PROTOTYPE-GUIDED DYNAMIC DOMAIN MODULATION AND A LARGE-SCALE MULTI-SKU BENCHMARK FOR SHAMPOO BOTTLE SURFACE DEFECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Industrial shampoo-bottle surface defect detection must operate across many stock keeping units (SKUs), where differences in bottle geometry, material, label layout, reflectance, and imaging conditions can substantially alter the appearance of the same defect. Existing industrial defect datasets rarely provide both instance-level defect annotations and product-variant identities, making it difficult to study SKU-aware detection under realistic manufacturing conditions. We introduce Bottle-SKU, a large-scale multi-SKU benchmark for shampoo bottle surface defect detection. Bottle-SKU contains 13,119 images, 20,419 instance-level bounding boxes, seven defect categories, and 48 bottle SKUs, enabling evaluation of both shared defect recognition and SKU-specific appearance variation. To address the challenge of SKU-specific appearance variation, we propose Prototype-Guided Dynamic Domain Modulation (ProDMM). ProDMM constructs a visual prototype for each SKU from training images and predicts a soft SKU distribution for each input. The resulting prototype-conditioned representation dynamically generates low-rank Query/Value modulation for the visual backbone and residual conditioning for defect-text-to-visual attention, while preserving the pretrained Grounding DINO detection pathway. The model is trained jointly with detection, SKU-domain classification, and modulation-regularization objectives. Experiments on Bottle-SKU and the public GC10-DET-SKU metallic surface defect dataset demonstrate the effectiveness of prototype-guided domain conditioning for industrial surface defect detection.

1 INTRODUCTION

Industrial visual inspection promises scalable quality control, yet models that perform well in a controlled production setting can degrade when the product, material, illumination, or imaging configuration changes. Reviews of surface-defect detection identify appearance variation, subtle defect boundaries, and strong dependence on acquisition conditions as recurring obstacles to deployment Bhatt et al. (2021). Industrial domain-adaptation studies likewise report that a detector trained on one inspection distribution requires explicit adaptation when transferred to another Li et al. (2023a). This problem is particularly visible on shampoo-bottle production lines, where one system serves many Stock Keeping Units (SKUs). Changes in bottle geometry, label layout, surface reflectance, viewpoint, and camera elevation can alter the texture, contrast, and geometric context of the same semantic defect.

This deployment problem is more specific than a shortage of labeled images. Adding samples from already observed products can improve interpolation, but it does not by itself establish transfer to a newly introduced SKU. A random image-level split can also place nearly identical product appearances in both training and test sets, thereby measuring interpolation rather than cross-product generalization. The desired detector must therefore separate two factors. Defect semantics should be shared across products, whereas the appearance prior should change with the current product variant. This motivates an evaluation protocol in which product variants are explicit domains and held-out variants are used to test generalization.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

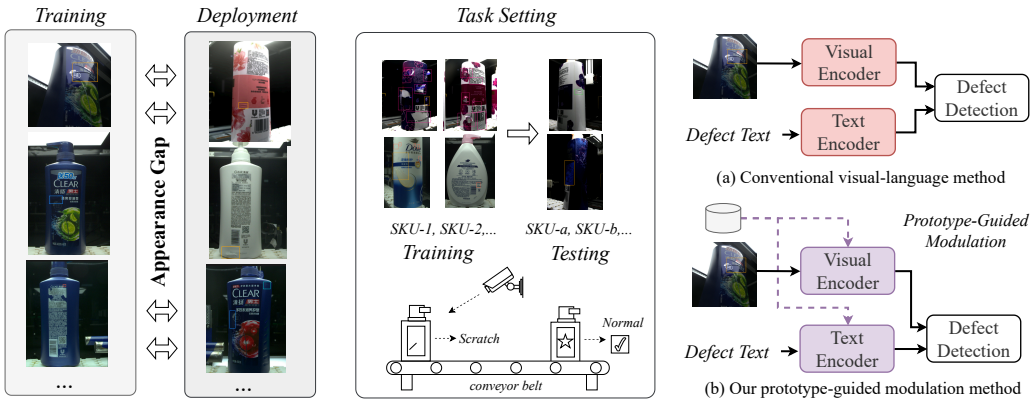


Figure 1: **Overview of the proposed cross-SKU defect detection benchmark and method.** (**Left**) *Appearance gap* between training and deployment. Shampoo bottles from different SKUs vary substantially in appearance and material, while defects such as scratches and wrinkles remain visually similar. (**Middle**) *Task setting*. We adopt an SKU-level deployment protocol, training on a set of seen SKUs (*SKU-1, SKU-2, ...*) and testing on unseen SKUs (*SKU-a, SKU-b, ...*), which emulates the real-world scenario where a production line switches to a new product. (**Right**) (a) Conventional visual-language methods feed visual and text features directly into the detection head. (b) Our prototype-guided modulation injects domain shift into the text and visual branches, enabling robust detection on unseen SKUs.

Domain adaptation provides an important starting point. Adversarial alignment methods learn representations that make domains difficult to distinguish Ganin et al. (2016). Detection-specific methods align image-level and instance-level features Chen et al. (2018). Multi-source adaptation combines information from several labeled domains Yao et al. (2021). These approaches are not a direct solution to the multi-SKU setting. They commonly formulate a source–target transfer problem or seek a domain-invariant representation, while a production line contains many related variants and the appearance differences themselves can provide useful evidence for resolving ambiguous defects. Recent open-set defect recognition also broadens the range of defect appearances that a model can recognize Kim et al. (2026), but does not establish an instance-level, SKU-disjoint detection protocol with explicit product-variant conditioning.

Vision-language detectors offer a complementary source of transferable defect semantics. Grounding DINO Liu et al. (2024) and subsequent systems use text-conditioned visual interaction to localize concepts beyond a fixed detector vocabulary. Recent work improves this paradigm through large-language-model supervision Fu et al. (2025); Zhou et al. (2026), retrieval-based open-vocabulary detection Fu et al. (2026a), semantic augmentation Cao et al. (2026), visual-token prompt refinement Qin et al. (2026), richer category descriptions Yang et al. (2026), and hierarchical objectness calibration Lee et al. (2026). These methods strengthen the semantic interface between a defect prompt and visual evidence, but the text prompt for "scratch" or "fold" is still shared across SKUs. Thus, language conditioning alone does not tell the detector how the same defect should appear on a particular product variant, nor does it provide a mechanism for soft transfer among visually related SKUs.

Existing defect benchmarks leave this product-variant setting under-specified. NEU-DET and GC10-DET focus on defects in steel-strip imagery Song & Yan (2013); Lv et al. (2020), HRIPCB targets PCB defects Huang et al. (2020a), and MVTec AD emphasizes unsupervised anomaly localization across object and texture categories Bergmann et al. (2019). These benchmarks are valuable for measuring defect localization, but they generally do not jointly expose full-product appearance variation, explicit product-variant identities, instance-level defect boxes, and a product-disjoint test protocol. As a result, they cannot directly answer whether a detector can preserve defect semantics while adapting to the appearance of an unseen product variant.

To study this problem, we introduce Bottle-SKU, a multi-SKU shampoo-bottle surface-defect benchmark with instance-level annotations and explicit SKU identities, and propose Prototype-Guided Dynamic Domain Modulation (ProDMM). ProDMM extracts a prototype for each training

SKU, predicts a soft SKU posterior for an input image, and uses the resulting mixture of prototypes to condition low-rank visual modulation and defect-text cross-attention. This design retains a shared detection pathway for defect semantics while allowing the appearance-dependent computation to vary smoothly across related products. We evaluate the proposed setting on Bottle-SKU and on a pseudo-SKU split of GC10-DET-SKU. Our contributions are threefold:

- We introduce Bottle-SKU, a large-scale full-object surface-defect benchmark with explicit SKU identities, professional instance-level annotations, and an SKU-disjoint evaluation protocol.
- We propose ProDMM, which converts a soft mixture of SKU prototypes into image-conditioned low-rank QLer/VLer residuals for the frozen visual backbone, preserving the shared pretrained detection pathway.
- We introduce Domain Cross Attention (DCA), which injects prototype-conditioned visual evidence into defect-text interaction, and demonstrate the resulting framework on Bottle-SKU and GC10-DET-SKU under product-aware evaluation.

2 Related Work

Industrial Surface Defect Detection. Industrial inspection includes supervised instance detection and unsupervised anomaly localization. NEU-DET Song & Yan (2013) established a widely used hot-rolled steel benchmark with six visually distinct defect types. GC10-DET Lv et al. (2020) broadened this setting to ten metallic-surface categories and high-resolution strip images. HRIPCB Huang et al. (2020a) emphasizes fine-grained PCB defects, while MPCG Tao et al. (2018) studies cover-glass defects under reflective imaging. PVMulti-Defect Li et al. (2023b) and APDDD Alibaba Tianchi (2023) represent photovoltaic and aluminum-profile inspection, respectively, and MVTec AD Bergmann et al. (2019) provides a multi-object benchmark for unsupervised anomaly detection. Severstal Severstal (2019) contributes large-scale production steel imagery, whereas DAGM Wieler et al. (2007) isolates textured optical inspection with controlled backgrounds.

In parallel, anomaly methods learn normal appearance rather than enumerating every defect type. PaDiM Defard et al. (2021) fits a multivariate Gaussian to deep patch features, PatchCore Roth et al. (2022) stores representative normal patches in a memory bank for nearest-neighbor scoring, CutPaste Li et al. (2021) creates synthetic anomalies through self-supervised transformations, and DRAEM Zavrtanik et al. (2021) trains a reconstruction-and-discrimination network on augmented defects. MVTec LOCO AD Bergmann et al. (2020) further tests logical and structural anomalies that cannot be detected from local texture alone. These datasets and methods usually assume a fixed product identity or remove object-level context through local crops. In contrast, our proposed Bottle-SKU dataset evaluates full-object, instance-level detection across explicit SKU identities, exposing a cross-domain generalization problem not addressed by standard benchmark protocols.

