GLANCING, FOCUSING, REASONING: A COGNITION-INSPIRED FRAMEWORK FOR FINE-GRAINED FOOD RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Since ancient times, understanding food has been essential to human survival, development, and cultural formation, relying on visual perception and reasoning grounded in prior knowledge. Currently, multimodal large language models (MLLMs) exhibit remarkable capabilities in visual-semantic understanding. However, enabling MLLMs to understand fine-grained food categories poses three food-specific challenges: discovering non-part-aligned discriminative cues under interference from shared superclass patterns, recognizing ingredients under inter-ingredient similarity and intra-ingredient appearance variation, and reasoning over ingredient roles and relations. Inspired by cognitive science, we decompose food understanding into three progressive cognitive actions to address these challenges: *glancing* establishes coarse food context for discriminative cue discovery; *focusing* localizes regions to recognize ingredient identity and state; and *reasoning* integrates high-level semantics such as cooking methods, ingredient roles, and inter-ingredient relations. Accordingly, we propose **GFRFood**, a cognition-inspired MLLM framework for fine-grained food recognition (FGFR) that instantiates these three cognitive actions through two complementary modules. **Cognition-Inspired Progressive Semantic Reasoning (CIPSR)** generates food attributes by progressively integrating discriminative ingredient cues and high-level semantics; **Taxonomy-Conditioned Semantic State-Space Model (TC-S3M)** uses a selective state-space model to refine fine-grained category understanding. The two modules model visual and taxonomy semantics, respectively, and jointly enhance the MLLM's fine-grained understanding. Extensive evaluations demonstrate that GFRFood outperforms general, fine-grained, and food-specific MLLMs, achieving state-of-the-art performance. Ablation studies and visualizations demonstrate the advantages of GFRFood in attribute quality and taxonomic understanding. Dataset and code will be released upon publication.

1 INTRODUCTION

Fine-grained food recognition (FGFR) is a challenging visual understanding task due to high inter-class similarity, large intra-class variation, and variable discriminative cues associated with ingredients, textures, and cooking states. Limited FGFR capability constrains food-related applications such as nutrition assessment (Yan et al., 2025; Fridolfsson et al., 2025) and food recommendation (Yang et al., 2024; Soekamto et al., 2025). Building on the vision-language understanding, reasoning, and generation capabilities of multimodal large language models (MLLMs), recent studies have explored food-domain adaptation (Yin et al., 2025), enhanced local visual perception (Jiao et al., 2025), and organized food analysis hierarchically (Lee et al., 2026). Despite these advances, current methods still struggle with FGFR due to insufficient cognitive guidance for modeling discriminative visual semantics. Specifically, we identify three food-specific challenges, as illustrated in Figure 1.

1. Shared superclass patterns hinder discriminative cue discovery. Fine-grained food subclasses within the same superclass share global visual patterns, leaving discriminative evidence confined to sparse local cues. These cues are not consistently part-aligned and vary in identity, appearance, and location, while intra-class variation further hinders discriminative cue discovery. In Figure 1(a,b), *shrimp pizza* and *shrimp fried rice* are predicted as *roast chicken pizza* and *egg fried rice*, respectively.

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

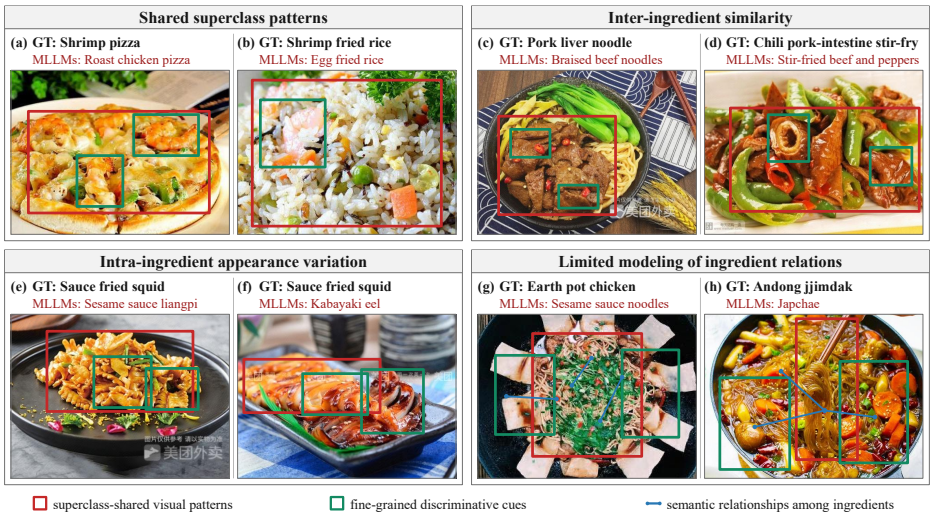

Figure 1: Food-specific challenges for MLLMs. Examples are selected from erroneous predictions by several general MLLMs, including Qwen3-VL, GLM-4.1V, and Idefics3. Red, green, and blue annotations denote superclass-shared visual patterns, fine-grained discriminative cues and ingredient relationships.

