Towards Open-Vocabulary Industrial Defect Understanding with a Large-Scale Multimodal Dataset

Abstract（CVPR2026 TsaiChing Ni, ZhenQi Chen, YuanFu Yang）

We present IMDD-1M, the **first large-scale Industrial Multimodal Defect Dataset comprising 1,000,000 aligned image-text pairs**, designed to advance multimodal learning for manufacturing and quality inspection. IMDD-1M contains high-resolution real-world defects spanning over 60 material categories and more than 400 defect types, each accompanied by expert-verified annotations and fine-grained textual descriptions detailing defect location, severity, and contextual attributes. **This dataset enables a wide spectrum of applications, including classification, segmentation, retrieval, captioning, and generative modeling.**

Building upon IMDD-1M**, we train a diffusion-based vision-language foundation model from scratch,** specifically tailored for industrial scenarios. The model serves as a generalizable foundation that can be efficiently adapted to specialized domains through lightweight fine-tuning. **With less than 5% of the task-specific data required by dedicated expert models,** it achieves comparable performance, highlighting the potential of data-efficient foundation model adaptation for industrial inspection and generation, paving the way for scalable, domain-adaptive, and knowledge-grounded manufacturing intelligence.

If a model can paint a defect, it understands it.

1 · UNDERSTAND

TEACHER → *distill* STUDENT

2 · DISTILL

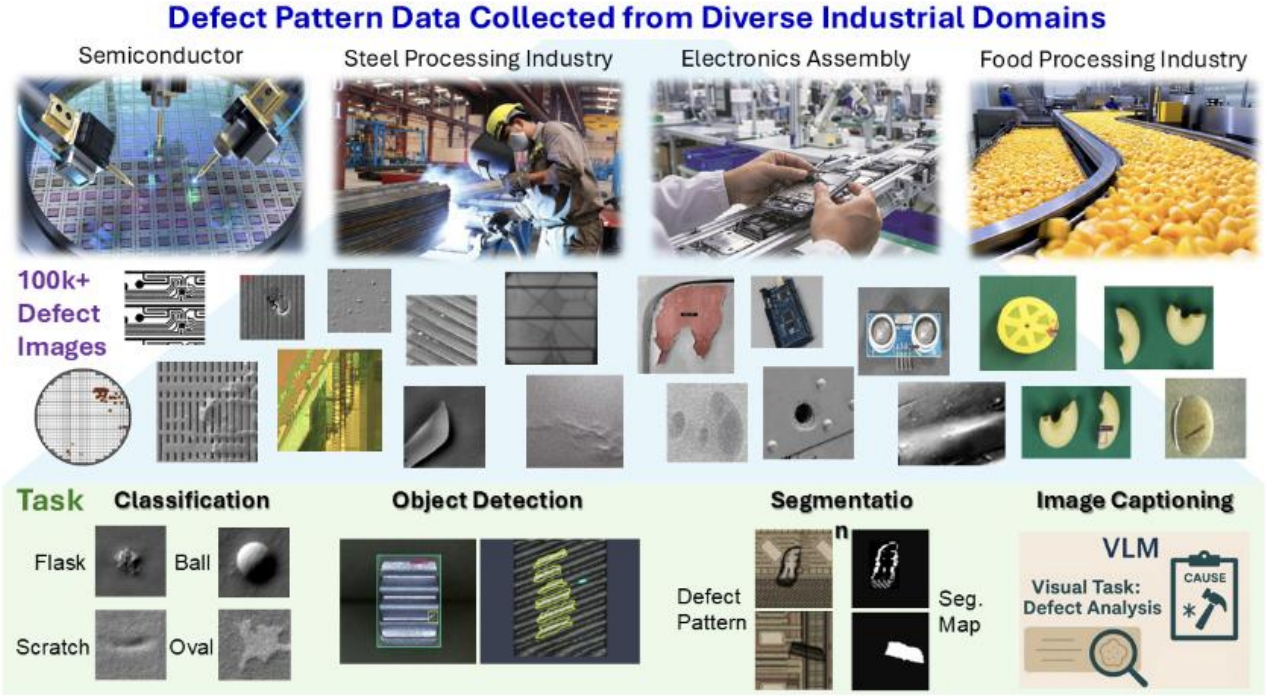
100–200 images
→ ~90% accuracy

3 · DEPLOY

From lab to line. From rules to understanding.

Research Overview （CVPR2026 Tsai-Ching Ni）

Towards Open-Vocabulary Industrial Defect Understanding with a Large-Scale Multimodal Dataset



IMDD-1M Dataset

- 1.24M image-text pairs
- 63 industrial domains
- 421 defect types

Goal

Build a **multimodal foundation model** for industrial defect understanding.

Method & Results

Diffusion-based multimodal framework

- Diffusion U-Net encoder
- Implicit captioner
- Mask2Former generator

Performance

- Classification: **96.7% accuracy**
- Object detection, Segmentation: SOTA
- Only **<5% labeled data required**