Prototype-based adaptation and conditional modulation. Prototype-based learning Snell et al. (2017) and conditional modulation Perez et al. (2018); Yang et al. (2019) adapt representations using representative embeddings or auxiliary signals. Domain-adaptation methods Ganin et al. (2016); Chen et al. (2018) align source and target features, while parameter-efficient transfer Houlsby et al. (2019); Hu et al. (2022) injects lightweight trainable modules into frozen backbones. Recent works explore dynamic modulation for domain adaptation and generalization. Feature-level modulation adapts visual features to domain shifts via semantic entropy guidance, channel-wise compensation, content-style decomposition, and concept-aware modulation Li et al. (2025); Song et al. (2026); Guo et al. (2026); Sultana et al. (2026). Conditional modulation generates input-dependent parameters or adapters to modulate frozen backbones Khan et al. (2026); Aiello et al. (2025); Zhang et al. (2025). Prototype-guided and representation-level modulation adjusts statistical properties or activation patterns for cross-domain robustness Cai & Ye (2026); Gan et al. (2026); Kundu et al. (2026). Unlike these approaches, our Prototype-Guided Dynamic Domain Modulation generates domain shift information. See Appendix A for a comprehensive review.

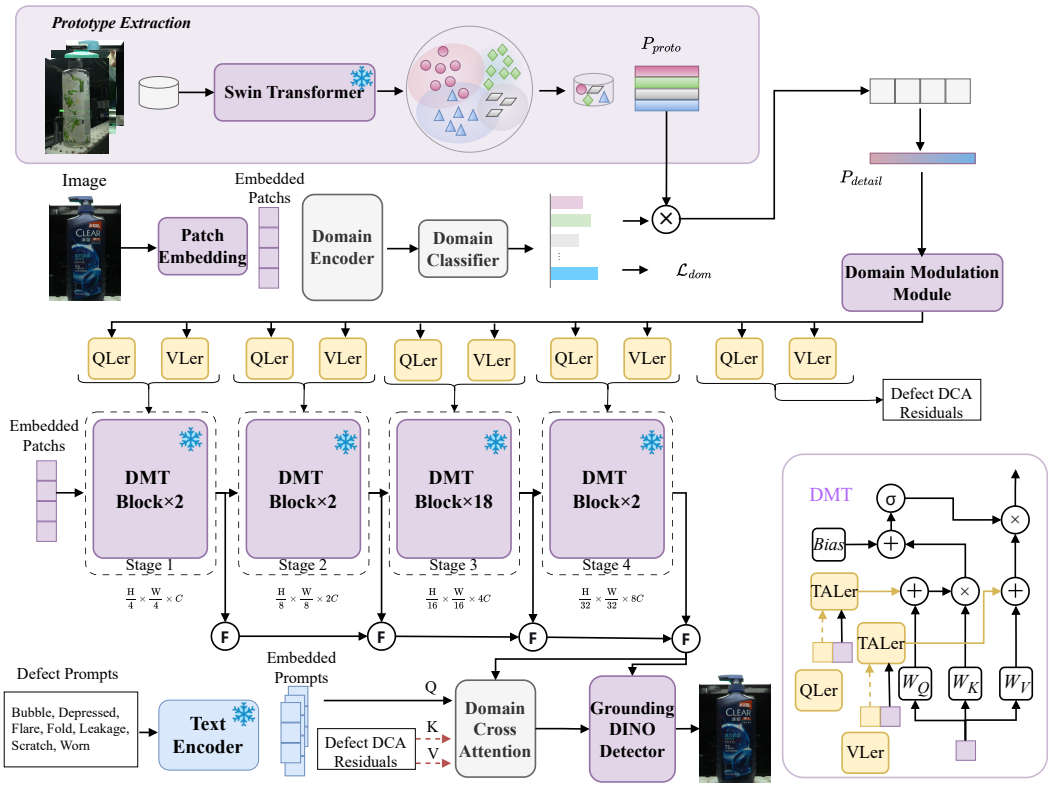


Figure 2: The overall architecture of ProDMM. Offline SKU prototypes condition dynamic QLer/VLer residuals in the Swin backbone and residuals for defect Domain Cross Attention (DCA). DCA updates only the text memory before the original GroundingDINO Feature Enhancer. The bottom-right inset illustrates a DMT block.

3 Our Proposed Approach

3.1 Overview

We propose a prototype-conditioned domain modulation framework for multi-SKU defect detection. In the main frozen-backbone setting, the pretrained Swin-B backbone and BERT text encoder are frozen, while the dynamic modulation parameters in DMT, the Domain Cross Attention (DCA), and the GroundingDINO interaction layers are trained. We first construct a fixed SKU prototype matrix by offline. Then, the framework extracts patch tokens of input image and uses a Domain Encoder to predict a soft SKU probability vector. The predicted probabilities softly fuse all SKU prototypes into a domain detail representation, then, this representation drives Domain Modulation Module to generate QLer/VLer modulation coefficients for all Swin-B blocks. The DCA is used to enhance text and visual tokens, then processed by the frozen GroundingDINO Detector to produce defect detection outputs. The whole model is optimized with the detection loss, an SKU-domain classification loss, and regularization terms for the generated modulation parameters. The overall framework is shown in Fig. 2.

3.2 SKU Prototype Construction

Given input images $I \in \mathbb{R}^{B\times C\times H\times W}$ and defect text prompts, our goal is to perform defect detection and SKU domain classification in a multi-SKU setting. For the j-th training image $I_{u,j}$ of the u-th SKU, we use the official GroundingDINO Swin-B weights to extract the final visual stage feature for each image to obtain $f_{u,j}$. The SKU prototype is computed as the mean of the image

features:

$$P_u = (\frac{1}{M_u}\sum_{j=1}^{M_u} f_{u,j}), \tag{1}$$

where $M_u \leq M_{\max}$ at most $M_{\max}$ training images are used for each SKU. Stacking the prototypes of all U SKUs gives the prototype features

$$P_{proto} = [P_1; \ldots; P_u], \tag{2}$$

where $P_{proto} \in \mathbb{R}^{U \times d_p}$ is computed offline from training images and kept fixed during detector optimization. No image from a held-out test SKU is used to construct this table. The table provides a compact appearance prior while leaving the defect categories shared across SKUs.

3.3 Prototype-conditioned Domain Modulation

The Domain Encoder operates on the patch tokens produced by the frozen Swin-B Patch Embedding. The input images is embedded as:

$$Z_0 = \mathrm{PatchEmbed}(I) \in \mathbb{R}^{B \times N \times d_e}, \tag{3}$$

where B is the batch size, N is the number of patch tokens, and d_e is the patch embedding dimension. A 1×1 convolution is then applied to Z_0 to extract gate features $G \in \mathbb{R}^{B \times N \times d_g}$. These features are concatenated with Z_0 along the channel dimension, yielding $[Z_0; G] \in \mathbb{R}^{B \times N \times (d_e + d_g)}$. A single-layer MLP followed by global average pooling over tokens projects the concatenated features to a domain vector g, the Domain Encoder is defined as follows:

$$g = AvgPool(\mathrm{MLP}\left([Z_0; G]\right)), g \in \mathbb{R}^{B \times d_0}, \tag{4}$$

$$h_{dom} = \mathrm{GELU}(\mathrm{MLP}(g)), h_{dom} \in \mathbb{R}^{B \times d_{dom}}, \tag{5}$$

where $[\cdot;\cdot]$ denotes channel concatenation and h_{dom} is the domain feature. The SKU domain classifier consists of a linear layer followed by softmax and outputs the soft SKU probability p:

$$p = \mathrm{Classifier}(h_{dom}), p \in \mathbb{R}^{B \times U}. \tag{6}$$

Instead of using a hard-selected prototype, ProDMM softly fuses all SKU prototypes using the predicted probabilities, $P_{detail} = p \cdot P_{proto}$, so the model receives a continuous domain representation containing information from all SKU prototypes. Domain Modulation Module (DMM) then maps P_{detail} to the global domain representation $\mathbf{c}$:

$$\mathbf{c} = \mathrm{DMM}(P_{detail}), \mathbf{c} \in \mathbb{R}^{B \times d_{dom}}, \tag{7}$$

where DMM is a two-layer MLP with GELU. Thus, all subsequent dynamic parameters are conditioned on global representation $\mathbf{c}$. This soft aggregation retains uncertainty among visually similar SKUs instead of committing to a hard prototype assignment. It is beneficial for inference when it can recouple appearance information from the same product but different appearance.

3.4 Prototype-Guided Dynamic Modulation with QLer/VLer

Let $\mathbf{h}$ denote the input tokens of each Swin Transformer block and the block index is omitted below for readability. Since $\mathbf{c}$ is derived from DMM for each image, the corresponding model parameters are computed as follows:

$$\theta_Q = \mathrm{QLer}(\mathbf{c}), \qquad \theta_V = \mathrm{VLer}(\mathbf{c}), \tag{8}$$

where QLer and VLer are single layer MLPs in implementation. For clarity, we describe θ_Q and θ_V as θ. Then, θ are split to initialize Task Aligned Layer (TALer) Khan et al. (2026) as follows:

$$\{W_\downarrow, b_\downarrow, W_\uparrow, b_\uparrow, s\} = \mathrm{Split}(\theta), \tag{9}$$

$$\Delta = Scale \cdot \sigma(W_\downarrow h + b_\downarrow) \odot (W_\uparrow \sigma(W_\downarrow h + b_\downarrow) + b_\uparrow), \tag{10}$$

where $W_\downarrow$ and $W_\uparrow$ are the down- and up-projection matrices, $b_\downarrow$ and $b_\uparrow$ are the corresponding biases, $Scale$ is a scaling matrix. The θ are zero-initialized so that training starts close to the frozen pretrained model.

