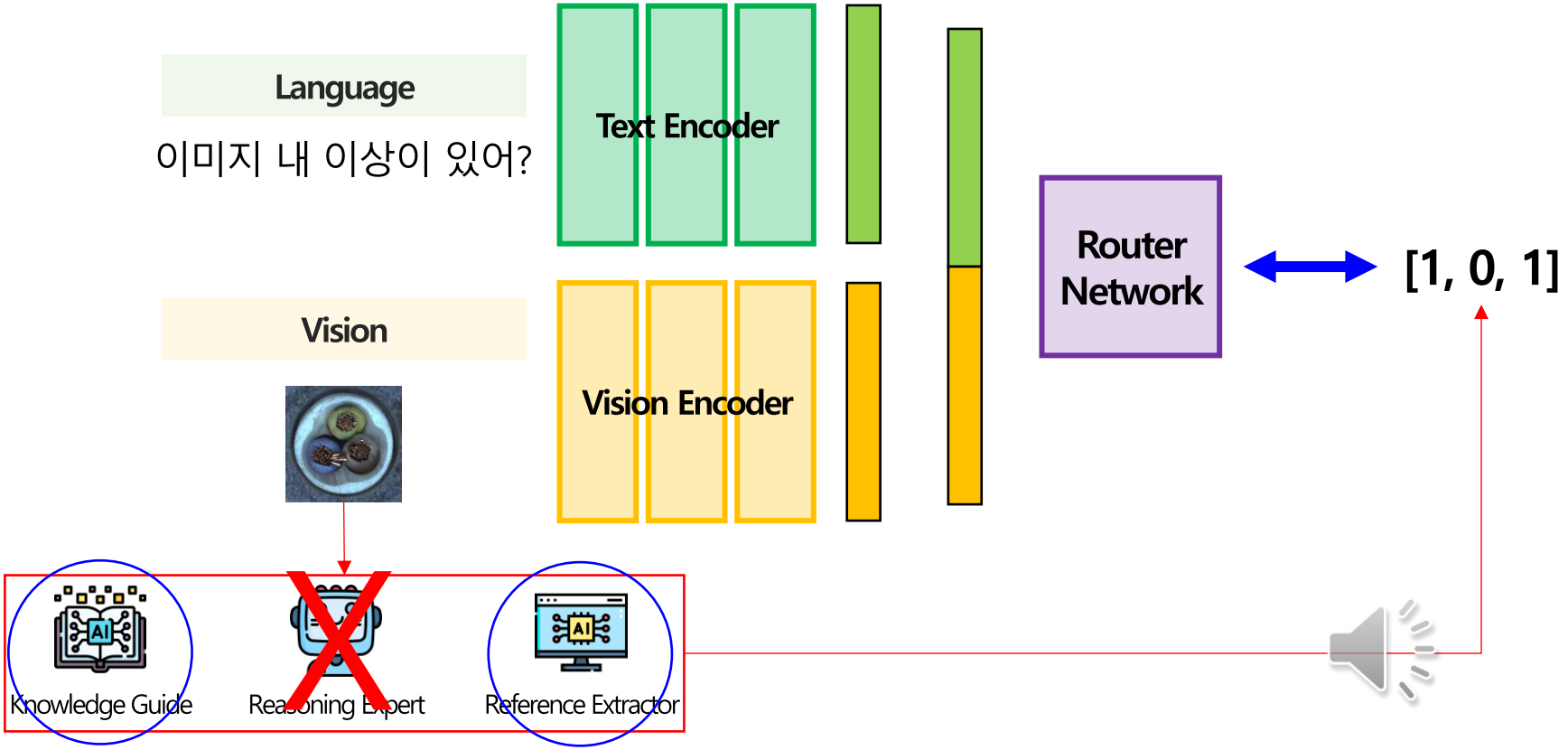


Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR) – (5) 전문가 선정

- (이미지, 질문)에 대해 인간이 직접 전문가들의 최적 조합 레이블링 (ex. [1, 0, 1] → 8개 中 Best)
- X: {이미지 Feature + 질문에 대한 텍스트 Feature} ↔ Y: 전문가들의 최적 조합



