SyMFood: Synergistic Multi-Modal Prompting for Fine-Grained Zero-Shot Food Detection

Xinlong Wang, Weiqing Min, *Senior Member, IEEE,*, Shoulong Liu, Guorui Sheng, Shuqiang Jiang, *Senior Member, IEEE*

***Abstract*—Fine-grained object detection in food computing is severely constrained by the vast diversity of food items and the high cost of data annotation. Existing Zero-Shot Food Detection (ZSFD) methods attempt to solve this by leveraging semantic information, but they suffer from two critical bottlenecks: (1) a "Semantic Dilemma" (SD) where textual descriptions are too ambiguous to distinguish visually similar food categories, and (2) an "Architectural Bottleneck" (AB) due to the granularity mismatch between high-level semantics and low-level visual features. In this article, we propose SyMFood (Synergistic Multi-modal Framework for Food Generalization), a novel ZSFD framework designed to systematically overcome these challenges. To resolve the SD, SyMFood employs a multi-modal prompt system, which combines rich descriptions from Large Language Models (LLMs) with unambiguous visual exemplars to provide precise semantic grounding. To break the AB, SyMFood introduces a "Refine-then-Fuse" architecture. This design first utilizes a Context-Aware Spatial-Channel Refinement (CaSC) block to enhance visual features independently. Subsequently, a Progressive Food Knowledge Fusion (ProFus) module performs bi-directional, iterative co-refinement between the enhanced visual features and multi-modal prompts across all scales. Extensive experiments across four challenging datasets, including food-specific (UEC FOOD 256, FOWA) and general-purpose (PASCAL VOC, MS COCO) benchmarks, validate our approach. The proposed method outperforms baselines, yielding a notable 8.5% improvement in Harmonic Mean on the FOWA dataset in the genearl ZSD (GZSD) setting. The source code will be available at https://github.com/Niko000202/SymFood0202.**

***Index Terms*—Food Computing, Zero-Shot Learning, Zero-Shot Detection, Cross-Modal Fusion, Food Detection.**

I. Introduction

FOOD computing, an interdisciplinary frontier merging computer vision and food science, plays an increasingly vital role in analyzing and understanding food-related visual data [1]. As a core enabling technology, food detection is critical for numerous applications, including intelligent nutritional assessment [2] and personalized dietary guidance systems [3]. This challenge is particularly acute for real-world systems, such as mobile nutritional assistants or automated checkout counters, where the computational and data-storage overhead of constantly retraining models for new food items is prohibitive. Zero-Shot Detection (ZSD) offers a promising solution. Technically, ZSD follows two main paradigms: generative-based and embedding-based methods. The former aims to synthesize pseudo-visual features for unseen classes [4], [5], thereby converting the problem into a fully supervised one. However, the quality and stability of this synthesis process are difficult to guarantee. For food images with complex textures and structures, the generated features often lack authenticity and discriminative power, thus limiting the model's performance ceiling. In contrast, embedding-based methods are more direct, seeking to learn a direct mapping function from visual to semantic space [6], [7]. While avoiding unstable feature generation, the efficacy of such methods is almost entirely contingent upon the quality and richness of the semantic information they rely on.

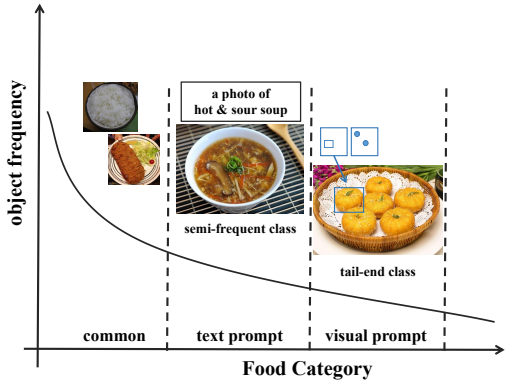


Fig. 1. Challenge of the long-tail distribution in food detection. While supervised methods cover high-frequency "head" categories and text-based ZSD extends to the "shoulder," both struggle with rare "tail" categories. Our multi-modal prompt approach specifically enhances recognition robustness for these challenging tail-end classes.

This reliance on semantics exposes the first core challenge in ZSFD, a "Semantic Dilemma" (SD) that we identify as two-fold. First, the inherent long-tail distribution of food datasets results in a sparse and biased semantic space (Fig. 1). Second, language itself is fundamentally inadequate for achieving precise, fine-grained visual alignment. This limitation is particularly pronounced in Fine-Grained Visual Classification tasks, where high intra-class variance and low inter-class similarity are inherent challenges that simple semantic labels cannot resolve [8], [9]. A simple class name like "cake" cannot distinguish between "tiramisu" and "cheesecake", while even detailed Large Language Models (LLMs)-generated descriptions for specific dishes, such as "fried spring rolls", often fail to capture intra-class visual diversity caused by differences in cooking or plating styles. This reveals a critical bottleneck that any textual description is merely a generalized concept, incapable of encompassing the full spectrum of a dish's visual

This work was supported by the Beijing Natural Science Foundation (JQ24021) and the National Natural Science Foundation of China (62125207 and 62472411)

Manuscript received April 19, 2021; revised August 16, 2021.

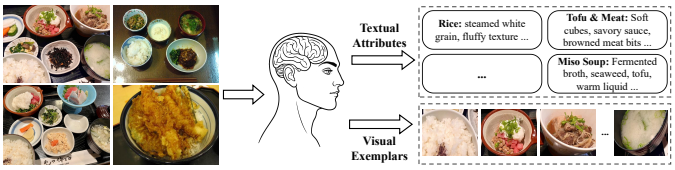


Fig. 2. Conceptual framework for synergistic multi-modal food representation. The proposed framework emulates human perception by integrating textual attributes and visual exemplars to establish robust semantic anchors and mitigate linguistic ambiguity for fine-grained alignment.

manifestations.

While building ever-larger pre-training datasets is a common approach to mitigate this issue [10], our work proposes a more fundamental solution through a multi-modal prompting system. Inspired by human perception, this framework synergizes attribute-rich textual descriptions with representative visual exemplars, as conceptually illustrated in Fig. 2. Specifically, it integrates textual descriptions from LLMs representing the textual aspect with unambiguous visual exemplars constituting the visual aspect. These visual exemplars provide deterministic anchors for specific visual instances, which circumvents the uncertainties of linguistic ambiguity and intra-class variance, thereby enabling a precise alignment between abstract concepts and their diverse visual manifestations.

However, even with these synergistic multi-modal prompts, a critical "Architectural Bottleneck" (AB) remains in the cross-modal fusion stage. Prevailing frameworks [11], [12] typically perform a one-shot fusion between high-level semantic attributes and raw multi-scale visual features. This direct alignment mechanism overlooks the fundamental problem of semantic-granularity mismatch [13], [14]. This challenge, characterized by the inconsistency between macroscopic concepts and microscopic features, is a well-documented bottleneck in cross-modal learning that has spurred targeted research into hierarchical and multi-grained alignment methodologies [15]. For instance, a holistic concept describing a "whole apple" is mathematically difficult to align with a high-resolution feature vector containing only the local texture of the "apple peel," which can lead to information loss and gradient confusion. This observation strongly motivates a superior fusion mechanism that is both scale-aware and iteratively optimized.

To systematically break through the aforementioned dual dilemmas, this paper proposes the Synergistic Multi-modal Framework for Food Generalization (SyMFood), a novel framework specifically designed for ZSFD. To the best of our knowledge, SyMFood is the first framework dedicated to solving the ZSFD problem by combining a "Refine-then-Fuse" architectural philosophy with a multi-source, multi-modal prompting system. To address the SD in the food domain, SyMFood integrates attribute-rich text generated by LLMs with unambiguous visual exemplars. More critically, to overcome the AB of multi-modal fusion, SyMFood first employs the Context-Aware Spatial-Channel Refinement (CaSC) block for focused intra-modal visual feature pre-processing to create a better "canvas" for fusion. Subsequently, through a Progressive Food Knowledge Fusion (ProFus) module, it performs bi-directional, iterative co-refinement of visual features and concept prompts across multiple scales, thereby achieving precise concept anchoring at all visual levels. Extensive experiments conducted on multiple datasets have demonstrated the effectiveness of our proposed method.

The main contributions of this research can be summarized as follows:

- We are the first to identify and systematically address the "dual dilemmas" in ZSFD: a SD and an AB. We propose a novel framework, SyMFood, as a holistic solution.
- To address the SD, we construct a powerful multi-modal prompt system that synergizes attribute-rich text from LLMs with unambiguous visual exemplars, providing high-quality and robust semantic guidance.
- To break through the AB, we propose an innovative "Refine-then-Fuse" architecture. This architecture employs a CaSC block for visual pre-refinement and a ProFus module for bi-directional, scale-aware co-refinement, effectively resolving the semantic-granularity mismatch.
- Extensive experiments on both food-specific (UEC FOOD 256, FOWA) and general-purpose (PASCAL VOC, MS COCO) datasets validate that our SyMFood achieves state-of-the-art (SOTA) performance and demonstrates robust generalization capabilities.

II. Related Work

A. Text Prompt

Text-prompted object detection has achieved significant progress, with these works leveraging large-scale Vision-Language Pre-training to realize impressive ZSD capabilities. Among them, GLIP [16] reformulates object detection as a phrase grounding task, learning aligned semantics from massive image-text pairs. Building on this foundation, Grounding DINO [17] combines this grounding paradigm with a DETR-like architecture and achieves SOTA performance through early-stage fusion. Other works, such as RegionCLIP [18] and DetCLIP [10], focus on enhancing region-level knowledge using image-text pairs with pseudo-generated boxes. Furthermore, [19] explore learning diversified primitive prompts to capture complex semantic compositions, which demonstrates the potential of prompt-based knowledge transfer in zero-shot scenarios. The concept of leveraging multi-modal and multi-grained semantics has also been successfully explored in related vision-language tasks, such as zero-shot video classification and temporal action detection, where prompting techniques are employed to enhance cross-modal alignment [20]. Specifically, [21] leverage multi-grained embedding to handle semantic ambiguity in video classification, which aligns with our motivation of addressing fine-grained food concepts. However, the reliance of many models on purely textual descriptions reveals their inherent limitations when faced with the problem of conceptual ambiguity in fine-grained domains. This challenge has directly spurred research into introducing more explicit and unambiguous visual prompts as a supplement or alternative, in pursuit of more reliable semantic grounding.