We obtain Δ_Q and Δ_V from above equation. Let Q, K, and V be the original attention projections. DMT modifies only Query and Value:

$$\widehat{Q} = Q + \Delta_Q, \qquad \widehat{K} = K, \qquad \widehat{V} = V + \Delta_V, \tag{11}$$

where Δ_Q and Δ_V are produced by the corresponding QLer and VLer. All adapters are conditioned on the same global domain feature $\mathbf{c}$, while different backbone depths learn different appearance-dependent visual relations without introducing a separate detector for every SKU. Keeping K unchanged preserves the pretrained key space and restricts the adaptation to a controlled residual update.

3.5 Domain Cross Attention

Let T denote the defect-prompt tokens produced by the text encoder and let F_{neck} denote the concatenated original neck tokens. In DCA, T acts as the query, while F_{neck} provides the keys and values. Unlike freezing the backbone network as a hypernetwork to generate model parameters for modulating modules, the DMM applies QLer and VLer to the same global condition to generate domain shift residuals for the DCA Query and Value:

$$\delta_Q = \mathrm{QLer}(\mathbf{c}), \quad \delta_V = \mathrm{VLer}(\mathbf{c}), \tag{12}$$

Therefore, the query, key and value are computed as follows:

$$\hat{Q} = TW_Q + \delta_Q, \quad \hat{K} = F_{neck}W_K, \quad \hat{V} = F_{neck}W_V + \delta_V. \tag{13}$$

The residual vectors are broadcast over the corresponding text or visual token dimension. Thus, QLer/VLer change the product-conditioned content of the Query and Value projections without changing the source sequence used to form the keys. The DCA output is computed by multi-head attention:

$$\mathrm{DCA}(T, V) = \mathrm{softmax}\left(\frac{\hat{Q}\hat{K}^\top}{\sqrt{d}} + \hat{b}\right)\hat{V}, \tag{14}$$

where b is a learnable SKU-specific bias. DCA allows defect-text tokens to retrieve image evidence from the multi-level visual representation. This design introduces SKU-conditioned visual evidence into defect-text interaction without replacing the pretrained visual processing pathway. Consequently, the model can preserve shared defect semantics while adapting the text-visual interaction to product-dependent appearance changes.

3.6 Detection Head and Training Objectives

The resulting visual features and DCA-conditioned text memory are processed by the GroundingDINO detection pathway to produce class scores and bounding boxes. The native detection objective is denoted by $\mathcal{L}_{det}$. For a ground-truth SKU label y, domain classifier is supervised by

$$\mathcal{L}_{dom} = \mathrm{CE}(p, y). \tag{15}$$

We use one regularizer for the generated DMT parameters, DCA residuals and DCA learnable parameters:

$$\mathcal{L}_{reg} = \mathrm{Mean}(\widehat{\theta}^2) + \mathrm{Mean}(\hat{\delta}^2) + \mathrm{Mean}(\hat{b}^2), \tag{16}$$

where $\widehat{\theta}$ denotes the generated DMT parameter and $\hat{\delta}$ denotes the DCA Q/V residuals. This regularizer constrains the magnitude of the conditional updates and helps preserve the pretrained representation. The final objective is

$$\mathcal{L} = \mathcal{L}_{det} + \beta\mathcal{L}_{dom} + \gamma\mathcal{L}_{reg}. \tag{17}$$

The SKU loss makes the product-appearance code discriminative enough to produce meaningful prototype mixtures, while the modulation regularizer limits the magnitude of the conditional updates. Together with $\mathcal{L}_{det}$, these objectives preserve shared defect semantics while allowing controlled adaptation to SKU-dependent appearance.

Table 1: Comparison of public datasets for surface defect detection. Appearance complexity is qualitatively assessed according to surface texture, background pattern, object geometry, material reflectance, and intra-class appearance variation.

Dataset	Application	Number	Category	Appearance Complexity	Inspection Level	Image Resolution Pixels
NEU-DET Song & Yan (2013)	Hot-rolled steel strip	1,800	6	Low	Local	120×120
GC10-DET Lv et al. (2020)	Steel strip	2,294	10	Medium	Local	2048×1000
PVMulti-Defect Li et al. (2023b)	Solar photovoltaic panels	307	5	Medium	Local	5800×3504
HRIPCB Huang et al. (2020a)	HRIPCB	1,386	6	High	Full object	600×600
Severstal Severstal (2019)	Steel strip	18,074	4	Low	Local	1600×256
DAGM Wieler et al. (2007)	Textured background	16,100	6	High	Local	800×800
MPCG Tao et al. (2018)	Mobile phone glass	11,808	-	Medium	Local	-
MT Huang et al. (2020b)	Magnetic tile	329	5	Low	Local	-
APDDD Alibaba Tianchi (2023)	Aluminum	3,005	10	Medium	Local	2560×1920
MVTec AD Bergmann et al. (2019)	Industrial products	5,354	-	Varies	Full object	700×700 // 1024×1024
Ours	Bottle surface	13,119	7	High	Full object	1240×1624 mainly

4 THE BOTTLE-SKU DATASET

Bottle-SKU contains images of shampoo bottles captured on factory production lines under real-world inspection conditions. The dataset will be publicly released upon publication to facilitate research on SKU-aware surface defect detection. Table 1 compares our Bottle-SKU dataset with existing public surface defect detection datasets. While prior datasets such as NEU-DET, GC10-DET, and MT focus on simple industrial surfaces with low appearance complexity and mostly local inspection, our dataset targets shampoo-bottle surfaces with high appearance complexity and full-object inspection. Compared with conventional defect datasets that focus on a single product or a small number of fixed appearances, bottles in the dataset vary in brand, label design, net weight, viewing side, viewing angle, and camera elevation. These variations substantially change the appearance of both the product surface and the same defect category. All images were annotated by six professional annotators recruited and organized by us, using bounding boxes for seven defect categories: *scratch*, *leakage*, *fold*, *bubble*, *worn*, *depressed*, and *flare*. More details can be found in Appendix C.

5 EXPERIMENTS

5.1 DATASETS AND EVALUATION METRICS

Bottle-SKU. Bottle-SKU is a full-object shampoo-bottle surface-defect benchmark with 13,119 images, 20,419 instance-level bounding boxes, seven defect categories, and 48 real product SKUs. We use a SKU-disjoint hold-out split rather than a random image split. The training set contains 10,495 images and 16,194 defect instances from 38 SKUs, whereas the test set contains 2,624 images and 4,225 instances from the remaining 10 SKUs. Therefore, no product SKU is shared between the two partitions, while all seven defect categories occur in both. This protocol evaluates whether a detector can recognize shared defect semantics under the appearance changes introduced by unseen bottle variants.

GC10-DET-SKU. GC10-DET Lv et al. (2020) is a public high-resolution metallic-surface defect dataset with ten defect categories, comprising 2,292 images of resolution 2048×1000 and 3,563 bounding boxes. GC10-DET does not provide a predefined train/test split. We therefore construct GC10-DET-SKU by clustering features from all usable images and holding out complete clusters as pseudo-SKUs. For each image, we extract a 768-dimensional L2-normalized CLS feature from a ViT-B/16 model pretrained on ImageNet-21k, and cluster the features using K-means with a fixed seed of 3407. We search $K \in \{4, \ldots, 12\}$ and select $K = 7$, which gives the highest cosine-silhouette score among partitions satisfying the minimum-cluster-size, class-coverage, and test-ratio constraints. Two complete clusters (pseudo-SKUs SKU-1 and SKU-2) are held out for testing, while the other five clusters (SKU-3–SKU-7) form training. This yields 1,773 training images with 2,762 boxes and 519 test images with 801 boxes (a 77.4%/22.6% image split).

Evaluation Metrics. We use the standard COCO bounding-box evaluation protocol for all methods. The primary metric is AP, namely the mean AP averaged over IoU thresholds from 0.50 to 0.95 with a step of 0.05. We additionally report $\mathrm{AP50}$, $\mathrm{AP75}$, and AP for small, medium, and large

Table 2: Comparison on Bottle-SKU and the GC10-DET-SKU pseudo-domain protocol. All values are percentages.

