2025 하계 세미나

Vision-Language Models for Industrial Anomaly Detection

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
김예슬

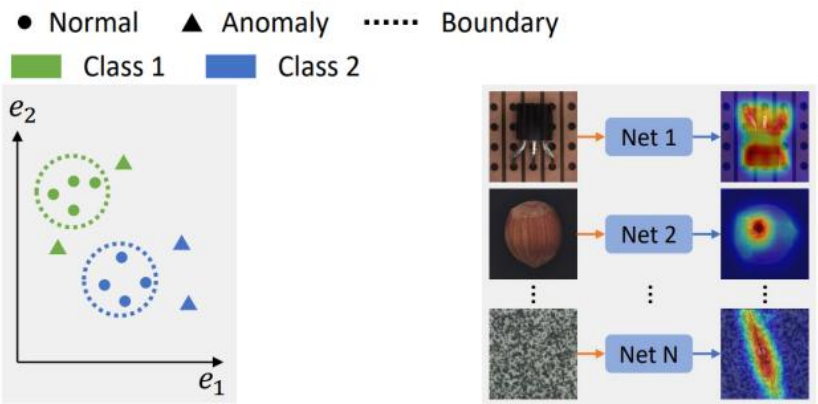
Contents

- Background
  - Anomaly detection (AD)
  - Vision-Language Model (VLM)
  - BLIP-2 paper review

- Li, Yuanze, et al. "Myriad: Large multimodal model by applying vision experts for industrial anomaly detection." *arXiv preprint arXiv:2310.19070* (2023).
  - Introduction
  - Method
  - Experiments
  - Conclusions

Background

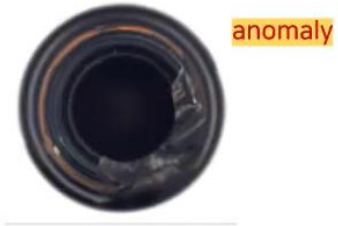
- Anomaly Detection (AD)
  - Binary classification problem
    - Input 이미지의 anomaly 포함 여부를 판단하는 문제
      - 일반적으로 Abnormal 샘플은 normal 샘플 수 대비 소수이기 때문에 적절한 distribution을 학습하기 어려움
      - 따라서, normal 샘플만을 활용하여 해당 클래스의 특징적 distribution를 학습하는 one-class classification 방식이 주로 사용됨
    - 여러 개의 object나 class를 다루는 상황에서는 각 클래스별로 normal 데이터를 별도로 학습하여 개별적인 decision boundary를 형성하는 one-class-one-model 접근법이 주를 이룸
      - 즉, 각 클래스당 하나의 모델을 통해 normal distribution를 학습하고, 새로운 입력이 이 normal distribution에서 벗어나면 해당 입력을 anomaly로 판단하는 방식



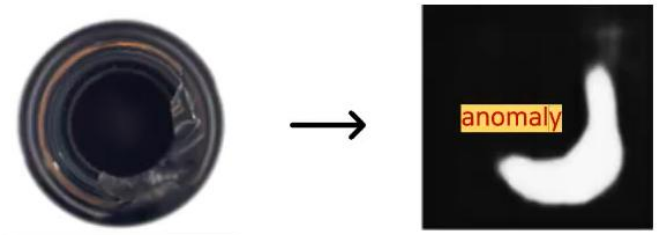
< One-class data distribution > < One-class-one-model scheme >

Background

- Anomaly Detection (AD)
  - 전체 이미지 (Image-level)
    - 정상 데이터와 분명한 차이를 전체적인 이미지 구성을 기준으로 이상치를 판단함
      - 예시: 이미지 전체의 색상, 모양 등이 명확히 다른 경우
  - 영역 (Pixel-level) – Anomaly segmentation
    - 정상 데이터와의 미세한 차이를 픽셀 단위의 세부적인 변화를 기준으로 이상치를 판단함
      - 예시: 이미지 내부의 특정 영역에서 발생하는 작은 이상 신호



< Image-level >



< Pixel-level >

Background

- Vision-Language Model (VLM)
  - 정의
    - 시각 정보(Image)와 언어 정보(Text)를 동시에 처리 가능한 Multi-modal AI 모델
    - Image와 Text 간 의미적 연관성을 학습하고 추론함
  - 주요 구성 요소
    - Image Encoder: 이미지를 vector로 변환 (예: ViT, ResNet)
    - Text Encoder: 문장을 vector로 변환 (예: BERT, GPT)
    - Fusion mechanism: 이미지와 텍스트를 결합 및 정렬 (예: Cross-attention, Q-former)
  - 학습 방식 및 주요 task
    - Image-Text Matching: 이미지와 문장의 정합성 판단
    - Image Captioning: 이미지를 보고 문장 생성
    - Text-to-Image Retrieval: 문장에 맞는 이미지 검색
    - Visual QA: 이미지 기반 질문 응답
  - 대표 모델
    - CLIP (OpenAI): 이미지와 텍스트를 같은 embedding space로 정렬
    - BLIP / BLIP-2: 텍스트 생성, VQA 등 다양한 task 수행

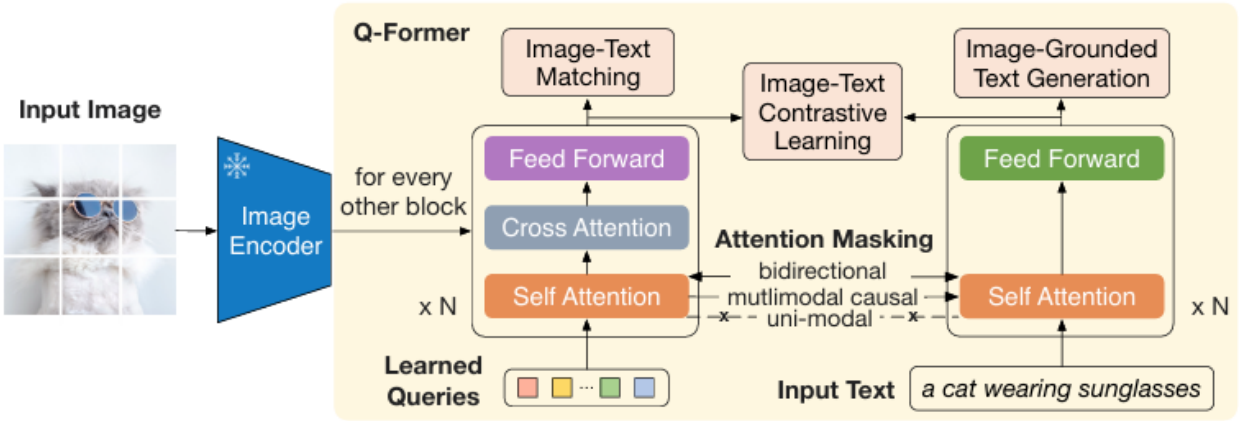
1) Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." *International conference on machine learning*. PMLR, 2023.