A plausible explanation is that superclass-shared *crust-and-cheese* and *fried-rice* patterns hinder the discovery of discriminative *shrimp* cues. In summary, MLLMs show effective coarse-grained recognition while remaining limited in discovering discriminative fine-grained cues.

2. Inter-ingredient similarity and intra-ingredient appearance variation disrupt visual-semantic correspondence. Different ingredients often share highly similar visual appearances, which may obscure fine-grained category boundaries. In Figure 1(c,d), *pork liver noodle* and *chili pork-intestine stir-fry* are mistaken for *braised beef noodles* and *stir-fried beef and peppers*, respectively, potentially due to the visual similarity of pork liver and pork intestine to beef. Furthermore, different processing methods such as cutting and cooking lead to intra-ingredient appearance variation, which disrupts visual-semantic consistency. For *sauce fried squid*, squid appears as pale, short curled pieces in Figure 1(e), while it appears as glossy brown sauce-coated slices in (f). Such variation can lead MLLMs to misclassify the two images as *sesame sauce liangpi* and *kabayaki eel*, respectively.

3. Limited modeling of semantic relationships among ingredients constrains fine-grained understanding. FGFR depends not only on which ingredients are present, but also on their semantic relationships, including ingredient roles, co-occurrence, interactions, and organization within the dish. In Figure 1(g), MLLMs appear to treat noodles as the core ingredient in *earth-pot chicken*, which they misclassify as *sesame sauce noodles*, without inferring their role in relation to the chicken stew and surrounding flatbreads. In (h), *Andong jjimdak* is misclassified as *Japchae*, likely due to insufficient modeling of the co-occurrence relationships among chicken, vegetables, and dark sauce.

Addressing these challenges requires **1)** discriminative visual cue discovery, **2)** ingredient recognition under inter-ingredient similarity and intra-ingredient appearance variation, and **3)** reasoning over semantic relationships among ingredients. Recent attribute-based MLLM methods are effective in fine-grained visual recognition (FGVR) (He et al., 2025; Kim & Ji, 2024). However, their attributes lack ingredient identities and exhibit poor ingredient coverage and low classification value, as shown in Figure 2(a). Neuroscientific and cognitive studies suggest that human visual recognition combines coarse-to-fine processing with top-down guidance, using global context to direct attention toward relevant local evidence and prior knowledge to interpret that evidence (Schyns & Oliva, 1994; Torralba et al., 2006; Bar et al., 2006). Inspired by this mechanism, we decompose food understanding into three progressive cognitive actions: ***glancing*** establishes global context to discover discriminative visual cues and infer a coarse food cluster. ***Focusing*** localizes discriminative regions to recognize ingredient identity and state and infer the relevant ingredient family. ***Reasoning*** combines visual cues with high-level semantics such as cooking methods and ingredient relations to form multi-dimensional visual attributes, infer the processing form, and ultimately identify the fine-grained category. As

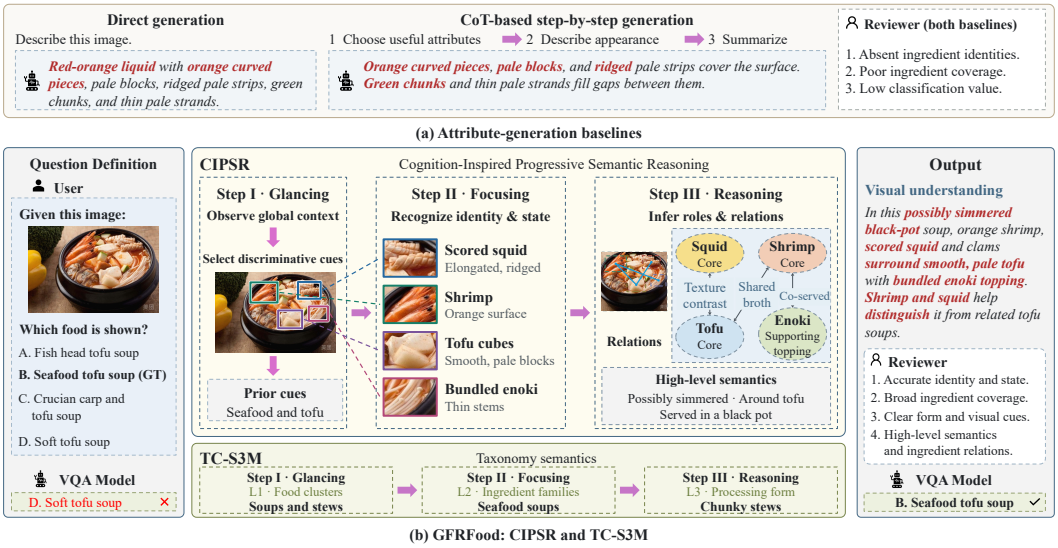

Figure 2: (a) Comparison of attribute quality across attribute-based generation methods. GPT-based comments on ingredient accuracy, coverage, and classification value serve only as qualitative illustrations; Tables 3, S2, and S3 report quantitative ablations. (b) Our cognition-inspired *glancing–focusing–reasoning* process for food understanding.

