*Journal of Food Science*

**ORIGINAL ARTICLE**
*Artificial Intelligence in Food Science and Nutrition*

# Visual Food Object Detection via Open-Vocabulary Learning

Shoulong Liu[1] | Guorui Sheng[1] | Shaojie Zhou[1] | Bolin Yang[1] | Weiqing Min[2,3] | Shuqiang Jiang[2,3]

[1]School of Computer Science and Artificial Intelligence, Ludong University, Yantai, China | [2]State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China | [3]University of Chinese Academy of Sciences, Beijing, China

**Correspondence:** Weiqing Min (minweiqing@ict.ac.cn)

**Received:** 15 February 2026 | **Revised:** 4 July 2026 | **Accepted:** 11 August 2026

**Keywords:** food image analysis | food quality inspection | multiple instance learning | open-vocabulary food detection | weakly supervised learning

## ABSTRACT

The rapid diversification of food products, frequent packaging updates, and seasonal variations pose significant challenges for vision-based food inspection and dietary monitoring in real-world food systems. To address the need for scalable and flexible food analysis, this study proposes a domain-adaptive open-vocabulary food object detection (OVFD) framework that enables the identification of both existing and previously unseen food categories without category-specific box annotations for new items. By leveraging open-vocabulary representations, the framework allows dynamic expansion of detectable food categories as new items emerge, supporting continuous adaptation to evolving products, recipes, and dietary patterns while reducing manual annotation requirements. The proposed method integrates a dynamic prompt distribution network (DPDN) to improve region-text alignment, along with a density-aware multiple instance learning (DA-MIL) strategy that uses weakly supervised image-level tags to improve generalization and suppress background-driven detections. To reduce semantic leakage, Food2K labels are filtered by normalized matching, synonym auditing, and overlap-based rules before training. The framework was evaluated on multiple benchmark food datasets under open-vocabulary settings, demonstrating consistent improvements in detection performance for both Base (seen) and Novel (unseen) food categories, with Novel-category gains also observed under AP75 (average precision at intersection-over-union = 0.75) and COCO-style AP. Repeated-run reporting, error analysis, and limited external evaluation support stability and practical plausibility. By enabling dynamic category expansion with reduced annotation overhead, the proposed approach provides an assistive localization module for downstream dietary assessment, sorting review, and quality-inspection workflows, while task-specific validation and human verification remain necessary before operational food-safety or production-line use.

### Practical Applications

This study provides an OVFD approach that can be adapted to new food items with reduced category-specific box annotation. Bounding-box localization can support upstream perception for portion estimation, ingredient-level dietary logging, assisted sorting review, and quality-inspection workflows by identifying where food items are located before secondary human or automated assessment. The present evidence supports assistive use only; human verification and validation under pilot-plant or real production conditions remain necessary before operational use. The detector is therefore intended to provide spatial evidence for subsequent review, not to replace validated food-inspection, quality-assessment, or production-control procedures.

**Abbreviations:** OVD, open-vocabulary detection; OVFD, open-vocabulary food object detection; VLM, vision-language model; DA-MIL, density-aware multiple instance learning; DPDN, dynamic prompt distribution network; RPN, region proposal network; KL, Kullback–Leibler; $AP_{50}$, average precision at IoU = 0.50; mAP, mean average precision.

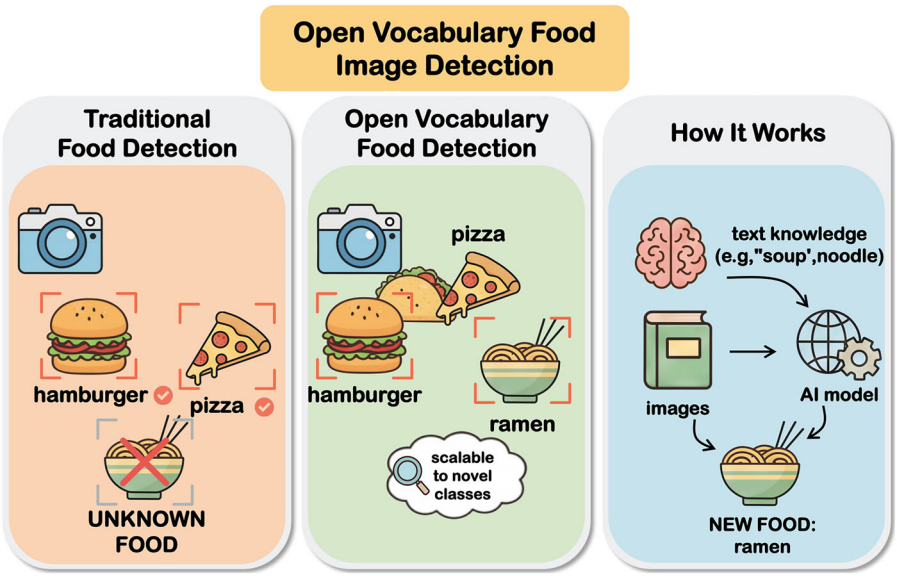


**FIGURE 1** | Illustration of OVFD enabling recognition of newly introduced food categories beyond predefined labels.

## 1 | Introduction

Open-vocabulary detection (OVD) offers a promising direction for scalable food object detection by localizing categories specified by natural-language prompts rather than a predefined label set. As illustrated in Figure 1, open-vocabulary food object detection (OVFD) enables the recognition of newly introduced food categories beyond predefined labels. Vision-language models (VLMs) such as CLIP and ALIGN enable this paradigm via joint image-text embedding spaces (Radford et al. 2021; Jia et al. 2021). Nevertheless, transferring OVD to fine-grained food imagery is challenging: visually similar foods may differ only in subtle texture and shape cues, and background elements (plates, packaging, utensils, conveyor belts) can dominate model attention and yield background-driven predictions, especially when prompt tuning overfits to spurious context (K. Zhou et al. 2022a, 2022b).

Most existing OVD studies focus on improving region-text alignment through distillation or region-level pretraining, and on scaling to large label spaces with open corpora or weak supervision (Zareian et al. 2021; Gu et al. 2021; Zhong et al. 2022; X. Zhou et al. 2022c). Related formulations include zero-shot detection and weakly supervised detection, which treat image-level labels as multiple instance learning (MIL) problems to mine object regions when box annotations are unavailable (Bilen and Vedaldi 2016; Tang et al. 2017, 2020). In food computing, multimodal learning has been widely studied for recognition, recipe and ingredient understanding, and dietary assessment (Min et al. 2019; Jiang and Min 2020; Marin et al. 2021; Myers et al. 2015), but scalable localization of novel food categories in cluttered scenes remains underexplored. Knowledge distillation provides a general paradigm for transferring supervision in such settings (Hinton et al. 2015). Zero-shot and open-world object detection further study unseen-category recognition with semantic supervision and unknown-class handling (Bansal et al. 2018; Z. Li et al. 2019; Rahman et al. 2019; Joseph et al. 2021).

For food-science and food-engineering applications, localization is a necessary intermediate step rather than the final decision. Bounding boxes provide spatial evidence that can be passed to portion-size estimation, nutrient-intake logging, defect or foreign-material review, package-food separation, and human-in-the-loop inspection interfaces. Therefore, the objective of this work is limited to improving open-vocabulary food localization under cluttered scenes, while downstream safety or quality decisions still require task-specific validation.

Within this scope, we propose an OVFD framework with two complementary components: (i) a dynamic prompt distribution network (DPDN) that adapts attribute-oriented prompts according to proposal-level uncertainty and (ii) a hybrid supervision strategy that combines box annotations for Base categories with large-scale image-level tags, optimized via a density-aware MIL objective with objectness reweighting. The overall framework is summarized in Figure 3. The objectives of this study were to evaluate whether instance-adaptive prompting (DPDN) and objectness-reweighted DA-MIL reduce background-driven predictions and improve discrimination among visually similar foods, and to assess the resulting detector under leakage-controlled zero-shot, cross-domain, and limited external food-scene settings relevant to practical food inspection. Recent open-vocabulary detection methods also leverage localized vision-language matching, retrieval augmentation, region-language pretraining, and grounded pretraining to strengthen region-text alignment (Bravo et al. 2022; Kim et al. 2024; L. H. Li et al. 2021; Liu et al. 2023).