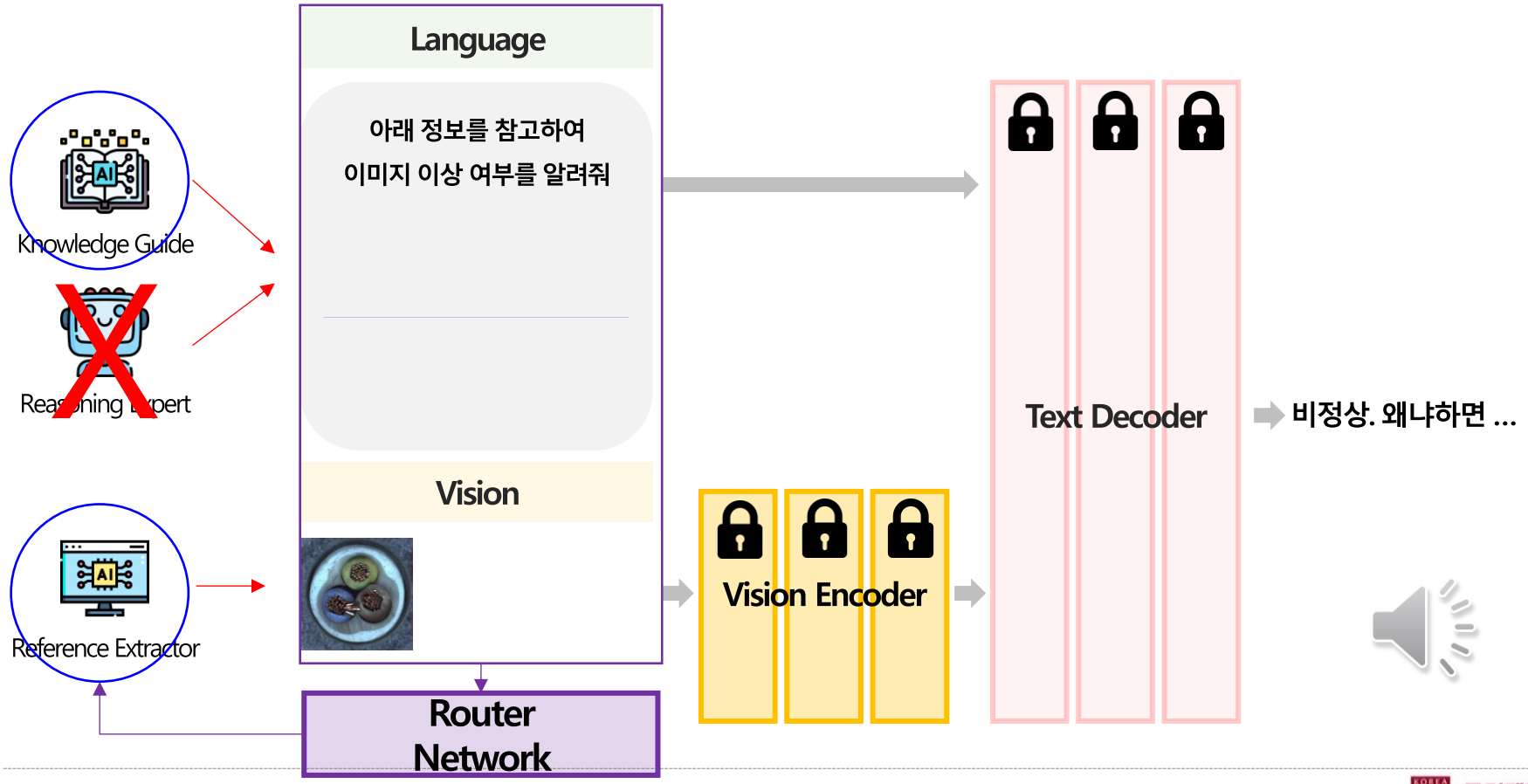
Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR) – (6) 최종 파이프 라인

① 주어진 텍스트 & 이미지를 기반으로 Router Network에서 전문가 선정

② 전문가를 기반으로 입력 텍스트 & 이미지 보완 → 예측 수행



Algorithm

- VLM-based Anomaly Detection (3/4)

❖ MoXpert (2026/PR) – 실험결과

- 벤치마크 데이터에 대해 이상 탐지 성능 크게 향상

Table 1
Accuracy score (%) on MVTec-AD for five tasks: Anomaly Discrimination, Defect Classification, Defect Localization, Defect Description, and Defect Analysis, with dataset-specific average.

