

Research trends in Anomaly Detection, Generation, and MLLM



RILS LAB

Robot Intelligence Learning & System

유호현

2026. 08. 21

Papers

- **(CVPR 2026) MMR-AD: A Large-Scale Multimodal Dataset for Benchmarking General AD with MLLMs**
- **(CVPR 2026) One-to-More: High-Fidelity Training-Free Anomaly Generation with Attention Control**
- **(CVPR 2026 Oral) ANTS: Adaptive Negative Textual Space Shaping for OOD Detection**
- **(CVPR 2026) AnomalyVFM: Transforming Vision Foundation Models into Zero-Shot Anomaly Detectors**
- **(CVPR 2026) IMDD-1M: Towards Open-Vocabulary Industrial Defect Understanding**

06

MMR-AD: A Large-Scale Multimodal Dataset for Benchmarking General Anomaly Detection with Multimodal Large Language Models

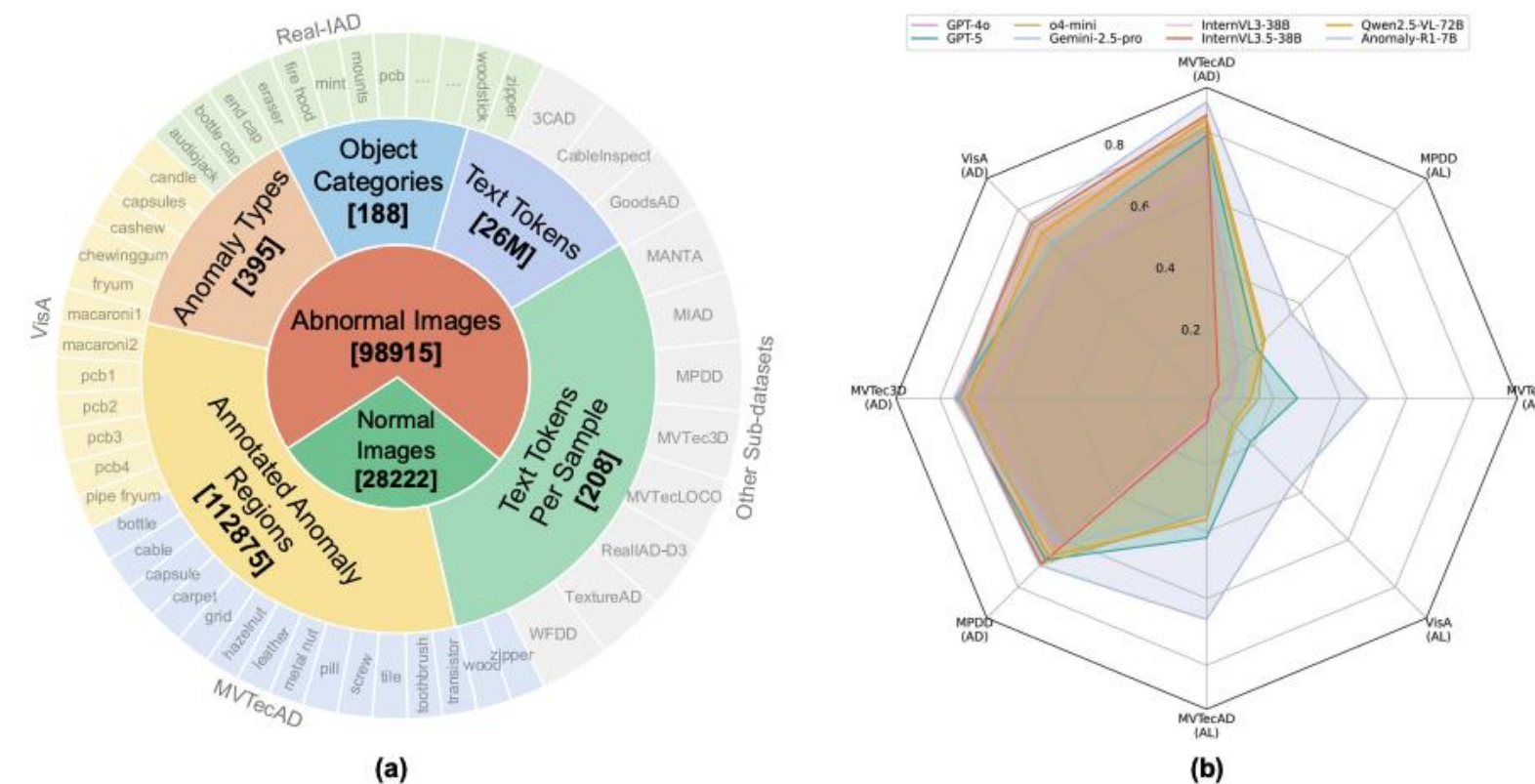
Xincheng Yao, Zefeng Qian, Chao Shi, Jiayang Song, Chongyang Zhang

CVPR 2026

MMR-AD: A Large-Scale Multimodal Dataset for Benchmarking General AD with MLLMs

Problem

- General AD (GAD) targets arbitrary novel classes without retraining; MLLMs are the natural candidate, yet their GAD ability is unmeasured.
- Web-scale pretraining leaves a large domain gap to industrial AD, and the image-text pairs are not AD-oriented.
- Mainstream AD datasets are image-only, hence unusable for MLLM post-training.