Index	Algorithm	Publish	Bottle-SKU						GC10-DET-SKU						Params
			AP	AP50	AP75	P	R	F1	AP	AP50	AP75	P	R	F1	
Lightweight Detector															
1	Faster R-CNN Ren et al. (2017)	TPAMI 2017	9.6	28.2	3.6	32.6	60.4	42.3	20.8	43.6	19.0	29.7	60.4	39.8	41.38M
2	Cascade R-CNN Cai & Vasconcelos (2018)	CVPR 2018	10.6	29.8	4.7	42.2	55.9	48.1	26.0	51.5	23.1	44.5	62.6	52.0	69.17M
3	FCOS Tian et al. (2019)	ICCV 2019	9.5	28.2	4.1	56.7	40.9	47.5	18.3	41.7	15.1	59.3	35.3	44.2	32.13M
4	GFL Li et al. (2020)	NeurIPS 2020	11.1	31.3	5.0	30.4	63.2	41.1	24.8	50.6	23.0	31.4	61.2	41.5	32.27M
5	Deformable DETR Zhu et al. (2021)	ICLR 2021	11.0	31.3	4.9	38.0	56.4	45.4	29.0	57.5	27.0	49.2	65.1	56.1	40.10M
6	TOOD Feng et al. (2021)	ICCV 2021	11.2	31.8	4.4	38.1	59.4	46.4	21.5	44.1	19.4	44.1	49.8	46.8	32.03M
7	YOLOX-S Ge et al. (2021)	arXiv 2021	6.3	18.7	2.5	29.7	43.2	35.2	23.4	50.3	19.5	51.3	51.3	51.3	8.94M
8	RTMDet-X Lyu et al. (2022)	arXiv 2022	6.1	17.1	2.6	20.0	48.7	28.4	27.6	54.8	27.0	32.2	65.0	43.0	94.78M
9	DINO Zhang et al. (2023)	ICLR 2023	11.0	31.3	5.0	39.9	56.2	46.7	27.6	52.6	26.4	49.8	50.4	50.1	47.55M
10	RT-DETR-L Zhao et al. (2024b)	CVPR 2024	7.9	24.8	2.7	36.3	31.3	33.6	29.4	59.7	27.6	66.9	60.6	63.6	32.82M
11	RT-DETR-X Zhao et al. (2024b)	CVPR 2024	7.7	24.6	2.9	36.1	35.6	35.8	27.0	58.9	20.7	66.9	59.0	62.7	67.32M
12	YOLO26-N Jocher et al. (2026)	arXiv 2026	11.2	30.0	5.5	43.5	35.9	39.4	29.3	56.0	29.4	63.7	56.4	59.8	2.51M
13	YOLO26-S Jocher et al. (2026)	arXiv 2026	12.0	31.4	5.7	44.4	34.4	38.8	28.9	55.8	26.3	67.9	54.0	60.2	9.95M
14	YOLO26-M Jocher et al. (2026)	arXiv 2026	12.1	31.6	5.8	40.3	36.9	38.5	29.6	55.9	27.2	61.4	61.7	61.5	21.78M
15	YOLO26-L Jocher et al. (2026)	arXiv 2026	12.1	31.8	5.8	43.8	34.0	38.3	29.4	58.1	23.4	63.1	58.6	60.8	26.19M
16	YOLO26-X Jocher et al. (2026)	arXiv 2026	12.7	33.7	6.0	41.9	38.3	40.0	28.9	26.4	27.0	65.7	58.8	62.1	58.82M
17	YOLOv8-M Jocher et al. (2023)	2023	11.9	32.1	5.8	37.7	36.2	37.0	28.0	57.2	24.8	62.8	57.8	60.2	25.84M
VLM-based															
18	WeDetect Fu et al. (2026a)	CVPR 2026	13.9	34.7	7.5	53.4	52.2	52.8	32.8	62.8	30.4	53.7	67.9	59.9	386.79M
19	DOSOD He et al. (2024)	arXiv 2024	13.1	33.7	6.6	49.9	53.5	51.6	28.6	55.2	26.8	56.4	61.1	58.7	108.24M
20	PET-DINO Fu et al. (2026b)	CVPR 2026	14.6	38.9	7.0	41.4	64.8	50.5	32.3	61.3	29.9	46.3	69.7	55.7	176.15M
21	MM-GDINO-T Zhao et al. (2024a)	arXiv 2024	14.9	38.2	8.0	39.8	65.8	49.6	35.8	68.1	35.0	49.7	73.5	59.3	172.91M
22	YOLO-World Cheng et al. (2024)	CVPR 2024	13.1	32.9	6.7	55.7	49.5	52.5	31.3	59.8	30.0	65.6	56.3	60.6	110.26M
23	Ours	-	16.0	40.8	8.9	38.3	67.7	49.0	37.8	72.1	37.2	60.5	72.0	65.8	252.68M

objects ($\mathrm{AP_S}$, $\mathrm{AP_M}$, and $\mathrm{AP_L}$). For an interpretable fixed operating point, we also report precision, recall, and F1 at a confidence threshold of 0.25 and an IoU threshold of 0.50. All results are reported as percentages, and all compared methods are evaluated nder the same dataset configuration. Predictions are processed in descending confidence order and matched one-to-one with ground-truth boxes of the same class. Thus, with true positives (TP), false positives (FP), and false negatives (FN),

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad F_1 = \frac{2PR}{P + R}. \tag{18}$$

The fixed-threshold P, R, and F_1 characterize an operating point and are reported separately from COCO mAP.

5.2 Comparison on Public Benchmark Datasets

Table 2 compares ProDMM with representative lightweight and vision-language-based detectors. The parameter column denotes the total number of model parameters.

Results on Bottle-SKU. On the SKU-disjoint Bottle-SKU test set, ProDMM obtains the highest overall AP (16.0), AP50 (40.8), and AP75 (8.9) among the evaluated methods. The strongest baseline scores on these three measures are 14.9, 38.9, and 8.0, respectively, giving gains of 1.1, 1.9, and 0.9 percentage points. The improvement at AP75 is particularly relevant because the defects frequently occupy only a small portion of a full bottle image and require accurate localization under previously unseen product appearances. At the fixed operating point, ProDMM also achieves the highest recall (67.7), showing that the model recovers more annotated defects. Its precision (38.3) and F1 (49.0) are lower than those of several comparison methods, which indicates a precision–recall trade-off at the fixed score threshold rather than a contradiction with the ranking by COCO AP. These threshold-dependent measures should therefore be interpreted together with the full precision–recall curve summarized by AP.

Results on GC10-DET-SKU. On the generated GC10-DET-SKU held-out protocol, ProDMM reaches 37.8 AP, 72.1 AP50, and 37.2 AP75. These values exceed the strongest baseline results by 2.0, 4.0, and 2.2 percentage points, respectively. ProDMM further obtains 72.0 recall and an F1 score of 65.8. The latter is the highest among the vision-language baselines in Table 2. The result on this public metallic-surface dataset suggests that prototype-conditioned modulation is not restricted to the bottle imagery. Since GC10-DET-SKU is a constructed pseudo-domain protocol, this comparison is a controlled cross-domain evaluation on a shared split.

Table 3: Component Analysis.

No.	Prototype	DMT	DCA	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
01	✗	✗	✗	15.1	39.8	7.9	2.5	9.1	16.6
02	✗	✓	✗	15.5	40.7	8.1	2.2	8.9	17.1
03	✓	✓	✗	15.8	41.6	8.0	2.2	8.4	17.4
04	✓	✓	✓	16.0	40.8	8.9	2.3	8.7	17.9

Table 4: Analysis of Layer Dimension in QLer and VLer.

No.	Dimension	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
01	16	15.5	39.9	8.1	2.7	9.9	16.9
02	64	15.0	39.0	7.9	3.8	10.5	16.6
03	128	15.5	39.8	8.5	2.6	10.7	17.1
04	256	16.0	40.8	8.9	2.3	8.7	17.9

Table 5: Analysis of the Number of Extracted Features for Prototypes.

No.	Features	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
01	100	16.0	40.8	8.9	2.3	8.7	17.9
02	300	15.8	40.5	8.4	2.9	9.6	17.5
03	500	15.5	40.2	8.2	2.7	10.1	17.1

5.3 Ablation Study

Component analysis. Table 3 isolates the contributions of the three main components. Starting from the frozen GroundingDINO baseline, adding dynamic modulation without prototypes raises AP from 15.1 to 15.5. Conditioning the modulation on SKU prototypes provides a further 0.3-point AP gain and improves AP50 from 40.7 to 41.6, showing that the prototype representation is useful beyond an image-only dynamic update. Adding Defect-DCA yields the best overall AP of 16.0 and increases AP75 from 8.0 to 8.9. Although AP50 changes from 41.6 to 40.8, the gains in overall AP and stricter-localization AP75 indicate that DCA changes the score and localization trade-off rather than uniformly improving every operating point.

Modulation width. Table 4 varies the latent width used to generate the QLer and VLer updates. The 256-dimensional setting achieves the best overall AP, AP50, and AP75. Smaller widths retain competitive results but do not provide a monotonic advantage. For example, the 64-dimensional setting has a higher AP_S but lower overall AP. This behavior is consistent with a capacity trade-off. The generator must encode both shared defect semantics and SKU-dependent appearance cues, while excessive compression can limit the fidelity of the dynamic updates.

Prototype support size. Table 5 studies the number of training-image features used to form each SKU prototype. Increasing the support size from 100 to 300 results in a modest 0.2-point AP decrease, whereas using 500 features does not yield further gains. The small spread across these settings suggests that the prototype estimate becomes stable with a limited number of samples. A very large support set is therefore unnecessary for obtaining a useful domain feature.

6 Conclusion

We introduced Bottle-SKU, a full-object shampoo-bottle surface-defect benchmark designed to evaluate detection under product-variant appearance shifts. The dataset contains 13,119 images, 20,419 instance-level bounding boxes, seven defect categories, and 48 SKUs. Its SKU-disjoint protocol trains on 38 product variants and evaluates on 10 unseen variants, separating shared defect recognition from memorization of a particular product appearance. We further presented Prototype-Guided Dynamic Domain Modulation (ProDMM). ProDMM summarizes training images with offline SKU prototypes, predicts a soft product-domain distribution for each input and uses it to generate dynamic residual updates for the Query and Value projections of the frozen backbone. Experiments on the Bottle-SKU and GC10-DET-SKU datasets demonstrate that prototype-conditioned modulation is suitable for the multi-SKU task. Component studies show complementary benefits from dynamic modulation, prototype conditioning, and DCA, while the prototype-support and modulation-width experiments indicate that the method does not require a large per-SKU image support set to be effective.

AI USE STATEMENT

In this work, we used generative AI tools to improve the quality of written English and for stylistic refinement, with attentive supervision by the authors. No scientific content, experimental design or data analysis were generated through LLMs. All AI-assisted writing was reviewed for correctness by 2 authors.