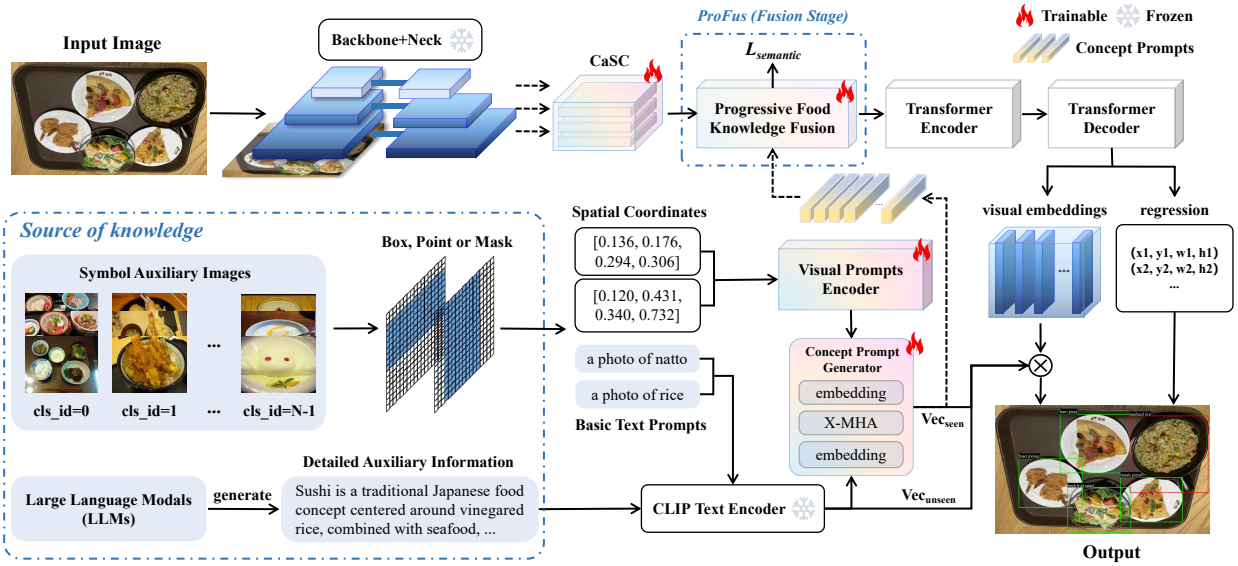


Fig. 3. Overall architecture of our proposed SyMFood, illustrating a structured left-to-right workflow. First, the Concept Prompts Generator fuses text descriptions with coordinate information (e.g., box, point, or mask). Then, image features refined by CaSC are integrated with these prompts within the ProFus stage (dashed box), where an intermediate $L_{semantic}$ ensures semantic consistency to guide the final object detection.
Acronyms: CaSC: Context-Aware Spatial-Channel Refinement; X-MHA: Cross-modal Multi-head Attention; ProFus: Progressive Food Knowledge Fusion; $L_{semantic}$: Semantic Alignment Supervision.

B. Visual Prompt

To address the limitations of text prompts, introducing visual prompts has become an emergent research direction. These methods provide a direct and unambiguous visual "anchor" for object concepts. One mainstream approach uses visual exemplars for reference-based localization at inference time; for instance, models like T-REX2 [22] and DINOv [23] locate objects in a target image based on a given visual reference, without updating the model's parameters. Concurrently, aiming for a deeper integration, another major paradigm, represented by CP-DETR [24] and MQ-DET [25], is dedicated to learning a shared embedding space during the training phase, which can unify or fuse visual instructions (e.g., points, bounding boxes, masks) with text prompts. In parallel, [26] emphasize the significance of learning visual attribute representations to bridge the gap between abstract concepts and concrete visual instances. However, a key issue commonly overlooked by methods based on either text or visual prompts is their tendency to interact semantic prompts with raw visual features directly extracted from the backbone. This can lead to sub-optimal alignment when processing fine-grained tasks, potentially due to inferior feature quality and interference from complex background information [27].

C. Zero-Shot Learning

Zero-Shot Learning (ZSL) aims to recognize novel classes without training samples via a shared semantic space constructed from word vectors/attributes. Early ZSL relied on embedding-based methods, mapping visual and semantic domains, which are categorized into three paradigms: projecting visual features into the semantic space [28], mapping semantic vectors into the visual space [29], or projecting both modalities into a common latent space [30]. While foundational, their performance was constrained by small dataset scales and low-quality attribute annotations. To improve the fidelity of semantic-visual mapping, [31] propose a progressive feature reconstruction network, highlighting the efficacy of iterative refinement in overcoming limited supervised data. The emergence of large-scale VLP models (e.g., CLIP [32]) reshaped the embedding paradigm, enabling modern research to focus on optimizing semantic alignment via Prompt Learning [33]. Generative ZSL (using diffusion models [34]) synthesizes pseudo-visual features but introduces complex pipelines. Our framework follows the embedding paradigm, focusing on semantic-visual alignment over feature synthesis. Regardless of the paradigm, reducing seen class prediction bias in GZSL is a critical challenge, motivating our work at the feature fusion level.

III. Method

We detail the novel framework we propose to address the ZSD problem in this chapter. Our methodology is presented in three core parts: first, a formal Problem Formulation; second, the Overall Architecture of our model, as illustrated in Fig. 3; and finally, the Key Innovative Components that constitute our framework.

A. Problem Formulation

Given a training dataset $\mathcal{D}_{\text{train}} = \{(I_i, A_i)\}_{i=1}^{M}$, where all annotations $A_i = \{(\mathbf{b}_j, c_j)\}_{j=1}^{K_i}$ belong to a set of N_{s} seen categories, $\mathcal{C}_{\text{s}}$. The primary challenge is to train a model that can generalize to detect objects from a disjoint set of N_{u} unseen categories, $\mathcal{C}_{\text{u}}$, where $\mathcal{C}_{\text{s}} \cap \mathcal{C}_{\text{u}} = \emptyset$. To bridge this gap, knowledge is transferred via a shared semantic space, where each category $c \in \mathcal{C}_{\text{s}} \cup \mathcal{C}_{\text{u}}$ is represented by a semantic embedding vector $\mathbf{w}_c$. These vectors form a

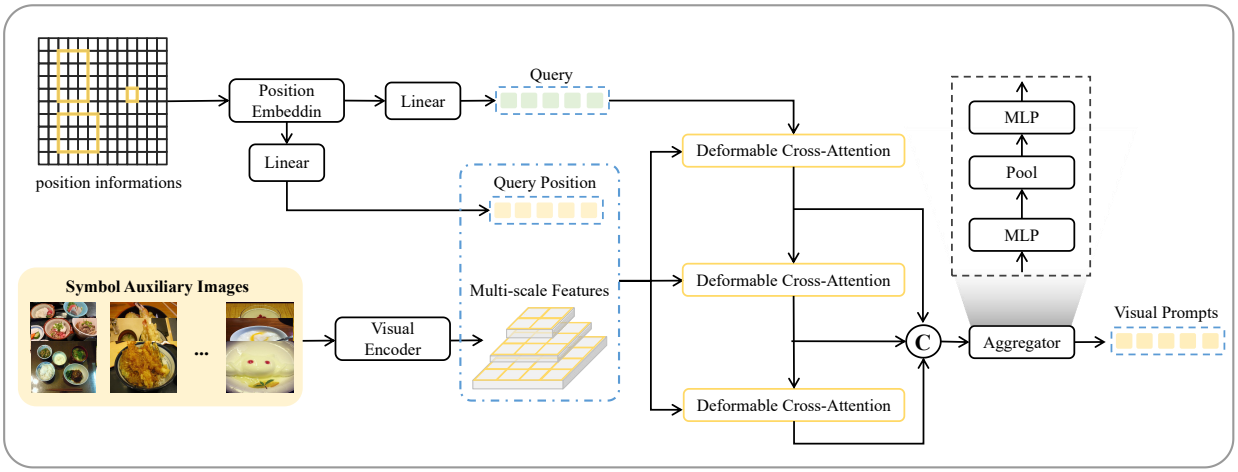


Fig. 4. Architecture of Visual Prompts Encoder. It first transforms spatial cues, such as 2D box coordinates, into a set of learnable queries and positional embeddings. These queries then extract contextual information from multi-scale features by employing a three-layer deformable cross-attention mechanism to generate the final visual prompts.

knowledge base $\mathbf{W}$, composed of seen ($\mathbf{W}_\text{s} \in \mathbb{R}^{N_\text{s} \times d_\text{sem}}$) and unseen ($\mathbf{W}_\text{u} \in \mathbb{R}^{N_\text{u} \times d_\text{sem}}$) embedding matrices, where d_sem is the dimensionality of the semantic space. The model's performance is then evaluated under two distinct protocols: the standard ZSD setting, which tests only on $\mathcal{C}_\text{u}$, and the more challenging GZSD setting. The GZSD setting requires the model to distinguish between both seen and unseen classes during inference, placing a higher demand on its generalization capability.

B. Overall Architecture Overview

The architecture of our proposed SyMFood framework, illustrated in Fig. 3, adheres to a "Refine-then-Fuse" philosophy to address the core challenges of cross-modal alignment in ZSFD. The framework first prepares two enhanced, modality-specific inputs in parallel. For semantics, our Concept Prompts Generation stage fuses attribute-rich text with visual exemplars to create a powerful initial concept vector. Concurrently, for vision, the raw multi-scale image features are processed by our CaSC block for intra-modal refinement, enhancing critical details like food textures and gloss. The core of our method lies in the subsequent ProFus stage, which ingests both the refined visual features and the initial concept prompt. To resolve the fundamental semantic-granularity mismatch (e.g., aligning a "whole apple" concept with "apple peel" features), ProFus performs a synergistic co-refinement: it iteratively and bi-directionally updates both the visual features and the concept prompt across each level of the feature pyramid. This scale-aware process ensures a deep, hierarchical alignment, yielding a final set of co-refined features and prompts for detection.