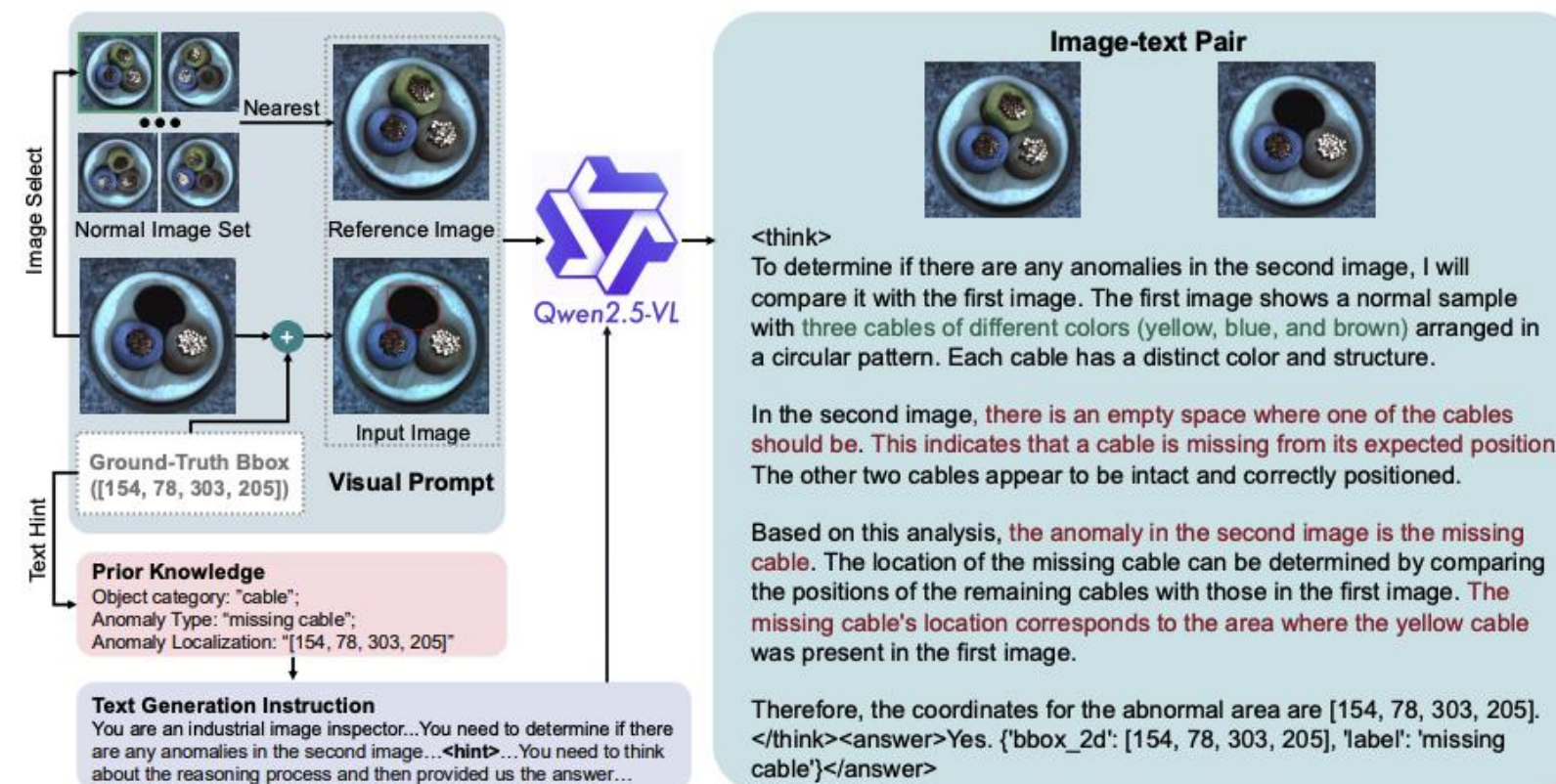


<Overview of the MMR-AD dataset>

MMR-AD: A Large-Scale Multimodal Dataset for Benchmarking General AD with MLLMs

Method

- MMR-AD: 127K+ industrial samples curated from 14 public AD datasets (MVTec AD, VisA, LOCO, Real-IAD, MANTA, 3CAD, ...).
- Each item pairs a normal reference with a query image; CoT reasoning text is auto-generated by Qwen2.5-VL-72B.
- Anomaly-R1: a reasoning AD baseline — CoT supervised fine-tuning followed by reinforcement learning.

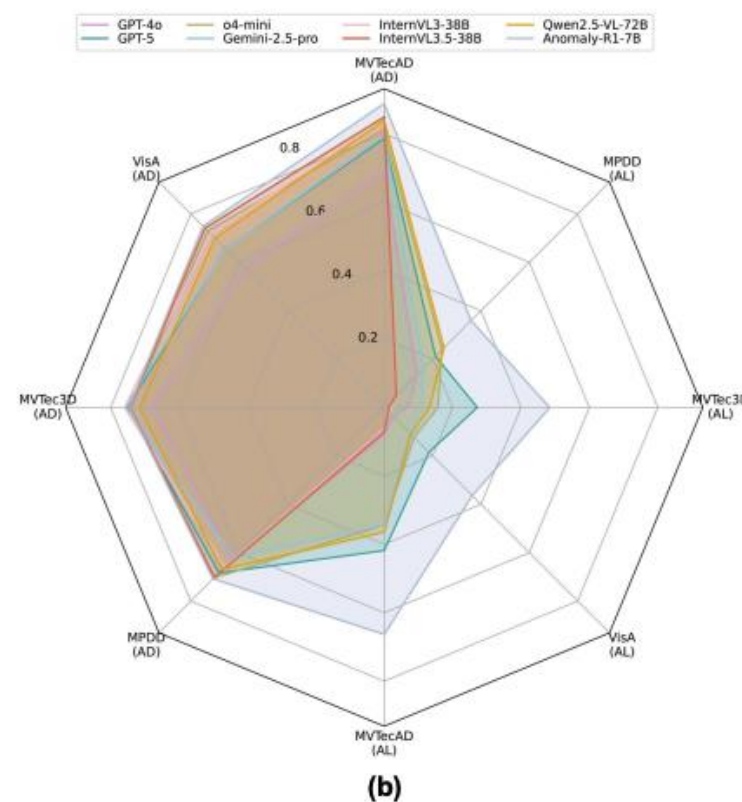


<The illustration of the text generation pipeline>

MMR-AD: A Large-Scale Multimodal Dataset for Benchmarking General AD with MLLMs

Results & Contribution

- SOTA generalist MLLMs still fall far short of industrial requirements, especially in localization.
- Anomaly-R1-7B yields remarkable gains over generalist MLLMs in both detection and localization.
- First large-scale multimodal corpus for both post-training and evaluation of MLLM-based general AD.



Anomaly-Instruct-125K



Are there any anomalies in this cable image?

- Yes, the green wire is missing its internal copper conductors, unlike the blue and brown wires, which contain complete sets of bundled copper strands.



Are there any significant anomalies detected on this cookie?

- Yes, there are two significant anomalies on the cookie: a central defect featuring a noticeable hole and an edge defect showing a gap or crack.

MMAD



Is there any defect in the object?

A: Yes. ✓
B: No.

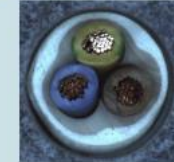
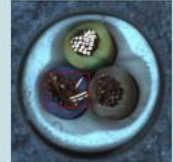
There is a defect in the object. What is the type of the defect?

A: A mold growth.
B: A crack. ✓
C: A stain.
D: A deformation.

There is a defect in the object. Where is the defect?

A: On the tip.
B: Inside the nut.
C: On the base.
D: Around the center. ✓

MMR-AD

<think>
To determine if there are any anomalies in the second image, I will compare it with the first image. The first image shows a normal sample of three cables with their internal wires neatly arranged and no visible defects.

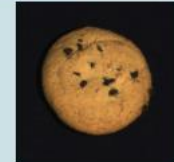
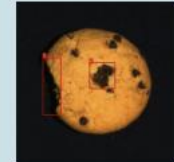
In the second image, upon close inspection:

- The green cable appears normal.
- The brown cable also appears normal.
- The blue cable has some wires that seem bent or misaligned compared to the neat arrangement in the first image.

Based on this analysis, the blue cable in the second image contains an anomaly where the wires are bent. The location of this abnormal area is [91, 199, 217, 296].