REFERENCES

Emanuele Aiello, Umberto Michieli, Diego Valsesia, Mete Ozay, and Enrico Magli. Dreamcache: Finetuning-free lightweight personalized image generation via feature caching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12480–12489, June 2025.

Alibaba Tianchi. Aluminum profile surface defects dataset. Public dataset, 2023. URL `https://www.kaggle.com/datasets/weihaoreal/aluminum-profile-surface-defects-data-set`. Also referred to as APDDD in industrial surface defect detection literature.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9592–9600, 2019. doi: 10.1109/CVPR.2019.00982.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec loco ad dataset: Advancing unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 395–403, 2020.

Prahar M. Bhatt, Rishi K. Malhan, Pradeep Rajendran, Brual C. Shah, Shantanu Thakar, Yeo Jung Yoon, and Satyandra K. Gupta. Image-based surface defect detection using deep learning: A review. *Journal of Computing and Information Science in Engineering*, 21(4), 2021. doi: 10.1115/1.4049535.

Haorun Cai and Han-Jia Ye. Feature-aware modulation for learning from temporal tabular data. *Advances in Neural Information Processing Systems*, 38:151014–151037, 2026.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, 2018.

Weihao Cao, Yunqiu Xu, Jianhao Li, Chunchao Li, Yicong Chen, and Kai Liu. Parameter-efficient semantic augmentation for enhancing open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20367–20376, 2026.

Junbum Cha, Sangjun Chun, Kyungwoo Lee, et al. Swad: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shoufa Chen, Chongyang Ge, Zhan Tong, et al. Adaptformer: Parameter-efficient transfer learning for vision transformers. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yuhua Chen, Wen Li, Christos Sakaridis, Deng Xin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3339–3348, 2018.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pp. 475–489, 2021.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3510–3519, 2021.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14987–14997, 2025.

Shenghao Fu, Yukun Su, Fengyun Rao, Jing Lyu, Xiaohua Xie, and Wei-Shi Zheng. Wedetect: Fast open-vocabulary object detection as retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Weifu Fu, Jinyang Li, Bin-Bin Gao, Jialin Li, Yuhuan Lin, Hanqiu Deng, Wenbing Tao, Yong Liu, and Chengjie Wang. Pet-dino: Unifying visual cues into grounding dino with prompt-enriched training. *arXiv preprint arXiv:2604.00503*, 2026b.

Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. *Advances in Neural Information Processing Systems*, 38:114432–114462, 2026.

Yaroslav Ganin, Evgenia Ustinova, Hana Ajakan, et al. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Ishaan Gulrajani and David Lopez-Paz. Domainbed: A benchmark for domain generalization. In *International Conference on Learning Representations*, 2021.

Guanting Guo, Shenglong Hu, Kaihua Zhang, Guangcan Liu, and Min Xia. Generalizable co-salient object detection via mixed content-style modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32173–32183, June 2026.

Yonghao He, Hu Su, Haiyong Yu, Cong Yang, Wei Sui, Cong Wang, and Song Liu. A light-weight framework for open-set object detection with decoupled feature alignment in joint space. *arXiv preprint arXiv:2412.14680*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, et al. Parameter-efficient transfer learning for nlp. In *Proceedings of the International Conference on Machine Learning*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Weibo Huang, Peng Wei, Manhua Zhang, and Hong Liu. Hripcb: a challenging dataset for pcb defects detection and classification. *The Journal of Engineering*, 2020(13):303–309, 2020a. doi: 10.1049/joe.2019.1183.

Yibin Huang, Congying Qiu, and Kui Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 36:85–96, 2020b. doi: 10.1007/s00371-018-1588-5.

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rex2: Towards generic object detection via text-visual prompt synergy. In *European Conference on Computer Vision*, pp. 38–57, 2024.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolo, 2023. URL `https://github.com/ultralytics/ultralytics`.

Glenn Jocher, Jing Qiu, Mengyu Liu, Shuai Lyu, Fatih Cagatay Akyon, and Muhammet Esat Kalfaoglu. Ultralytics yolo26: Unified real-time end-to-end vision models, 2026. URL `https://arxiv.org/abs/2606.03748`.

Ufaq Khan, Umair Nawaz, Massimo Caputo, Muhammad Bilal, Junaid Qadir, and Muhammad Haris Khan. Keep it frozen: Domain-routed conditional residual modulation for multi-domain vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21016–21025, June 2026.

Prannay Khosla, Piotr Teter Wakem, Yun Xiong, et al. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Geonuk Kim, Minhoi Kim, Kangil Lee, Minsu Kim, Hyeonseong Jeon, Jeonghoon Han, Hyoungjoon Lim, and Junho Yim. Unispector: Towards universal open-set defect recognition via spectral-contrastive visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6261–6270, June 2026.

Amit Kumar Kundu, Vaishnavi Patil, and Joseph Jaja. Boosting open set recognition performance through modulated representation learning. In *International Conference on Learning Representations*, volume 2026, pp. 135212–135234, 2026.

Sanghoon Lee, Geon Lee, Hyekang Park, and Bumsub Ham. Exploring hierarchical consistency and unbiased objectness for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 6819–6828, 2026.

Chen Li, Haoxin Yan, Xiang Qian, Shidong Zhu, Peiyuang Zhu, Chengwei Liao, Haoyang Tian, Xiu Li, Xiaohao Wang, and Xinghui Li. A domain adaptation YOLOv5 model for industrial defect inspection. *Measurement*, 213:112725, 2023a. doi: 10.1016/j.measurement.2023.112725.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9664–9674, 2021.

Haochen Li, Rui Zhang, Hantao Yao, Xin Zhang, Yifan Hao, Xinkai Song, Shaohui Peng, Yongwei Zhao, Chen Zhao, Yanjun Wu, et al. Seen-da: Semantic entropy guided domain-aware attention for domain adaptive object detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25465–25475. IEEE, 2025.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

Longlong Li, Zhifeng Wang, and Tingting Zhang. Gbh-yolov5: Ghost convolution with bottleneckcsp and tiny target prediction head incorporating yolov5 for pv panel defect detection. *Electronics*, 12(3):561, 2023b. doi: 10.3390/electronics12030561.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21002–21012, 2020.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024.

Xiaoming Lv, Fajie Duan, Jiajia Jiang, Xiao Fu, and Lin Gan. Deep metallic surface defect detection: The new benchmark and detection network. *Sensors*, 20(6):1562, 2020. doi: 10.3390/s20061562.

Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yitong Qin, Lihua Zhou, Jiwei Wei, Ran Ran, Shiyuan He, Zeyu Ma, Shuaifeng Li, Nianxin Li, and Heng Tao Shen. Vitprompt: Training-free prompt refinement with visual tokens for open-vocabulary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3111–3121, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

Tianhe Ren, Qing Chen, Shilong Liu, Ailing Zeng, Jing Lin, Kun Li, Hao Tang, Yiding Jiang, Hang Huang, Yuxiao Gao, Hao Yan, Xiangyu Zeng, Yao Zhang, Feng Xiao, Hui Huang, Zhaoyang Liu, Wenqiang Zhang, Honggang Zhang, Jianhui Gao, Bohan Jiang, and Meng Wang. Grounding dino 1.5: Advance the edge of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024.

Karsten Roth, Latha Pemula, Joaquin Zepeda, et al. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Severstal. Severstal: Steel defect detection. Kaggle Competition, 2019. URL `https://www.kaggle.com/c/severstal-steel-defect-detection`. Accessed: 2026-06-29.

Noam Shazeer, Quoc V. Mirzadeh, Krzysztof Maziarz, et al. Sparsely-gated mixture-of-experts. In *International Conference on Learning Representations*, 2017.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Kechen Song and Yunhui Yan. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science*, 285:858–864, 2013. doi: 10.1016/j.apsusc.2013.09.002.

Youngjun Song, Hyeongyu Kim, and Dosik Hwang. Cd-buffer: Complementary dual-buffer framework for test-time adaptation in adverse weather object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15050–15059, June 2026.

Nurjahan Sultana, Moi Hoon Yap, Xinqi Fan, and Wenqi Lu. Cofida-m: Concept-aware feature modulation for cross-domain adaptation with image-only inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15060–15069, June 2026.

Xian Tao, Dapeng Zhang, Wenzhi Ma, Xilong Liu, and De Xu. Vision-based defect detection for mobile phone cover glass using deep neural networks. *International Journal of Precision Engineering and Manufacturing*, 19:801–810, 2018. doi: 10.1007/s12541-018-0096-x.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9627–9636, 2019.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, and Xiaodan Liang. Ov-dino: Unified open-vocabulary detection with language-aware selective fusion. *arXiv preprint arXiv:2407.07844*, 2024.

Matthias Wieler, Tobias Hahn, and Fred A. Hamprecht. Weakly supervised learning for industrial optical inspection. DAGM 2007 benchmark dataset, 2007. URL https://hci.iwr.uni-heidelberg.de/content/weakly-supervised-learning-industrial-optical-inspection.

Brandon Yang, Gabriel Bender, Quoc V. Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Li Yang, Boyu Cai, Wei Liu, Yan Wang, Chunfeng Yuan, Bing Li, and Weiming Hu. Sra-det: Learning omni-grained open-vocabulary detection beyond category names. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27611–27620, 2026.

Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. Detclipv3: Towards versatile generative open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5610–5619, 2024.

Xingxu Yao, Sicheng Zhao, Pengfei Xu, and Jufeng Yang. Multi-source domain adaptation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3253–3262, 2021. doi: 10.1109/ICCV48922.2021.00326.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14388–14397, 2021. doi: 10.1109/CVPR46437.2021.01416.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. Draem – a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2023.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19513–19524. IEEE, 2025.

Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024a.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2024b. doi: 10.1109/CVPR52733.2024.01605.

Xingyi Zhou, Rohit Girdhar, Brian Jou, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 2022.

Yang Zhou, Shiyu Zhao, Yuxiao Chen, Zhenting Wang, Can Jin, Mingyu Zhao, and Dimitris N. Metaxas. Led: Llm enhanced open-vocabulary object detection without human curated data generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9643–9653, 2026.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021.

A Extended Related Work

A.1 Vision-Language and Open-Vocabulary Object Detection

Vision-language pre-training aligns visual representations with natural-language concepts. CLIP Radford et al. (2021) demonstrates that contrastive learning over web-scale image-text pairs can produce transferable embeddings whose textual prompts act as a flexible classifier. GLIP Li et al. (2022) extends this paradigm from image-level matching to region-level grounding, showing that phrases can supervise both semantic recognition and localization. Caption-driven open-vocabulary detection uses image-level descriptions to transfer supervision beyond the detector label set Zareian et al. (2021), while Detic Zhou et al. (2022) demonstrates that image-level supervision can scale detection to tens of thousands of categories through a large vocabulary classifier.

Transformer detectors form an effective backbone for language-conditioned localization. Deformable DETR Zhu et al. (2021) replaces dense global attention with reference-point sampling over multi-scale features, making transformer detection practical for small objects and high-resolution inputs. DINO Zhang et al. (2023) improves DETR convergence by combining denoising queries, improved anchor initialization, and contrastive query selection, and consequently provides a strong closed-set detection foundation. Grounding DINO Liu et al. (2024) couples this detector design with grounded vision-language pre-training. Text tokens and image features interact in the feature enhancer and language-guided query selection, enabling open-set localization from natural-language prompts. Grounding DINO 1.5 Ren et al. (2024) further focuses on stronger open-set accuracy and deployment-oriented efficiency.

Recent detectors explore different ways to inject language at inference. YOLO-World Cheng et al. (2024) distills vision-language knowledge into a YOLO-style architecture and uses prompt- or vocabulary-embeddings for real-time open-vocabulary prediction. DetCLIPv3 Yao et al. (2024) treats open-vocabulary detection as a more general generative recognition problem, improving the ability to transfer from diverse textual descriptions. T-Rex2 Jiang et al. (2024) studies the synergy between visual exemplars and text prompts, which is useful when category names alone are insufficient. OV-DINO Wang et al. (2024) introduces language-aware selective fusion so that visual features receive category-relevant semantic information without uniformly mixing all language tokens. LLMDet Fu et al. (2025) uses large-language-model supervision to enrich detector semantics, while MM-GroundingDINO Zhao et al. (2024a) presents a broader grounding-and-detection pipeline. These methods mainly target category expansion. ProDMM instead addresses SKU-dependent appearance variation while retaining the grounded detection pathway.

A.2 Prototype-Based Domain Adaptation and Conditional Modulation

Prototype-based learning represents a class or domain with a representative embedding. Prototypical Networks Snell et al. (2017) introduced episodic learning in which a query is classified by its distance to the mean embedding of each support class, providing a simple mechanism for few-shot generalization. N-pair loss Sohn (2016) replaces pairwise sampling with multiple negative examples in one objective, improving metric separation and reducing the instability of contrastive training. Supervised contrastive learning Khosla et al. (2020) generalizes this idea by pulling together all samples sharing a label while pushing apart samples from other labels, which is useful for modeling intra-class variation. For domain shift, DANN Ganin et al. (2016) introduces a gradient-reversal domain classifier to learn features that confuse source and target domains. ADDA Tzeng et al. (2017) decouples source representation learning from target mapping, allowing an unlabeled target encoder to be aligned adversarially. Domain Adaptive Faster R-CNN Chen et al. (2018) transfers these principles to object detection by aligning image-level and instance-level features. DomainBed Gulrajani & Lopez-Paz (2021) standardizes domain-generalization evaluation, while SWAD Cha et al. (2022) seeks flat minima that are empirically less sensitive to unseen domain changes.

Conditional modulation changes a representation as a function of an auxiliary signal. FiLM Perez et al. (2018) predicts feature-wise scale and bias parameters from the conditioning input, allowing the same visual backbone to implement different computations for different contexts. CondConv Yang et al. (2019) extends this idea to convolution by generating input-dependent mixtures of kernel experts. Sparsely gated MoE Shazeer et al. (2017) demonstrates that a learned gate can activate only a small subset of experts, increasing model capacity without proportional computation, and

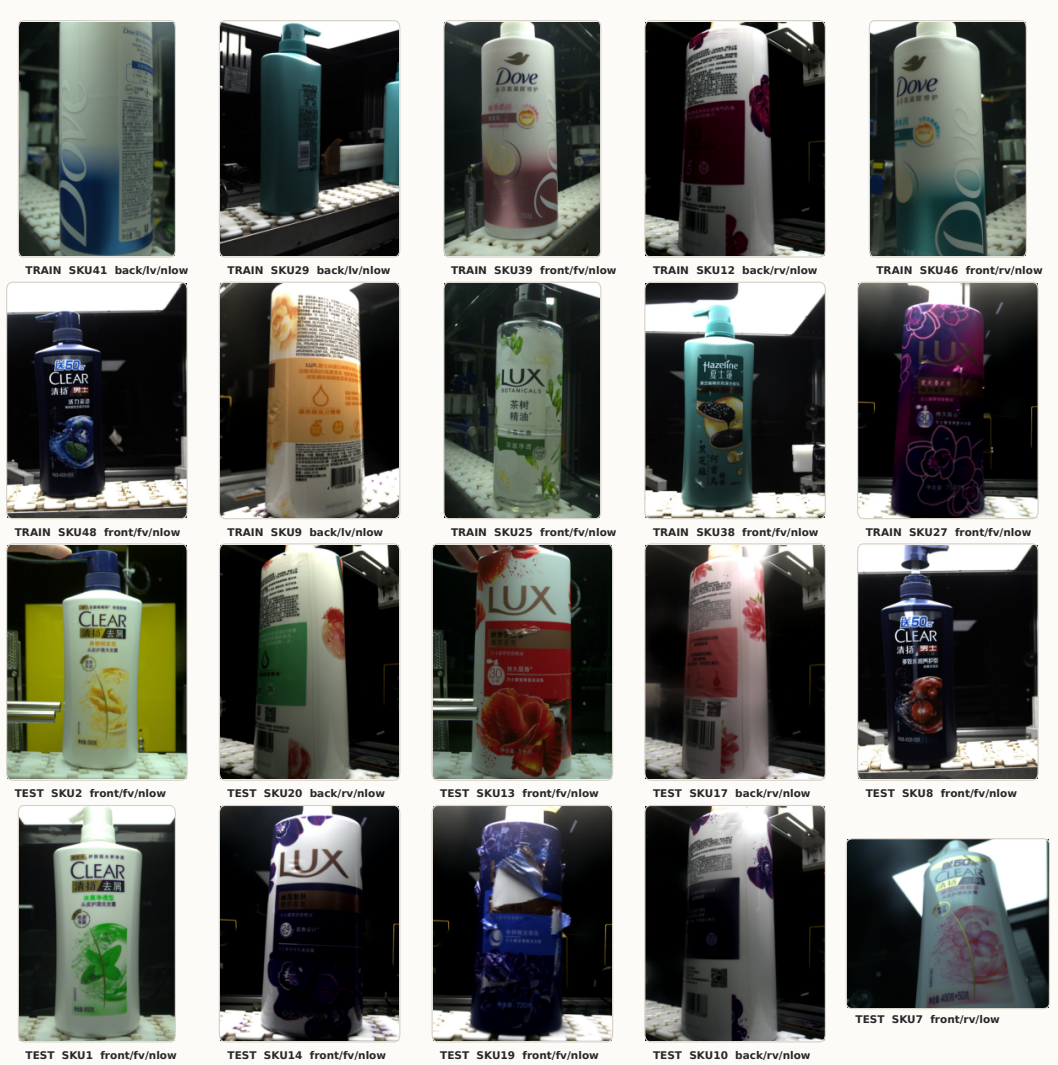


Figure 3: Examples of bottle images from different SKUs in our dataset.

Switch Transformer Fedus et al. (2022) simplifies this expert-selection mechanism to one expert per token for efficient large-scale training. Parameter-efficient transfer uses bottleneck adapters Houlsby et al. (2019) to isolate task-specific updates, LoRA Hu et al. (2022) to inject trainable low-rank matrices into frozen linear layers, and AdaptFormer Chen et al. (2022) to adapt vision transformers with lightweight residual branches.

B Implementation Details

All experiments are conducted on NVIDIA A800 80GB GPUs and initialized from the official GroundingDINO Swin-B checkpoint. For the main frozen ProDMM setting, the Swin-B backbone and BERT encoder are frozen, while the dynamic modulation modules and detection adaptation pathway remain trainable. We build 38 offline SKU prototypes from at most 100 training images per SKU. Each prototype feature is obtained by global average pooling the final Swin stage and applying feature normalization before prototype aggregation.

The Domain Encoder uses the initial patch-embedding tokens. A 1×1 spatial gate is fused with the original token feature, followed by global pooling and an MLP domain classifier. The DMM encoder is a two-layer GELU MLP that projects the 1024-dimensional prototype mixture to a 256-

dimensional condition. It drives 48 independent Swin Q/V heads and two 256-dimensional Defect-DCA Q/V residual heads. QLer and VLer use rank $r = 1$ and are implemented with low-rank down/up projections, biases, and a scalar modulation factor. The dynamic output heads use zero-output initialization with a fixed adapter basis, so the initial residual update is zero.