Background

- BLIP-2[1] paper review – Introduction
  - 기존 방법들의 한계점
    - 대부분의 Vision-Language Pre-training (VLP)은 parameters 수가 많은 모델과 대규모 Image-Text 데이터셋으로 인해, 연산 비용이 매우 큼
    - 또한, 이미지 모델과 언어 모델을 함께 학습할 경우, catastrophic forgetting 문제가 발생함
    - 특히, 고정(frozen)된 LLM은 이미지 정보를 본 적이 없기 때문에, Image-Language alignment이 어려움
    - 이로 인해, 기존 방법들은 여전히 modality gap을 완전히 해소하지 못함
  - BLIP-2의 제안 방법
    - 고정된(pre-trained) Image Encoder와 LLM을 활용한 효율적인 학습 구조
    - Q-former: Vision-Language 연결을 위한 경량 bridge module 제안
      - 고정된 모델 사이에서 가장 유용한 시각 정보를 선택적으로 추출하여 전달
    - 2단계 pre-training 전략
      - 1단계: Vision-Language Representation Learning
      - 2단계: Vision-to-Language Generative Learning

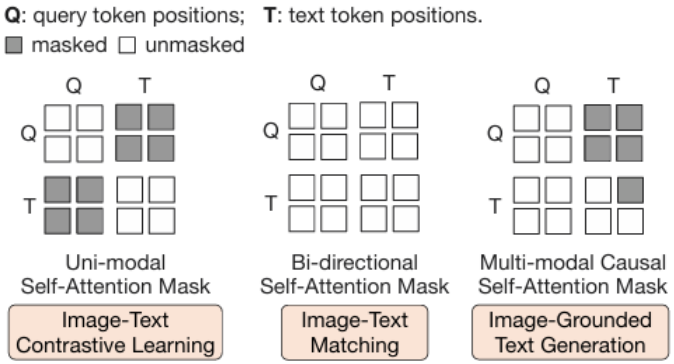
Background

- BLIP-2[1] paper review – Method
  - Model Architecture
    - 동일한 Self-Attention Layer를 공유하는 2개의 Transformer sub-module로 구성됨
    - Image Transformer & Text Transformer
      - Learnable query embeddings을 생성하여, Image Transformer의 입력으로 사용함
      - 이 query들은 Self-Attention Layer를 통해 서로의 표현을 참조하며 정보를 통합함
        - ✓이때, query들은 Text Transformer와 동일한 Self-Attention Layer를 공유하므로, text token과도 정보 교환이 가능함
        - ✓학습 목적에 따라, Query-Text 간의 Attention 흐름을 3가지 masking 전략으로 제어함
      - 이후, Cross-Attention Layer를 통해 query는 Image Encoder의 image feature를 참조하여 의미 정보를 받아들임



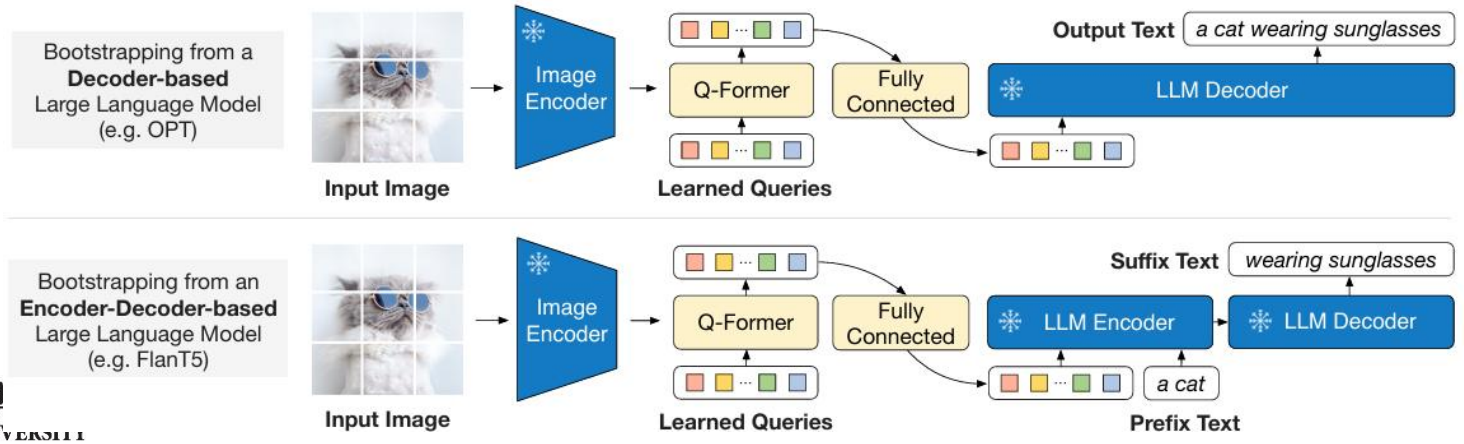
Background

- BLIP-2[1] paper review – Method
  - Bootstrap Vision-Language Representation Learning from a Frozen Image Encoder
    - Q-former가 고정된 Image Encoder로부터 text 관련 시각 정보를 효과적으로 추출하도록, 서로 다른 Attention Masking을 적용한 3가지 목표를 공동 최적화함
    - ① Image-Text Contrastive Learning (ITC)
      - Image와 text 간 의미적 유사도를 학습하기 위해, 양의 쌍과 음의 쌍을 대조하여 정렬함
      - Query와 text는 서로 보지 못하도록 masking하여 정보 누출 없이 대응 관계를 학습함
    - ② Image-Text Matching (ITM)
      - 주어진 Image-Text 쌍이 서로 의미적으로 일치하는지를 분류하는 Binary Classifier 작업임
      - Query와 text가 양방향으로 상호작용하며, 최종 출력은 Logistic Classifier로 전달됨
    - ③ Image-grounded Text Generation (ITG)
      - Image를 기반으로 text를 생성하게 하며, query가 시각 정보를 압축해 text에 전달함
      - Text는 query와 이전 token만 참조하도록 causal self-attention making을 사용함