## 2 | Materials and Methods

### 2.1 | Datasets and Preprocessing

The proposed framework was evaluated under an open-vocabulary setting using a zero-shot detection (ZSD) protocol. The detector was trained with box annotations for Base categories

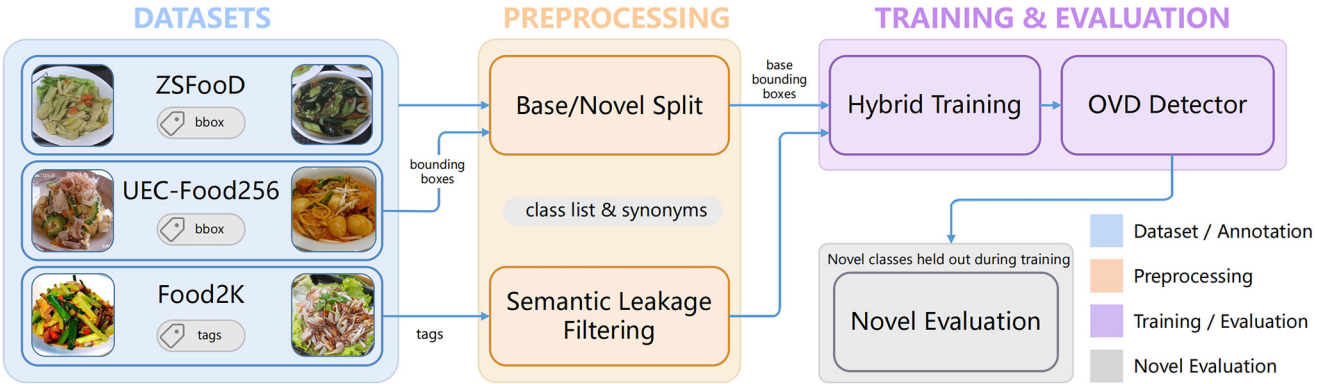


**FIGURE 2** | Food2K leakage-filtering workflow to prevent semantic overlap with the Base/Novel splits under hybrid supervision.

and evaluated on disjoint Novel categories not observed during training, testing whether the model can localize and recognize previously unseen food categories without additional box-level annotations.

#### 2.1.1 | ZSFooD

This publicly available dataset was used as the primary benchmark for fine-grained food object detection and contains 228 food categories annotated with bounding boxes. We followed the standard split, partitioning categories into observed Base classes and disjoint Novel classes.

#### 2.1.2 | UEC-Food256

To assess cross-dataset robustness, we additionally used UEC-Food256 (Kawano and Yanai 2014), which contains 256 food categories with bounding box annotations. This dataset was partitioned into Base and Novel subsets following the same ZSD principle.

#### 2.1.3 | Food2K (Auxiliary Weak Supervision)

To introduce large-scale semantic diversity, we incorporated Food2K (Min et al. 2023), which in the released training subset used here contains 1944 food categories and 1,036,594 images with image-level tags but without spatial localization. In practical food-image applications, image-level tags are often more readily available than bounding-box annotations because they can be derived from menu records, retail product-image repositories, dietary photo logs, and production or inventory documentation. Food2K therefore serves as a realistic weak-supervision source for expanding category coverage without requiring box annotation for every new food item. To reduce potential label leakage, we applied a strict semantic filtering protocol to exclude Food2K classes whose tags overlap with the predefined Base/Novel category splits of ZSFooD or UEC-Food256. The filtering workflow is illustrated in Figure 2.

To make the weak-supervision source auditable, we define a split-specific filtering protocol before any Food2K image-level tags are used. Food2K labels and benchmark category names are first normalized by lowercasing, removing punctuation, replacing hyphens and slashes with spaces, collapsing whitespace, and applying singular–plural normalization. The remaining labels are then screened by a manually reviewed synonym list, keyword-set inclusion, keyword Jaccard overlap, and string-similarity matching. Any Food2K class that overlaps with a Base or Novel benchmark class under these rules is removed from the weak-supervision pool.

The filtering is performed independently for ZSFooD and UEC-Food256 because their class inventories differ. For ZSFooD, the strict audit removes 249 Food2K classes and 139,191 images, leaving 1695 classes and 897,403 images for weak supervision. For UEC-Food256, the audit removes 225 classes and 134,655 images, leaving 1719 classes and 901,939 images. The removed-class list and retained-class list are released as supplementary material so that the filtering can be independently inspected. The Food2K semantic-leakage audit and strict filtering summart is provided in Table 8.

All experiments reported in this study use the strict-filtered Food2K subset described above. This protocol provides auditable rule-based criteria and corresponding class-image statistics so that Novel-category performance can be interpreted under an explicitly leakage-controlled weak-supervision setting.

### 2.2 | Overall Framework Architecture

We propose an OVFD framework that integrates object-level supervision (bounding boxes) with image-level weak supervision (tags). The framework follows a two-stage detection pipeline (Faster R-CNN; Ren et al. 2017) and leverages vision-language representations (e.g., CLIP; Radford et al. 2021) to align region features with text descriptions. A region proposal network (RPN) generates class-agnostic candidate regions, and two components are introduced to improve fine-grained discrimination, as summarized in Figure 3.

The framework consists of (1) a DPDN, which estimates instance ambiguity and adaptively reweights attribute-oriented prompt templates (e.g., emphasizing texture or shape) for visually similar food categories, and (2) a hybrid-supervised detection head with density-aware MIL (DA-MIL), which learns from both box-

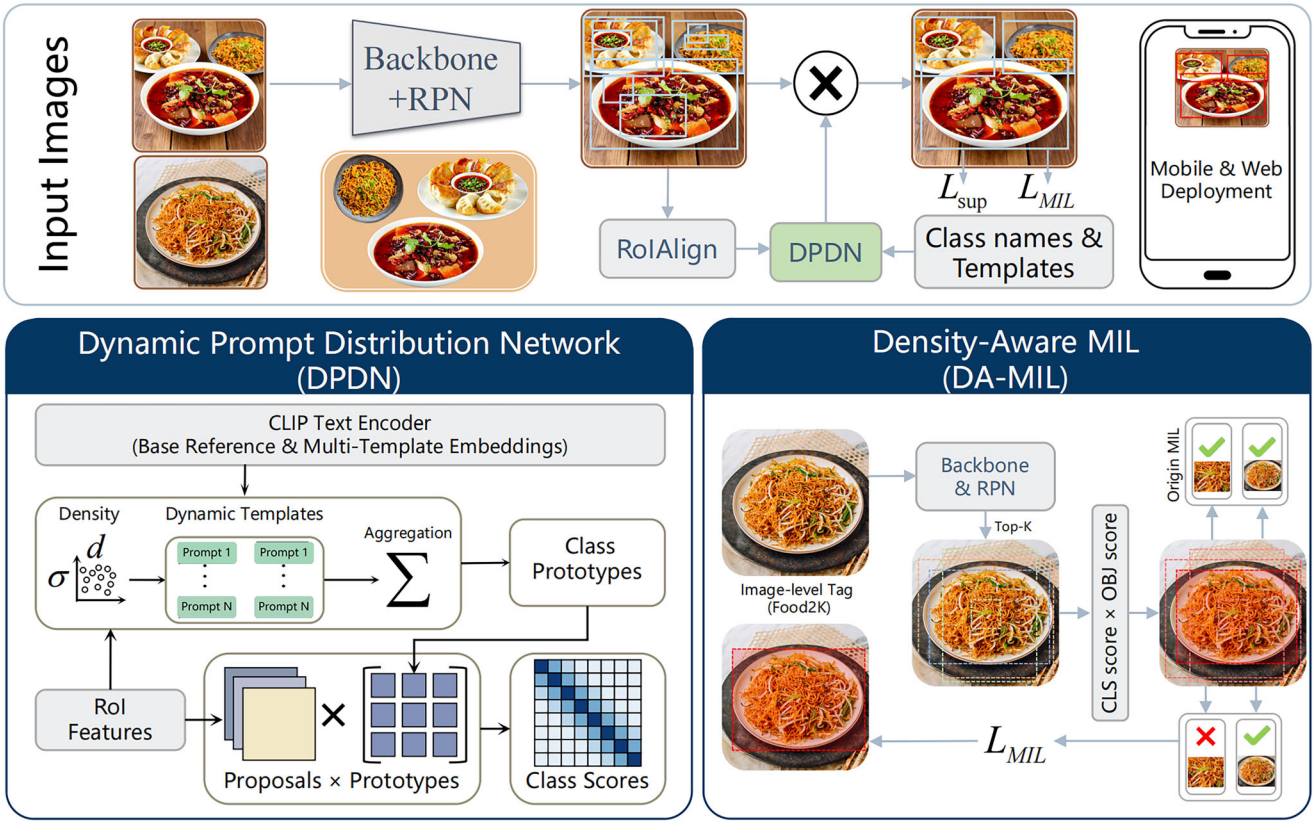


**FIGURE 3** | Overall architecture of the proposed OVFD framework with hybrid supervision, DPDN, and density-aware MIL (DA-MIL).

annotated Base data and tag-labeled Food2K data by mining object regions from image-level supervision while suppressing background responses commonly induced by tableware and clutter.