shown in Figure 2(b), *glancing* establishes a soup context and discovers potential shrimp and tofu cues, inferring *Soup dishes*. *Focusing* recognizes scored squid and shrimp, leading to *Seafood soups*. *Reasoning* combines ingredient cues with high-level semantics such as simmering and seafood-tofu co-occurrence, infers *Chunky stew soups*, and ultimately identifies *seafood tofu soup*.

Motivated by the above analysis, we propose **GFRFood**, a cognition-inspired MLLM framework for FGFR that progressively captures food-specific discriminative semantics through this cognitive process. Our framework consists of two complementary modules modeling visual semantics and taxonomy semantics, respectively. **Cognition-Inspired Progressive Semantic Reasoning (CIPSR)** progressively models fine-grained food attributes by glancing to capture category-level ingredient priors, focusing on discriminative regions to recognize ingredient identity and state, and reasoning over visual cues and high-level semantics to generate multi-dimensional semantic attributes. Inter-stage conditioning propagates these priors and ingredient-level evidence, while multi-agent verification and self-correction improve the visual grounding and semantic consistency of generated attributes. **Taxonomy-Conditioned Semantic State-Space Model (TC-S3M)** models taxonomy semantics through selective state transitions across taxonomy levels, progressively refining representations from coarse food clusters through ingredient families to processing form. Furthermore, we construct FoodFG, a large-scale FGFR benchmark comprising 284,935 images across 1,622 categories, with reduced noise, standardized category names, and consistent category granularity. Extensive experiments on FoodFG demonstrate that GFRFood outperforms evaluated MLLMs, surpassing Qwen3-VL, Fine-R1, and FoodLMM in relaxed accuracy (R-Acc.) by 33.03%, 18.99%, and 32.50%, respectively. Ablation studies and visualizations verify the effectiveness of both modules.

In summary, our contributions are as follows: **(i)** We systematically analyze the food-specific challenges faced by MLLMs in FGFR and propose **GFRFood**, a cognition-inspired framework for progressive food understanding. **(ii)** We introduce **CIPSR** to model discriminative food semantics and construct multi-dimensional attributes, and **TC-S3M** to refine coarse-to-fine category understanding through taxonomy semantic modeling. **(iii)** Extensive experiments demonstrate the superiority of GFRFood over general, fine-grained, and food-specific MLLMs.

2 Related Work

Multimodal Large Language Models. MLLMs aim to support scene understanding, visual reasoning, and vision-language generation by integrating visual perception with large language models.

Early MLLMs connect pretrained visual encoders and language models through cross-attention or query-based modules (Alayrac et al., 2022; Li et al., 2023). Beyond cross-modal fusion, subsequent research emphasizes visual detail preservation and vision-language representation alignment. Qwen2-VL and Qwen3-VL enhance visual perception through dynamic-resolution processing and multilevel feature integration, respectively (Wang et al., 2024; Bai et al., 2025). InternVL further combines visual encoder scaling with contrastive and generative training (Chen et al., 2024). Despite these advances, MLLMs still struggle with FGVR, particularly in detecting subtle visual differences and generating discriminative attributes (Tong et al., 2024; Kim & Ji, 2024). Some studies address fine-grained recognition through local-region search for discriminative cues (Li et al., 2025). Other studies improve category discrimination through attribute-based alignment and comparative reasoning (He et al., 2025; 2026a), and prediction consistency across category levels through taxonomy-aware representations (He et al., 2026b). However, without explicit food-domain guidance, these methods alone do not ensure the discovery of class-discriminative visual patterns, reliable ingredient recognition, or accurate interpretation of inter-ingredient semantic relationships.

Fine-grained Food Recognition. Early studies addressed FGFR through ingredient-guided attention, local feature learning, and taxonomy-based classification (Min et al., 2019; 2023; Mao et al., 2021). Recently, MLLM-based studies have explored three directions in food understanding: domain adaptation, local visual perception, and multimodal reasoning. For domain adaptation, FoodLMM and LLaVA-Chef use food-domain instruction tuning for food recognition and recipe generation (Yin et al., 2025; Mohbat & Zaki, 2024), while RoDE employs diverse LoRA experts to reduce conflicts among food tasks (Jiao et al., 2026). For local visual perception, visual cue enhancement complements Dual-LoRA by aggregating multilevel local features (Jiao et al., 2025). For multimodal reasoning, MultiFoodChat refines food predictions through multi-agent visual–textual dialogue (Hu & Zhuang, 2025), whereas FoodCHA organizes category, subcategory, and cooking-style recognition into a taxonomy-guided process (Lee et al., 2026). However, existing methods remain limited in discriminative visual cue discovery, accurate ingredient recognition, and reasoning over ingredient roles and relations. **GFRFood** addresses these limitations through cognition-inspired guidance for progressive food understanding.