Model	Scale	MVTec-AD					Average
		Discrimination	Classification	Localization	Description	Analysis	
Random Chance	-	50.00	25.00	25.00	25.00	25.00	30.00
GPT-4o	-	77.52	**84.94**	**88.55**	**92.17**	**95.18**	**87.67**
GPT-4o-mini	-	74.22	71.78	62.62	79.72	90.62	75.79
Gemini-2-flash	-	83.80	72.37	73.18	77.82	90.37	79.49
Gemini-2-flash-lite	-	82.79	70.87	69.57	77.25	89.46	78.02
AnomalyGPT	7B	82.84	27.80	28.33	34.62	34.36	54.78
InternVL2	4B	70.96	44.81	66.97	59.52	87.39	65.93
InternVL2	8B	76.88	51.04	59.18	64.55	85.73	67.48
MiniCPM-V2.6	8B	72.50	64.07	68.65	79.55	90.04	74.96
LLaVA-NeXT	7B	78.42	45.23	64.96	68.18	87.14	68.79
LLaVA-OneVision	7B	**94.09**	79.59	78.12	83.18	91.29	85.25
Qwen2-VL	2B	73.21	60.08	65.46	74.86	88.63	72.45
Qwen2-VL	7B	82.26	68.46	76.11	82.19	92.28	80.26
Qwen2-VL (+MoXpert)	7B	89.65 (+7.39)	72.86 (+4.40)	76.11 (=)	85.57 (+3.38)	93.20 (+0.92)	83.48 (+3.22)

Table 2
Accuracy score (%) on VisA for five tasks: Anomaly Discrimination, Defect Classification, Defect Localization, Defect Description, and Defect Analysis, with dataset-specific average.

Model	Scale	VisA					Average
		Discrimination	Classification	Localization	Description	Analysis	
Random Chance	-	50.00	25.00	25.00	25.00	25.00	30.00
GPT-4o	-	67.16	61.51	59.82	70.59	76.10	67.04
GPT-4o-mini	-	69.13	59.24	60.15	70.08	75.15	66.75
Gemini-2-flash	-	74.45	58.49	50.79	67.56	77.04	65.67
Gemini-2-flash-lite	-	69.87	54.71	52.38	68.99	70.42	63.27
AnomalyGPT	7B	74.88	27.23	28.91	38.66	32.42	40.42
InternVL2	4B	63.29	20.17	53.71	58.15	70.25	53.11
InternVL2	8B	68.85	35.04	55.81	59.24	75.24	58.84
MiniCPM-V2.6	8B	64.41	50.67	57.73	69.50	68.44	62.15
LLaVA-NeXT	7B	56.98	40.08	58.90	62.86	68.87	57.54
LLaVA-OneVision	7B	76.46	52.44	60.40	68.49	75.92	66.74
Qwen2-VL	2B	59.79	37.31	58.15	66.81	68.53	58.12
Qwen2-VL	7B	73.00	57.06	**63.41**	**74.71**	77.82	69.20
Qwen2-VL (+MoXpert)	7B	**76.79** (+6.79)	**62.35** (+5.29)	**63.41** (=)	74.03 (-0.68)	**79.11** (+1.29)	**71.14** (+1.94)

Algorithm

- VLM-based Anomaly Detection (3/4)

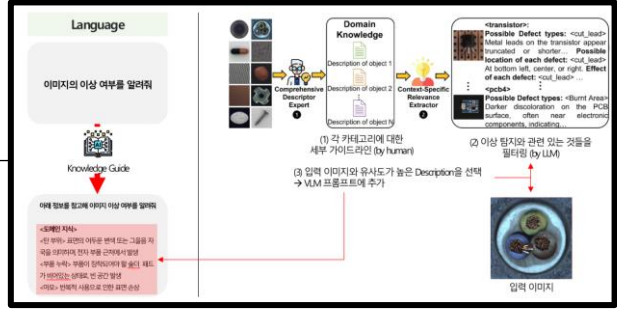
❖ MoXpert (2026/PR) – 실험결과

- **모델 크기 별 Reference Extraction 효과:** 7B급 똑똑한 모델일 때 우수한 성능을 보임
  - ➢ 2B급 모델에서는 Reference 이미지를 잘 구분하지 못함 → 오히려 노이즈로 작용