## 2.3 | Dynamic Prompt Distribution Network

Food categories often exhibit both high intra-class variation (e.g., presentation differences) and high inter-class similarity (e.g., fine-grained variants). Consequently, the amount of attribute detail needed in the text description can vary across instances. The DPDN module addresses this by estimating instance ambiguity from visual features and producing an adaptive distribution over prompt templates.

### 2.3.1 | Variational Latent Modeling

Rather than using fixed text-template heuristics, instance ambiguity is modeled as a probabilistic latent distribution conditioned on the visual input. Let $x$ denote the region-of-interest (RoI)-pooled feature vector of a candidate region. A probabilistic encoder predicts the parameters of a latent Gaussian distribution:

$$\mu = f_{\mu}(E(x)), \log \sigma^2 = f_{\sigma}(E(x)) \tag{1}$$

where $E(\cdot)$ is a shared MLP encoder and $\mu$ and $\log \sigma^2$ denote the mean and log-variance of the latent distribution, respectively. Here, $\sigma$ reflects instance-level visual ambiguity: larger uncertainty typically corresponds to visually similar, fine-grained cases that benefit from more attribute-focused prompts.

### 2.3.2 | Dynamic Template Weighting via Sampling

To obtain adaptive prompt weights, a latent variable $z$ is sampled using the reparameterization trick:

$$z = \mu + \sigma \odot \varepsilon, \varepsilon \sim \mathrm{N}(0, I) \tag{2}$$

This latent variable encodes instance-specific attribute cues. The latent code $z$ is then transformed into modulation parameters $\gamma$ (scaling) and $\beta$ (shifting) via two projection heads:

$$\gamma = sigmoid\left(W_{\gamma} z\right), \beta = W_{\beta} z \tag{3}$$

These parameters modulate the template logits, and the final prompt-weight distribution is computed as follows:

$$w_{\text{gate}} = \mathrm{Softmax}\left(\mathrm{Decoder}(z) \odot \gamma + \beta\right) \tag{4}$$

where Decoder$(\cdot)$ outputs a vector of template logits and Softmax$(\cdot)$ is applied across prompt templates.

DPDN outputs a normalized weight vector over multiple prompt templates for each detected instance. For ambiguous, fine-grained cases, the distribution tends to emphasize attribute-rich templates (e.g., texture- or shape-oriented prompts), whereas for

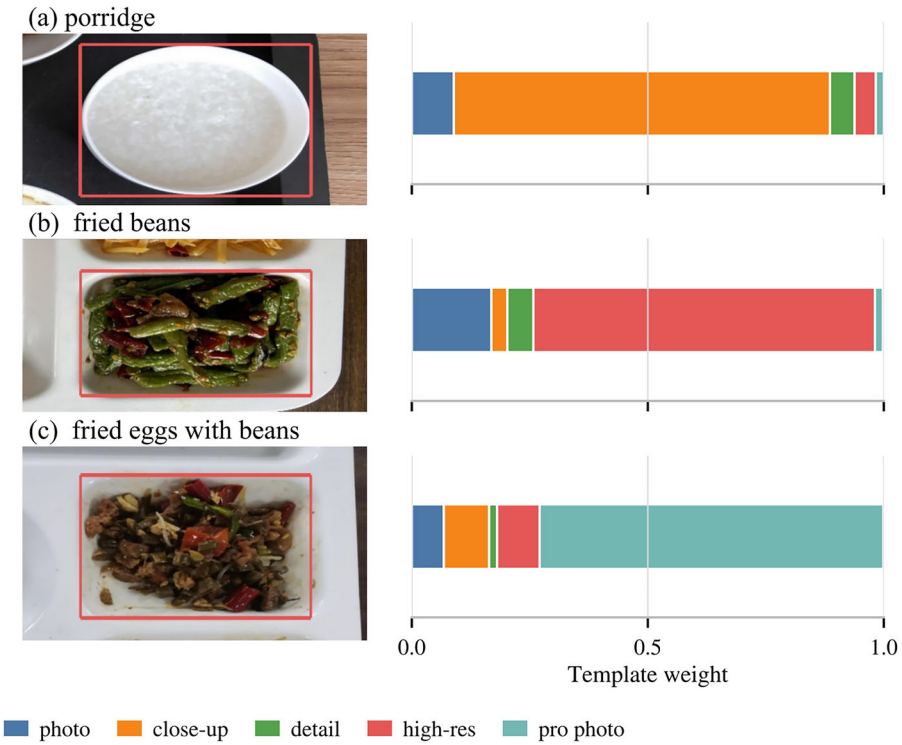


**FIGURE 4** | Examples of instance-adaptive prompt-template weighting. Left: detection visualization; right: normalized prompt-template weights predicted for the corresponding instance.

visually distinct cases, it remains less concentrated. Representative template-weight distributions are visualized in Figure 4.

## 2.4 | Hybrid Learning Strategy

Model training uses a hybrid objective that combines fully supervised detection on box-annotated Base categories and weakly supervised learning on tag-labeled Food2K images.

To ensure that the latent distribution approximates a standard Gaussian and to mitigate overfitting, a Kullback–Leibler (KL) divergence loss is included for the DPDN module (Kullback and Leibler 1951):

$$\mathcal{L}_{\mathrm{KL}} = -\frac{1}{2}\sum_{j}\left(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2\right) \tag{5}$$

This regularization encourages a continuous and smooth latent space for fine-grained attributes.

### 2.4.1 | Supervised Branch (Base Classes)

For the Base classes of ZSFooD and UEC-Food256 with bounding box annotations, the standard Faster R-CNN detection losses are minimized, including classification loss and bounding box regression loss. This branch provides core localization capability for food objects under complex visual backgrounds (e.g., plates and tableware).

### 2.4.2 | Weakly Supervised Learning via Density-Aware MIL

To exploit the semantic diversity of Food2K without spatial annotations, learning on tag-labeled images is formulated as a MIL problem. This weak-supervision setting is practically relevant because large-scale food data collected in real application scenarios are more likely to provide image-level category tags than instance-level box annotations. A weakly labeled image containing category $c$ is treated as a bag of candidate regions generated by the RPN. The objective is to assign higher contribution to regions that correspond to the target food item and to propagate learning signals to the detector without ground-truth boxes.

Traditional MIL often uses a max-score update, where only the single highest-scoring region is used. In fine-grained food object detection, this can overemphasize a salient part (e.g., a garnish) rather than the whole dish. To mitigate this, we adopt a density-aware top-K soft-selection strategy that aggregates supervision signals across multiple high-confidence regions. Related weakly supervised detection lines study instance mining and weakly supervised open-vocabulary detection, which motivates selecting multiple informative regions under weak labels (J. Lin et al. 2024).

Specifically, for a given category $c$, the top-$K$ regions are retrieved based on the classification logits $s_{i,c}$. To reduce learning from background regions that may receive spuriously high scores (e.g., tableware patterns), each region is reweighted using the objectness prior from the RPN. The contribution weight for each

selected region is computed as follows:

$$w_i = \frac{\exp\left(s_{i,c}/\tau\right) \cdot sigmoid\left(O_i\right)}{\sum_{j \in \mathcal{P}_{topk}} \exp\left(s_{j,c}/\tau\right) \cdot sigmoid\left(O_j\right)} \tag{6}$$

where $\tau$ is a temperature hyperparameter controlling the smoothness of the distribution, $\mathcal{P}_{topk}$ denotes the selected top-$K$ proposal set, $O_i$ is the objectness logit of proposal $i$, and $sigmoid(\cdot)$ is the logistic function. The final MIL loss is defined as the objectness-weighted binary cross-entropy over these selected instances, where $p_{i,c}$ is the predicted probability for category $c$:

$$\mathcal{L}_{\text{mil}} = -\sum_{i \in \mathcal{P}_{topk}} w_i \log\left(p_{i,c}\right) \tag{7}$$

By integrating weak supervision into the density-aware formulation, the model learns more stable region representations from tag-only images and reduces the localization instability commonly observed in standard MIL.