3 METHOD

Inspired by human food cognition, GFRFood decomposes FGFR into *glancing*, *focusing*, and *reasoning*. Figure 3 shows two complementary modules: (1) **Cognition-Inspired Progressive Semantic Reasoning (CIPSR)** constructs discriminative food attributes through progressive multi-agent reasoning; (2) **Taxonomy-Conditioned Semantic State-Space Model (TC-S3M)** models taxonomy semantics through selective state updates. The two modules jointly model complementary visual and taxonomy semantics to enhance the MLLM's fine-grained understanding.

3.1 COGNITION-INSPIRED PROGRESSIVE SEMANTIC REASONING

In FGFR, discriminative visual attributes effectively align complex visual patterns with simple category names despite the large semantic gap between them (He et al., 2025). CIPSR generates discriminative, verifiable, multidimensional visual attribute descriptions through self-correcting multi-agent collaboration across three cognition-inspired stages: *glancing*, *focusing*, and *reasoning*.

Glancing: Prior-Guided Discriminative Pattern Discovery. Human glancing draws on food knowledge to establish coarse global context and discover potentially discriminative cues amid superclass-shared visual patterns. Accordingly, we introduce category-level ingredient priors. For training class y, **Proposer** generates ingredient candidates from representative images R_y and the taxonomic ancestor structure T_y (Section 3.2). It then generates visually similar confusers for each candidate and blindly combines both into the prior to guide cue discovery and limit bias toward class-typical ingredients. A **Verifier** assesses candidate plausibility and visual confusability, yielding

$$P_y = \mathbf{V}(\mathbf{P}(R_y, T_y)), \tag{1}$$

where $P_y = \{p_{y,1}, \ldots, p_{y,N_p}\}$ contains N_p ingredient candidates and confusers; $\mathbf{P}$, $\mathbf{V}$ denote the **Proposer** and **Verifier** agents, respectively.

Focusing: Collaborative Ingredient Identity–State Recognition and Correction. Focusing aims to discover ingredients and their states from local visual evidence. To address visual–semantic