C. Concept Prompts Generation

The generalization capability of a ZSD model is highly contingent on the quality of its class semantic representations. Traditional methods relying on sparse word vectors from class names are insufficient for fine-grained domains like food, as a single vector for "cake" cannot distinguish between "tiramisu" and "cheesecake." To address this, we propose a Concept Prompts Generator that constructs a robust multi-modal representation for each category by fusing two complementary information sources: attribute-rich textual prompts and representative visual prompts.

1) Attribute-Rich Textual Prompts: To mitigate the high intra-class visual variance within food classes, we leverage LLMs to generate a set of detailed descriptive sentences for each category, covering core attributes like ingredients, cooking methods, and appearance. To ensure high-fidelity semantic alignment and prevent representational degradation, we strictly regulate this process through a deterministic system prompt. Specifically, the LLM is instructed to focus exclusively on objective visual properties (e.g., color, texture, and shape) while explicitly excluding subjective descriptors or non-visual elements like taste and mouthfeel. To align with the architectural limitations of the CLIP text encoder, we impose a strict length constraint by limiting each description to under 60 words. This heuristic effectively ensures that the generated text remains within the 77-token context window, thereby avoiding attention dispersion and information loss caused by sequence truncation. By encoding multiple such prompts via CLIP's text encoder and aggregating their embeddings, we obtain a textual concept representation $\mathbf{P}_t^c$:

$$\mathbf{P}_t^c = \frac{1}{M} \sum_{m=1}^{M} \mathcal{E}_T(p_{c,m}) \tag{1}$$

where M denotes the number of descriptive prompts generated for category c. This robust textual representation is more stable and attribute-aware than a single word vector, serving as the semantic anchor for subsequent visual grounding. A mechanistic analysis and qualitative comparison, including the specific prompt templates and their impact on semantic purity, are provided in Appendix A.

2) Visual Prompt Grounding with Textual Supervision: While rich text provides discriminative features, its abstract nature limits its ability to capture specific visual variations. We therefore introduce visual prompts to "ground" these semantic

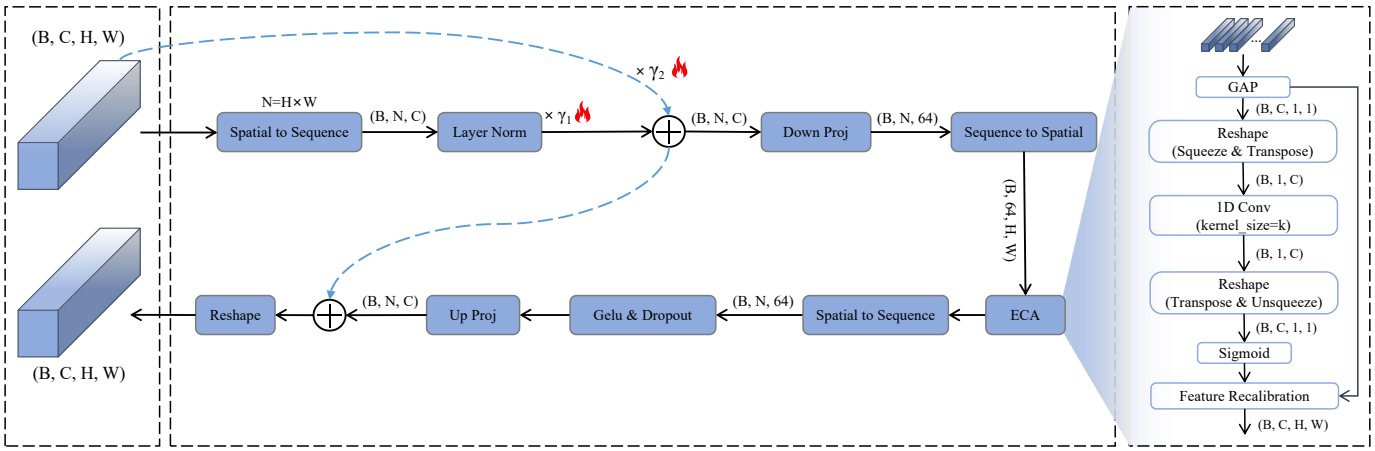


Fig. 5. Architecture of our proposed CaSC block. The module utilizes a bottleneck structure centered on an ECA layer, complemented by a weighted normalization input and a residual connection.

concepts in the visual world. For this, we propose a Visual Prompts Encoder (VPE), whose architecture is shown in Fig. 4. The VPE takes the bounding box coordinates of a visual exemplar as a query to interact with the corresponding image's multi-scale features via a stack of Deformable Cross-Attention layers. The aggregated output forms the visual prompt embedding, $\mathbf{P}_v$.

Crucially, the VPE is not trained in an unsupervised manner. To ensure semantic consistency, we employ the category-specific text prompt embedding $\mathbf{P}_t$ formulated in Section III-C1 as the primary supervision target. A Visual-Text Prompt Alignment Loss, denoted as $\mathcal{L}_{V\_T\_Align}$, is introduced to maximize the cosine similarity between the two prompts:

$$\mathcal{L}_{V\_T\_Align} = \frac{1}{K}\sum_{i=0}^{K-1}\left(1 - \frac{\mathbf{P}_v^i \cdot \mathbf{P}_t^i}{\|\mathbf{P}_v^i\|_2 \|\mathbf{P}_t^i\|_2}\right) \tag{2}$$

where K is the number of positive categories in a training batch, and $\mathbf{P}_t^i$ represents the pre-computed semantic anchor for the i-th category. This supervision paradigm compels the VPE to extract generalizable, core visual features that are most relevant to the class's textual description, creating a semantically calibrated and information-dense visual embedding.

D. Context-Aware Spatial-Channel Refinement

To extract generalizable visual features for ZSFD, we propose the CaSC block, a block designed to enhance multi-scale features prior to cross-modal fusion. Our design is inspired by Mona [35], but we posit that its focus on spatial structures can overfit to the layouts of seen classes. We argue that for food items, channel-wise features like color and texture represent more transferable attributes than variable spatial arrangements. Therefore, our key innovation is replacing the spatial operator with an Efficient Channel Attention (ECA) layer to adaptively recalibrate channel features for better cross-category generalization. As illustrated in Fig. 5, the CaSC architecture first stabilizes input features using a weighted normalization layer, which adaptively balances a normalization branch and an identity branch via two learnable scaling vectors (γ_1, γ_2). The resulting features are then processed through an ECA-centric bottleneck and integrated with a final residual connection.

$$\mathbf{X}_{\text{norm}} = \gamma_1 \odot \text{LN}(\mathbf{X}_{\text{in}}) + \gamma_2 \odot \mathbf{X}_{\text{in}} \tag{3}$$

The normalized features $\mathbf{X}_{\text{norm}}$ are then fed into an efficient "down-project, enhance, up-project" bottleneck structure. First, a down-projection MLP, MLP_{down}, reduces the channel dimension of $\mathbf{X}_{\text{norm}}$ to create an intermediate feature map $\mathbf{X}_{\text{enhanced}} = \text{MLP}_{\text{down}}(\mathbf{X}_{\text{norm}})$.

The core feature enhancement is then performed by the ECA module on this intermediate feature map $\mathbf{X}_{\text{enhanced}} \in \mathbb{R}^{B\times C'\times H\times W}$. The process begins by aggregating spatial information from each channel of $\mathbf{X}_{\text{enhanced}}$ into a channel descriptor vector $\mathbf{z}$ via Global Average Pooling (GAP):

$$z_c = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{\text{enhanced},c}(i,j) \tag{4}$$

Subsequently, a 1D convolution Conv1D_k with an adaptive kernel size (k) is employed to efficiently capture local cross-channel interactions:

$$k = \psi(C') = \left|\frac{\log_2(C'+b)}{\alpha}\right|_{\text{odd}} \tag{5}$$

where α and b are hyperparameters, and $|\cdot|_{\text{odd}}$ denotes rounding to the nearest odd integer. A channel attention weight vector η is then generated through a Sigmoid function σ:

$$\eta = \sigma(\text{Conv1D}_k(\mathbf{z})) \tag{6}$$

After obtaining the weights, the refined output is produced by recalibrating the intermediate features, where the learned weights η are multiplied channel-wise with the feature map that entered the ECA module, $\mathbf{X}_{\text{enhanced}}$:

$$\mathbf{X}_{\text{recalibrated}} = \mathbf{X}_{\text{enhanced}} \odot \eta \tag{7}$$

Finally, the ECA-enhanced features $\mathbf{X}_{\text{recalibrated}}$ are passed through GeLU activation and Dropout (GD), then restored to their original dimension via an up-projection MLP (MLP_{up}).

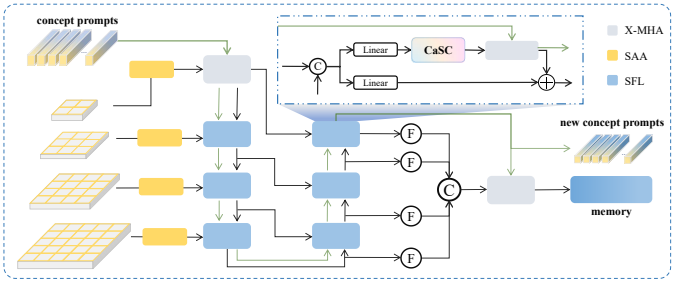


Fig. 6. Architecture of our proposed ProFus module, which resolves the semantic-granularity mismatch by implementing a bi-directional pathway via SAA.. This allows for the progressive co-refinement of multi-scale visual features and concept prompts by propagating semantics downwards and localizing details upwards.