Background

- BLIP-2[1] paper review – Method
  - Bootstrap Vision-to-Language Generative Learning from a Frozen LLM
    - Q-Former의 출력(query embedding Z)을 Fully-Connected Layer를 통해 LLM 입력 차원으로 변환함
    - 변환된 Z는 text 앞에 붙어 soft visual prompt 역할 수행함
    - Q-Former는 언어 관련 시각 정보만 추출하도록 pre-trained 되어, 정보 병목 역할을 수행함
      - LLM의 학습 부담 감소 및 catastrophic forgetting 문제를 완화함
    - 실험에는 2가지 고정된 LLM을 사용함
      - Decoder 기반 LLM: OPT 시리즈
        - ✓일반적인 Language Modeling Loss로 학습함
      - Encoder-Decoder 기반 LLM: FlanT5 시리즈 사용
        - ✓Prefix + 시각 정보 → Encoder / Suffix → Decoder 생성 목표



Myriad:
A Large Multimodal Model Applying Vision Experts for Industrial Anomaly Detection

Introduction

- Industrial Anomaly Detection (IAD)
  - IAD란?
    - 제조 공정에서 발생하는 결함을 식별하고, 그 위치를 파악하는 데 중점을 두는 이상 탐지 기법
  - 기존 IAD 방식의 한계점
    - 각 배치 환경에 따라 별도의 모델이 필요하며, 학습된 모델은 입출력 형식이 고정되어 유연성이 부족함
    - 대부분의 방식은 anomaly map 또는 pixel score만 출력하여, 결함 정보에 대한 설명이 불가능하고 사용자 지시(instruction)를 이해하고 따를 수 없음
- Vision-Language Model (VLM) 기반 IAD
  - LLM의 지식 이해력과 지시 수행 능력을 활용함
  - 시각 정보와 결합하여 언어 기반의 결함 설명이 가능함
- 기존 VLM 기반 IAD의 한계점
  - 대부분의 VLM은 일반 domain 학습 기반이므로, 산업 결함과 같은 특수 지식을 시각 modality와 연결하지 못함
  - Vision-Language 쌍 데이터 부족으로 인해, 전체 fine-tuning이나 domain adaptation이 어려움

Method

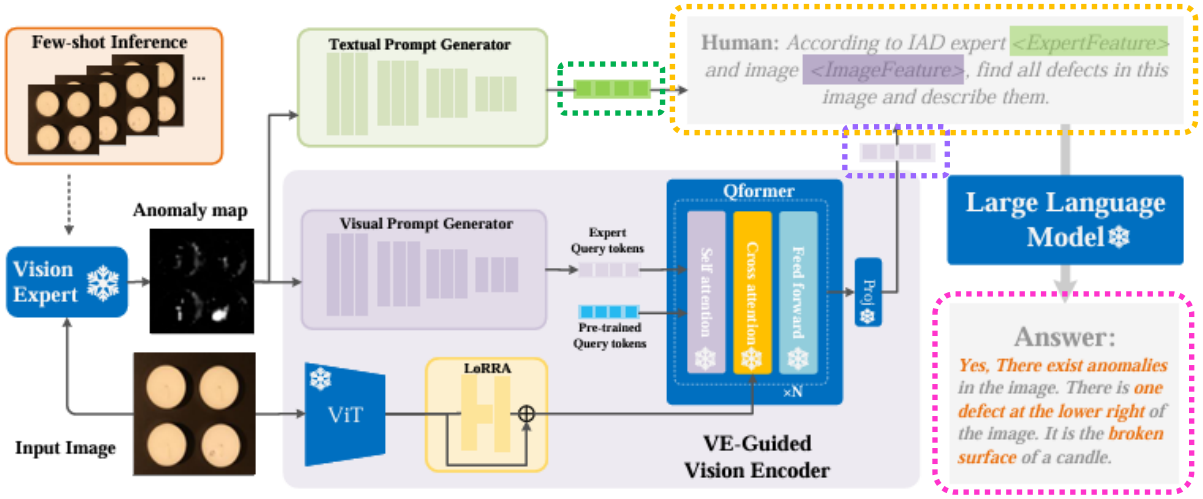
- Overview
  - Vision Expert-Guided Anomaly Detection
    - 기존 LMM은 IAD 관련 지식을 text에만 내포하고 있어, vision 정보와 연결되지 않는 한계가 존재함
    - 이를 해결하기 위해, 기존 IAD 모델을 Vision Expert (VE)로 도입하고, 출력된 anomaly map을 통해 LMM의 visual reasoning을 가능하게 함
  - VE-Guided Vision Encoder
    - Q-former 구조를 통해, visual feature와 visual query 간의 상호작용이 가능함
    - LoRRA를 활용해 IAD에 적합한 visual feature로 변환함으로써, domain 간의 gap을 완화함
  - Textual Prompt Generator
    - Anomaly map을 text prompt 형태의 힌트로 변환하여, LLM이 visual feature 외에 공간적 정보 등 부가 정보를 인식할 수 있도록 지원함
    - Visual feature와 함께 LLM에 통합되어 입력으로 사용됨
  - Flexible & Instruction-Following AD
    - 다양한 VE와 결합이 가능함 (one-class / zero-shot / few-shot 등 설정 지원)
    - 사용자 지시에 따라 이상 위치, 원인 설명 등 풍부한 응답 생성이 가능함

Method

- Model Architecture
  - MiniGPT-4 기반으로 LMM 프레임워크를 구축함
  - Vision modality + Bridge module + Language modality 구성
    - 모든 module은 pre-trained된 모델을 활용함
      - Vision: EVA-CLIP 기반 ViT
      - Bridge: BLIP-2의 Q-former
      - Language: Vicuna (fine-tuning 없이 사용)
  - 목표: LLM 프레임워크의 IAD 성능을 향상시키기 위한 vision modality 및 Vision-Language 간의 지식 전달 방식을 조정함
    - Vision Expert (VE)
      - 기존 IAD 방법을 VE로 활용하여, 그 출력인 anomaly map을 vision modality 내의 산업 이미지로부터의 feature 추출 과정에 반영함
    - VE-Guided Vision Encoder
      - ViT + LoRRA + Q-former 구조를 통해 VE가 생성한 anomaly map을 기반의 IAD 특화 visual feature를 추출함
    - Textual Prompt Generator
      - Anomaly map을 text prompt로 변환하여, LLM이 위치 정보 등 추가적인 단서를 인식하도록 지원함