Model	Setting	Defect Classification	Defect Description	Mean
Qwen2-VL 2B	0 shot	**48.70**	**70.84**	**66.16**
	1 shot	46.95 (*–1.75*)	70.11 (*–0.73*)	65.31 (*–0.85*)
	1 shot *	46.87 (*–1.83*)	70.07 (*–0.77*)	65.34 (*–0.82*)
Qwen2-VL 7B	0 shot	62.76	78.45	75.97
	1 shot	67.12 (*+4.36*)	79.18 (*+0.73*)	78.79 (*+2.82*)
	1 shot *	**67.61** (*+4.85*)	**79.80** (*+1.35*)	**79.20** (*+3.23*)

Algorithm

- VLM-based Anomaly Detection (3/4)



❖ MoXpert (2026/PR) – 실험결과

- **Knowledge Guide 효과:** 추가적인 지식이 활용될 때 우수한 성능을 보임
  - ➢ 특히, 정제된 지식을 활용 시 성능이 더욱 향상

Model	Method	Defect Classification	Defect Description	Mean
Qwen2-VL 2B	w/o EK	44.82	67.21	56.02
	+ DK	45.10 (+0.28)	**70.70** (+3.49)	57.90 (+1.88)
	+ CSK	**46.87** (+2.05)	70.07 (+2.86)	**58.47** (+2.45)
Qwen2-VL 7B	w/o EK	63.55	78.75	71.15
	+ DK	65.98 (+2.43)	79.31 (+0.56)	72.65 (+1.50)
	+ CSK	**67.61** (+4.06)	**79.80** (+1.05)	**73.71** (+2.56)

Algorithm

- VLM-based Anomaly Detection (4/4)

❖ VERA (2025/CVPR)

- Training 없이 텍스트 질문을 강화하여 VLM 기반 비디오 이상 탐지 성능 개선

VERA: Explainable Video Anomaly Detection via Verbalized Learning of Vision-Language Models

Muchao Ye[1*] Weiyang Liu[2] Pan He[3]
[1]The University of Iowa [2]Max Planck Institute for Intelligent Systems, Tübingen [3]Auburn University
[1]muye@uiowa.edu [2]weiyang.liu@tuebingen.mpg.de [3]pan.he@auburn.edu *Corresponding Author
https://vera-framework.github.io

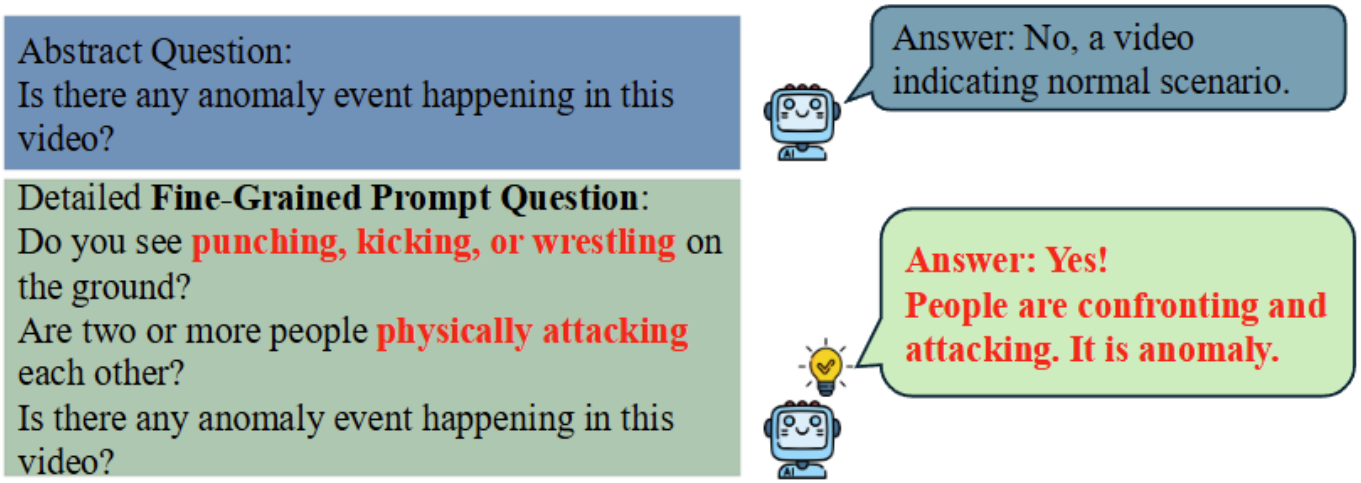
[9] Ye, M., Liu, W., & He, P. (2025). Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In CVPR.