## 2.5 | Implementation Details

The framework was implemented in PyTorch using Detectron2. Unless otherwise stated, a Faster R-CNN detector with a ResNet-50 backbone was used (He et al. 2016). Optimization used stochastic gradient descent(SGD) with momentum 0.9 and a step learning-rate schedule. ResNet backbones are commonly pretrained on ImageNet (Russakovsky et al. 2015), and detector implementations are often benchmarked on MS COCO (T. Y. Lin et al. 2014). Multi-scale feature pyramids are also widely used in object detection (T. Y. Lin, Dollar, et al. 2017).

For hybrid supervision, mini-batches were constructed from both box-annotated (Base) data and tag-labeled Food2K data using a multi-dataset sampling strategy. Unless otherwise noted, DA-MIL used $k = 5$ and $\tau = 1.0$ with RPN objectness reweighting, while the DPDN used $k = 10$ nearest neighbors, $\sigma = 0.1$, $\gamma = 0.8$, and $\beta = 0.5$ for density-aware prompt selection.

For reproducibility, the experiments use the following training protocol unless otherwise specified. The detector is trained with SGD, momentum 0.9, weight decay 0.0001, base learning rate 0.02, 180,000 iterations, 1000 warm-up iterations, and learning-rate decay at 60,000 and 80,000 iterations. For mixed supervised/weakly supervised updates, each multi-dataset batch uses four box-annotated images and 16 Food2K tag-labeled images, corresponding to a 1:4 supervised-to-weak sampling ratio. The main detection branch uses ResizeShortestEdge with short side 800 and max size 1333, while the weakly supervised Food2K branch uses short side 400 and max size 667. Test-time evaluation uses short side 800, max size 1333, and class-agnostic NMS threshold 0.5; benchmark AP computation uses score threshold 0.0.

The prompt templates used by DPDN are listed as follows. DPDN dynamically selects three to eight templates from the following 10 candidates: “a photo of a {},” “a close-up of {},” “a detailed view of {},” “a high resolution image of {},” “a professional photograph of {},” “an image focusing on the texture of {},” “a visual highlighting the color of {},” “a shot emphasizing the shape of {},” “a picture capturing the unique appearance of {},” and “a photo showing the distinctive features of {}.” All repeated-run experiments used random seeds 0, 1, and 2 with identical data splits.

## 2.6 | Application Prototype and Deployment

To illustrate potential use in practice, we developed a lightweight mobile client and web interface that allow users to upload images or capture photos for on-demand analysis. The trained detector is deployed on a server and accessed through these clients, returning predicted categories together with bounding-box visualizations. The web interface also provides a structured view (class/score/bounding box) to support review and reporting. This prototype is intended to demonstrate practical integration of open-vocabulary detection into food-image analysis workflows (Figure 5).

Prototype latency was measured server-side with batch size 1 on an NVIDIA A800 80GB PCIe GPU. Using short side 800 and max size 1333, mean preprocessing time was 59.8 ms, model-inference time was 153.8 ms, post-processing time was 0.56 ms, and total latency was 214.2 ms per image, corresponding to 4.67 FPS over 280 measured iterations after 20 warm-up runs.

# 3 | Results and Discussion

## 3.1 | Comparison With Representative Methods on ZSFooD

To evaluate the proposed framework, comparisons were made with representative OVD methods on the ZSFooD dataset. Quantitative results are summarized in Table 1.

As shown in Table 1, the proposed method achieves 15.1% Novel AP50 while maintaining 90.0% Base AP50 (72.7% All AP50) on ZSFooD. Compared with representative baselines such as CLIFF and Detic, the higher Novel performance indicates improved generalization to previously unseen food items without sacrificing localization for familiar items. From a food engineering perspective, this behavior is desirable because new stock-keeping units (SKUs) and fine-grained variants can be introduced without immediate category-specific re-annotation and retraining.

We also compare with LLMDet, which augments open-vocabulary detector training using large language model (LLM)-generated captions to enrich language supervision (Fu et al. 2025). Although LLMDet achieves 57.3% Base AP50 on ZSFooD, its Novel performance is 5.3% AP50 (Table 1). This suggests that generic caption-style supervision may not directly translate to fine-grained food categories, where discriminative differences are subtle and inference-time text cues are often limited to short class names. In contrast, the proposed DPDN explicitly strengthens class-name prompts with attribute-oriented cues and adapts them per proposal, which may better match practical text-prompt usage at inference. Transformer-based detector architectures such as DETR provide an alternative design for detection (Carion et al. 2020).

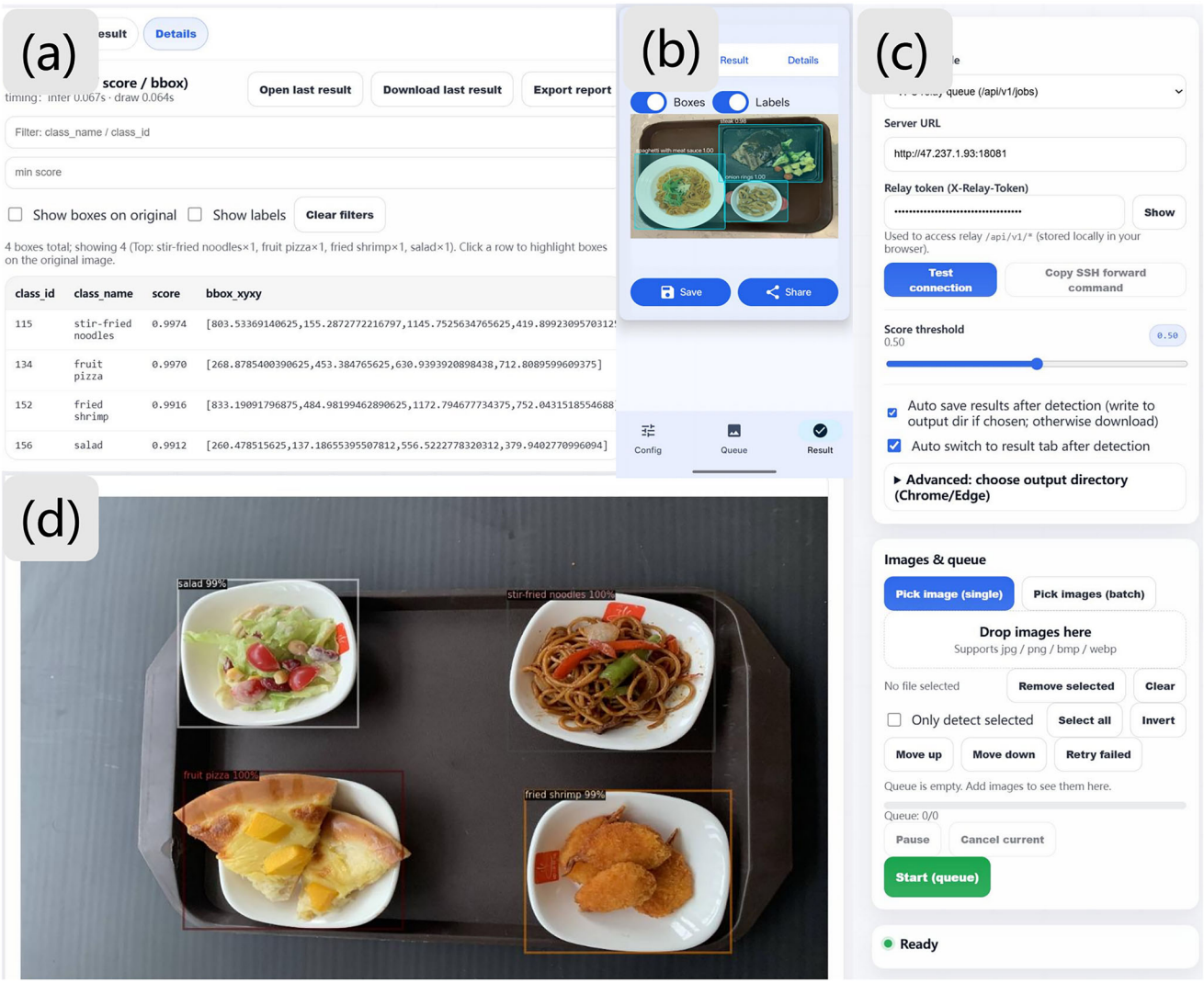


**FIGURE 5** | Proof-of-concept client–server deployment of the proposed OVFD system. (a) Detailed web interface summarizing detected food categories, confidence scores, and bounding boxes. (b) Mobile detection results with food item localization. (c) Web-based interface. (d) Web interface result visualization.