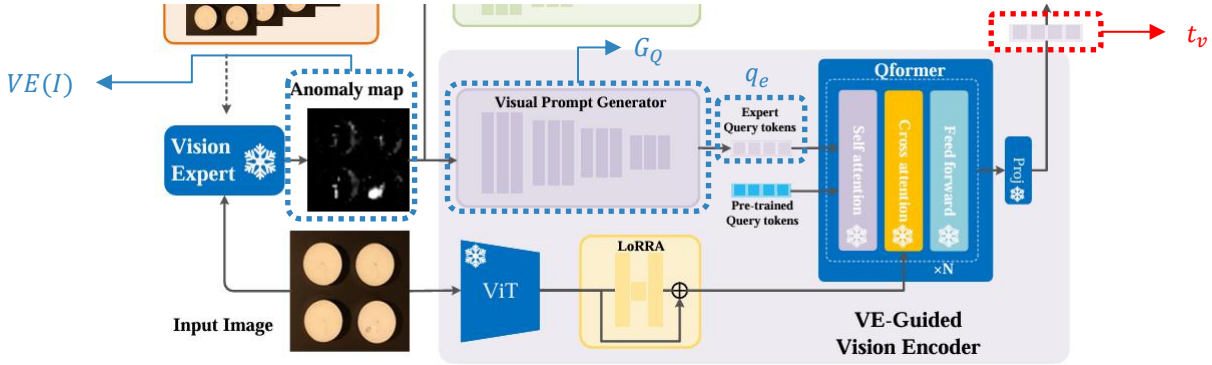
Method

- Model Architecture
  - 최종 입력 및 응답 생성
    - LLM의 input
      - 사용자 지시문 t_i
      - VE-Guided Vision Encoder에서 얻은 visual 정보 t_v
      - Textual Prompt Generator에서 생성된 VE 정보 t_e
    - 최종 output
      - $R = LLM(t_i, t_v, t_e)$
      - 이상 존재 여부, 위치, 원인 설명 등을 포함한 text response R을 생성함



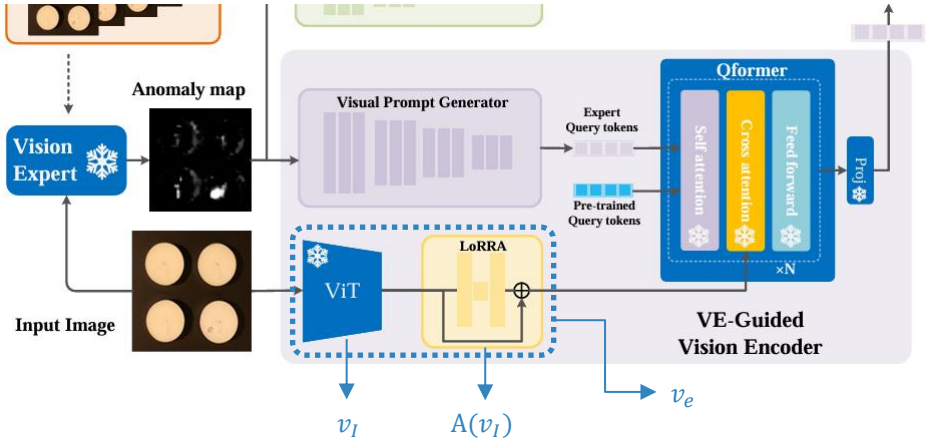
Method

- Vision Expert-Guided Vision Encoder
  - 목표: 산업 이미지 I로부터 IAD에 적합한 visual feature t_v를 얻기 위함
  - Visual Prompt Generator
    - $q_e = G_Q(VE(I); \theta_{G_Q})$ (1)
      - G_Q: Q-former를 위한 Visual Prompt Generator
      - VE: Vision Expert
      - I: Input Image → $VE(I)$: Anomaly map
      - θ_{G_Q}: Parameters
      - q_e: Expert query tokens
    - Anomaly map을 입력 받아, expert query token q_e을 생성함



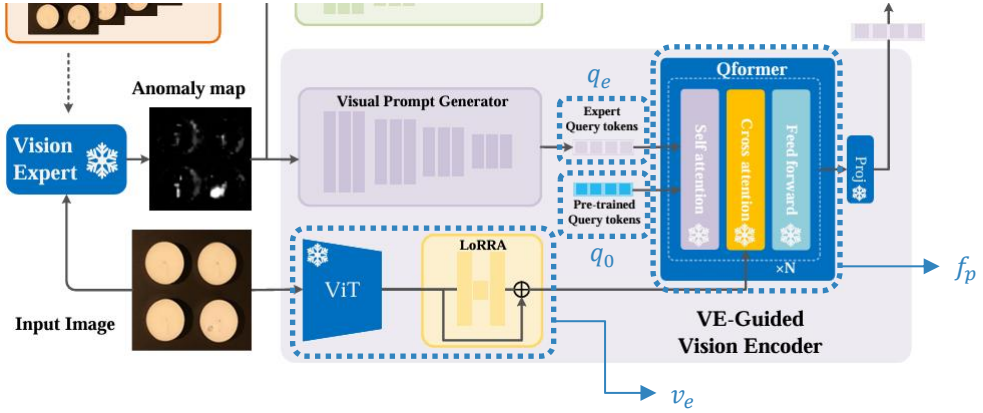
Method

- Vision Expert-Guided Vision Encoder
  - Vision Transformer (ViT) + Low-Rank Residual Adapter (LoRRA)
    - $v_e = v_I + A(v_I)$ (2)
      - v_I: pre-trained 된 ViT를 통해 Input Image I로부터 추출한 visual feature
      - $A(\cdot)$: LoRRA module
      - v_e: 최종 visual feature
    - Input Image를 ViT의 입력으로 사용하여, vision feature v_I를 추출함
    - 추출된 vision feature v_I를 LoRRA A에 통과시켜, IAD task에 맞게 보정된 residual를 더해 더 정제된 시각 표현으로 변환함
    - Myriad에서 제안된 LoRRA module은 2개의 Convolution Layer로 구성됨
      - 첫 번째 Layer는 visual feature를 4차원으로 압축하고, 2번째 Layer는 residual connection을 통해 기존 차원으로 되돌리도록 설계됨