</think><answer>Yes. {'bbox_2d': [91, 199, 217, 296], 'label': 'bent wire'}</answer>

MMR-AD

<think>
To determine if there are any anomalies in the second image, I will compare it with the first image which is considered normal.

- Shape and Structure**: The second image shows a cookie that appears to have an irregular shape compared to the first one. There seems to be a part where the structure of the cookie is broken or cracked.
- Surface Texture**: Upon closer inspection, there is a noticeable area on the surface of the cookie in the second image where the texture is disrupted. This disruption looks like a hole or a missing piece of the cookie.
- Chocolate Chips**: The distribution of chocolate chips seems consistent between both images, so this is not a point of concern.

Based on these observations, the areas that appear abnormal are:

- A crack on the left side of the cookie.
- A hole near the center-right of the cookie.

The coordinates for these areas can be estimated as follows:

- Crack: [72, 154, 123, 314]
- Hole: [203, 167, 275, 241]

</think><answer>Yes. {'bbox_2d': [72, 154, 123, 314], 'label': 'crack'}, {'bbox_2d': [203, 167, 275, 241], 'label': 'hole'}</answer>

MMR-AD




<think>
To determine if there are anomalies in the second image, I'll compare the contents of both images:

- First Image Analysis**:
 - The left side contains three fruits: a peach and two oranges.
 - The right side contains granola with nuts and dried fruit.
- Second Image Analysis**:
 - The left side now contains granola with nuts and dried fruit.
 - The right side contains the three fruits (peach and two oranges).
- Comparison**:
 - In the first image, the fruits were on the left and the granola was on the right.
 - In the second image, the positions of these components have been swapped.

Based on this analysis, the content has been rearranged, which indicates an anomaly in the second image. The locations of the abnormal areas can be identified as follows:

- The area that originally contained fruits is now filled with granola: [50, 28, 256, 395].
- The area that originally contained granola is now filled with fruits: [222, 33, 515, 395].

</think><answer>Yes. {'bbox_2d': [50, 28, 256, 395], 'label': 'Component Swapped'}, {'bbox_2d': [222, 33, 515, 395], 'label': 'Component Swapped'}</answer>

<AD and AL accuracy of Anomaly-R1 vs. generalist MLLMs>

07

One-to-More: High-Fidelity Training-Free Anomaly Generation with Attention Control

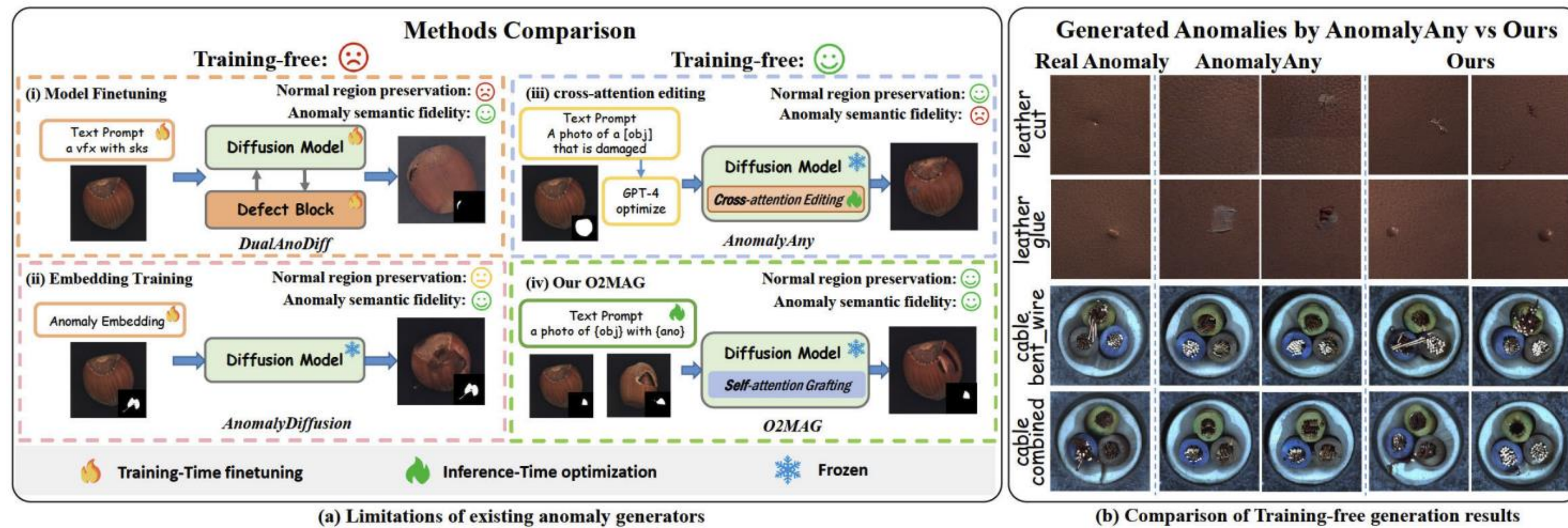
Haoxiang Rao, Zhao Wang, Chenyang Si, Yan Lyu, Yuanyi Duan, Fang Zhao,
Caifeng Shan

CVPR 2026

One-to-More: High-Fidelity Training-Free Anomaly Generation with Attention Control

Problem

- Few-shot anomaly generation needs per-category training (defect blocks, textual inversion), which overfits and scales poorly.
- Existing training-free generators stay far from the real anomaly distribution, yielding low-fidelity defects.
- Goal: synthesize many faithful anomalies from one reference anomalous image, with no training at all.