This result is then added to $\mathbf{X}_{\text{norm}}$ through a residual connection to form the module's final output $\mathbf{X}_{\text{out}}$. The entire operation can be summarized as:

$$\mathbf{X}_{\text{out}} = \mathbf{X}_{\text{norm}} + \text{MLP}_{\text{up}}(\text{GD}(\mathbf{X}_{\text{recalibrated}})) \tag{8}$$

In summary, through its weighted normalization, ECA-centric channel attention bottleneck, and residual learning framework, the CaSC block efficiently recalibrates and enhances channel-wise feature responses at a low computational cost. This improves the model's ability to discern and highlight visual cues that are critical for the precise recognition and localization of food items, especially for unseen categories, thus providing a more solid and generalizable visual foundation for subsequent cross-modal analysis and final zero-shot detection performance.

E. Progressive Food Knowledge Fusion

Even after feature refinement by the CaSC block, a fundamental challenge remains: the "Semantic-Granularity Mismatch" across multi-scale features. This issue creates a dilemma: high-resolution features (e.g., C3 layer) capture fine details like the texture of a "fried spring roll" but lack the global context to distinguish it from a "fried dough stick," leading to semantic ambiguity. Conversely, low-resolution features (e.g., C5 layer) grasp holistic concepts like "sushi" but suffer from severe spatial degradation, causing significant localization errors. To resolve this contradiction, we propose the ProFus framework. Unlike unidirectional methods like FPN, ProFus implements a bi-directional (top-down and bottom-up) information flow (see Fig. 6). This architecture enables a deep, iterative co-refinement between multi-scale visual features and the concept prompts, simultaneously enriching high-resolution features with semantic context and refining the localization details of low-resolution features.

1) Single-Scale Fusion Layer: The fundamental building block of our bi-directional fusion framework is the Single-Scale Fusion Layer (SFL), which is designed to achieve efficient visual-semantic feature interaction. The SFL takes the CaSC-refined visual features of the current scale $\mathbf{F}_{\text{curr}} \in \mathbb{R}^{B\times C\times H\times W}$, and the feature from an adjacent scale $\mathbf{F}_{\text{adj}}$ as input. As illustrated in Fig. 6, these two visual streams are integrated via bilinear resizing and concatenation firstly:

$$\mathbf{F}_{\text{concat}} = \text{Concat}(\mathbf{F}_{\text{curr}}, \text{Resize}(\mathbf{F}_{\text{adj}})) \tag{9}$$

Subsequently, the concatenated feature $\mathbf{F}_{\text{concat}}$, after being reshaped into a sequence by operator $\mathcal{R}(\cdot)$, is fed into two parallel linear projection paths, ϕ_{main} and ϕ_{sc}, to generate the main path feature $\mathbf{X}_{\text{main}}$ and the shortcut feature $\mathbf{X}_{\text{sc}}$, respectively.

$$\mathbf{X}_{\text{main}}, \mathbf{X}_{\text{sc}} = \phi_{\text{main}}(\mathcal{R}(\mathbf{F}_{\text{concat}})), \phi_{\text{sc}}(\mathcal{R}(\mathbf{F}_{\text{concat}})) \tag{10}$$

On the main path, the visual query then interacts with projected concept prompts $\mathbf{P}_{\text{proj}}$ within a cross-modal multi-head attention (X-MHA) [36] block to inject semantic knowledge. Let $\mathbf{X}_{\text{sum}} = \mathbf{X}_{\text{sc}} + \mathbf{X}_{\text{main}}$. The output of this path is added to the shortcut feature via a residual connection, and the result is passed through an output projection layer ϕ_{out} to produce the final fused feature $\mathbf{F}_{\text{SFL}}$:

$$\mathbf{F}_{\text{SFL}} = \phi_{\text{out}}(\text{Norm}(\mathbf{X}_{\text{sum}} + \text{X-MHA}(\mathbf{X}_{\text{main}}, \mathbf{P}_{\text{proj}}))) \tag{11}$$

Crucially, this bi-directional interaction facilitates an implicit modality calibration. Through the X-MHA mechanism, the model adaptively assigns higher attention weights to shared invariant attributes (e.g., structural layering and texture) while suppressing contradictory instance-level noise (e.g., mismatched colors between an image and its textual prior), thereby ensuring robust alignment even under high intra-class visual variance.

2) Bi-directional Fusion Pathway and Feature Aggregation: Leveraging the SFL, our ProFus module constructs a complete bi-directional fusion pathway. This architecture incorporates Scale-Aware Attention (SAA) to facilitate the progressive co-refinement of multi-scale visual features and concept prompts. The process is initiated with a distinct semantic injection strategy at the top feature level. Specifically, the top-level feature map $\mathbf{F}_{C_L}$, which possesses the largest receptive field,

TABLE I
ZSD AND GZSD PERFORMANCE ON UEC FOOD 256(%). † INDICATES GENERATIVE-MODEL-BASED METHODS; OTHERS USE EMBEDDING FUNCTIONS.

Metric	Method	Split	ZSD	GZSD		
				S	U	HM
Recall@100	CZSD [7]	205/51	60.7	57.6	45.5	50.8
	SU† [4]	205/51	61.9	52.5	52.8	52.6
	RRFS† [5]	205/51	64.8	54.9	55.1	55.0
	SeeDS† [37]	205/51	74.0	55.2	61.4	58.1
	ZSFDet† [38]	205/51	74.4	57.0	61.8	59.3
	SA [30]	205/51	92.9	71.2	86.9	78.3
	Ours	205/51	**95.6**	**83.4**	**89.3**	**86.2**
mAP	CZSD [7]	205/51	22.0	20.8	16.2	18.2
	SU† [4]	205/51	22.4	19.3	20.1	19.7
	RRFS† [5]	205/51	23.6	20.1	22.9	21.4
	SeeDS† [37]	205/51	27.1	20.2	26.0	22.7
	ZSFDet† [38]	205/51	27.3	21.9	26.1	23.8
	SA [30]	205/51	24.2	18.6	24.9	21.3
	Ours	205/51	**31.7**	**29.6**	**30.7**	**30.1**

is directly fused with the concept prompts $\mathbf{P}_{\text{proj}}$ via X-MHA. This initial step yields a feature $\mathbf{F}'_{C_L}$ that is rich in contextualized semantics:

$$\mathbf{F}'_{C_L} = \mathbf{F}_{C_L} + \text{X-MHA}(\mathbf{F}_{C_L}, \mathbf{P}_{\text{proj}}) \tag{12}$$

This semantically calibrated top-level feature then serves as the starting point for the top-down pathway, which propagates high-level contextual information downwards. This recursive process can be formulated as:

$$\mathbf{F}'_i = \mathbf{F}_{\text{SFL}}(\mathbf{F}'_{i+1}, \mathbf{F}_i), \quad i = L-1, \ldots, 1 \tag{13}$$

Subsequently, the bottom-up pathway mitigates localization imprecision. This path takes the outputs from the top-down fusion $\mathbf{F}'$ and propagates precise spatial details upwards, creating a new set of features $\mathbf{F}''$. This process can be defined as:

$$\mathbf{F}''_j = \mathbf{F}'_j + \mathbf{F}_{\text{SFL}}(\mathbf{F}''_{j-1}, \mathbf{F}'_j), \quad j = 2, \ldots, L \tag{14}$$

where the process is initialized with $\mathbf{F}''_1 = \mathbf{F}'_1$. After this comprehensive bi-directional optimization, we selectively aggregate the feature maps. Specifically, we collect the final outputs from the bottom-up path, namely $\mathbf{F}''_1, \mathbf{F}''_2, \ldots, \mathbf{F}''_L$. These chosen feature maps are then flattened and concatenated to form a unified feature sequence $\mathbf{M}_{\text{fused}}$:

$$\mathbf{M}_{\text{fused}} = \text{Concat}(\text{Flatten}(\mathbf{F}''_1), \text{Flatten}(\mathbf{F}''_2), \ldots, \text{Flatten}(\mathbf{F}''_L)) \tag{15}$$

This resulting sequence $\mathbf{M}_{\text{fused}}$ encapsulates a rich, multi-scale representation where features at each level are informed by both global semantics and fine-grained spatial details. It subsequently serves as the input memory for our Transformer decoder, which is tasked with generating the final detection results.

3) Semantic Alignment Supervision: To further enhance the model's discriminative power during the detection phase, we introduce $L_{semantic}$ as a task-oriented intermediate supervision. Distinct from the prompt-level $\mathcal{L}_{V\_T\_Align}$, which calibrates category knowledge offline, $L_{semantic}$ dynamically constrains the multi-modal instance features output by the ProFus module. Specifically, it forces the fused query embeddings $\hat{f}_i$ to strictly align with their corresponding ground-truth semantic anchors e_{y_i} before entering the global Transformer encoding stage:

$$L_{semantic} = 1 - \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \frac{\hat{f}_i \cdot e_{y_i}}{\|\hat{f}_i\| \cdot \|e_{y_i}\|} \tag{16}$$

where N_{pos} is the number of matched positive samples. By imposing this instance-level consistency, $L_{semantic}$ prevents the complex fusion process from distorting critical food attributes, ensuring that the visual tokens remain highly representative for zero-shot knowledge transfer.

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: Consistent with prior work [37], [38], we evaluate our proposed model on two food-specific datasets: UEC FOOD 256, which is split into 205 seen and 51 unseen classes, and Food Objects With Attributes (FOWA), which is split into 184 seen and 44 unseen classes. To validate the generalization capability of our model, we further conduct experiments on two general-purpose detection datasets: PASCAL VOC and MS COCO. The class splits for these two datasets strictly adhere to the settings in HRE [6].

TABLE II
ZSD AND GZSD PERFORMANCE ON FOWA (%). † INDICATES GENERATIVE-MODEL-BASED METHODS; OTHERS USE EMBEDDING FUNCTIONS.