Algorithm

- VLM-based Anomaly Detection (4/4)

❖ VERA (2025/CVPR)

- 단순한 질문만으로는 VLM 기반 비디오 이상 탐지는 어려움
- **Motivation:** 구체적인 질문을 한다면 VLM 성능을 끌어올릴 수 있지 않을까?

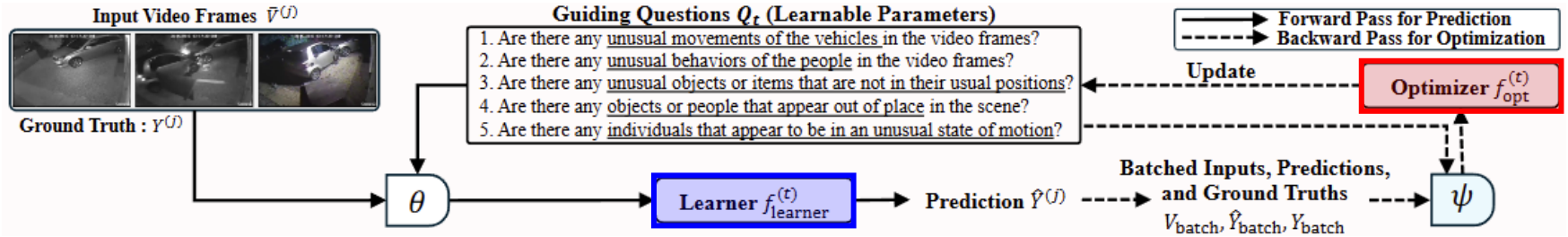


Algorithm

- VLM-based Anomaly Detection (4/4)

❖ VERA (2025/CVPR)

- Learner와 Optimizer라는 VLM Agent를 정의하여 최적 질문을 산출 (파라미터 튜닝X)
- Validation 데이터로 최적 질문 산출 → 해당 질문을 기반으로 Test 데이터를 모두 추론
- **Learner:** 질문에 대하여 각 동영상의 이상치 점수 산출
- **Optimizer:** 기존 질문을 수정하여 개선된 질문을 제안



[입력]
1) 동영상
2) 질문세트
[출력] Anomaly 여부

VS

[입력]
1) 동영상
2) Learner에서 예측한 Anomaly Score
3) Ground Truth
[출력] 새로운 질문 세트

질문을 개선해도 Valid Accuracy가 N번 개선되지 않으면 Stop

Algorithm

- VLM-based Anomaly Detection (4/4)

❖ VERA (2025/CVPR) - 실험결과

- 기존 방법론들 대비 성능이 크게 향상
- 인간이 직접 쓴 질문과 비교했을 때도 5%이상 성능이 크게 향상

Method	AUC
Non-Explainable VAD Methods	
Hasan et al. [13]	50.32
Lu et al. [25]	53.56
BODS [38]	57.32
GODS [38]	61.56
RareAnom [34]	**68.33**
Explainable VAD Methods	
LAVAD [52]	85.36
ZS CLIP [52]	38.21
ZS IMAGEBIND-I [52]	58.81
ZS IMAGEBIND-V [52]	55.06
LLAVA-1.5 [20]	79.62
VERA	**88.26**