**TABLE 1** | Comparison with representative methods on the ZSFooD dataset.

| Method | Backbone | Detector | Novel AP50 | Base AP50 | All AP50 |
|---|---|---|---|---|---|
| OVR-CNN | RN50 | Faster R-CNN | 1.2 | 31.8 | 24.8 |
| Object-Centric-OVD | RN50 | Faster R-CNN | 6.2 | 91.1 | 71.5 |
| CORA+ | RN50 | Faster R-CNN | 5.2 | 8.5 | 7.7 |
| Detic | RN50 | Faster R-CNN | 5.4 | 90.6 | 71.0 |
| CoDet | RN50 | Faster R-CNN | 7.0 | 77.6 | 61.3 |
| VLDet | RN50 | Faster R-CNN | 8.4 | 79.1 | 62.8 |
| CLIFF | RN50 | Faster R-CNN | 10.4 | 89.7 | 71.4 |
| SIA-OVD | RN50 | Faster R-CNN | 2.8 | 23.2 | 18.5 |
| RALF-Centric | RN50 | Faster R-CNN | 6.6 | 82.8 | 65.2 |
| OV-DQUO | RN50 | Faster R-CNN | 9.5 | 12.6 | 11.9 |
| LLMDet | RN50 | DETR | 5.3 | 57.3 | 45.3 |
| Ours | RN50 | Faster R-CNN | 15.1 | 90.0 | 72.7 |

*Note:* Metrics are AP50 for Novel, Base, and All categories.

**TABLE 2** | Comparison with representative methods on the UEC-Food256 dataset to evaluate cross-domain robustness.

| Method | Backbone | Detector | Novel AP50 | Base AP50 | All AP50 |
|---|---|---|---|---|---|
| Object-Centric-OVD | RN50 | Faster R-CNN | 4.7 | 12.3 | 9.6 |
| Detic | RN50 | Faster R-CNN | 3.5 | 6.1 | 5.2 |
| CoDet | RN50 | Faster R-CNN | 4.6 | 13.9 | 10.5 |
| VLDet | RN50 | Faster R-CNN | 5.4 | 6.0 | 5.8 |
| CLIFF | RN50 | Faster R-CNN | 15.2 | 22.5 | 19.9 |
| LLMDet | RN50 | DETR | 8.5 | 3.6 | 5.3 |
| Ours | RN50 | Faster R-CNN | 16.5 | 22.6 | 20.4 |

### 3.2 | Cross-Domain Generalization on UEC-Food256

To assess robustness in varying environments (e.g., lighting conditions, plating styles, and cultural food varieties), the model was evaluated on UEC-Food256. As presented in Table 2, the proposed framework achieves 16.5% Novel AP50 and 22.6% Base AP50, providing higher Novel-category performance than the best-performing comparison method (CLIFF). Cross-domain robustness can also be influenced by backbone and loss design and weakly supervised mining strategies (T. Y. Lin, Goyal, et al. 2017b; J. Lin et al. 2024).

This cross-domain performance trend may be partly attributable to DA-MIL. When transferring across environments, background appearance (e.g., tableware, tablecloth patterns, packaging) and imaging conditions can shift, and weak supervision can amplify spurious correlations if only the single highest-scoring proposal is updated. By aggregating the top-$K$ proposals and down-weighting low-objectness regions, DA-MIL reduces the influence of background-dominant regions and encourages learning from more complete food regions, which may improve robustness in practical deployment.

### 3.3 | Additional Evaluation Rigor

Because AP50 alone may overstate localization quality, we additionally report AP75 and COCO-style AP averaged over IoU thresholds from 0.50 to 0.95 (Table 9). On ZSFooD, the proposed method achieves Novel AP50/AP75/AP of 15.1/12.8/9.4, compared with 10.4/8.6/6.1 for CLIFF and 5.4/4.3/3.1 for Detic. On UEC-Food256, the proposed method achieves 16.5/13.2/8.7 on Novel categories, compared with 15.2/12.0/7.9 for CLIFF. These results show that the gains remain visible under stricter localization criteria rather than appearing only at AP50.

Table 10 reports repeated-run stability over three random seeds on ZSFooD. The full model reaches Novel AP50 = 14.89 ± 0.50, Base AP50 = 89.98 ± 0.30, and All AP50 = 72.69 ± 0.35. The ablations show that DPDN primarily improves Novel-category generalization, whereas DA-MIL contributes more directly to recovering Base-category stability under weak supervision. These results indicate stable but limited overall gains, with the largest improvement observed for Novel categories.

For error analysis, predicted boxes were first filtered using a confidence score threshold of 0.30. This threshold was selected to retain sufficient low- and medium-confidence predictions for diagnostic error counting while excluding very low-confidence noise. The remaining predictions were assigned to ground-truth boxes by one-to-one greedy IoU-based matching. A prediction was counted as a true-positive match only when it had IoU ≥0.50 with a ground-truth box of the same class, and each prediction could be matched to at most one ground-truth instance. Fine-grained class confusion was counted when a prediction overlapped a ground-truth food item with IoU ≥0.50 but was assigned to a different food class. Background false positives were predictions with IoU <0.50 against all ground-truth boxes, whereas missed detections were ground-truth instances without any matched prediction above the score threshold. Localization errors were recorded when a correct-class prediction localized the object at IoU ≥0.50 but failed the stricter IoU ≥0.75 criterion, and duplicate/overlap errors were recorded for additional predictions assigned to already matched instances or fragmented predictions in dense tray scenes. Because the error distribution depends on the score threshold and matching protocol, the results in Table 11 should be interpreted as protocol-specific diagnostic statistics rather than evidence that background-driven detections have been fully resolved.

Table 11 summarizes validation-set error analysis on the full ZSFooD validation set at score threshold 0.30. Fine-grained food confusion is the dominant failure mode, accounting for 92.0% of all false positives and 82.1% of all false negatives. By contrast, background false positives account for only 1.1% of all false positives, and localization errors account for only 0.6% of all error items. This pattern indicates that the main remaining bottleneck is discrimination among visually similar food categories rather than background leakage or box regression.

### 3.4 | Ablation Studies

Ablation studies were performed on the ZSFooD dataset to quantify the contribution of each component, including DPDN, DA-MIL, and the hybrid supervision strategy. Beyond the core ablations reported in Table 3, further analyses assessed (i) the necessity of KL regularization in DPDN, (ii) sensitivity to the key DA-MIL hyperparameters (top-k and tau), (iii) the contribution of objectness reweighting, and (iv) deterministic mean-only

**TABLE 3** | Ablation study on the ZSFooD dataset quantifying the contribution of each component in the proposed framework.

| Configuration | Novel AP50 | Base AP50 | All AP50 |
|---|---|---|---|
| Baseline | 10.73 | 89.92 | 71.68 |
| +DPDN | 11.68 | 89.46 | 71.54 |
| +Food2K (max-MIL) | 13.56 | 70.21 | 57.16 |
| +Food2K (DA-MIL) | 13.35 | 89.48 | 71.94 |
| +DPDN +Food2K (DA-MIL) | 15.03 | 90.02 | 72.75 |

**TABLE 4** | Ablation on KL regularization in DPDN on the ZSFooD dataset.

| Setting | Novel AP50 | Base AP50 | All AP50 |
|---|---|---|---|
| DPDN w/o KL | 10.78 | 89.45 | 71.33 |
| DPDN w/ KL | 11.68 | 89.46 | 71.54 |

prompting versus stochastic sampling. Results are summarized in Tables 4–7.

Unless otherwise stated, each ablation changes only the corresponding component or hyperparameter and keeps the remaining settings identical to the main training protocol.

Table 3 summarizes the contribution of each component. DPDN improves Novel performance over the Baseline (+0.95 points), indicating that adapting prompts toward texture/shape cues benefits discrimination among visually similar foods. Incorporating Food2K with naive max-score MIL improves Novel AP50 but markedly degrades Base performance, consistent with unstable instance selection under weak labels. Replacing max-score MIL with DA-MIL recovers Base performance while improving overall AP50, suggesting that top-k aggregation with objectness reweighting stabilizes learning from tagged images. The full model (DPDN + Food2K with DA-MIL) achieves the best overall performance.

*DPDN KL regularization*: The KL term was ablated by removing the KL loss and compared against the default setting with KL regularization while keeping all other training settings fixed (Table 4). Removing KL reduces Novel AP50 from 11.68 to 10.78 while leaving Base AP50 nearly unchanged (89.46 vs. 89.45), decreasing the overall AP50 from 71.54 to 71.33. This regularization corresponds to the KL divergence term that encourages a smooth latent distribution (Kullback and Leibler 1951).