Metric	Method	Split	ZSD	GZSD		
				S	U	HM
Recall@100	ConSE [39]	184/44	39.7	58.0	38.1	46.4
	BLC [40]	184/44	41.2	55.3	40.5	46.8
	CZSD [7]	184/44	48.0	86.1	44.8	58.9
	SU† [4]	184/44	45.3	82.3	44.1	57.4
	RRFS† [5]	184/44	48.8	86.6	47.6	61.4
	SeeDS† [37]	184/44	52.9	87.0	49.8	63.3
	ZSFDet† [38]	184/44	53.5	87.0	50.1	63.6
	SA [30]	184/44	89.3	97.5	65.5	78.3
	Ours	184/44	**91.0**	**97.7**	**72.1**	**83.0**
mAP	ConSE [39]	184/44	0.8	54.3	0.7	1.4
	BLC [40]	184/44	1.1	51.1	0.9	1.8
	CZSD [7]	184/44	4.0	81.2	2.1	4.1
	SU† [4]	184/44	3.9	79.1	2.3	4.5
	RRFS† [5]	184/44	4.3	82.7	2.7	5.2
	SeeDS† [37]	184/44	5.9	82.8	3.5	6.7
	ZSFDet† [38]	184/44	6.1	82.8	3.6	6.9
	SA [30]	184/44	7.7	90.5	5.4	10.1
	Ours	184/44	**10.3**	**91.3**	**10.4**	**18.6**

2) Evaluation Protocols: We evaluate our method on UEC FOOD 256, FOWA, PASCAL VOC and MS COCO datasets for both ZSD and GZSD settings, following prior works [37], [38]. For ZSD, we report mean Average Precision (mAP) and Recall@100, using an Intersection over Union (IoU) threshold of 0.5 for UEC FOOD 256, FOWA, and VOC, and IoU thresholds of 0.4, 0.5, and 0.6 for MS COCO. For GZSD, we report mAP@0.5 for seen (S) and unseen (U) classes across all datasets; for MS COCO, GZSD evaluation also includes Recall@100 for S and U classes. The Harmonic Mean (HM) of Seen and Unseen is the key GZSD performance metric.

3) Implementation Details: Our framework is built upon a ResNet-50 [41] backbone and a frozen CLIP-ViT-B/32 text encoder, though a Swin-T [42] backbone is used on PASCAL VOC for fair comparison. We train all models for 30,000 iterations using the AdamW [43] optimizer with an initial learning rate of 1e-4, which is decayed by a factor of 0.1 at the 80% and 90% training marks. The attribute-rich textual prompts are generated via the Gemini 2.5 Pro API (version: gemini-2.5-pro-preview-06-05) with a sampling temperature of 0.7. Following established practices in LLM-guided prompting research [44], we set $M = 5$ and average five distinct descriptive sentences per category to construct a stable semantic prior $\mathbf{P}_t^c$. This ensemble-based strategy effectively neutralizes potential linguistic fluctuations, ensuring that our textual prior remains a robust semantic anchor. The exact prompt templates used for this generation process are detailed in Appendix A. For visual prompt construction, we follow recent findings [22]

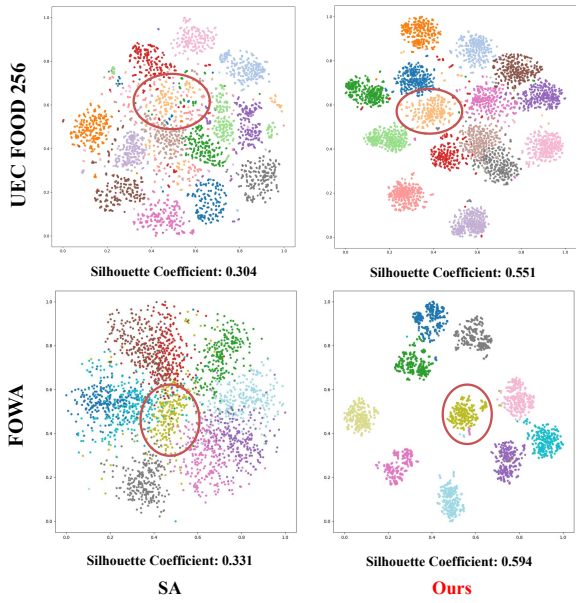


Fig. 7. t-SNE visualizations of the UEC Food 256 and FOWA datasets. Left: baseline SA; Right: our proposed SyMFood. The silhouette coefficient is provided for each distribution to quantitatively assess the clustering quality.

and select 32 exemplars per category. All experiments were conducted on four A800 80GB GPUs.

B. Experiments on ZSFD datasets

We evaluated our SyMFood framework on two food-specific datasets, UEC FOOD 256 (Table I) and FOWA (Table II), to validate its effectiveness. On the UEC FOOD 256 benchmark, SyMFood demonstrates a notable advantage over existing methods, particularly when compared to the representative generative-based baseline, ZSFDet [38]. In the challenging GZSD setting, our method achieves a 6.3% higher HM mAP and, more remarkably, a significant 26.9% lead in HM Recall@100, indicating a lower miss detection rate. We attribute this advantage to our prompt-fusion-based paradigm, which learns more robust and discriminative representations by directly fusing high-quality multi-modal prompts, thus bypassing the instability and potential mode collapse issues inherent in generative feature synthesis. This strong performance trend was consistently replicated on the FOWA dataset, where SyMFood also outperformed all evaluated baselines, firmly corroborating the effectiveness and robustness of our proposed framework for the ZSFD task.

C. Experiments on general ZSD datasets

To rigorously assess the cross-domain generalization capability of our SyMFood framework, we extended our evaluation to two widely used general-purpose benchmarks: PASCAL VOC and MS COCO. The results on PASCAL VOC, presented in Table III, demonstrate that our method achieves superior performance, notably improving the GZSD HM mAP from 56.0% of the baseline [30] to 61.1%, primarily by boosting performance on seen classes. This strong performance trend continues on the more complex MS COCO dataset. As detailed

TABLE III
ZSD AND GZSD PERFORMANCE ON PASCAL VOC (%). † INDICATES GENERATIVE-MODEL-BASED METHODS; OTHERS USE EMBEDDING FUNCTIONS.

Method	Split	ZSD	GZSD		
			S	U	HM
ConSE [39]	16/4	52.1	59.3	22.3	32.4
SAN [45]	16/4	59.1	48.0	37.0	41.8
HRE [6]	16/4	54.2	62.4	25.5	36.2
PL [46]	16/4	62.1	-	-	-
BLC [40]	16/4	55.2	58.2	22.9	32.9
TCB [47]	16/4	59.3	61.0	29.8	40.0
SU† [4]	16/4	64.9	-	-	-
CZSD [7]	16/4	65.7	63.2	46.5	53.6
RRFS† [5]	16/4	65.5	47.1	49.1	48.1
SeeDS† [37]	16/4	68.5	48.4	50.2	49.3
ZSFDet† [38]	16/4	69.2	48.5	**50.8**	49.6
SA [30]	16/4	68.7	64.8	49.3	56.0
Ours	**16/4**	**70.0**	**78.9**	45.2	**57.5**

in Table IV, our model consistently surpasses all baselines in the ZSD setting across both "48/17" and "65/15" splits. The GZSD results, summarized in Table V, show that SyMFood establishes a clear advantage over the representative generative method ZSFDet [38], increasing the HM mAP up to 1.5%. In comparison with the strong embedding-based baseline SA, while a slight performance trade-off is observed on unseen classes in the "48/17" split, our method achieves a superior HM mAP in the more demanding "65/15" split, elevating it from 26.9% to 28.0%. This consistent and compelling performance across diverse, general-purpose datasets validates that our direct embedding-fusion strategy, which avoids the computational overhead of generative processes, provides a powerful and transferable generalization capability beyond its primary food-specific domain.

D. Ablation Study

1) Effectiveness of Core Architectural Components: To quantitatively dissect the contribution of each proposed component, we conducted a series of ablation studies, with the results presented in Table VI. Our analysis begins with a baseline model that incorporates none of our innovations; this baseline establishes an initial performance of 21.3% in GZSD HM mAP on UEC FOOD 256. The introduction of our multi-modal Concept Prompts system yields the first performance gain, increasing the HM to 23.2%. Building on this, the incremental integration of the CaSC module (boosting performance to 28.2%) and the ProFus module (reaching a peak of 30.1% with the full SyMFood framework) demonstrates clear, progressive improvements. This step-wise gain is consistently replicated on the FOWA dataset. This comprehensive analysis compellingly validates that each component is indispensable and that their synergy, guided by our "Refine-then-Fuse" philosophy, is critical to achieving the final SOTA performance.

TABLE IV
ZSD PERFORMANCE ON MS COCO (%). † DENOTES GENERATIVE MODEL-BASED METHODS, WHILE OTHERS ARE EMBEDDING FUNCTION-BASED.

Method	Split	Recall@100			mAP
		IoU=0.4	IoU=0.5	IoU=0.6	IoU=0.5
CZSD [7]	48/17	56.1	52.4	47.2	12.5
GRAN [48]	48/17	58.5	55.0	50.3	11.4
RRFS† [5]	48/17	58.1	53.5	47.9	13.4
TCB [47]	48/17	55.5	52.4	48.1	11.4
SeeDS† [37]	48/17	59.2	55.3	48.5	14.0
ZSFDet† [38]	48/17	58.6	54.7	48.3	14.0
SA [30]	48/17	76.7	73.0	68.8	19.5
Ours	**48/17**	**79.6**	**75.9**	**71.3**	**21.7**
SU†	65/15	54.4	54.0	47.0	19.0
CZSD	65/15	62.3	59.5	55.1	18.6
GRAN [48]	65/15	65.3	62.7	58.3	14.9
RRFS† [5]	65/15	65.3	62.3	55.9	19.8
TCB [47]	65/15	62.5	59.9	55.1	13.8
SeeDS† [37]	65/15	66.4	63.8	56.5	20.1
ZSFDet† [38]	65/15	66.5	64.2	56.7	20.3
SA [30]	65/15	88.0	85.3	81.9	24.0
Ours	**65/15**	**89.4**	**86.6**	**82.6**	**25.4**

TABLE V
GZSD PERFORMANCE ON MS COCO (%). † DENOTES GENERATIVE MODEL-BASED METHODS, WHILE OTHERS ARE EMBEDDING FUNCTION-BASED.