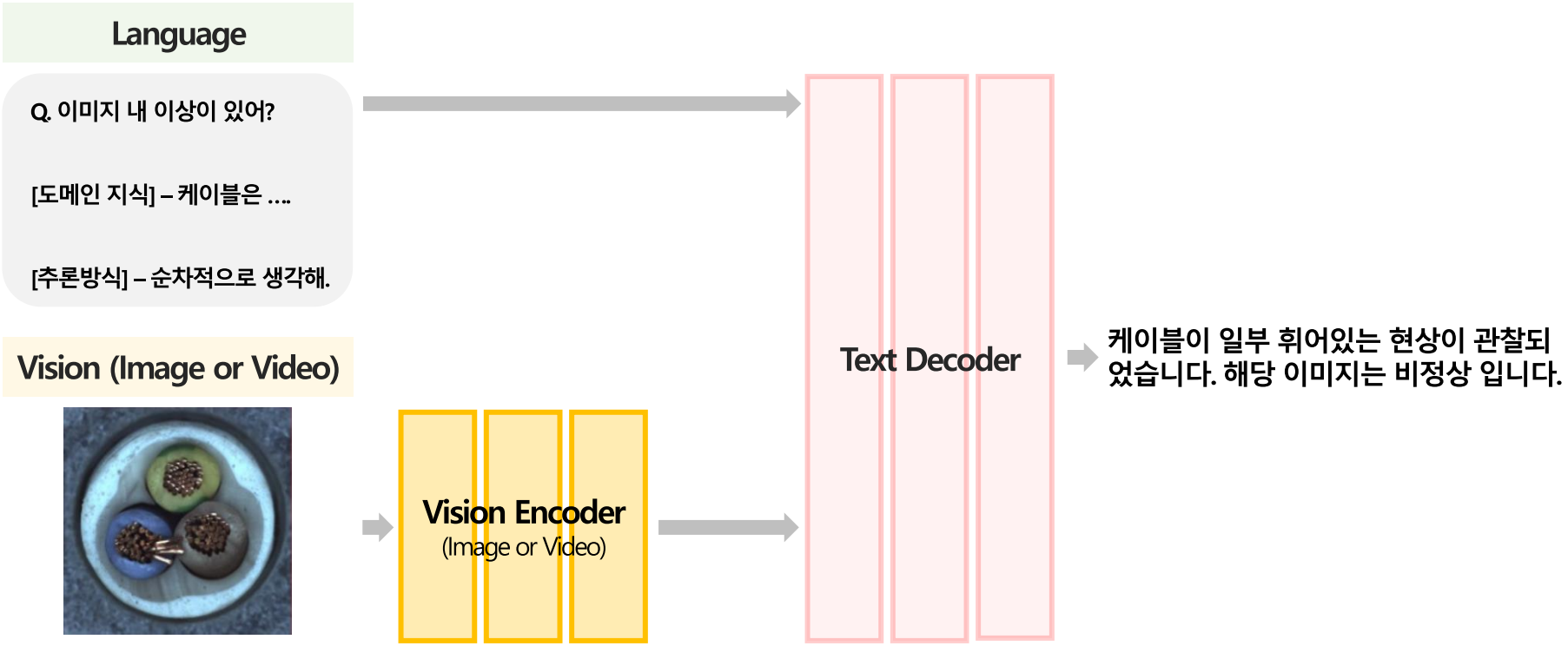
Table 3. AUC (%) on XD-Violence.

Question Type	AUC (%)
No questions	78.81
Manually written questions by human	81.15
Learned questions w/o iteratively inputting V_{batch} in Eq. (2)	78.06
Iteratively learned questions (used in VERA)	**86.55**

Conclusion

Conclusion

Vision-Language Model-based Anomaly Detection



	입력 형태	모델 추가 학습	입력 이미지 개선	입력 텍스트 개선
① Triad (2025/ICCV)	이미지	O	중요 지역 Crop	도메인 지식과 추론방식 개선
② MoXpert (2026/PR)	이미지	△	정상 이미지 활용	도메인 지식과 추론방식 개선
③ VERA (2025/CVPR)	비디오	X	X	질문 개선

Reference

1. Abdalla, M., Javed, S., Al Radi, M., Ulhaq, A., & Werghi, N. (2025). Video anomaly detection in 10 years: A survey and outlook. Neural Computing and Applications.
2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In ICML.
3. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in neural information processing systems, 36, 34892-34916.
4. Wei-Lin, C., Zhuohan, L., Lin, Z., Ying, S., Wu, Z., Hao, Z., ... & Ion, S. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. LMSYS.
5. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., & Yuan, L. (2024, November). Video-llava: Learning united visual representation by alignment before projection. In EMNLP.
6. Jiang, X., Li, J., Deng, H., Liu, Y., Gao, B. B., Zhou, Y., ... & Zheng, F. (2025) MMAD: A Comprehensive Benchmark for Multimodal Large Language Models in Industrial Anomaly Detection. In ICLR.
7. Li, Y., Yuan, S., Wang, H., Li, Q., Liu, M., Xu, C., ... & Zuo, W. (2025). Triad: Empowering LMM-based Anomaly Detection with Expert-guided Region-of-Interest Tokenizer and Manufacturing Process. In ICCV.
8. Chen, Z., & Imani, F. (2026). A multi-expert framework for enhancing multimodal large language models in industrial anomaly detection. Pattern Recognition, 112752.
9. Ye, M., Liu, W., & He, P. (2025). Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In CVPR.