*Hyperparameter sensitivity*: The top-k selection and temperature tau were varied in DA-MIL to assess the sensitivity of weak-supervision mining under different sparsity/smoothness levels (Table 5). The setting top-k = 5 and tau = 1.0 achieves the best All AP50 (71.94) among the tested configurations, while other choices lead to only modest changes (71.59–71.81), suggesting that the method is relatively insensitive within a reasonable range of top-k and tau.

*Objectness reweighting in DA-MIL*: The proposal objectness prior was removed (i.e., objectness reweighting was disabled) to isolate its role in suppressing background-dominant proposals and stabilizing weakly supervised learning (Table 6). Disabling objectness reweighting decreases All AP50 from 71.94 to 71.26 and reduces Novel AP50 from 13.35 to 12.63, suggesting that objectness priors help suppress background-dominant proposals during weak supervision.

*Mean-only versus sampling*: Stochastic sampling was replaced with a deterministic mean-only variant (without sampling) to evaluate whether variational sampling contributes to generalization beyond deterministic prompting (Table 7). The mean-only variant yields lower Novel AP50 (10.85 vs. 11.68) and a lower All AP50 (71.32 vs. 71.54) than sampling, supporting the use of stochastic sampling during training.

### 3.5 | Qualitative Analysis

To complement the quantitative evaluation, qualitative examples produced by the deployed prototype are shown. Figure 5 presents representative detection outputs obtained through the mobile application and the web interface, illustrating multi-item localization and consistent labeling in user-captured images. Figure S1 provides a qualitative comparison with a baseline under complex multi-item scenes. In Figure S1, the proposed method (bottom) often yields more complete multi-item localization and more specific food labels than the baseline (top), while reducing background-driven false positives in cluttered meal-tray scenes.

In deployment-oriented scenarios, the proposed framework tends to localize dominant food items while suppressing many background distractors (e.g., tableware and clutter), without category-specific retraining for new categories. These properties may support assistive quality screening, sorting review for new product lines, and more consistent human-in-the-loop visual assessment. Computer vision has been used for food quality inspection and measurement in food engineering (Brosnan and Sun 2004; Russ 2015).

A manually annotated external pilot set of 195 web-collected household and restaurant food-scene images with 530 bounding boxes was used as a limited real-world plausibility check. On this external set, the deployed model achieved AP50 = 40.7, AP75 = 30.5, and AP = 25.2. These results are reported to assess transfer under realistic clutter and domain shift, not to claim production-line readiness.

The external evaluation set was assembled from web-collected food-scene images associated with recipe and food-content

**TABLE 5** | Sensitivity to DA-MIL hyperparameters on the ZSFooD dataset.

| Top-k | tau | Novel AP50 | Base AP50 | All AP50 |
|---|---|---|---|---|
| 3 | 1.0 | 12.55 | 89.31 | 71.63 |
| 5 | 0.5 | 13.10 | 89.10 | 71.59 |
| 5 | 1.0 | 13.35 | 89.48 | 71.94 |
| 5 | 2.0 | 13.05 | 89.18 | 71.64 |
| 10 | 1.0 | 13.24 | 89.34 | 71.81 |

**TABLE 6** | Contribution of objectness reweighting in DA-MIL on the ZSFooD dataset.

| Setting | Novel AP50 | Base AP50 | All AP50 |
|---|---|---|---|
| DA-MIL without objectness reweight | 12.63 | 88.81 | 71.26 |
| DA-MIL with objectness reweight | 13.35 | 89.48 | 71.94 |

**TABLE 7** | Ablation on deterministic (mean-only) prompting versus stochastic sampling in DPDN on the ZSFooD dataset.

| Setting | Novel AP50 | Base AP50 | All AP50 |
|---|---|---|---|
| Mean-only (without sampling) | 10.85 | 89.42 | 71.32 |
| Sampling | 11.68 | 89.46 | 71.54 |

**TABLE 8** | Food2K semantic-leakage audit and strict filtering summary.

| Benchmark | Base classes | Novel classes | Removed classes | Removed images | Retained classes | Retained images |
|---|---|---|---|---|---|---|
| ZSFooD | 184 | 44 | 249 | 139,191 | 1695 | 897,403 |
| UEC-Food256 | 205 | 51 | 225 | 134,655 | 1719 | 901,939 |
| Rules | — | — | Normalized exact match | Keyword inclusion | Jaccard $\geq 0.60$ | String similarity $\geq 0.86$ |

**TABLE 9** | Stricter localization metrics (AP50/AP75/AP) beyond AP50.

| Dataset | Method | Novel AP50/AP75/AP | Base AP50/AP75/AP | All AP50/AP75/AP | Note |
|---|---|---|---|---|---|
| ZSFooD | Ours | 15.1/12.8/9.4 | 90.0/87.2/69.1 | 72.7/69.4/56.8 | Strict-filtered Food2K |
| ZSFooD | CLIFF | 10.4/8.6/6.1 | 89.7/86.1/68.0 | 71.4/68.1/55.7 | Best comparable baseline |
| ZSFooD | Detic | 5.4/4.3/3.1 | 90.6/87.4/69.0 | 71.0/68.6/55.8 | Reference baseline |
| UEC-Food256 | Ours | 16.5/13.2/8.7 | 22.6/19.3/13.4 | 20.4/17.2/11.8 | Strict-filtered Food2K |
| UEC-Food256 | CLIFF | 15.2/12.0/7.9 | 22.5/18.9/13.0 | 19.9/16.5/11.1 | Best comparable baseline |

**TABLE 10** | Repeated-run stability over three random seeds on ZSFooD.

| Configuration | Novel AP50 mean $\pm$ std | Base AP50 mean $\pm$ std | All AP50 mean $\pm$ std |
|---|---|---|---|
| Baseline | $10.81 \pm 0.26$ | $89.96 \pm 0.22$ | $71.74 \pm 0.23$ |
| + DPDN | $11.58 \pm 0.32$ | $89.42 \pm 0.24$ | $71.50 \pm 0.26$ |
| + Food2K (DA-MIL) | $13.23 \pm 0.45$ | $89.52 \pm 0.30$ | $71.93 \pm 0.33$ |
| Full model | $14.89 \pm 0.50$ | $89.98 \pm 0.30$ | $72.69 \pm 0.35$ |

**TABLE 11** | ZSFooD validation-set error analysis at score threshold 0.30.

| Error type | Count | Share | Interpretation |
|---|---|---|---|
| Fine-grained class confusion (FP) | 22,591 | 92.0% of all FP | Primary false-positive source; the model more often predicts the wrong food class than a non-food region. |
| Fine-grained class confusion (FN) | 12,721 | 82.1% of all FN | Ground-truth instances are frequently absorbed by semantically adjacent food categories. |
| Missed detections | 1479 | 9.5% of all FN | Secondary difficulty driven by small, low-contrast, or occluded food items. |
| Background false positives | 277 | 1.1% of all FP | Background leakage is present but is not the dominant bottleneck after DA-MIL. |
| Localization errors | 252 | 0.6% of all error items | AP degradation is driven far more by class confusion than by box regression. |
| Other overlap/duplicate errors | 2747 | 6.9% of all error items | Dense tray scenes can still cause box fragmentation or incorrect merging. |

**TABLE 12** | Composition and benchmark-overlap audit of the external food-scene evaluation set.

| Coarse food type | No. of fine categories | Images | Annotations | Benchmark overlap |
|---|---|---|---|---|
| Dumplings | 14 | 30 | 190 | No overlap |
| Egg Tarts | 4 | 15 | 56 | No overlap |
| Fried Chicken Cutlets | 2 | 15 | 52 | No overlap |
| Onion Rings | 2 | 15 | 48 | No overlap |
| Fried Shrimp | 9 | 15 | 45 | Base-class overlap |
| Roast Chicken | 7 | 15 | 31 | Base-class overlap |
| Pumpkin Pie | 2 | 15 | 25 | No overlap |
| French Fries | 2 | 15 | 22 | Base-class overlap |
| Pork Chops | 13 | 15 | 16 | No overlap |
| Fried Rice | 15 | 15 | 15 | Base-class overlap |
| Spaghetti with Meat Sauce | 6 | 15 | 15 | No overlap |
| Tomato Soup | 12 | 15 | 15 | No overlap |
| Total | 88 | 195 | 530 | |

*Note:* Benchmark overlap was audited at the class-name level against ZSFooD and UEC-Food256 Base/Novel classes and at the image-source level against the benchmark distribution pages and repositories.

websites, including recipe.rakuten.co.jp, cookpad.com, tudogostoso.com.br, taste.com.au, recetasgratis.net, douguo.com, russianfood.com, and xiangha.com. Specific source URLs are not listed in the manuscript, but image metadata and source information will be released with the external-set metadata subject to source-site terms.