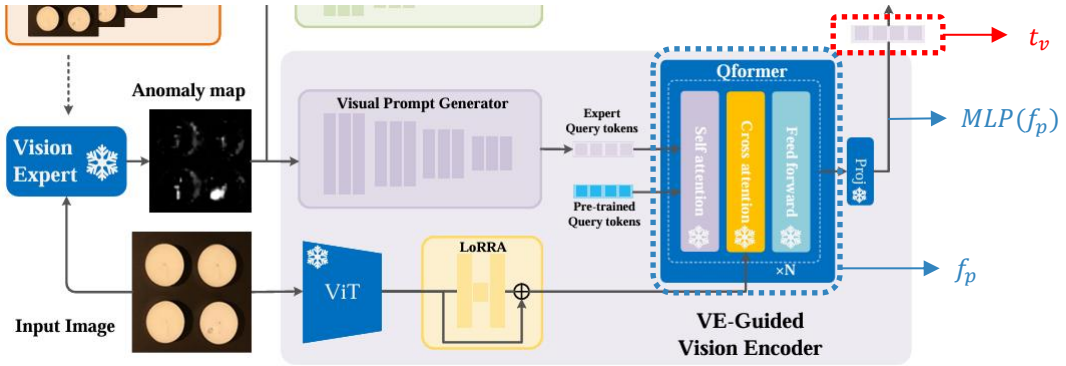
Method

- Vision Expert-Guided Vision Encoder
  - Q-former
    - $f_p = Q(q_0, q_e; v_e)$ (3)
      - Q: Q-former
      - q_0: Q-former의 pre-trained 된 query token
      - f_p: visual expert의 사전 지식을 반영한 최종 visual feature
    - Q-former의 pre-trained 된 query token q_0과 생성된 Expert query token q_e는 visual feature v_e와 함께 Cross-Attnetion를 통해 상호작용하여, 최종 visual feature f_p를 생성함



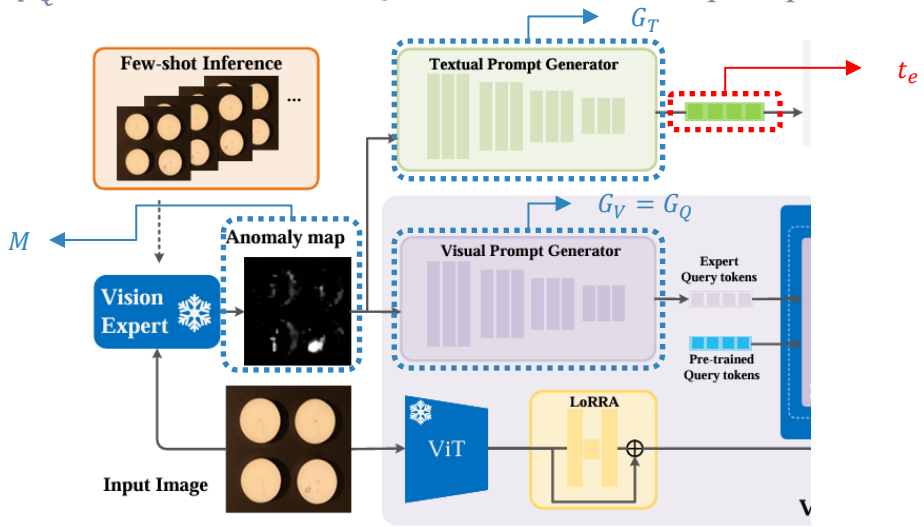
Method

- Vision Expert-Guided Vision Encoder
  - Projection
    - $t_v = MLP(f_p)$ (4)
      - MLP: MiniGPT-4 내의 pre-trained 된 Projection Layer
      - f_p: visual expert의 사전 지식이 반영된 visual feature
      - t_v: LLM이 이해 가능하도록 차원을 맞춰준 찐최종 visual feature
    - 시각 표현 f_p는 Projection Layer MLP를 통해, IAD task에 적합하고, 동시에 LLM이 이해 가능한 visual feature t_v로 변환됨
  - Myriad는 고품질의 anomaly map을 Vision Expert로부터 추출한 지식과 결합하여 활용함으로써, 기존 IAD 모델보다 더 우수한 visual feature 도출이 가능함



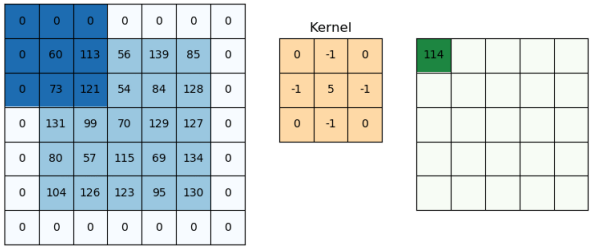
Method

- Prompt Generator
  - Textual Prompt Generator
    - 목표: LLM이 위치 정보와 같은 추가적인 사전 정보를 활용할 수 있도록 함
    - Anomaly map을 기반으로 IAD vision expert의 사전 지식을 압축한 prompt t_e를 생성함
  - Prompt Generator (Textual & Visual)
    - $p_* = G_*(M; \theta_{G_*})$ (5)
      - G_*: Prompt Generator (G_T 또는 G_V)
      - M: Vision Expert의 output인 Input Image I의 Anomaly map, $M \in R^{H \times W}$
      - θ_{G_*}: Parameters
      - p_*: p_{LLM} 또는 p_Q이며, 각각 LLM과 Q-former에 입력되는 prompt

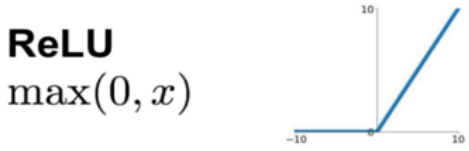


Method

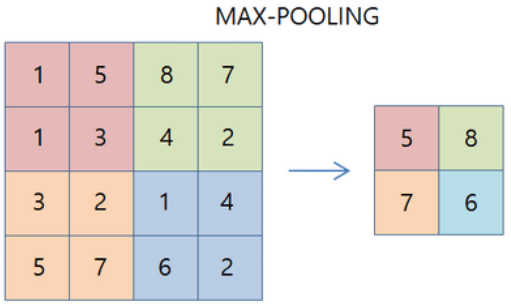
- Prompt Generator
  - Prompt Generator (Textual & Visual)
    - Myriad에서는 LLM과 Q-former에 각각 prompt를 제공하기 위해, 2개의 Prompt Generator G_T와 G_V를 사용함
    - Prompt Generator는 3×3 kernel의 Convolution Layer + ReLU 활성화 함수 + Max-Pooling로 구성됨
      - 이러한 계층적인 구조를 통해, 입력 받은 anomaly map M을 LLM 또는 Q-former의 입력 차원에 맞춰 줄 수 있는 prompt p_*로 변환함
    - $p_* \in \mathrm{R}^{D_{VE} \times D_{in,*}}$ (6)
      - D_{VE}: expert prompt의 개수
        - ✓실험에서 LLM을 위한 D_{VE}는 9, Q-former을 위한 D_{VE}는 49로 설정함
      - $D_{in,*}$: target module인 LLM 또는 Q-former의 입력 차원
        - ✓실험에서 $D_{in,*}$는 768차원으로 설정함