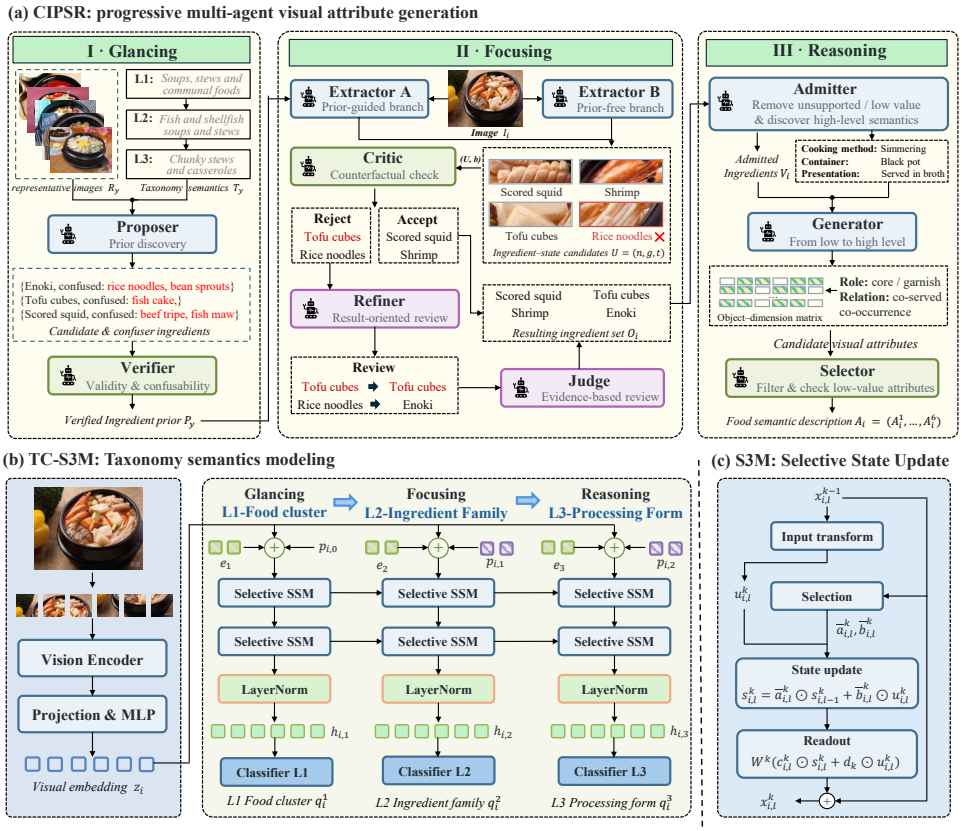

Figure 3: Overview of GFRFood. (a) CIPSR generates food attributes through prior discovery, dual-branch ingredient extraction, collaborative verification and recovery, and semantic reasoning. (b) TC-S3M models taxonomy semantics across L1 food clusters, L2 ingredient families, and L3 processing form. (c) The Selective SSM block in (b) updates its state across taxonomy depths and produces its output through readout and residual addition.

inconsistencies caused by form variation, we first decompose each ingredient into a state-aware triplet $U = (n, g, t)$, where n, g, and t denote the ingredient name, base identity, and visible physical form (e.g., cubes or strips), respectively. To complement prior-guided pattern discovery with image-specific exploration, **Extractor** processes food image I_i of class y_i through two branches: the prior-guided branch uses P_{y_i} to localize and identify discriminative ingredients, while the prior-free branch discovers discriminative ingredients absent from the prior. Merging the two branches yields ingredient–region pairs (U, b), where b denotes the bounding box grounding the ingredient.

Extractor may misjudge ingredient identity, semantic granularity, physical form, or presence. Drawing on counterfactual reasoning (Kim et al., 2024), a **Critic** assumes each ingredient candidate is incorrect and revises the ingredient identity or state when an alternative has stronger local visual support, aiming to mitigate confirmation bias in self-verification. **Refiner** and **Judge** provide two-stage verification of **Critic**'s counterfactual judgments: **Refiner** performs result-oriented reassessment of rejected candidates, blinded to their source agents, while **Judge** conducts evidence-based adjudication of **Critic** and **Refiner**'s reassessments, blinded to reviewer identities. The process is formulated as

$$\begin{aligned}(C_i^+, C_i^-) &= \mathbf{C}(I_i, \mathbf{E}(I_i, P_{y_i}))\,, \\ O_i &= C_i^+ \cup \mathbf{J}\big(I_i, C_i^-, \mathbf{R}(I_i, C_i^-)\big)\,,\end{aligned} \tag{2}$$

where C_i^+ and C_i^- are **Critic**'s accepted and rejected sets; $O_i = \{(U_{i,j}, b_{i,j})\}_{j=1}^{m_i}$ pairs m_i verified ingredient triplets $U_{i,j}$ with their bounding boxes $b_{i,j}$; $\mathbf{E}$, $\mathbf{C}$, $\mathbf{R}$, and $\mathbf{J}$ denote the **Extractor**, **Critic**, **Refiner**, and **Judge** agents, respectively.

Reasoning: High-Level Semantic and Relational Attribute Generation. Reasoning integrates basic visual features with high-level semantics of cooking, presentation, and ingredient roles and relations. We organize these attributes into six complementary dimensions: discriminative cues; surface texture and state; color contrast; geometric form and portion; spatial arrangement and containers; and toppings and garnish. Specifically, **Admitter** first filters O_i by visual support and classification value, adding missed ingredients or other distinctive visual cues, such as containers and decorative elements, that satisfy both criteria to form $V_i = \mathbf{A}(I_i, O_i)$. **Generator** builds an object–dimension correspondence table over V_i and combines whole-image context with object crops to generate attributes. **Selector** filters low-value attributes, checks subject consistency, and removes cross-dimensional duplicates through object–predicate–relation comparisons, retaining distinct properties of the same object and discriminative evidence. The resulting attributes are

$$A_i = \mathbf{S}(\mathbf{G}(I_i, \mathbf{A}(I_i, O_i))), \tag{3}$$

where $A_i = (A_i^1, \ldots, A_i^6)$ denotes the generated food attributes and groups them by these six dimensions; $\mathbf{A}$, $\mathbf{G}$, and $\mathbf{S}$ denote the **Admitter**, **Generator**, and **Selector** agents, respectively.

3.2 Taxonomy-Conditioned Semantic State-Space Model

Alongside visual semantic understanding, human food cognition involves coarse-to-fine inference of taxonomy semantics. Glancing infers a coarse-grained food cluster from global context. Focusing narrows this cluster to an ingredient family through local discriminative cues. Reasoning further infers the processing form using high-level semantics such as cooking methods and ingredient relations.