The set contains 12 visually distinct coarse food types and 88 fine-grained food labels, with 195 images and 530 bounding-box annotations in total. The 12 reported coarse types are Dumplings, Egg Tarts, Fried Chicken Cutlets, Onion Rings, Fried Shrimp, Roast Chicken, Pumpkin Pie, French Fries, Pork Chops, Fried Rice, Spaghetti with Meat Sauce, and Tomato Soup. The mean number of annotations was 2.72 boxes per image. The external-set composition and overlap status are summarized in Table 12.

External-set annotation was completed by two annotators. Annotator 1 drew all bounding boxes and assigned fine-grained and coarse labels for the 195 images; Annotator 2 independently reviewed every image and annotation, identified missed food items, and added missing annotations. Bounding boxes were drawn tightly around visible food items using LabelImg in Pascal VOC format and then converted to COCO JSON for detector evaluation. Of the final 530 boxes, 498 instances (94.0%) were confirmed by Annotator 2, and 32 instances (6.0%) were added during review; minor annotation inconsistencies were identified and resolved during the review process.

To evaluate potential overlap with benchmark data, all 88 fine-grained external-set category names were compared with the Novel and Base class lists of ZSFooD and UEC-Food256 by

**TABLE 13** | External-set baseline comparison on the 195-image food-scene evaluation set.

| Method | Backbone | Detector | Base-overlap AP50/AP75/AP/AP95 | No-overlap AP50/AP75/AP/AP95 | Overall AP50/AP75/AP/AP95 |
|---|---|---|---|---|---|
| Detic | RN50 | Faster R-CNN | 17.2/13.1/10.6/4.9 | 14.0/10.3/8.3/3.6 | 14.7/10.9/8.8/3.9 |
| CoDet | RN50 | Faster R-CNN | 18.9/14.3/11.7/5.4 | 15.5/11.4/9.2/4.0 | 16.2/12.0/9.7/4.3 |
| Object-Centric-OVD | RN50 | Faster R-CNN | 19.6/14.9/12.1/5.7 | 16.0/11.8/9.6/4.3 | 16.8/12.5/10.1/4.6 |
| VLDet | RN50 | Faster R-CNN | 21.4/16.2/13.2/6.0 | 17.6/12.9/10.3/4.5 | 18.4/13.6/10.9/4.8 |
| LLMDet | RN50 | DETR | 25.6/18.9/15.3/6.7 | 21.2/15.2/12.0/4.9 | 22.1/16.0/12.7/5.3 |
| CLIFF | RN50 | Faster R-CNN | 32.2/24.8/20.5/10.0 | 27.0/20.5/16.8/7.7 | 28.1/21.4/17.6/8.2 |
| Ours | RN50 | Faster R-CNN | 43.7/34.0/29.1/15.2 | 39.9/29.6/24.1/12.8 | 40.7/30.5/25.2/13.3 |

*Note:* AP denotes COCO-style AP averaged over IoU thresholds from 0.50 to 0.95. Base-overlap contains four coarse types and 113 annotations; no-overlap contains eight coarse types and 417 annotations; the overall set contains 12 coarse types and 530 annotations.

exact string matching and normalized substring matching. No exact or near-exact matches were found with the Novel classes. Four coarse types (Fried Shrimp, Roast Chicken, French Fries, and Fried Rice) contained fine-grained names that overlapped with Base-class semantics, whereas the remaining eight coarse types had no benchmark-class overlap. The 195 source URLs were also checked to confirm that the images did not originate from ZSFooD, UEC-Food256, or Food2K distribution pages or associated dataset repositories.

To address external-set comparability, we additionally evaluated Detic, CoDet, Object-Centric-OVD, VLDet, LLMDet, and CLIFF on the same 195-image external food-scene set using the method configurations evaluated on UEC-Food256 (RN50 backbone; Faster R-CNN detector except LLMDet, which uses DETR). As summarized in Table 13, the proposed method achieved the highest overall AP50/AP75/AP/AP95 of 40.7/30.5/25.2/13.3, compared with 28.1/21.4/17.6/8.2 for the strongest baseline CLIFF. The advantage was observed on both the Base-class-overlap subset (43.7/34.0/29.1/15.2 vs. 32.2/24.8/20.5/10.0 for CLIFF) and the no-benchmark-overlap subset (39.9/29.6/24.1/12.8 vs. 27.0/20.5/16.8/7.7 for CLIFF). These results strengthen the domain-shift evaluation while still representing external image-level validation rather than pilot-plant or production-line deployment evidence.

Failure cases primarily occur under heavy occlusion, extreme illumination changes, or among visually near-duplicate categories where fine-grained appearance differences are subtle. The error analysis further shows that semantically adjacent food categories are more problematic than plate, tray, or packaging distractors, which is consistent with the remaining difficulty of fine-grained food discrimination in open-vocabulary settings.

### 3.6 | Implications for Food Inspection and Engineering Applications

From a food engineering perspective, the proposed open-vocabulary detection framework addresses several practical challenges commonly encountered in real-world food inspection and monitoring systems. In food inspection, retail, and dietary-monitoring environments, food items are rarely presented in isolation and are often accompanied by diverse background elements such as tableware, trays, packaging materials, reflective surfaces, or complex serving containers. By combining instance-adaptive prompting (DPDN) with objectness-reweighted weak supervision (DA-MIL), the proposed method reduces many background-driven responses while preserving the ability to expand to new food categories without immediate box-level re-annotation.

Accordingly, references to inspection, sorting, dietary assessment, and production environments should be understood as potential downstream contexts for upstream object localization. The present study does not evaluate defect detection, food-safety classification, quality grading, process-control decisions, or pilot-plant/production-line deployment.

Bounding-box localization should therefore be interpreted as an upstream perception step for downstream portion estimation, ingredient-level dietary assessment, assisted sorting review, and quality-inspection workflows rather than a final food-safety or quality decision. The stricter AP metrics, three-seed stability analysis, and validation-set error analysis characterize model performance more consistently, while also showing that the gains are modest and that fine-grained class confusion remains the dominant limitation.

The limited external evaluation on 195 manually annotated food-scene images provides additional evidence that the approach can transfer beyond the benchmark split, but the results should still be treated as assistive-use evidence only. The study does not yet constitute pilot-plant or production-line validation, and it does not replace human review in quality-inspection or dietary-monitoring workflows.

The method should therefore be interpreted as a human-in-the-loop assistive perception tool. In quality-inspection or safety-related workflows, detections should be verified by trained personnel or by downstream task-specific sensors or assays before operational decisions are made.

## 4 | Conclusion

These results suggest that OVFD is primarily limited by semantic misalignment between generic prompts and fine-grained

food attributes. By modeling proposal-level ambiguity, DPDN improves disambiguation among visually similar categories, and DA-MIL enables effective use of image-level tags by suppressing background-dominant proposals through objectness reweighting. Under strict Food2K leakage filtering, the resulting detector improves Novel-category performance on both ZSFooD and UEC-Food256 while retaining competitive Base-category localization, and the gains remain visible under AP75 and COCO-style AP.

Remaining limitations include proposal quality, residual ambiguity among highly similar foods, sensitivity to severe occlusion and illumination changes, and the absence of full pilot-plant or production-line scale validation. The limited external evaluation supports practical plausibility, but human verification remains necessary before using the system for food safety, quality grading, or scale-up decisions. Future work will evaluate the model under controlled pilot-plant conditions, expand real-world data collection, and integrate stronger vision-language backbones with task-specific quality and safety validation modules.

### Author Contributions

**Shoulong Liu**: conceptualization, methodology, software, writing – original draft. **Guorui Sheng**: writing – review and editing. **Shaojie Zhou**: visualization. **Bolin Yang**: data curation. **Weiqing Min**: supervision, writing – review and editing. **Shuqiang Jiang**: supervision, writing – review and editing.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Data Availability Statement

The code for this work is available at https://github.com/LTaiQin/dpdn-det. The datasets used in this study are available from their original providers under the corresponding terms of use. The external evaluation annotations, class labels, image metadata, and source-domain information will be released in the project GitHub repository upon publication; image redistribution will follow the terms of the original source websites.