Method	Split	Recall@100			mAP		
		S	U	HM	S	U	HM
PL [46]	48/17	38.2	26.3	3.2	35.9	4.1	7.4
BLC [40]	48/17	57.6	46.4	51.4	42.1	4.5	8.1
CZSD [7]	48/17	65.7	52.4	58.3	45.1	6.3	11.1
GRAN [48]	48/17	66.7	54.5	60.0	43.9	4.7	8.5
RRFS† [5]	48/17	59.7	58.8	59.2	42.3	13.4	20.4
TCB [47]	48/17	71.9	52.4	60.6	47.3	4.9	8.8
SeeDS† [37]	48/17	60.1	60.8	60.5	42.5	14.5	21.6
ZSFDet† [38]	48/17	60.1	60.7	60.4	42.5	14.3	21.4
SA [30]	48/17	78.4	49.7	**68.2**	34.0	**17.0**	**22.7**
Ours	**48/17**	**83.5**	**54.0**	65.6	**47.9**	14.6	22.4
PL [46]	65/15	36.4	37.2	36.8	34.1	12.4	18.2
BLC [40]	65/15	56.4	51.2	53.9	36.0	13.1	19.2
SU† [4]	65/15	57.7	53.9	55.7	36.9	19.0	25.1
CZSD [7]	65/15	62.9	58.6	60.7	**40.2**	16.5	23.4
RRFS† [5]	65/15	58.6	61.8	60.2	37.4	19.8	26.0
TCB [47]	65/15	69.3	59.8	64.2	39.9	13.8	20.5
SeeDS† [37]	65/15	59.3	**62.5**	60.9	37.5	20.3	26.3
ZSFDet† [38]	65/15	59.3	63.1	61.1	37.5	20.5	26.5
SA [30]	65/15	79.7	55.8	65.7	35.5	21.7	26.9
Ours	**65/15**	**82.5**	58.1	**68.2**	39.0	**21.9**	**28.0**

TABLE VI
ABLATION ON MULTI-MODAL PROMPTS

Dataset	Core Components			ZSD	GZSD		
	Concept Prompts	CaSC	ProFus		S	U	HM
UEC FOOD 256				24.2	18.6	24.9	21.3
	✓			27.8	19.6	28.4	23.2
	✓	✓		30.0	26.1	30.6	28.2
	✓	✓	✓	**31.7**	**29.6**	**30.7**	**30.1**
FOWA				7.7	90.5	5.4	10.1
	✓			8.7	90.1	8.0	14.7
	✓	✓		9.3	89.4	8.8	16.0
	✓	✓	✓	**10.3**	**91.4**	**9.9**	**17.9**

TABLE VII
ABLATION ON MULTI-MODAL PROMPTS. TEXT (S) AND TEXT (R) DENOTE SIMPLE CATEGORY NAMES AND ATTRIBUTE-RICH TEXTUAL PROMPTS, RESPECTIVELY. VISUAL REFERS TO THE VISUAL PROMPT EXTRACTED FROM EXEMPLARS.

Dataset	Input Components			ZSD	GZSD		
	Text (S)	Text (R)	Visual		S	U	HM
UEC FOOD 256	✓			24.2	18.6	24.9	21.3
		✓		28.2	20.5	29.5	24.2
			✓	28.4	28.2	27.8	28.0
		✓	✓	**31.7**	**29.6**	**30.7**	**30.1**

2) Analysis of Multi-modal Prompting Strategy: We conduct an ablation study on different prompt configurations to validate our multi-modal design, with results detailed in Table VII. The experiments reveal a consistent performance trend: while both attribute-rich textual prompts (Text (R)) and standalone visual prompts (Visual) outperform the simple class-label baseline (Text (S)), our final configuration that synergistically fuses both modalities delivers the best results. Specifically, in the GZSD setting, the fused multi-modal prompt achieves a final HM of 30.1%, surpassing the performance of the rich-text-only setting (24.2%) and the visual-only setting (28.0%). This compellingly demonstrates that the progressive enrichment and fusion of multi-modal prompts systematically enhance the model's generalization capabilities in complex ZSD scenarios.

To further evaluate the robustness of our alignment strategy, we explored alternative optimization objectives, including InfoNCE and Triplet losses. Due to the fine-grained nature of food categories, these contrastive objectives resulted in a noticeable performance degradation compared to our proposed framework. A comprehensive comparative analysis and the corresponding hyperparameter configurations are provided in Appendix B.

TABLE VIII
MULTI-MODAL PROMPTING STRATEGY

Dataset	Fusion Method			ZSD	GZSD		
	One-Shot	One-way	Bi-directional		S	U	HM
FOWA	✓			9.0	88.5	8.0	14.7
		✓		9.7	90.3	9.3	16.9
			✓	**10.3**	**91.4**	**9.9**	**17.9**

3) Sensitivity Analysis of Textual Constraints: To further justify the necessity of the proposed semantic constraints, we conduct a sensitivity analysis on various textual prompt strategies. As summarized in Table XIII, while unconstrained long-form descriptions introduce significant semantic noise, yielding only marginal improvements (+1.3% in ZSD) over simple category names, our constrained prompts achieve a substantial performance leap (+7.5% in ZSD). This empirical evidence indicates that proper constraints are vital for maintaining representational purity. A deeper mechanistic analysis is provided in Appendix A.

4) Robustness to LLM Architectures and Variations: To assess the stability of SyMFood across different knowledge sources and evaluate its resilience against the inherent stochas-

ticity of LLMs, we conduct an extensive robustness study. As presented in Table IX, we evaluate our framework using a diverse suite of state-of-the-art LLMs, ranging from closed-source architectures (e.g., GPT-5, Grok-4) to prominent open-source vision-language models (e.g., Qwen2.5-VL series). This experiment investigates whether the proposed synergistic fusion mechanism effectively captures fundamental visual knowledge rather than being biased toward the linguistic patterns of a specific model.

The empirical results demonstrate that SyMFood is remarkably resilient to architectural variations and linguistic fluctuations. Notably, the performance gap among top-tier models is negligible, with Gemini 2.5 Pro (18.6% HM) and GPT-5 (18.4% HM) achieving nearly identical results. This consistency indicates that our synergistic fusion mechanism effectively extracts the underlying semantic invariants of food attributes. Furthermore, the ensemble-based stabilization strategy with $M = 5$ successfully neutralizes instance-level semantic noise, ensuring that SyMFood provides a stable semantic anchor regardless of the specific LLM employed.

TABLE IX
GENERALIZATION ANALYSIS OF SYMFOOD ACROSS DIFFERENT LLMS ON FOWA (%).

LLM Version	ZSD	GZSD		
		S	U	HM
Grok-4-fast-reasoning	10.3	73.6	8.4	15.0
Gemini 2.5 flash	9.6	73.0	9.2	16.3
Qwen2.5-VL-32B-Instruct	9.7	62.0	10.0	17.2
GPT-5	10.1	90.7	10.3	18.4
Qwen2.5-VL-72B-Instruct	10.0	81.7	10.4	18.5
Gemini 2.5 Pro (Ours)	**10.3**	**91.3**	**10.4**	**18.6**

5) Impact of Progressive Co-Refinement: To dissect the contributions of our ProFus module's key designs, we conducted an ablation study comparing three fusion strategies: (1) a *One-shot Fusion* baseline; (2) a *One-way Fusion* that introduces a progressive pipeline; and (3) our full *Bi-directional Co-Refinement* mechanism. As presented in Table VIII, the results on the FOWA dataset exhibit a clear, step-wise improvement across these configurations. The GZSD HM mAP progressively increases from 14.7% with *One-shot Fusion* to 16.9% with *One-way Fusion*, and finally reaches a peak of 17.9% with our complete ProFus module. This progression compellingly validates our design: the initial +2.2% gain demonstrates the necessity of the scale-by-scale process for resolving semantic-granularity mismatch, while the final +1.0% gain indicates that our core innovation—bi-directional co-refinement where visual features dynamically update the prompt—is critical for achieving the most precise cross-modal alignment.

6) Impact of Visual Exemplar Configuration: The configuration of visual exemplars is pivotal for bridging linguistic attributes with visual distributions. As demonstrated in Table X, increasing the number of exemplars from 8 to 32 leads to a substantial performance gain, peaking at 18.6% HM. This improvement confirms that a sufficient quantity of exemplars is necessary to capture the high intra-class diversity of food appearances, such as varying textures and presentation styles. However, the performance decline at 64 exemplars (17.7% HM) reveals a critical trade-off: an excessive exemplar count introduces redundant background noise and irrelevant visual features, which ultimately dilutes the purity of the semantic prototypes.

TABLE X
ABLATION STUDY ON THE NUMBER OF VISUAL EXEMPLARS PER CATEGORY.

Number of exemplars	ZSD	GZSD		
		S	U	HM
8	7.7	87.0	8.5	15.5
16	8.8	62.9	9.6	16.6
32 (Ours)	**10.3**	**91.3**	**10.4**	**18.6**
64	7.8	72.6	10.1	17.7

TABLE XI
ABLATION STUDY ON THE VISUAL EXEMPLAR SELECTION STRATEGY ($N = 32$).

Selection Strategy	ZSD	GZSD		
		S	U	HM
Random Selection	9.2	84.1	8.7	15.8
Representative Selection (Ours)	**10.3**	**91.3**	**10.4**	**18.6**

Furthermore, the selection strategy evaluation in Table XI underscores the necessity of representative sampling. Compared to naive random selection (15.8% HM), our strategy achieves a 2.8% improvement. This gap suggests that random sampling is susceptible to noisy samples and domain bias, which can cause significant semantic drift in the established anchors. By identifying the representative feature prototypes for each food category, our selection strategy effectively filters out visual outliers and ensures that the resulting semantic anchor is both robust and discriminative for fine-grained zero-shot detection.

7) Feature Representation and Localization Analysis: To provide an intuitive validation of SyMFood, we present t-SNE visualizations in Fig. 7. While the baseline exhibits scattered and overlapping distributions, our method forms compact and well-separated class clusters. To quantitatively assess this separability, we employ the silhouette coefficient, defined as $s(i) = (b(i) - a(i))/\max\{a(i), b(i)\}$, where $a(i)$ and $b(i)$ represent the average intra-cluster and minimum inter-cluster distances, respectively. As labeled in Fig. 7, SyMFood achieves a substantial silhouette coefficient improvement (e.g., 0.594 vs. 0.331 on FOWA). This gain suggests that our synergistic fusion mechanism effectively compresses the intra-class variance while maximizing the inter-class margins, establishing a highly discriminative semantic anchor.