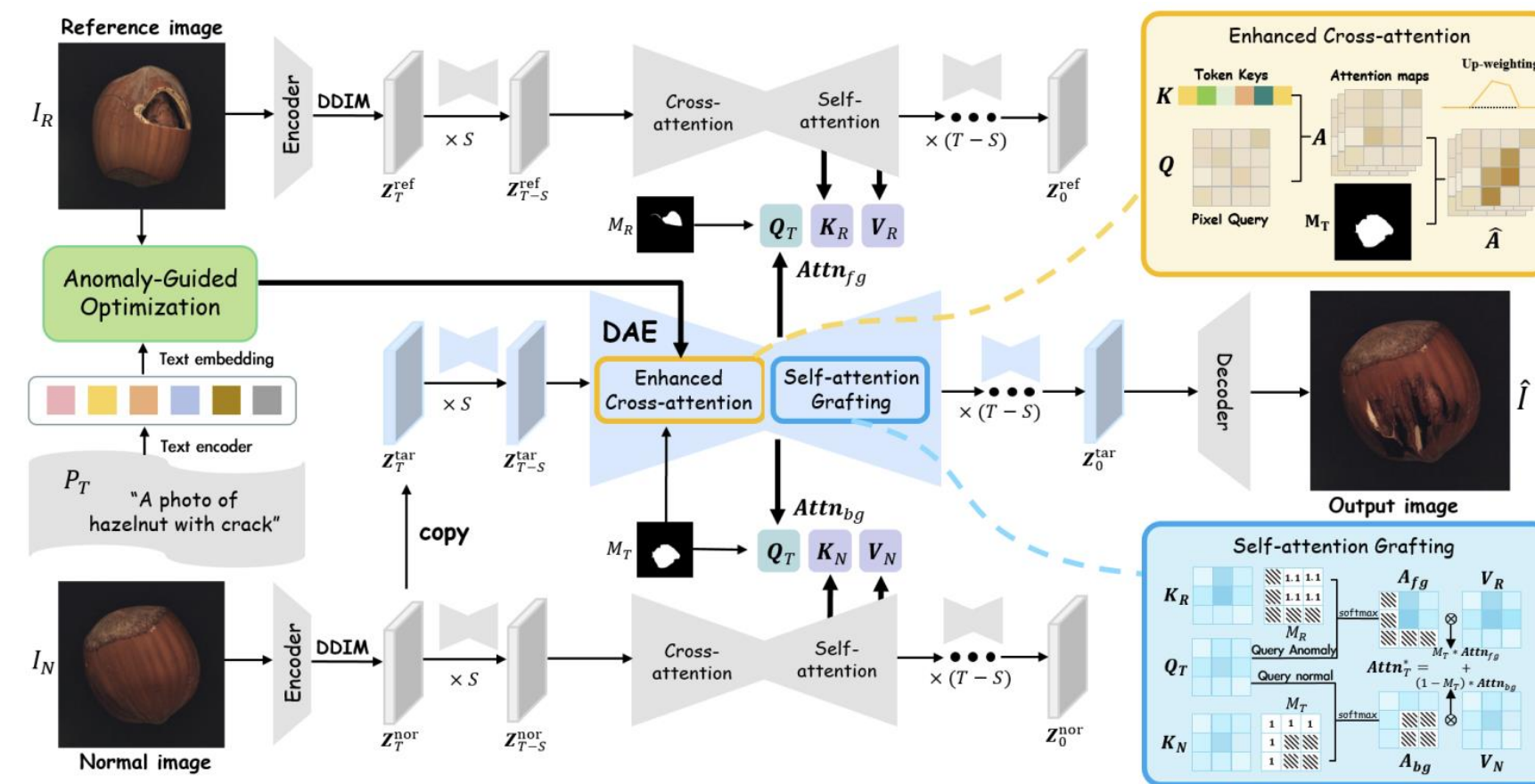


<Comparison of diffusion-based anomaly generation paradigms>

One-to-More: High-Fidelity Training-Free Anomaly Generation with Attention Control

Method

- O2MAG coordinates self-attention across three parallel diffusion processes, grafting reference attention into the synthesis branch.
- Dual-Attention Enhancement (DAE) applies the anomaly mask on self-/cross-attention to enforce full-mask filling and kill FG-BG query confusion.
- Anomaly-Guided Optimization (AGO) refines the anomaly text embedding toward the target anomalous distribution.

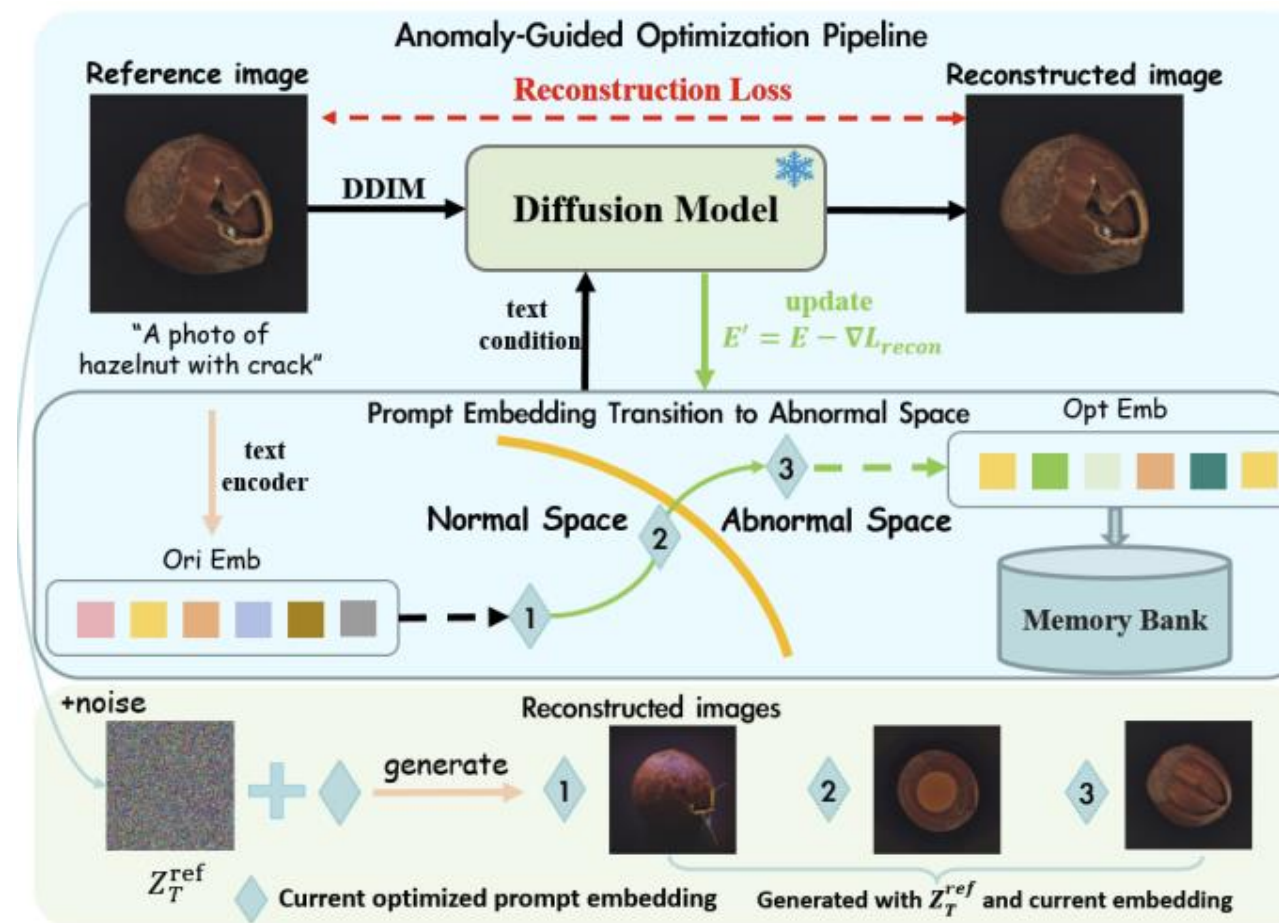


<Overview of the proposed O2MAG>

One-to-More: High-Fidelity Training-Free Anomaly Generation with Attention Control

Results & Contribution

- Higher-fidelity, text-consistent anomalies than prior few-shot / training-free generators, with zero category-specific training.
- Synthesized samples improve downstream anomaly detection and localization.
- Shows attention control alone suffices to bind defect semantics to the correct spatial support.