As shown in Figure 3(b), we formalize this coarse-to-fine progression as three taxonomy levels: (i) L1 food clusters, (ii) L2 ingredient families, and (iii) L3 processing form (details in Appendix A.2). Specifically, image I_i with fine-grained label y_i has a unique taxonomy path with ordered ancestors $T_{y_i} = (y_i^1, y_i^2, y_i^3)$, where y_i^l denotes its ancestor at level $l \in \{1, 2, 3\}$. We adopt a taxonomy-conditioned selective state-space model (S3M) (Gu & Dao, 2024) to model coarse-to-fine taxonomy evolution, owing to input-dependent selection that propagates ancestor semantics to finer-grained categories through taxonomy paths. As shown in Figure 3(b,c), we first extract the visual embedding of the input food image I_i as:

$$z_i = \mathbf{M}(\mathbf{P}_v(\mathbf{E}_v(I_i))) \in \mathbb{R}^{d_h}, \tag{4}$$

where $\mathbf{E}_v$, $\mathbf{P}_v$, and $\mathbf{M}$ denote the frozen visual encoder with pooling, input projection, and two residual MLP blocks, respectively. To explicitly represent the current granularity and ancestor taxonomy semantics, we form the taxonomy-conditioned feature $x_{i,l}^0 = z_i + e_l + p_{i,l-1}$, where e_l and $p_{i,l-1}$ denote learnable depth and parent-node embeddings, respectively, with $p_{i,0} = 0$. The parent-node embeddings correspond to ground-truth ancestors during training and predicted ancestors during inference. Greedy inference selects the highest-scoring valid node at L1 and L2 and incorporates its embedding into the next depth's taxonomy-conditioned feature. The taxonomy-conditioned feature $x_{i,l}^0$ is fed into two stacked S3M layers, where the input is first mapped through a learned transformation $u_{i,l}^k = F_k(x_{i,l}^{k-1})$ and then used to update the state across taxonomy depths as follows:

$$s_{i,l}^k = \bar{a}_{i,l}^k \odot s_{i,l-1}^k + \bar{b}_{i,l}^k \odot u_{i,l}^k, \tag{5}$$

where $k \in \{1, 2\}$ indexes the S3M layers, $s_{i,l}^k$ denotes the latent state initialized as $s_{i,0}^k = 0$, and $\odot$ denotes element-wise multiplication. $\bar{a}_{i,l}^k$ and $\bar{b}_{i,l}^k$ denote the discrete state-transition and input coefficients, respectively, and are computed from input-dependent parameters through matrix-exponential discretization of diagonal state dynamics:

$$\bar{a}_{i,l}^k = \exp\big(-\Delta_{i,l}^k \odot r_k\big), \qquad \bar{b}_{i,l}^k = \frac{1 - \bar{a}_{i,l}^k}{r_k} \odot b_{i,l}^k, \tag{6}$$

where $r_k > 0$ denotes the learnable decay vector governing state retention. $\Delta_{i,l}^k$ and $b_{i,l}^k$ denote the discretization step-size and input-to-state projection vectors, computed from the input through projections and activations, reflecting input-dependent selection. The updated state $s_{i,l}^k$ and input $u_{i,l}^k$ are then fed into a weighted readout with an output projection and a residual shortcut to produce the updated feature:

$$x_{i,l}^k = x_{i,l}^{k-1} + W^k\big(c_{i,l}^k \odot s_{i,l}^k + d_k \odot u_{i,l}^k\big), \tag{7}$$

where $c_{i,l}^k$ denotes the input-dependent readout vector computed through projection and activation. d_k and W^k denote learnable direct-input weights and output projection, respectively. The S3M output

$x^2_{i,l}$ is layer-normalized to obtain the final taxonomy semantic feature $h_{i,l} = \mathrm{LN}(x^2_{i,l})$ at each depth. For optimization, $h_{i,l}$ is mapped to level-l scores q^l_i through a learnable cosine classifier and softmax over the selected parent's child categories. The path loss optimizes ancestor-category prediction across taxonomy depths and is formulated as:

$$L_{\text{path}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{3}\omega_l \sum_j y^l_{i,j}\log q^l_{i,j}, \tag{8}$$

where N, ω_l, and $y^l_{i,j}$ denote the training-set size, depth weight, and one-hot indicator of ancestor y^l_i at category j, respectively. We concatenate z_i with the three depth-specific features $h_{i,1}$, $h_{i,2}$, and $h_{i,3}$ and feed the resulting vector to the learnable classifier W_f to predict fine-grained logits f_i. The equally weighted objective is $L_{\text{SSM}} = L_{\text{path}} + L_{\text{fine}}$, with $L_{\text{fine}} = \frac{1}{N}\sum_i \mathrm{CE}(f_i, y_i)$, where CE denotes cross-entropy. During inference, TC-S3M logits are combined with MLLM category scores through late logit fusion:

$$o_{i,j} = (1-w)g_{i,j} + wf_{i,j}, \qquad j \in S_i, \tag{9}$$

where S_i denotes the candidate category set, $g_{i,j}$ denotes the MLLM log-probability score for category j, and $w \in [0,1]$ is the fusion weight.