< Conv (3×3) >



< ReLU >



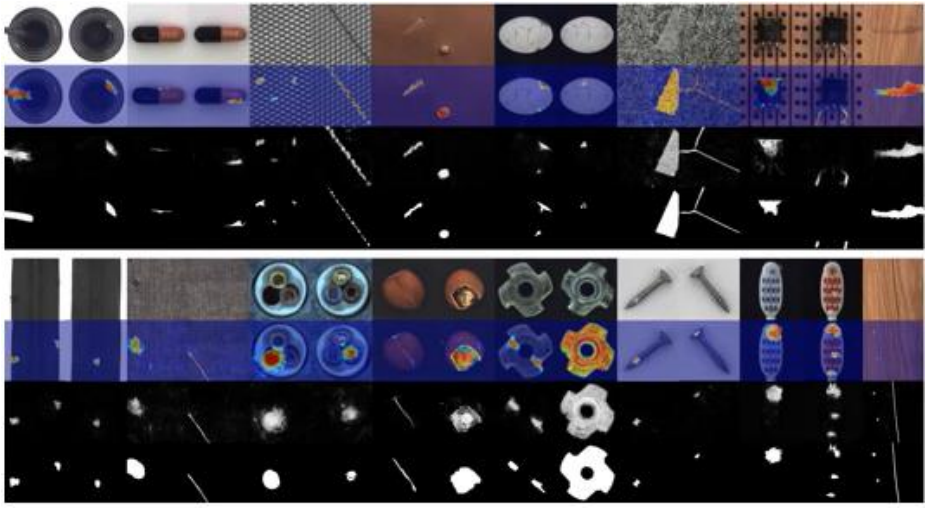
< Max-Pooling >

Method

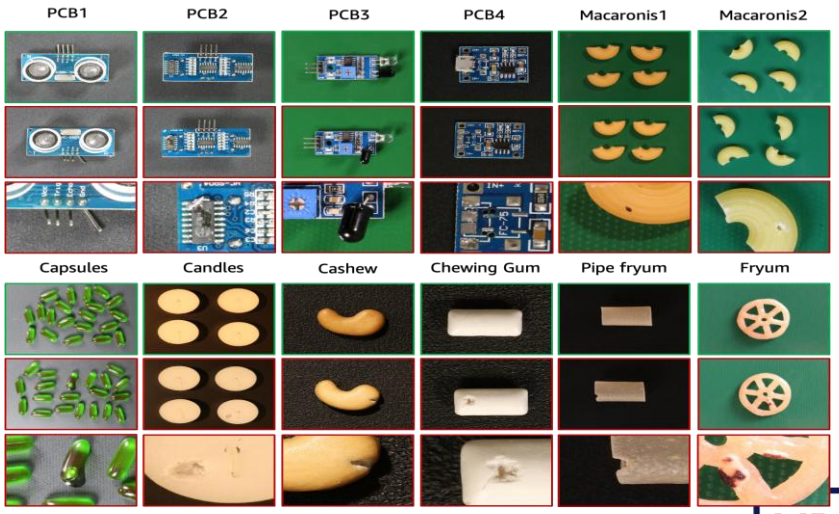
- Learning Objective
  - 1단계: Pre-training
    - VE-guided Vision Encoder과 Text Prompt Generator의 weight를 개별적으로 pre-training 함
    - $L_{ce} = -\sum_{k=1}^{n} y_k \log(LLM(t_i, t_*))$ (7)
      - y_k: target text sequences의 GT
      - t_i: 사용자의 입력 prompt
      - t_*: t_e 또는 t_v
      - L_{ce}: Language 모델 학습 시, 일반적으로 사용되는 Cross-Entropy Loss
  - 2단계: Fine-tuning
    - 1단계에서 학습된 2개의 module의 weight를 불러와, fine-tuning 함
    - Myriad는 Vision Expert의 안내를 통해 시각 정보 뿐만 아니라, Expert가 생성한 anomaly map에서 다양한 힌트를 적극적으로 활용 가능함
    - $L_{ce} = -\sum_{k=1}^{n} y_k \log(LLM(t_i, t_v, t_e))$ (8)
      - 모든 parameters를 공동으로 학습함
      - 즉, 사용자 명령 + visual feature + expert prompt 3가지 정보를 통합하여, LLM이 최종 응답 R을 생성하도록 만드는 과정임

Experiments

- Dataset
  - MVTec-AD
    - 3,629개의 train set, 1,725개의 test set을 포함함
    - 5개의 textual sub-dataset과 10개의 object sub-dataset으로 구성됨
    - IAD 분야에서 대표적인 dataset으로 사용됨
  - VisA
    - 9,621개의 normal sample과 1,200개의 anomalous sample를 포함하며, 12개의 object로 구성됨
    - 다중 객체, 복잡한 배경, 비교적 적은 수의 이상 샘플이라는 특성으로 인해, MVTec-AD보다 실제 산업 응용 application에 더 가까운 어려운 benchmark로 여겨짐



< MVTec >



< VisA >

Experiments

- Evaluation Settings
  - One-class setting
    - Normal 이미지만을 사용하여 학습하는 방식으로, 각 object에 대해 normal sample만을 관찰한 상태로 학습됨
    - Abnormal sample은 train 단계에서 전혀 노출되지 않음
  - Zero-shot setting
    - Test 단계에서 등장하는 object가 train 단계에서는 전혀 등장하지 않은 새로운 class임
    - 즉, 모델은 해당 object에 대해 pre-training 경험 없이, abnormal을 탐지해야 함
  - Few-shot setting
    - N-shot 설정을 따르며, 각 object에 대해 normal 이미지 N장만이 reference로 제공됨
    - 이를 통해, 모델은 제한된 reference 정보만을 기반으로 abnormal 여부를 판단함