<Anomaly-Guided Optimization (AGO) pipeline>

08

ANTS: Adaptive Negative Textual Space Shaping for OOD Detection via Test-Time MLLM Understanding and Reasoning

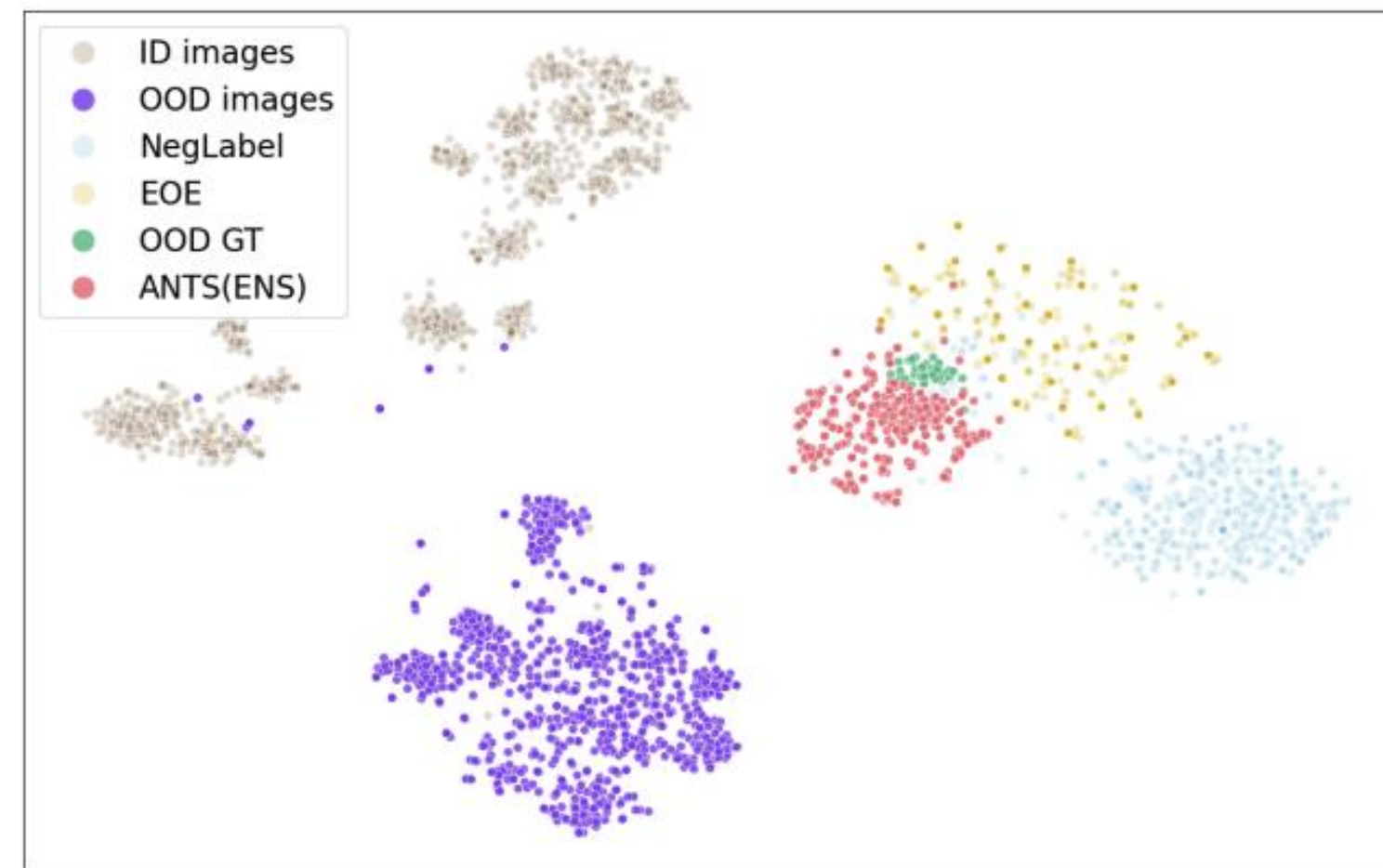
Wenjie Zhu, Yabin Zhang, Xin Jin, Wenjun Zeng, Lei Zhang

CVPR 2026 (Oral)

ANTS: Adaptive Negative Textual Space Shaping for OOD Detection

Problem

- Negative labels help OOD detection, but are built without understanding the actual OOD images, giving an inaccurate negative space.
- The absence of negative labels semantically close to ID classes structurally limits near-OOD detection.
- A single fixed negative space cannot serve near-OOD and far-OOD simultaneously.