3.3 Two-Stage Attribute Alignment and Instruction Tuning

Attribute Alignment Fine-Tuning. Inspired by FineDefics (He et al., 2025), we concatenate image, offline-constructed attribute, and category-name embedding sequences, each ending with an `[eos]` token, as input to the LLM. We extract the corresponding final-layer `[eos]` hidden states as v_i, a_i, and c_i, respectively, and align them to enhance the MLLM's fine-grained understanding. The visual–attribute and category–attribute losses are $L_{\text{VA}} = \frac{1}{2}\sum_{i\in B}[\mathrm{Align}(v_i, a_i) + \mathrm{Align}(a_i, v_i)]$ and $L_{\text{CA}} = \frac{1}{2}\sum_{i\in B}[\mathrm{Align}(c_i, a_i) + \mathrm{Align}(a_i, c_i)]$, respectively. The alignment function is defined as

$$\mathrm{Align}(v_i, a_i) = -\log\frac{\exp(\mathrm{Sim}(v_i, a_i))}{\sum_{j\in B}\exp(\mathrm{Sim}(v_j, a_i))}, \tag{10}$$

where B denotes the training-batch index set, and Sim denotes temperature-scaled cosine similarity.

Instruction Tuning for FGFR. Following attribute alignment fine-tuning, we fine-tune the MLLM to translate attribute understanding into fine-grained category prediction. We construct instructions from FGFR training data as open-set category questions and closed-set multiple-choice questions.

4 Experiments

4.1 Implementation Details

Datasets. To address inconsistent category granularity, alias conflicts, and image noise, we construct FoodFG from seven food datasets (Bossard et al., 2014; Chen & Ngo, 2016; Min et al., 2019; 2023; 2020; Kawano & Yanai, 2015; Hou et al., 2017). FoodFG contains 284,935 images from 1,622 categories (details in Appendix A.1). Following Geigle et al. (2024), we evaluate FGFR using four-way multiple-choice questions (MCQs) with hard negatives retrieved through OpenCLIP (Cherti et al., 2023). We also introduce prototype-based MCQs that select visually confusable categories using class prototypes. To prevent evaluation leakage, we construct representative image sets R_y, category-level ingredient priors P_y, and class prototypes from training data only.

Experimental Settings. All CIPSR agents are implemented using GPT-5 (OpenAI, 2025). We build GFRFood upon Qwen3-VL (Bai et al., 2025) and first train TC-S3M offline using its frozen visual features. 4-bit QLoRA (Dettmers et al., 2023) with rank $\gamma = 8$ and scaling factor $\alpha = 8$ is used for parameter-efficient two-stage fine-tuning (Section 3.3). We use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 2×10^{-4} and an effective batch size of 128. The depth weights ω_l and fusion weight w are set to $\{0.3, 0.3, 0.4\}$ and 0.3, respectively.

Evaluation Metrics. Following Liu et al. (2025), we evaluate GFRFood using four cyclic option orders to assess robustness to option positions, and report relaxed accuracy (R-Acc.), rotation accuracy (Rot-Acc.), and rotation hit rate (Rot-Hit) (details in Appendix A.3).

Table 1: Comparison with representative MLLMs on FoodFG four-way MCQs (%). #P denotes the parameter count; † denotes instruction tuning on open-set QA and closed-set MCQs.

		Text-based MCQs			Prototype-based MCQs		
Model	#P	R-Acc.	Rot-Acc.	Rot-Hit	R-Acc.	Rot-Acc.	Rot-Hit
LLaVA-1.5	7B	26.43	0.79	26.53	27.41	1.06	27.32
PaliGemma2-pt-224	10B	27.34	0.01	27.36	28.07	0.03	28.17
InstructBLIP-Vicuna	7B	30.80	10.77	30.88	38.57	15.50	38.46
Molmo2-O	7B	35.12	23.00	35.13	46.10	32.68	45.97
InternVL3.5	8B	35.92	25.78	36.06	37.67	27.85	39.03
InternVL2.5	8B	39.14	30.84	39.06	45.25	36.50	45.14
Idefics3-Llama3	8B	40.26	28.30	40.31	48.78	35.77	48.67
Qwen2.5-VL	7B	49.56	31.19	49.45	51.79	33.62	51.65
GLM-4.1V-Thinking	9B	53.10	44.07	52.59	54.29	46.32	54.64
Qwen3-VL	8B	57.21	49.40	57.11	59.27	51.50	59.15
FoodLMM†	7B	57.74	45.01	57.82	58.82	47.20	58.96
Qwen3-VL†	8B	70.75	70.06	70.82	71.93	71.68	72.05
FineDefics†	8B	70.89	70.12	70.93	72.08	71.75	72.17
Fine-R1†	7B	71.25	71.13	71.41	73.10	73.07	73.28
GFRFood (ours)	**8B**	**90.24**	**89.72**	**90.23**	**92.18**	**91.96**	**92.11**