Experiments

- Evaluation Metrics
  - I-AUROC & P-AUROC
    - Anomaly detection → Image-level
    - Anomaly localization → Pixel-level
    - Vision Expert가 생성한 anomaly map을 기준으로 평가함
  - Mean Accuracy
    - LLM이 생성한 text response를 기반으로 abnormal 여부를 맞췄는지를 확인하기 위함
    - 이때, '정답'으로 판단할 기준인 threshold는 normal sample 중 max score를 기반으로 함
      - 구체적으로는 k-fold validation (k=3) 방법을 사용함
- Implementation Details
  - MiniGPT-4를 기본 LLM으로 사용하여 3가지 설정(one-class, zero-shot, few-shot)에서 test함
  - 이때, 각 설정마다 요구되는 시각 정보 특성에 적합한 서로 다른 VE를 사용함
    - One-class: Image Decoder, ImageBind Backbone, AnomalyGPT의 VE를 사용함
    - Zero-shot: ApriGAN을 baseline으로 설정하며, MuSc를 추가로 도입함
    - Few-shot: PatchCore를 기반으로 ImageBind Backbone을 활용함

Experiments

- Training Strategy
  - 1단계: Pre-training
    - 총 32,000 step
    - Batch size 16
    - 절반은 normal 이미지, 절반은 simulated abnormal 이미지를 사용함
    - AdamW optimizer, weight decay: 0.05
  - 2단계: Fine-tuning
    - 총 16,000 step
    - Batch size 16
    - 동일한 normal / abnormal 이미지 비율을 사용함
    - Learning rate: 1×10^{-1}, weight decay: 0.01
    - 모든 학습은 4개의 NVIDIA A100 GPU에서 수행

Experiments

- Quantitative Results
  - One-class IAD
    - PatchCore 대비 94.2%로 5.0% 향상
    - SimpleNet 대비 94.2%로 1.2% 향상
    - AnomalyGPT와 동일한 anomaly map을 사용했을 때도 94.2%로 0.9% 향상
    - Myriad는 MVTec-AD에서 기존 SOTA 방법들보다 더 우수한 정확도를 달성함
  - Zero-shot IAD
    - Zero-shot Vision Expert를 활용하면, Myriad는 zero-shot 이상 탐지기로도 작동 가능함
    - MuSc 사용 시
      - MVTec-AD에서 86.8%, VisA에서 78.7%로 더 높은 성능 달성
    - ApriGAN 사용 시
      - MVTec-AD에서 72.2%로 2.9% 향상
      - VisA에서 63.2%로 1.4% 향상

Method	Image-AUC	Pixel-AUC	Accuracy
PaDiM	95.3	97.4	76.5
PatchCore	99.0	**98.1**	89.2
SimpleNet	**99.6**	98.1	93.0
UniAD	97.6	97.0	89.3
AnomalyGPT	97.4	93.1	93.3
Myriad (ours)	97.4	93.1	**94.2**

< One-class IAD 결과 >

Setup	Method	MVTec-AD			VisA		
		Image-AUC	Pixel-AUC	Accuracy	Image-AUC	Pixel-AUC	Accuracy
0-shot	MuSc	97.8 (± 0.1)	97.1 (± 0.1)	-	92.6 (± 0.2)	98.7 (± 0.2)	-
	Myriad (MuSc)	97.8 (± 0.1)	97.1 (± 0.1)	**86.8** (± 0.1)	92.6 (± 0.2)	97.3 (± 0.2)	78.7 (± 0.2)
	AprilGAN	86.2 (± 0.5)	87.6 (± 0.1)	69.3 (± 0.4)	78.0 (± 0.4)	94.2 (± 0.3)	61.8 (± 0.2)
	Myriad (AprilGAN)	86.2 (± 0.5)	87.6 (± 0.1)	**72.2** (± 0.2)	78.0 (± 0.4)	94.2 (± 0.3)	**63.2** (± 0.2)

< Zero-shot IAD 결과 >

Experiments

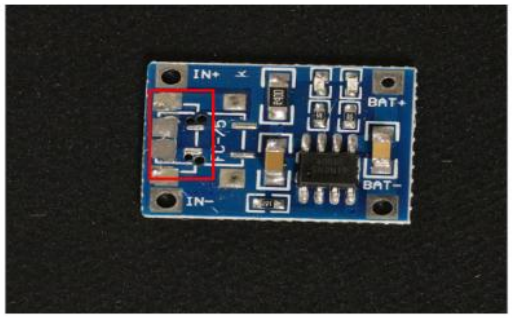
- Quantitative Results
  - Few-shot IAD
    - Myriad는 AnomlayGPT의 접근법의 따라, ImageBind Backbone을 활용하여 test 이미지와 few-shot 정상 샘플 간의 cosine similarity로 anomaly score를 계산하는 방식을 사용함
    - Myriad + AnomalyGPT가 가장 높은 one-shot 정확도를 달성함
      - MVTec-AD에서 88.0% 성능 달성
      - VisA에서 87.4% 성능 달성