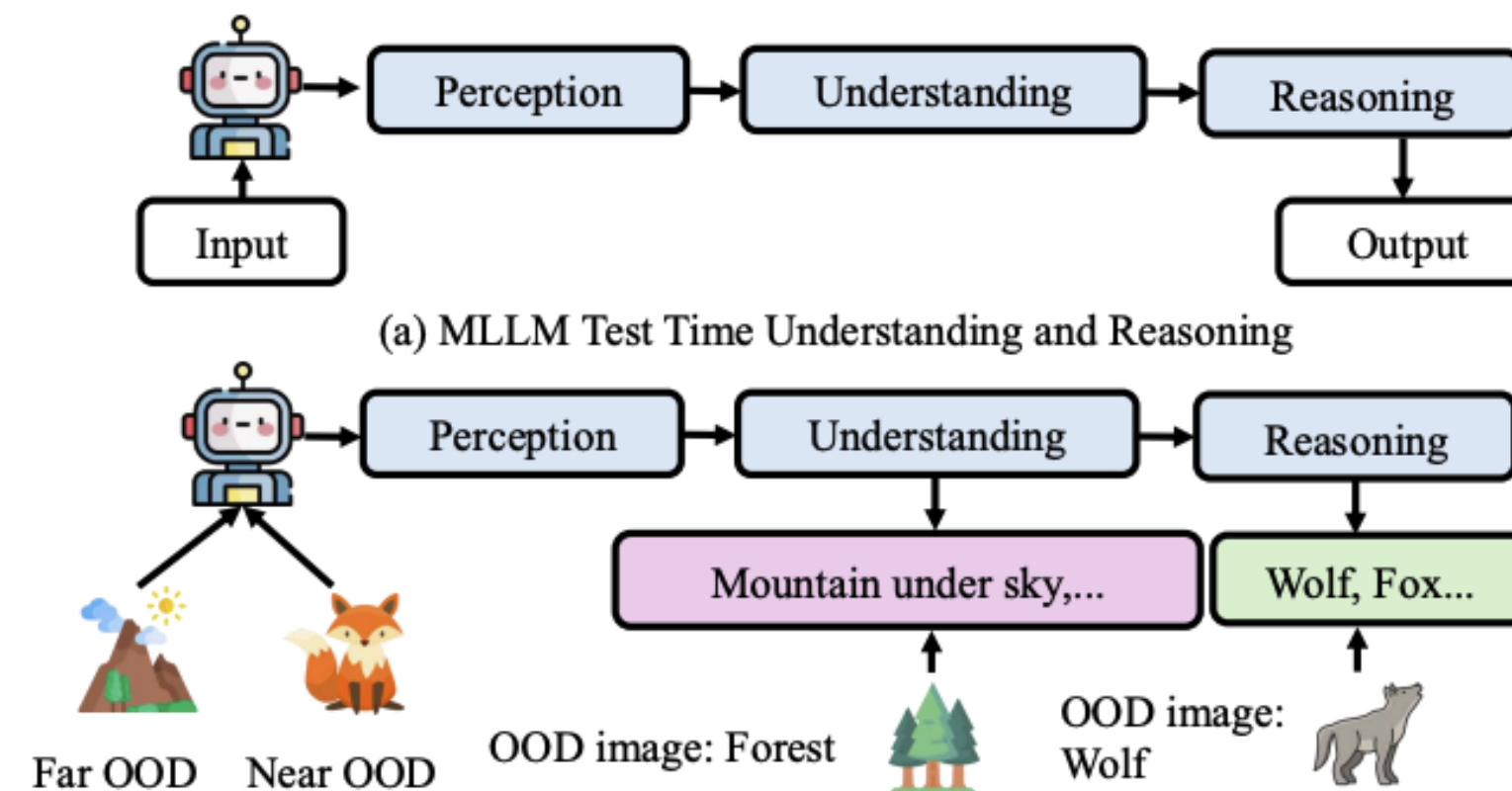


<Limitation of a fixed negative textual space>

ANTS: Adaptive Negative Textual Space Shaping for OOD Detection

Method

- Cache likely-OOD images from the test stream and prompt an MLLM to describe them: expressive negative sentences for far-OOD.
- Cache the ID subset visually similar to test images; MLLM reasoning then synthesizes similar negative labels for near-OOD.
- Fuse the two negative textual spaces with an adaptive weighted score; the whole procedure is training-free and zero-shot.



<Overview of the ANTS framework>

ANTS: Adaptive Negative Textual Space Shaping for OOD Detection

Results & Contribution

- Reduces FPR95 by 3.1% on the ImageNet benchmark, establishing a new state of the art.
- Adaptive weighting handles both near- and far-OOD without task-specific tuning.
- Training-free and zero-shot, hence highly scalable in open environments.

Methods	OOD datasets									
	INaturalist		SUN		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Training-required (or with Fine-tuning)										
MSP [13]	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
ZOC [11]	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
CLIPN [47]	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
LSN [34]	95.83	21.56	94.35	26.32	91.25	34.48	90.42	38.54	92.26	30.22
LoCoOp [33]	93.93	29.45	90.32	41.13	90.54	44.15	93.24	33.06	92.01	36.95
ID-Like [2]	98.19	8.98	91.64	42.03	90.57	44.00	94.32	25.27	93.68	30.07
NegPrompt [25]	90.49	37.79	92.25	32.11	91.16	35.52	88.38	43.93	90.57	37.34
SCT [52]	95.86	13.94	95.33	20.55	92.24	29.86	89.06	41.51	93.27	26.47
LAPT [58]	99.63	1.16	96.01	19.12	92.01	33.01	91.06	40.32	94.68	23.40
CMA [20]	99.62	1.65	96.36	16.84	93.11	27.65	91.64	33.58	95.13	19.93
SynOOD [24]	99.57	1.57	95.82	20.46	97.37	12.12	95.29	22.94	97.01	14.27
Zero Shot (No Training Required)										
MCM [30]	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
CoVer [55]	95.98	22.55	93.42	32.85	90.27	40.71	90.14	43.39	92.45	34.88
EOE [4]	97.52	12.29	95.73	20.40	92.95	30.16	85.64	57.63	92.96	30.09
NegLabel [19]	99.49	1.91	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
OODD [51]	99.36	2.22	95.01	21.49	87.10	44.76	93.27	30.69	93.69	24.79
AdaNeg [57]	99.71	0.59	97.44	9.50	94.55	34.34	94.93	31.27	96.66	18.92
CSP [5]	99.60	1.54	96.66	13.66	92.90	29.32	93.86	25.52	95.76	17.51
ANTS	99.75	0.54	98.77	5.43	96.10	20.21	96.38	18.52	97.75	11.20

<Comparison on the ImageNet OOD detection benchmark>

09

AnomalyVFM: Transforming Vision Foundation Models into Zero-Shot Anomaly Detectors

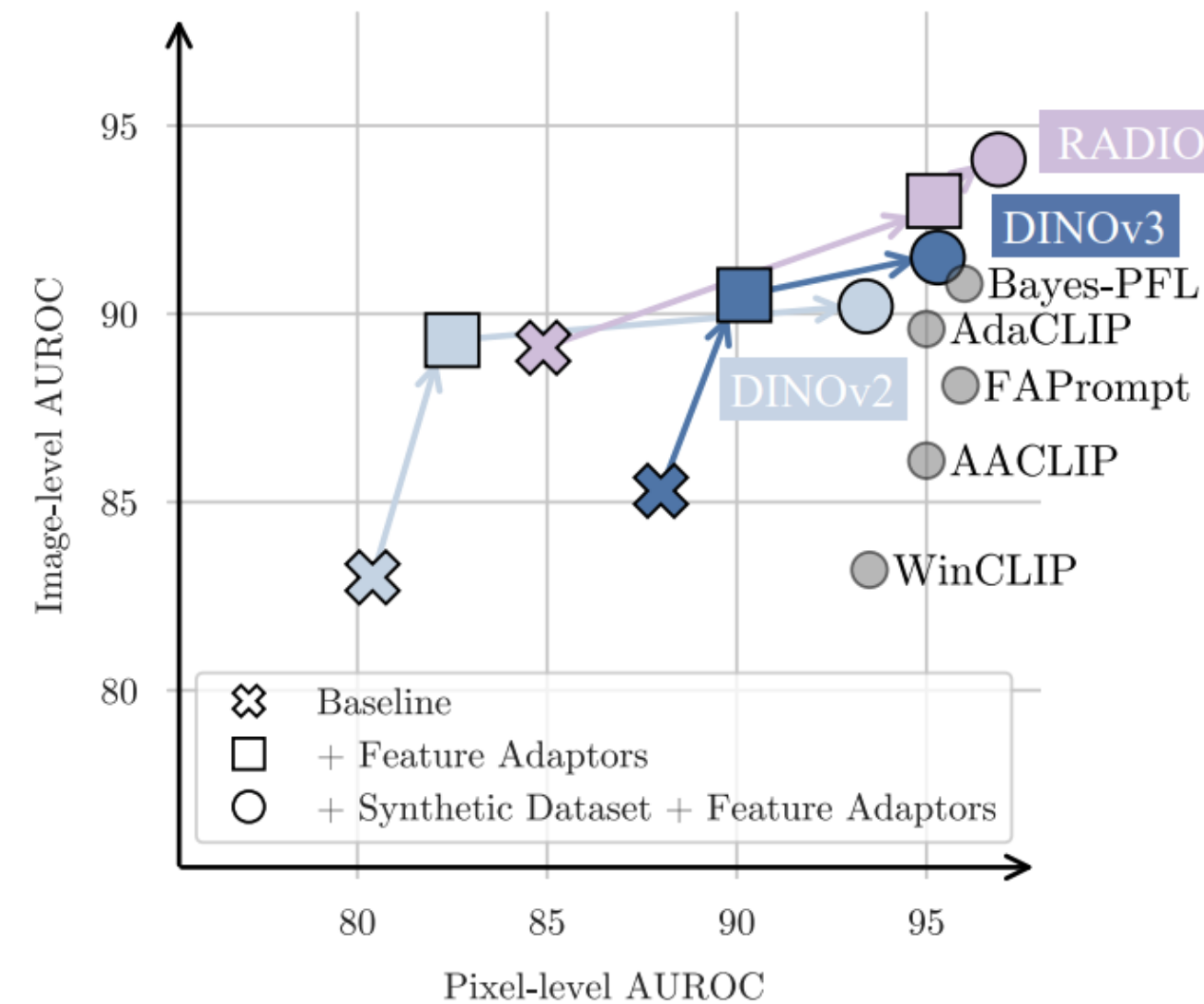
Matic Fucka, Vitjan Zavrtanik, Danijel Skocaj