Defect-only DCA uses eight attention heads over four 256-dimensional neck levels. The SKU-conditioned spatial prior is represented by a 16×16 grid for each domain, attention head, and neck level. The soft SKU posterior mixes these grids, which are bilinearly interpolated to the corresponding attention resolution. $\mathcal{L}_{reg}$ is the mean-square penalty over the generated QLer/VLer increments, DCA Q/V residuals, and spatial-prior parameters. We use $\beta = 1.0$ and $\gamma = 10^{-4}$. Training uses AdamW with learning rate 10^{-4}, weight decay 10^{-4}, gradient clipping at 0.1, a constant learning rate, and a fixed seed of 3407. We train for 30 epochs with training and validation batch sizes of 18 and 36, respectively.

C Bottle-SKU Dataset Details

C.1 SKU Definition and Product Grouping

Bottle-SKU was curated at the product-variant level rather than at the coarse brand or bottle-color level as shown in Fig. 3. During annotation, bottles with different label descriptions or package specifications were treated as distinct product variants. Each filename contains three groups of metadata separated by underscores. These groups include a *product descriptor* (`SKU-ID` and package weight), a *capture descriptor* (visible side, horizontal view, and camera elevation), and a unique image identifier.

For example, `SKU3_450g_back_fv_nlow_2925` denotes an image from the internal product family `SKU3` with a `450g` package specification. Its capture descriptor indicates the bottle back, frontal view (`fv`), and non-upward camera elevation (`nlow`). The final field `2925` is the image identifier. The first two fields jointly represent product identity, where `SKU-ID` denotes the curated product/label family and the weight distinguishes package variants within that family.

The acquisition fields are retained as metadata rather than defect labels. Specifically, `front` and `back` identify the visible bottle side. The tokens `fv`, `lv`, and `rv` represent frontal, left-oblique, and right-oblique views, respectively. The tokens `low` and `nlow` denote upward-looking and non-upward camera elevations. An additional suffix is used only when necessary to disambiguate duplicate image identifiers.

This organization preserves appearance factors that are relevant to industrial inspection. Images of the same product can differ substantially in reflections, label textures, bottle contours, and visible regions under different sides and viewpoints. Bottle-SKU therefore supports defect detection under SKU-related appearance variation while retaining the metadata needed for controlled analysis.

C.2 Dataset Annotation

Annotation follows the principle that a box should cover the complete visually observable defect region, rather than only the most salient local texture. The annotation samples are shown in Fig. 4. The detailed criteria are summarized below.

- **Scratch.** Scratches are generally elongated and strip-like. The box should cover the visible extent of the scratch.
- **Leakage.** Leakage usually exhibits a flowing pattern with irregular boundaries. A connected leakage region is annotated with one enclosing box. Separated regions are annotated independently when their connection is sufficiently weak or discontinuous.
- **Fold.** Folds often occupy a relatively broad area with uneven surfaces and blurred boundaries. The annotation covers the deformation and its associated shadow region. Folded bottle labels are also annotated as folds.
- **Bubble.** An isolated bubble is usually approximately circular and is annotated individually. Dense bubble clusters are annotated using one box covering the complete cluster rather than fragmented boxes for individual bubbles.

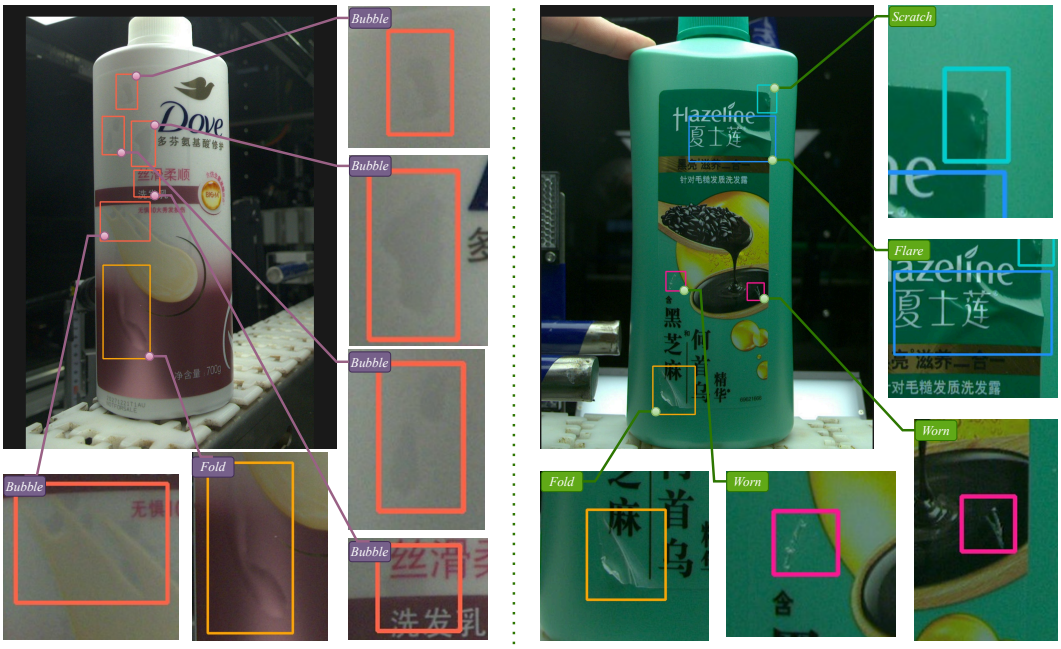


Figure 4: Examples of defects appearing on bottle surfaces.

Table 6: Statistics of the Bottle-SKU dataset.

Split	#SKU	#Images	#Instances	Ins./Img.
Train	38	10,495	16,194	1.54
Test	10	2,624	4,225	1.61
Total	48	13,119	20,419	1.56

- **Worn.** Worn defects include missing corners, chipping, cracking, and other visibly broken regions.
- **Depressed.** Depressions are primarily revealed by local illumination and shading changes. The box covers the full depressed surface. Pixel-level boundary precision is not required.
- **Flare.** Flare refers to a misaligned or bent label that forms a cone-like hollow region, typically caused by label attachment failure.

After annotation, the labels were inspected according to the above category definitions, spatial coverage rules, and box consistency rules. In particular, annotators avoided splitting one connected defect into multiple boxes and avoided assigning a defect label solely according to bottle color or product appearance.

C.3 Dataset Statistics

Table 6 summarizes the scale and split of Bottle-SKU. The dataset contains 13,119 images, 20,419 defect instances, 48 SKU groups, and seven defect categories. We adopt a SKU-disjoint evaluation protocol. The training set contains 38 SKU groups, whereas the test set contains 10 different SKU groups.

All seven defect categories occur in both splits. The instance distribution is naturally long-tailed. *Fold* is the most frequent category, accounting for 38.7% and 47.2% of the training and test instances, respectively, whereas *flare* is comparatively scarce. This distribution reflects the practical prevalence of different bottle-surface defects and makes the task more challenging than a class-balanced benchmark.