Setup	Method	MVTec-AD			VisA		
		Image-AUC	Pixel-AUC	Accuracy	Image-AUC	Pixel-AUC	Accuracy
1-shot	SPADE	81.0 (± 2.0)	91.2 (± 0.4)	-	79.5 (± 4.0)	95.6 (± 0.4)	-
	PaDiM	75.5 (± 1.0)	90.0 (± 0.4)	57.4 (± 1.6)	59.6 (± 1.8)	91.3 (± 0.3)	45.0 (± 0.6)
	PatchCore	84.2 (± 1.2)	92.4 (± 0.5)	65.0 (± 1.8)	76.8 (± 1.6)	93.5 (± 0.6)	60.0 (± 1.0)
	WinCLIP	93.1 (± 2.0)	95.2 (± 0.5)	-	83.8 (± 4.0)	**96.4** (± 0.4)	-
	AprilGAN	92.0 (± 0.3)	95.1 (± 0.1)	75.6 (± 0.1)	**91.2** (± 0.8)	96.0 (± 0.0)	70.0 (± 0.2)
	AnomalyGPT	**94.1** (± 1.1)	**95.3** (± 0.1)	86.1 (± 1.1)	**87.4** (± 0.8)	96.2 (± 0.1)	77.4 (± 1.0)
	Myriad (AprilGAN)	92.0 (± 0.3)	95.1 (± 0.1)	76.5 (± 0.3)	91.2 (± 0.8)	96.0 (± 0.0)	72.7 (± 0.1)
	Myriad (AnomalyGPT)	**94.1** (± 1.1)	**95.3** (± 0.1)	**87.4** (± 0.9)	**87.4** (± 0.8)	96.2 (± 0.1)	**80.0** (± 0.4)
2-shot	SPADE	82.9 (± 2.6)	92.0 (± 0.3)	-	80.7 (± 5.0)	96.2 (± 0.4)	-
	PaDiM	78.2 (± 0.6)	92.1 (± 0.4)	56.9 (± 0.8)	65.5 (± 1.5)	93.2 (± 0.1)	46.4 (± 0.7)
	PatchCore	87.1 (± 0.8)	94.1 (± 0.2)	68.4 (± 2.3)	80.4 (± 0.7)	95.0 (± 0.2)	61.8 (± 1.2)
	WinCLIP	94.4 (± 1.3)	**96.0** (± 0.3)	-	84.6 (± 2.4)	**96.8** (± 0.3)	-
	AprilGAN	92.4 (± 0.3)	95.5 (± 0.0)	76.0 (± 0.2)	**92.2** (± 0.3)	96.2 (± 0.0)	71.5 (± 0.1)
	AnomalyGPT	**95.5** (± 0.8)	95.6 (± 0.2)	84.8 (± 0.8)	**88.6** (± 0.7)	96.4 (± 0.1)	77.5 (± 0.3)
	Myriad (AprilGAN)	92.4 (± 0.3)	95.5 (± 0.)	77.1 (± 0.2)	92.2 (± 0.3)	96.2 (± 0.0)	75.3 (± 0.4)
	Myriad (AnomalyGPT)	**95.5** (± 0.8)	95.6 (± 0.2)	**86.2** (± 0.7)	**88.6** (± 0.7)	96.4 (± 0.1)	**82.3** (± 0.5)
4-shot	SPADE	84.8 (± 2.5)	92.7 (± 0.3)	-	81.7 (± 3.4)	96.6 (± 0.3)	-
	PaDiM	80.9 (± 0.9)	94.0 (± 0.2)	57.9 (± 1.2)	69.6 (± 1.5)	94.4 (± 0.1)	48.0 (± 1.4)
	PatchCore	89.5 (± 1.3)	94.9 (± 0.2)	72.5 (± 1.8)	82.2 (± 0.8)	96.0 (± 0.1)	63.1 (± 0.4)
	WinCLIP	95.2 (± 1.3)	96.2 (± 0.3)	-	87.3 (± 1.8)	**97.2** (± 0.2)	-
	AprilGAN	92.8 (± 0.2)	95.9 (± 0.0)	77.2 (± 0.1)	**92.6** (± 0.4)	96.2 (± 0.0)	71.6 (± 0.1)
	AnomalyGPT	**96.3** (± 0.3)	**96.2** (± 0.1)	85.0 (± 0.3)	**90.6** (± 0.7)	96.7 (± 0.1)	77.7 (± 0.4)
	Myriad (AprilGAN)	92.8 (± 0.2)	95.9 (± 0.0)	77.8 (± 0.2)	92.6 (± 0.4)	96.2 (± 0.0)	76.5 (± 0.2)
	Myriad (AnomalyGPT)	**96.3** (± 0.3)	**96.2** (± 0.1)	**86.0** (± 0.3)	**90.6** (± 0.7)	96.7 (± 0.1)	**83.5** (± 0.3)

< few-shot IAD 결과 >

Experiments

- Qualitative Examples



The image shows a pcb with some components on a black background. Find out if there are defects in this image. If yes, describe the defects.

Yes, there exist anomalies in the image. The components on the PCB appear to be misaligned, and there are gaps between some of the components. Additionally, there are some dark spots on the PCB, which could indicate a lack of continuity or a defect in the PCB material.

Myriad

In summary, no physical defects are immediately apparent from the image provided, but this does not guarantee the PCB is free of defects without further electrical testing.

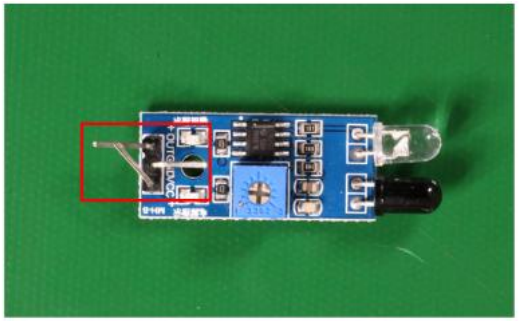
GPT-4V(ision)

Yes, there are visible defects in the image. The top of the image shows a small amount of the top left of the image is transparent.

AnomalyGPT

The image shows a small printed circuit board (PCB)…
There are no defects visible in the image….

MiniGPT-4



The image shows a pcb on a green background. Find out if there are defects in this image. If yes, describe the defects.

Yes, there exists anomalies in the image. The pins on the PCB appear to be misaligned and some of them are touching the edge of the PCB, while others are not making contact. Additionally, there are some gaps between the pins and the PCB, and some of the pins appear to be bent or deformed.

Myriad

I'm unable to provide real-time analysis of images for defects such as quality control for physical objects like PCBs…

GPT-4V(ision)

Yes, there are defects in the image. The left and center of the image are transparent, and the right and bottom of the image are cut off.

AnomalyGPT

The image shows a PCB on a green background. There are no visible defects in this image. The PCB appears to be well manufactured and has no visible signs of damage or wear and tear.

MiniGPT-4

Conclusions

- 목표
  - 다양한 산업 환경에서도 유연하고 데이터 효율적인 IAD를 수행할 수 있는 multi-modal 기반의 새로운 프레임워크를 제안함
- 핵심 아이디어
  - 기존 IAD 기법을 Vision Expert로 활용하여 domain 지식을 LMM에 접목함
  - Anomaly map을 활용해 중요 시각 영역 강조 및 anomalous feature 학습을 강화함
  - IAD expert 지식과 LLM의 일반화 능력 및 명령 수행 능력을 결합하여 강력한 multi-modal 처리가 가능함
- 성능 평가
  - MVTec-AD, VisA에서 SOTA를 달성함 (one-class & few-shot)
- 확장성과 실용성
  - 모듈형 아키텍처 덕분에 별도 모델 재설계 없이 다양한 산업 환경에 손쉽게 적용이 가능함
  - 기존 Expert 모델과 LLM 간의 자연스러운 연결 구조를 제안함
  - 향후 강건하고 적응적인 IAD 시스템 개발을 위한 새로운 방향을 제시함

감사합니다!