CVPR 2026

AnomalyVFM: Transforming Vision Foundation Models into Zero-Shot Anomaly Detectors

Problem

- Zero-shot AD is dominated by VLMs such as CLIP, while purely visual foundation models (DINOv2) lag behind.
- The gap is not representational: it stems from low-diversity auxiliary AD data and overly shallow VFM adaptation.
- Question: can any pretrained VFM become a strong zero-shot anomaly detector without language supervision?

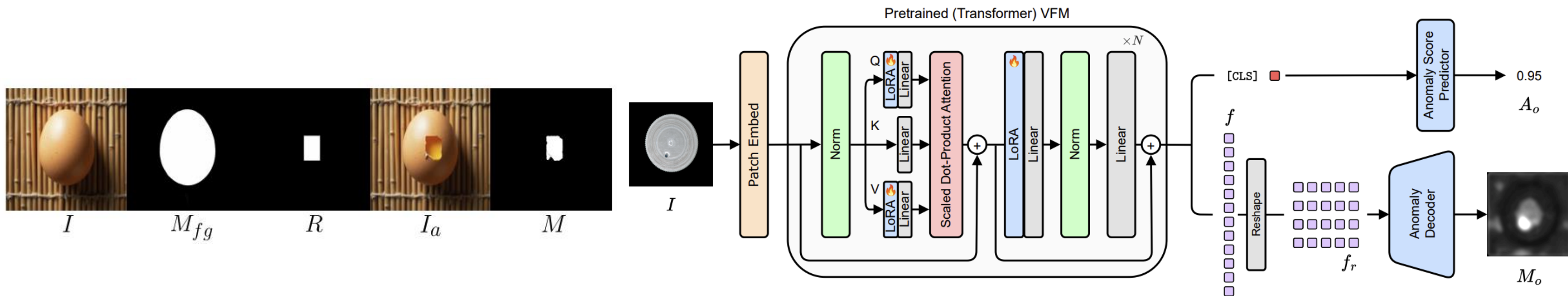


<Untapped potential of purely visual foundation models>

AnomalyVFM: Transforming Vision Foundation Models into Zero-Shot Anomaly Detectors

Method

- Three-stage synthetic data pipeline: generate anomaly-free images, sample a foreground region, synthesize anomalies with masks.
- Parameter-efficient adaptation: low-rank feature adapters (LoRA, rank 64) on the frozen VFM plus a light convolutional decoder.
- Confidence-weighted pixel loss suppresses gradients from unreliable synthetic mask pixels.



<Dataset generation pipeline and architecture of AnomalyVFM>

AnomalyVFM: Transforming Vision Foundation Models into Zero-Shot Anomaly Detectors

Results & Contribution

- With a RADIO backbone, 94.1% mean image-level AUROC over 9 diverse datasets, surpassing prior work by 3.3 pp.
- Backbone-agnostic: consistent gains across different vision foundation models.
- Demonstrates that language supervision is not required for state-of-the-art zero-shot AD.

Model	Additions		Image-level		Pixel-level	
	SD	FA	AUROC	F_1 -Max	AUROC	F_1 -Max
DINOv2 [50]			83.0	78.9	80.4	23.9
	✓		86.4 ↑ 3.4	79.3 ↑ 0.4	93.2 ↑ 12.8	41.2 ↑ 17.3
		✓	89.3 ↑ 6.3	82.7 ↑ 3.7	82.5 ↑ 2.1	27.1 ↑ 3.2
DINOv3 [62]	✓	✓	90.2 ↑ 7.2	83.4 ↑ 4.5	93.4 ↑ 13.0	41.7 ↑ 17.8
			85.3	80.1	88.0	32.5
	✓		89.0 ↑ 3.7	82.0 ↑ 1.9	95.0 ↑ 7.0	44.9 ↑ 12.4
RADIO [56]		✓	90.5 ↑ 5.2	83.9 ↑ 3.8	90.2 ↑ 2.2	39.3 ↑ 7.3
	✓	✓	91.5 ↑ 6.2	84.7 ↑ 4.6	95.3 ↑ 7.2	44.6 ↑ 12.1
			89.1	83.9	84.9	30.8
RADIO [56]	✓		92.1 ↑ 3.0	87.2 ↑ 3.3	95.9 ↑ 11.0	43.2 ↑ 12.6
		✓	93.0 ↑ 3.9	88.9 ↑ 5.0	95.2 ↑ 10.3	42.8 ↑ 12.0
	✓	✓	94.1 ↑ 5.0	87.6 ↑ 3.7	96.9 ↑ 12.0	44.3 ↑ 13.5

<Zero-shot anomaly detection results across 9 datasets>

10

Towards Open-Vocabulary Industrial Defect Understanding with a Large-Scale Multimodal Dataset (IMDD-1M)

TsaiChing Ni, ZhenQi Chen, YuanFu Yang

CVPR 2026

IMDD-1M: Towards Open-Vocabulary Industrial Defect Understanding

Problem

- Industrial defect understanding remains closed-set: every new defect type demands retraining.
- No large-scale image-text corpus exists for open-vocabulary defect understanding; benchmarks are 100x smaller and lack text.
- Expert-verified descriptions of defect location, severity and context are essentially unavailable.