Thank you!

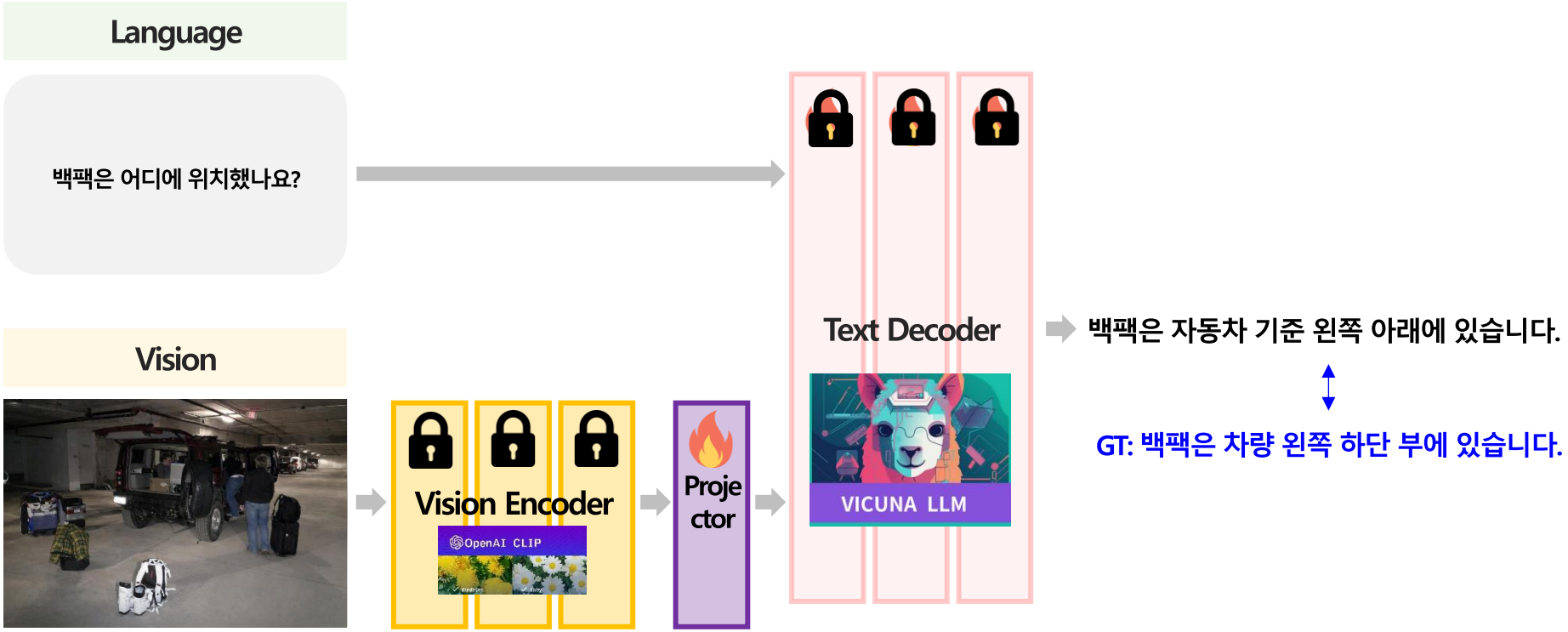
Appendix

- Vision Language Model (2/3)

❖ LLaVA (2023/NeurIPS – Microsoft)

① 이미지와 Caption만을 활용하여 Projector만 미세조정 → 이미지와 텍스트 간 정렬 학습

② Insturction을 활용하여 Projector + Text Decoder 학습 → 실제 대답할 수 있는 방법 학습



[4] Wei-Lin, C., Zhuohan, L., Lin, Z., Ying, S., Wu, Z., Hao, Z., ... & Ion, S. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. LMSYS.