### References


Bansal, A., K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran. 2018. "Zero-Shot Object Detection." In *Lecture Notes in Computer Science* 397–414. Springer. https://doi.org/10.1007/978-3-030-01246-5_24.

Bilen, H., and A. Vedaldi. 2016. "Weakly Supervised Deep Detection Networks." In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2846–2854. IEEE. https://doi.org/10.1109/CVPR.2016.311.

Bravo, M. A., S. Mittal, and T. Brox. 2022. "Localized Vision-Language Matching for Open-Vocabulary Object Detection." In *Lecture Notes in Computer Science* 393–408. Springer. https://doi.org/10.1007/978-3-031-16788-1_24.

Brosnan, T., and D. W. Sun. 2004. "Improving Quality Inspection of Food Products by Computer Vision——A Review." *Journal of Food Engineering* 61, no. 1: 3–16. https://doi.org/10.1016/S0260-8774(03)00183-3.

Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. 2020. "End-to-End Object Detection With Transformers." In *Lecture Notes in Computer Science* 213–229. Springer. https://doi.org/10.1007/978-3-030-58452-8_13.

Fu, S., Q. Yang, Q. Mo, et al. 2025. "LLMDet: Learning Strong Open-Vocabulary Object Detectors Under the Supervision of Large Language Models." arXiv. https://arxiv.org/abs/2501.18954.

Gu, X., T. Y. Lin, W. Kuo, and Y. Cui. 2021. "Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation." arXiv. https://arxiv.org/abs/2104.13921.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning For Image Recognition." In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. IEEE. https://doi.org/10.1109/CVPR.2016.90.

Hinton, G., O. Vinyals, and J. Dean. 2015. "Distilling the Knowledge in a Neural Network." arXiv. https://arxiv.org/abs/1503.02531.

Jia, C., Y. Yang, Y. Xia, et al. 2021. "Scaling up Visual and Vision-Language Representation Learning With Noisy Text Supervision." arXiv. https://arxiv.org/abs/2102.05918.

Jiang, S., and W. Min. 2020. "Food Computing for Multimedia." In Proceedings of the 28th ACM International Conference on Multimedia, 4782–4784. Association for Computing Machinery. https://doi.org/10.1145/3394171.3418544.

Joseph, K. J., S. Khan, F. S. Khan, and V. N. Balasubramanian. 2021. "Towards Open World Object Detection." In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5826–5836. IEEE. https://doi.org/10.1109/CVPR46437.2021.00577.

Kawano, Y., and K. Yanai. 2014. "Food Image Recognition With Deep Convolutional Features." In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, 589–593. Association for Computing Machinery. https://doi.org/10.1145/2638728.2641339.

Kim, J., E. Cho, S. Kim, and H. J. Kim. 2024. "Retrieval-Augmented Open-Vocabulary Object Detection." In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17427–17436. IEEE. https://doi.org/10.1109/CVPR52733.2024.01650.

Kullback, S., and R. A. Leibler. 1951. "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22, no. 1: 79–86. https://doi.org/10.1214/aoms/1177729694.

Li, L. H., P. Zhang, H. Zhang, et al. 2021. "Grounded Language-Image Pre-Training." arXiv. https://arxiv.org/abs/2112.03857.

Li, Z., L. Yao, X. Zhang, X. Wang, S. Kanhere, and H. Zhang. 2019. "Zero-shot Object Detection With Textual Descriptions." *Proceedings of the AAAI Conference on Artificial Intelligence* 33, no. 01: 8690–8697. https://doi.org/10.1609/aaai.v33i01.33018690.

Lin, J., Y. Shen, B. Wang, S. Lin, K. Li, and L. Cao. 2024. "Weakly Supervised Open-Vocabulary Object Detection." *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 4: 3404–3412. https://doi.org/10.1609/aaai.v38i4.28127.

Lin, T. Y., P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. "Feature Pyramid Networks for Object Detection." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944. IEEE. https://doi.org/10.1109/CVPR.2017.106.

Lin, T. Y., P. Goyal, R. Girshick, K. He, and P. Dollar. 2017. "Focal Loss for Dense Object Detection." *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007. IEEE. https://doi.org/10.1109/ICCV.2017.324.

Lin, T. Y., M. Maire, S. Belongie, et al. 2014. "Microsoft COCO: Common Objects in Context." In *Lecture Notes in Computer Science*, 740–755. Springer. https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, S., Z. Zeng, T. Ren, et al. 2023. "Grounding DINO: Marrying DINO With Grounded Pre-Training for Open-Set Object Detection." arXiv. https://arxiv.org/abs/2303.05499.

Marin, J., A. Biswas, F. Ofli, et al. 2021. "Recipe1M+: a Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, no. 1: 187–203. https://doi.org/10.1109/TPAMI.2019.2927476.

Min, W., S. Jiang, L. Liu, Y. Rui, and R. Jain. 2019. "A Survey on Food Computing." *ACM Computing Surveys* 52, no. 5: 1–36. https://doi.org/10.1145/3329168.

Min, W., Z. Wang, Y. Liu, et al. 2023. "Large Scale Visual Food Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, no. 8: 9932–9949. https://doi.org/10.1109/TPAMI.2023.3237871.

Myers, A., N. Johnston, V. Rathod, et al. 2015. "Im2calories: Towards an Automated Mobile Vision Food Diary." In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1233–1241. IEEE. https://doi.org/10.1109/ICCV.2015.146.

Radford, A., J. W. Kim, C. Hallacy, et al. 2021. "Learning Transferable Visual Models From Natural Language Supervision." arXiv. https://arxiv.org/abs/2103.00020.

Rahman, S., S. Khan, and N. Barnes. 2019. "Transductive Learning for Zero-Shot Object Detection." 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 6081–6090. IEEE. https://doi.org/10.1109/ICCV.2019.00618.

Ren, S., K. He, R. Girshick, and J. Sun. 2017. "Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, no. 6: 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

Russ, J. C. 2015. "Image Analysis of Foods." *Journal of Food Science* 80, no. 9: E1974–E1987. https://doi.org/10.1111/1750-3841.12987.

Russakovsky, O., J. Deng, H. Su, et al. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115, no. 3: 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Tang, P., X. Wang, S. Bai, et al. 2020. "PCL: Proposal Cluster Learning for Weakly Supervised Object Detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, no. 1: 176–191. https://doi.org/10.1109/TPAMI.2018.2876304.

Tang, P., X. Wang, X. Bai, and W. Liu. 2017. "Multiple Instance Detection Network With Online Instance Classifier Refinement." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3059–3067. IEEE. https://doi.org/10.1109/CVPR.2017.326.

Zareian, A., K. D. Rosa, D. H. Hu, and S. F. Chang. 2021. " Open-Vocabulary Object Detection Using Captions." In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14388–14397. IEEE. https://doi.org/10.1109/CVPR46437.2021.01416.

Zhong, Y., J. Yang, P. Zhang, et al. 2022. "RegionCLIP: Region-Based Language-Image Pretraining" In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16772–16782. IEEE. https://doi.org/10.1109/CVPR52688.2022.01629.

Zhou, K., J. Yang, C. C. Loy, and Z. Liu. 2022a. "Conditional Prompt Learning for Vision-Language Models." In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16795–16804. IEEE. https://doi.org/10.1109/CVPR52688.2022.01631.

Zhou, K., J. Yang, C. C. Loy, and Z. Liu. 2022b. "Learning to Prompt for Vision-Language Models." *International Journal of Computer Vision* 130, no. 9: 2337–2348. https://doi.org/10.1007/s11263-022-01653-1.

Zhou, X., R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra. 2022. "Detecting Twenty-Thousand Classes Using Image-Level Supervision." arXiv. https://arxiv.org/abs/2201.02605.

## Supporting Information

Additional supporting information can be found online in the Supporting Information section.

**Supporting Material**: jfds71414-sup-0001-SuppMat.xls. Supplementary Data S1 provides the removed and retained Food2K class lists, filtering rules, and protocol details for the semantic-leakage audit. **Supporting Material**: jfds71414-sup-0002-FigureS1.png. **Figure S1**. Qualitative comparison between a baseline (top row) and the proposed method (bottom row) on matched complex food-scene images. The proposed method more often yields complete item coverage and more specific labels while reducing background-driven responses; image pairs are aligned with consistent box colors and enlarged side labels for readability. A high-resolution supplementary version of this figure is provided with corrected text-label overlap and improved readability.