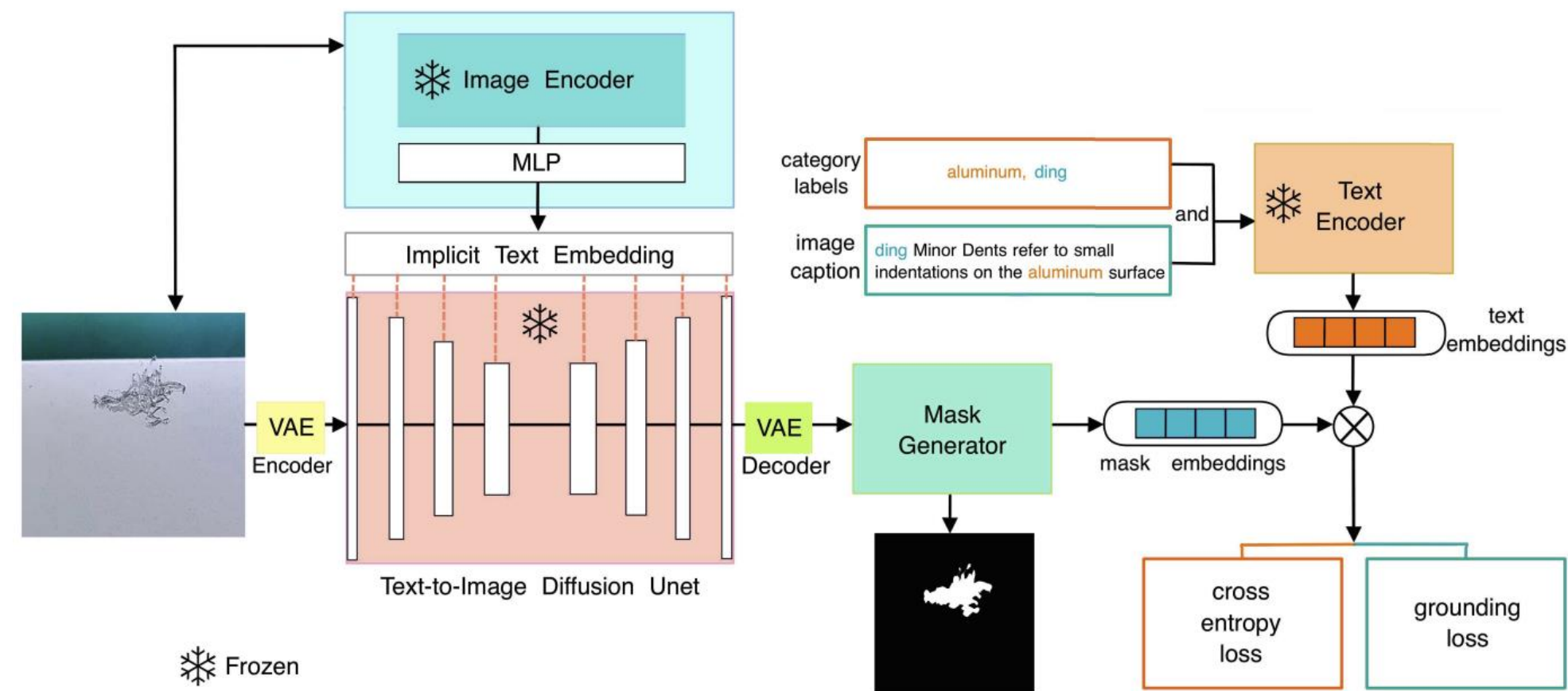


<Statistics of the IMDD-1M dataset>

IMDD-1M: Towards Open-Vocabulary Industrial Defect Understanding

Method

- IMDD-1M: 1.24M aligned image-text pairs, 421 defect types, 63 industrial domains, expert-verified fine-grained captions.
- A diffusion-based VL foundation model (890M) from scratch: implicit captioner + frozen diffusion U-Net (860M) + Mask2Former (45M).
- Two-stage training: diffusion pretraining on IMDD-1M (100 epochs), then freeze it and fine-tune only the mask generator.

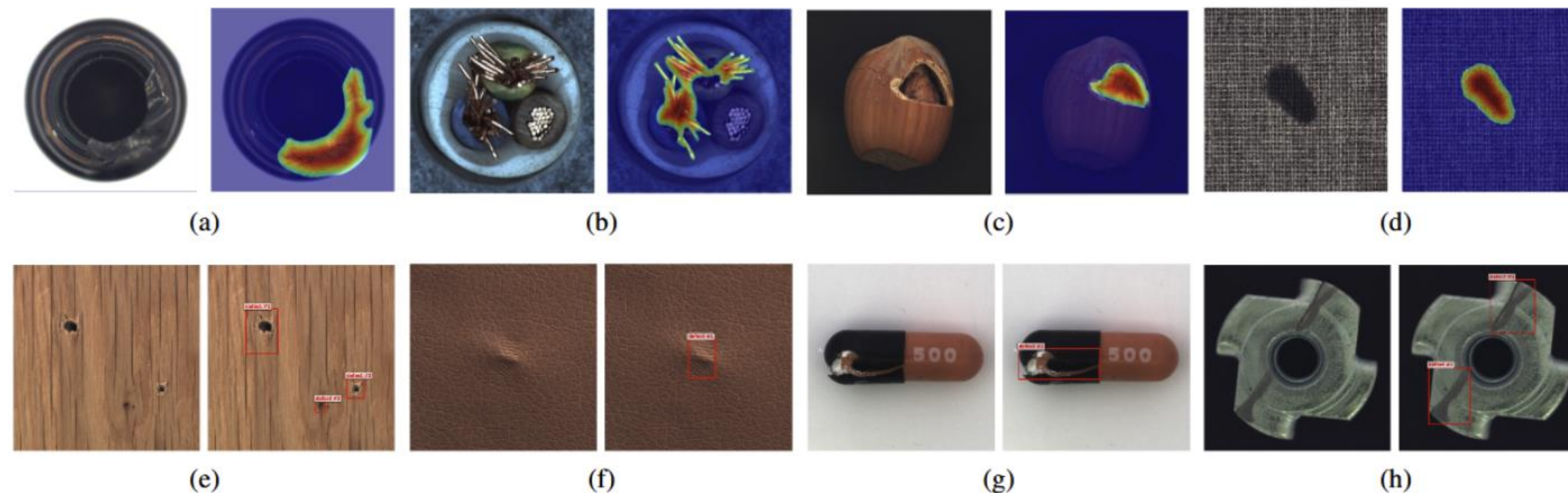


<Overview of the proposed framework>

IMDD-1M: Towards Open-Vocabulary Industrial Defect Understanding

Results & Contribution

- 96.1% accuracy with 200 samples/class (<5% of supervised data); 96.7% classification, 52.9% IoU, 74.6% mAP@0.5, FID 5.5-13.6.
- Ablation: removing diffusion conditioning costs 7.0 pp accuracy - large-scale pretraining is the key factor.
- One architecture covers classification, detection, segmentation and text-guided generation without task-specific modules.



<Qualitative visualization of multimodal results on various MVTec AD samples>

Thank you

Question & Discussion



RILS LAB