The source of this improved separability is further illustrated by the Grad-CAM visualizations in Fig. 8. These attention maps demonstrate that our model learns to concentrate its focus precisely on target food items, explaining how our framework constructs the high-quality representations crucial

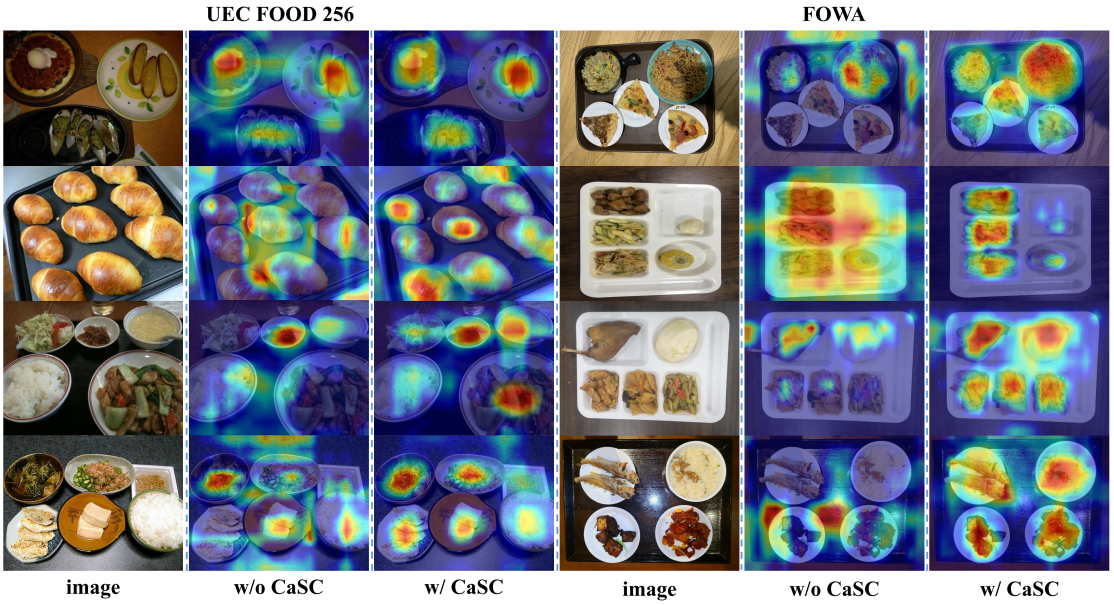


Fig. 8. Visualization of feature refinement effects via Grad-CAM. Each row contrasts the attention map of the model without CaSC (middle) with that of the model with CaSC (right) for a given original image (left) from the UEC FOOD 256 and FOWA datasets.

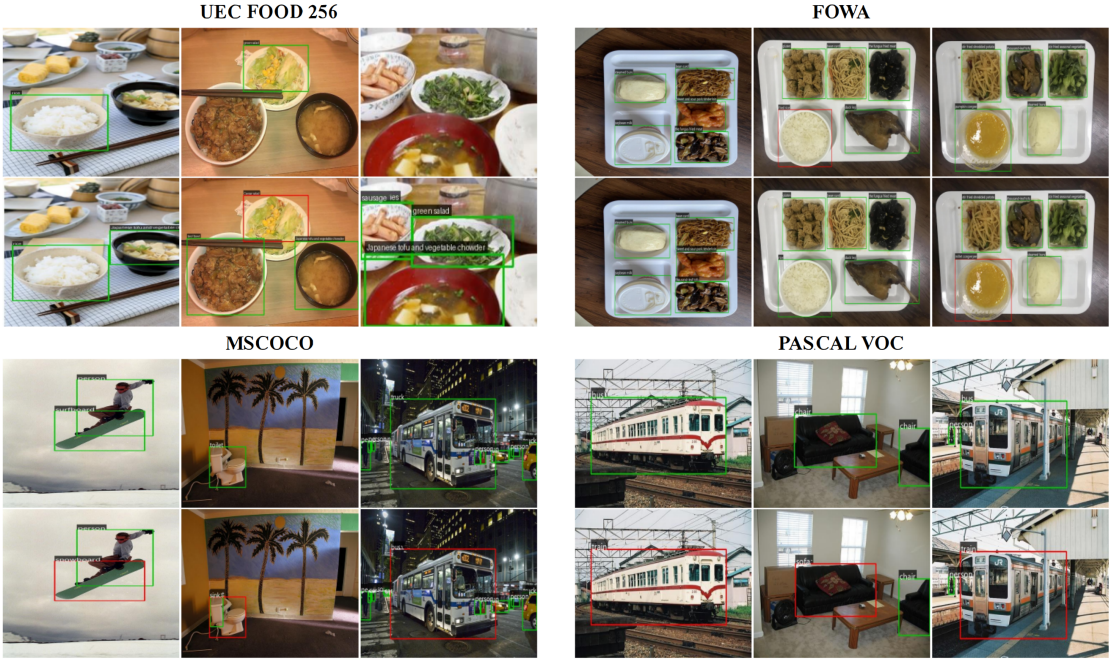


Fig. 9. Qualitative detection results comparison across four datasets under the GZSD setting. For each dataset: top row shows the baseline's results; bottom row presents our model's results.

for accurate zero-shot detection.

TABLE XII
MODEL COMPLEXITY ANALYSIS OF VARIOUS METHODS WITH THEIR MAP (%) ON UEC FOOD 256

Model	#Param. ↓	FLOPs ↓	FPS ↑	ZSD ↑	GZSD		
					S ↑	U ↑	HM ↑
SA [30]	50.1MB	193.5G	2.7	24.2	18.6	24.9	21.3
Ours	73.1MB	219.8G	2.7	31.7	29.6	30.7	30.1

E. Qualitative Analysis

To provide a visual evaluation of the detection results obtained by the proposed method, we generate visualizations across UEC FOOD 256, FOWA, PASCAL VOC and MS COCO datasets under GZSD setting, as depicted in Fig. 9. Red-colored bounding boxes are employed to mark unseen objects, while green-colored ones represent seen objects. Our method remarkably demonstrates precise simultaneous detection of both seen and unseen objects across all evaluated datasets and settings.

Fig. 10. Failure analysis for ZSFD, illustrating common errors like misclassification and false positives.

F. Discussion

Our experimental results validate that SyMFood establishes a new SOTA in ZSFD, and we attribute this success to our proposed "Refine-then-Fuse" architecture. The initial visual feature refinement by the CaSC module and the subsequent bi-directional co-refinement by the ProFus module prove critical in resolving the semantic-granularity mismatch. To assess its practicality, we also analyze its efficiency. As shown in Table IX, while SyMFood has a marginally higher parameter count and computational cost compared to the SA baseline, it achieves substantial performance gains while maintaining a competitive inference speed (e.g., 2.7 FPS on FOWA), demonstrating a superior performance-cost trade-off.

Beyond the scope of food detection, the core philosophy of the proposed method exhibits significant potential for broader multi-modal applications. Specifically, the hierarchical refinement and progressive learning strategies introduced in [49] and [50] emphasize the importance of structured feature inference and instance-aware learning, which are highly relevant to our architectural design. Furthermore, [51] demonstrates the efficacy of progressive alignment in action recognition, while the clustering-based prototype learning in [52] and cross-modal correspondence learning in [53] provide insights for extending our synergistic prompting mechanism to more complex tasks such as compositional zero-shot learning and audio-visual localization.

However, we acknowledge certain limitations. As illustrated in Fig. 10, common failure cases include misclassifications between highly similar food items and false positive detections in ambiguous regions, indicating that further optimization is still needed. Future work could thus focus on more advanced prompt generation techniques and on creating a lighter, more efficient version of the framework through methods like knowledge distillation.

V. Conclusion

Our work identified and addressed two fundamental challenges in ZSFD: the SD caused by the abstract and diverse nature of food concepts, and the AB arising from sub-optimal cross-modal fusion strategies. To this end, we proposed SyMFood, a novel framework built upon a "Refine-then-Fuse" philosophy. SyMFood integrates a powerful multi-modal prompting system, which combines attribute-rich text with representative visual exemplars, and a ProFus architecture. This architecture first refines visual features via our CaSC block and then performs a bi-directional, iterative fusion to achieve robust alignment between semantics and vision across multiple scales. Extensive experiments have validated that SyMFood not only sets a new SOTA on multiple food-specific datasets but also demonstrates strong generalization capabilities. While the current framework focuses on textual and visual modalities, future research will aim to expand this sensory scope. Specifically, the integration of non-visual attributes, such as olfactory (smell) and gustatory (taste) cues, represents a promising direction for achieving a more holistic and human-like representation of food concepts. We believe our work provides a promising new direction for tackling fine-grained ZSD problems.

References

[1] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *Acm Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–36, 2019.

[2] Y. Liu, W. Min, S. Jiang, and Y. Rui, "Convolution-enhanced bi-branch adaptive transformer with cross-task interaction for food category and ingredient recognition," *IEEE Transactions on Image Processing*, 2024.

[3] X. Wu, S. Yu, E.-P. Lim, and C.-W. Ngo, "Ovfoodseg: Elevating open-vocabulary food image segmentation via image-informed textual representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4144–4153.

[4] N. Hayat, M. Hayat, S. Rahman, S. Khan, S. W. Zamir, and F. S. Khan, "Synthesizing the unseen for zero-shot object detection," in *Proceedings of the Asian conference on computer vision*, 2020.

[5] P. Huang, J. Han, D. Cheng, and D. Zhang, "Robust region feature synthesizer for zero-shot object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7622–7631.

[6] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis, "Zero-shot object detection by hybrid region embedding," *arXiv preprint arXiv:1805.06157*, 2018.

[7] C. Yan, X. Chang, M. Luo, H. Liu, X. Zhang, and Q. Zheng, "Semantics-guided contrastive network for zero-shot object detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 3, pp. 1530–1544, 2022.