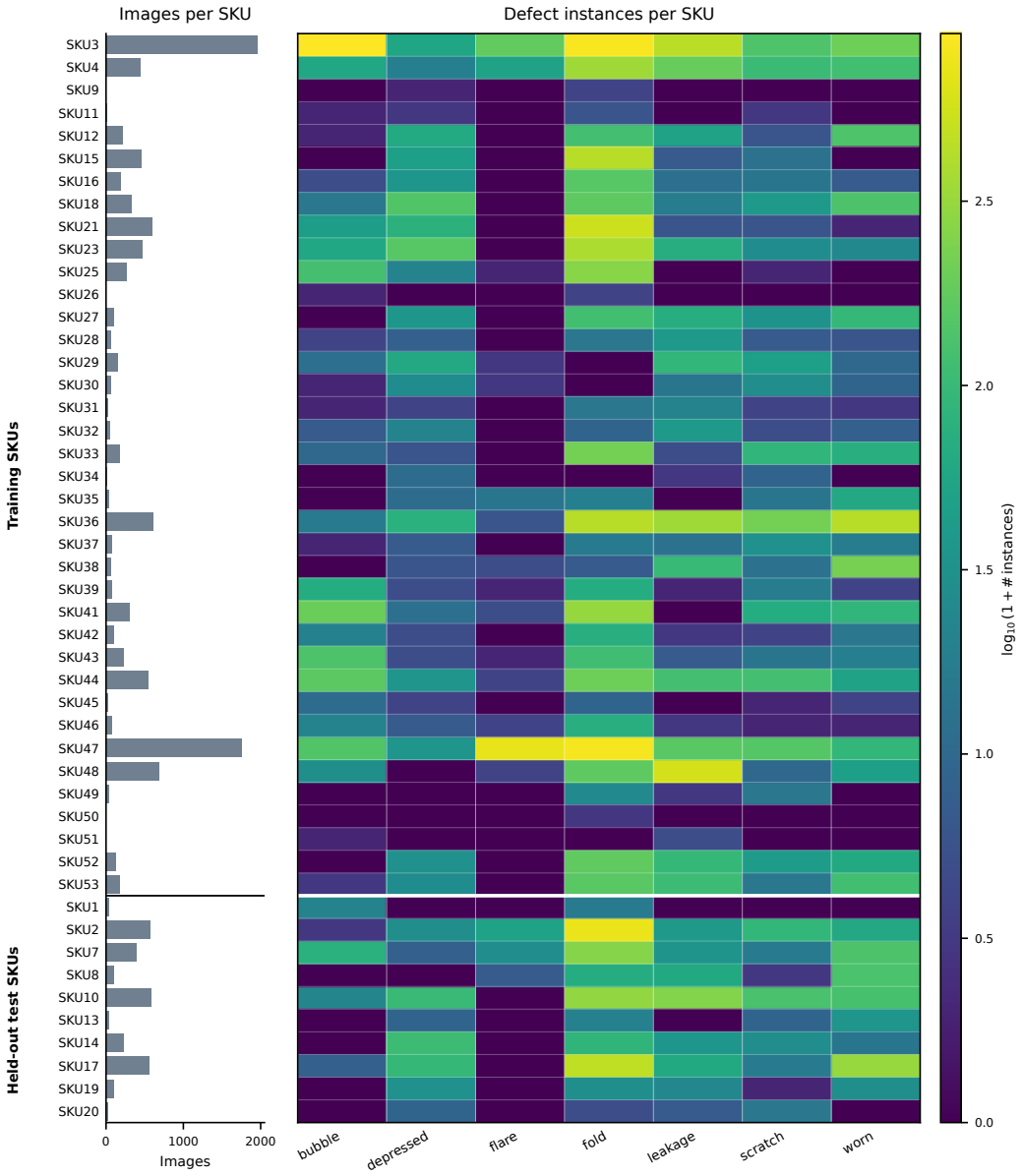


Figure 5: Distribution of annotated defect instances across Bottle-SKU product variants. The rows are divided into training SKUs and held-out test SKUs. Color intensity represents $\log_{10}(1 + \#\text{instances})$, and the bars on the left indicate the number of images for each SKU.

Figure 5 presents the distribution of defect instances across product variants. Each row corresponds to one SKU and each column corresponds to one defect category. The color indicates the log-scaled number of annotated instances, while the bars on the left show the number of images for each SKU. The distribution is heterogeneous, with different SKUs exhibiting different defect compositions and substantially different numbers of annotated instances. In addition, the training and held-out test SKUs show different category frequencies. This confirms that the evaluation protocol involves both product-appearance variation and domain-dependent defect imbalance.

Figure 6 visualizes the frozen visual features of training images from each of the 38 training SKU groups. The figure illustrates both intra-SKU appearance variation and partial overlap between visually similar SKU groups. Such overlap is expected in real production settings, where bottles with different labels, package specifications, or viewpoints share similar colors, reflections, and

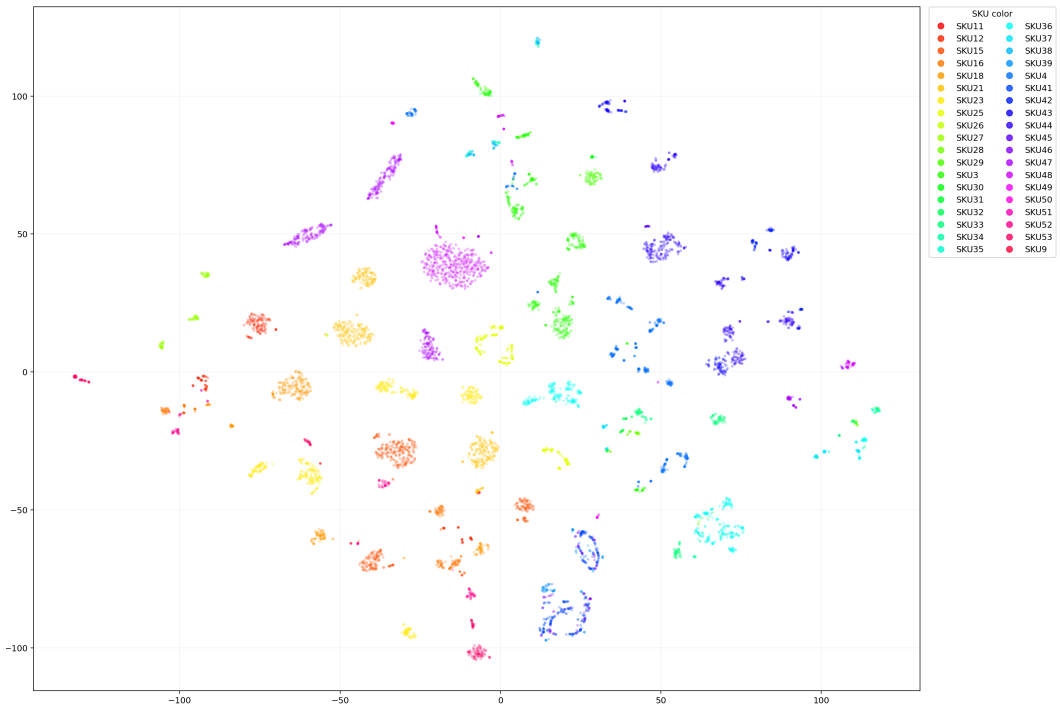


Figure 6: A t-SNE visualization of the features extracted from the training set for the 38 SKU classes.

Table 7: Comparison with static parameter-efficient fine-tuning baselines.

No.	Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
01	LoRA rank=1	15.5	39.8	8.1	2.8	9.8	17.2
02	LoRA rank=64	15.7	40.4	8.1	3.3	9.5	17.3
03	LoRA rank=128	15.8	39.7	8.8	3.2	9.2	17.3
04	AdaptFormer	15.7	40.4	8.3	3.8	9.1	17.5
05	Ours	16.0	40.8	8.9	2.3	8.7	17.9

local surface textures. This motivates the use of soft prototype-conditioned modulation instead of a hard SKU assignment.

D ADDITIONAL RESULTS

Comparison with static parameter-efficient updates. Table 7 compares the proposed prototype-conditioned updates with static LoRA and AdaptFormer baselines under the frozen-backbone setting. ProDMM achieves the highest AP (16.0) and AP75 (8.9), exceeding the best LoRA result by 0.2 and 0.1 points, respectively, and AdaptFormer by 0.3 and 0.6 points. AdaptFormer obtains a higher AP_S, so the comparison does not imply that ProDMM dominates every object-size regime. Instead, it supports the central claim that conditioning the update on the inferred product domain improves the aggregate cross-SKU detection result over a single static adapter.

Domain loss weight. Table 8 varies the weight of the SKU loss. Increasing this weight improves AP from 15.4 at $\beta = 0.25$ to 16.0 at $\beta = 1.0$, and improves AP75 from 7.6 to 8.9. AP50 is not monotonic and is largest at $\beta = 0.25$, again indicating that the auxiliary objective controls a localization and score-calibration trade-off. The result confirms that domain supervision must be sufficiently strong for the Domain Encoder to provide useful conditioning, while detection loss remains necessary to preserve defect-level discrimination.

Table 8: Parameter Analysis of the Domain Classifier Loss Weight.

No.	Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
01	0.25	15.4	41.0	7.6	2.0	8.8	17.1
02	0.5	15.6	39.9	8.4	3.0	10.0	17.2
03	1.0	16.0	40.8	8.9	2.3	8.7	17.9



Figure 7: Comparison of Ground Truth, ProDMM, DOSOD, PET-DINO, WeDetect and YOLO-World.

E VISUALIZATION

Figure 7 provides qualitative comparisons on Bottle-SKU testing images. All results use the same original-image geometry, class-color convention, and box-drawing rule, so differences reflect model predictions. In both examples, ProDMM detect multiple spatially separated annotated defects, including regions with weak contrast against labels or bottle surfaces. The comparison methods also detect several prominent instances, but miss or localize some of the smaller and visually ambiguous regions in these examples.

The visualization is intended to complement rather than replace the aggregate metrics. ProDMM is not error-free. Extra or imperfectly localized predictions can occur, consistent with its fixed-threshold precision-recall trade-off in Table 2. Together with the quantitative results, these cases illustrate the practical role of SKU-conditioned modulation. It preserves detection of subtle defect evidence while adapting to changes in bottle appearance and imaging conditions.

F LIMITATION ANALYSIS

Despite the consistent improvements over the evaluated baselines, several limitations remain. First, ProDMM relies on an offline prototype table and on the availability of reliable SKU assignments in the training data. A single mean prototype is a compact summary of a product variant, but it cannot fully represent multimodal appearance factors such as front and back labels, viewpoint changes, illumination, or different production batches. The soft SKU representation can combine several

prototypes to reduce hard assignment errors, yet its representation is still bounded by the coverage and quality of the stored prototypes. A newly introduced product, a substantially redesigned label, or strong prototype drift may therefore require prototype updating or an additional unknown-domain mechanism.

Second, the main setting freezes the Swin-B visual backbone and the BERT text encoder. This design isolates the effect of conditional modulation and limits the number of optimized pretrained parameters, but it also restricts the adaptation capacity under severe changes in material, camera configuration, or defect morphology. ProDMM modifies only the Query and Value pathways in the visual transformer and uses residual conditioning for defect-text interaction. It does not adapt every representation in the detector. The total model still contains the large pretrained GroundingDINO parameters. Dynamic parameter generation and DCA introduce additional memory and inference computation. Thus, the frozen-backbone setting should be viewed as conditional adaptation with a reduced optimization burden, rather than as a strict reduction of the deployed model size or latency.

Third, the current evaluation is limited to one factory dataset and one constructed public pseudo-domain protocol. Bottle-SKU provides a realistic SKU-disjoint split, but it represents one product family and one acquisition pipeline. GC10-DET-SKU is derived from GC10-DET by unsupervised visual clustering and is useful for controlled cross-domain analysis. Future evaluation on independently collected factories, unseen production batches, and temporally separated test sets is needed to quantify deployment robustness. Finally, the reported precision, recall, and F1 values correspond to one confidence threshold, while the class distribution is imbalanced. Threshold calibration, cost-sensitive evaluation, and defect-specific operating points are therefore important before making production-level claims.