[8] Z.-X. Ma, Z.-D. Chen, T. Zheng, X. Luo, and X.-S. Xu, "Btg-net++: Enhanced bi-directional task-guided network for few-shot fine-grained image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[9] N. Zheng, X. Song, W. T. Tang, S. K. Ng, L. Nie, R. Zimmermann, "Unsupervised Few-shot Food Recognition with Intra-Class Variation and Inter-Class Similarity Modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[10] L. Yao, R. Pi, J. Han, X. Liang, H. Xu, W. Zhang, Z. Li, and D. Xu, "Detclipv3: Towards versatile generative open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 391–27 401.

[11] J. Wu, Y. Jiang, Q. Liu, Z. Yuan, X. Bai, and S. Bai, "General object foundation model for images and videos at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3783–3795.

[12] Y. Shen, C. Fu, P. Chen, M. Zhang, K. Li, X. Sun, Y. Wu, S. Lin, and R. Ji, "Aligning and prompting everything all at once for universal visual perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 193–13 203.

[13] M. Liu, H. Bai, F. Li, C. Zhang, Y. Wei, M. Wang, T.-S. Chua, and Y. Zhao, "Psvma+: exploring multi-granularity semantic-visual adaption for generalized zero-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[14] W. Jun, W. Moon, C.-H. Cho, M. Jung, and J.-P. Heo, "Bridging the semantic granularity gap between text and frame representations for partially relevant video retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 4166–4174.

[15] R. Ran, J. Wei, S. He, Y. Zhou, P. Wang, Y. Yang, and H. T. Shen, "Fine-grained alignment and interaction for video grounding with cross-modal semantic hierarchical graph," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[16] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao, "Glipv2: Unifying localization and vision-language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 067–36 080, 2022.

[17] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 38–55.

[18] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "Regionclip: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 793–16 803.

[19] Z. Chen, X. Zhao, C. Lang, L. Wei, T. Wang, and Y. Li, "Learning diversified primitive prompts for compositional zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 10, pp. 10423–10436, 2025.

[20] A. Raza, B. Yang, and Y. Zou, "Zero-shot temporal action detection by learning multimodal prompts and text-enhanced actionness," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 11 000–11 012, 2024.

[21] M. Hong, X. Zhang, G. Li, and Q. Huang, "Multi-modal multi-grained embedding learning for generalized zero-shot video classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5959–5972, 2023.

[22] Q. Jiang, F. Li, Z. Zeng, T. Ren, S. Liu, and L. Zhang, "T-rex2: Towards generic object detection via text-visual prompt synergy," in *European Conference on Computer Vision*. Springer, 2024, pp. 38–57.

[23] F. Li, Q. Jiang, H. Zhang, T. Ren, S. Liu, X. Zou, H. Xu, H. Li, J. Yang, C. Li *et al.*, "Visual in-context prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 861–12 871.

[24] Q. Chen, W. Jin, J. Ge, M. Liu, Y. Yan, J. Jiang, L. Yu, X. Guo, S. Li, and J. Chen, "Cp-detr: Concept prompt guide detr toward stronger universal object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 2, 2025, pp. 2141–2149.

[25] Y. Xu, M. Zhang, C. Fu, P. Chen, X. Yang, K. Li, and C. Xu, "Multi-modal queried object detection in the wild," *Advances in Neural Information Processing Systems*, vol. 36, pp. 4452–4469, 2023.

[26] X. Ma, J. Yang, J. Lin, Z. Zheng, S. Li, B. Hu, and X. Tang, "LVAR-CZSL: Learning visual attributes representation for compositional zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 12, pp. 13311–13323, 2024.

[27] Y. Huang, Z. Hechen, M. Zhou, Z. Li, and S. Kwong, "An attention-locating algorithm for eliminating background effects in fine-grained visual classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[28] W. Hou, S. Chen, S. Chen, Z. Hong, Y. Wang, X. Feng, S. Khan, F. S. Khan, and X. You, "Visual-augmented dynamic semantic prototype for generative zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 23 627–23 637.

[29] S. Chen, D. Fu, S. Chen, S. Ye, W. Hou, and X. You, "Causal visual-semantic correlation for zero-shot learning," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 4246–4255.

[30] H. Liu, L. Zhang, J. Guan, and S. Zhou, "Zero-shot object detection by semantics-aware detr with adaptive contrastive loss," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 4421–4430.

[31] L. Hu, W. Cao, Z. Zhang, and Y. Liang, "Progressive feature reconstruction network for zero-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 6, pp. 5265–5278, 2025.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[33] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[34] Y. Wang, M. Hong, L. Huangfu, and S. Huang, "Data distribution distilled generative model for generalized zero-shot recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5695–5703.

[35] D. Yin, L. Hu, B. Li, Y. Zhang, and X. Yang, "5%> 100%: Breaking performance shackles of full fine-tuning on visual recognition tasks," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 071–20 081.

[36] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 965–10 975.

[37] P. Zhou, W. Min, Y. Zhang, J. Song, Y. Jin, and S. Jiang, "Seeds: Semantic separable diffusion synthesizer for zero-shot food detection," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8157–8166.

[38] P. Zhou, W. Min, J. Song, Y. Zhang, and S. Jiang, "Synthesizing knowledge-enhanced features for real-world zero-shot food detection," *IEEE Transactions on Image Processing*, vol. 33, pp. 1285–1298, 2024.

[39] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," *arXiv preprint arXiv:1312.5650*, 2013.

[40] Y. Zheng, R. Huang, C. Han, X. Huang, and L. Cui, "Background learnable cascade for zero-shot object detection," in *Proceedings of the Asian conference on computer vision*, 2020.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[44] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15691–15701, 2023.

[45] S. Rahman, S. Khan, and F. Porikli, "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts," in *Asian conference on computer vision*. Springer, 2018, pp. 547–563.

[46] S. Rahman, S. Khan, and N. Barnes, "Polarity loss: Improving visual-semantic alignment for zero-shot detection," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[47] H. Li, J. Mei, J. Zhou, and Y. Hu, "Zero-shot object detection based on dynamic semantic vectors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9267–9273.

[48] H. Nie, R. Wang, and X. Chen, "From node to graph: Joint reasoning on visual-semantic relational graph for zero-shot detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 1109–1118.

[49] R. Yan, L. Xie, J. Tang, X. Shu, T. Wang, and Q. Tian, "HiGCIN: Hierarchical graph-based cross inference network for group activity recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6955–6968, 2023.

[50] R. Yan, L. Xie, X. Shu, L. Zhang, and J. Tang, "Progressive instance-aware feature learning for compositional action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10317–10330, 2023.

[51] H. Qu, R. Yan, X. Shu, H. Gao, P. Huang, and G. Xie, "MVP-Shot: Multi-velocity progressive-alignment framework for few-shot action recognition," *IEEE Transactions on Multimedia*, vol. 27, pp. 6593–6605, 2025.

[52] H. Qu, J. Wei, X. Shu, and W. Wang, "Learning clustering-based prototypes for compositional zero-shot learning," *arXiv preprint arXiv:2502.06501*, 2025.

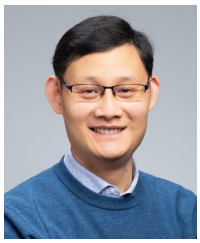
[53] L. Xing, H. Qu, R. Yan, X. Shu, and J. Tang, "Locality-aware cross-modal correspondence learning for dense audio-visual events localization," *arXiv preprint arXiv:2409.07967*, 2024.

Xinlong Wang received a Bachelor's degree in Information Management and Information Systems at Hebei North University. Currently, he is pursuing his Master's degree in the Computer Science Department at Ludong University. His research primarily revolves around zero-shot object detection, with a specific focus on applications in the domain of food.

Weiqing Min (Senior Member, IEEE) is currently an Associate Professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored or coauthored more than 50 peer-referenced papers in relevant journals and conferences, including Patterns (Cell Press), ACM Computing Surveys, Trends in Food Science and Technology, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON IMAGE PROCESSING, Food Chemistry, ACM MM, AAAI, and IJCAI. His research interests include multimedia content analysis and food computing. He was a Senior Member of CCF. He was a recipient of the 2016 ACM Transactions on Multimedia Computing, Communications, and Applications, the Nicolas D. Georganas Best Paper Award, and the 2017 IEEE Multimedia Magazine Best Paper Award. He was the Guest Editor for the special issues on international journals, such as IEEE TRANSACTIONS ON MULTIMEDIA, IEEE MULTIMEDIA, and Foods.

Shoulong Liu obtained a Bachelor's degree in Software Engineering from Shandong University of Science and Technology and is currently pursuing a Master's degree in Computer Science and Technology at Ludong University. His research interests include open-vocabulary object detection in food and related fields.

Guorui Sheng received the M.E. degree from Kunsan National University, South Korea in 2007, and received the Ph.D. degree from Nankai University, China, in 2017. He served as a research assistant to scholar Bruce Denby at the School of Computer Science and Technology, Tianjin University, from 2017 to 2018. He is currently a Lecturer at the Department of Information and Electrical Engineering, Ludong University, Yantai, China. He has authored or co-authored more than 20 peer-referenced papers in relevant journals and conferences, including ACM Transactions on Multimedia Computing, Communications, and Applications and Nutrients. His research interests include computer vision, deep learning and food computing.



Shuqiang Jiang (Senior Member, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, and a Professor with the University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. He has authored or coauthored more than 150 articles. He was supported by the National Science Fund for Distinguished Young Scholars in 2021, the NSFC Excellent Young Scientists Fund in 2013, and the Young Top-Notch Talent of Ten Thousand Talent Program in 2014. His research interests include multimedia analysis and multimodal intelligence. He is a Senior Member of CCF and a member of ACM. He has served as a TPC Member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, AAAI, ICME, ICIP, and PCM. He received the Lu Jiaxi Young Talent Award from CAS in 2012 and the CCF Award of Science and Technology in 2012. He is the Vice Chair of the IEEE CASS Beijing Chapter and the ACM SIGMM China Chapter. He was the General Chair of ICIMCS in 2015 and the Program Chair of the 2019 ACM Multimedia Asia and PCM in 2017. He is an Associate Editor of Multimedia Tools and Applications and ACM Transactions on Multimedia Computing, Communications, and Applications.