4.2 Main Results

Table 1 compares GFRFood with representative MLLMs on FoodFG using text-based and prototype-based MCQs. GFRFood outperforms Qwen3-VL in R-Acc. by 33.03% and 32.91%, consistently surpassing mainstream MLLMs and achieving state-of-the-art performance under both protocols. Among food-specific and fine-grained MLLMs, GFRFood outperforms FoodLMM† in R-Acc. by 32.50% and 33.36%, and Fine-R1† by 18.99% and 19.08%, respectively. GFRFood achieves 89.72% and 91.96% in Rot-Acc., close to Rot-Hit under both protocols. LLaVA-1.5 and PaliGemma2 achieve at most 1.06% in Rot-Acc. These results demonstrate GFRFood's robustness to option positions.

4.3 Ablation Studies

Table 2: Effects of two-stage fine-tuning (%).

Strategy	R-Acc.	Rot-Acc.	Rot-Hit
Baseline	57.21	49.40	57.11
Attribute-only	59.58	50.59	59.37
I.T.-only	70.75	70.06	70.82
Attribute + I.T.	**78.31**	**76.65**	**78.32**

Table 3: Effects of attribute generation (%).

Method	R-Acc.	Rot-Acc.	Rot-Hit
Baseline + I.T.	70.75	70.06	70.82
Direct	70.56	69.68	70.62
FineDefics	70.89	70.12	70.93
CIPSR	**78.31**	**76.65**	**78.32**

Do Attribute Alignment Fine-Tuning and Instruction Tuning Provide Complementary Gains in FGFR? Table 2 shows the ablation results of attribute alignment fine-tuning and instruction tuning (I.T.) on text-based MCQs, used for all FGFR ablations. Attribute alignment fine-tuning alone outperforms the baseline in R-Acc. by only 2.37%, while our two-stage fine-tuning achieves 78.31% in R-Acc., outperforming instruction tuning alone in R-Acc. by 7.56%. These results indicate that the two stages are complementary: attribute alignment fine-tuning enhances fine-grained food understanding, while instruction tuning develops question-answering ability.

Does Cognition-Inspired Attribute Generation Improve FGFR Performance? Direct generation uses the prompt *Describe the visual attributes of the image under* {*class prior*}, whereas FineDefics uses the paradigm of attribute discovery, visual extraction, and summarization. We reproduce the complete FineDefics method with Qwen3-VL replacing its original backbone. All attribute methods use two-stage fine-tuning. In Table 3, Direct generation slightly underperforms instruction tuning alone, while FineDefics provides only a marginal improvement. Our CIPSR outperforms FineDefics

Table 4: Analysis of zero-shot FGFR (%).

Strategy	R-Acc.	Rot-Acc.	Rot-Hit
Baseline	57.74	49.94	57.77
I.T.-only	69.38	66.49	69.15
FineDefics	68.85	65.92	68.62
CIPSR	**74.10**	**70.86**	**73.98**

Table 5: Effects of CIPSR and TC-S3M (%).

Configuration	R-Acc.	Rot-Acc.	Rot-Hit
Baseline	57.21	49.40	57.11
TC-S3M	66.89	64.24	66.75
CIPSR	78.31	76.65	78.32
GFRFood	**90.24**	**89.72**	**90.23**

in R-Acc. by 7.42%, as cognition-inspired *glancing–focusing–reasoning* captures discriminative food semantics. Appendix A.4.1 analyzes the effects of key agents in CIPSR.

Does Fine-Tuning with CIPSR Attributes Improve Zero-Shot FGFR Performance? We split the categories into disjoint training and evaluation sets at a 7:3 ratio, excluding all images, CIPSR attributes, and category-level ingredient priors from the held-out categories during both attribute alignment fine-tuning and instruction tuning. This experiment evaluates the MLLM without TC-S3M fusion. All query images and candidate categories are drawn from the held-out set. In Table 4, CIPSR outperforms instruction tuning alone and FineDefics in R-Acc. by 4.72% and 5.25%, respectively, indicating that CIPSR improves generalization to held-out food categories.

Are Visual Semantic Modeling and Taxonomy Modeling Complementary for FGFR? In Table 5, TC-S3M fuses its prediction logits with those of baseline Qwen3-VL without instruction tuning. GFRFood achieves 90.24% in R-Acc. by fusing TC-S3M with Qwen3-VL after two-stage fine-tuning, indicating that food semantics learned with CIPSR attributes complement taxonomy cues modeled by TC-S3M. Appendix A.4.2 analyzes the effects of TC-S3M components.

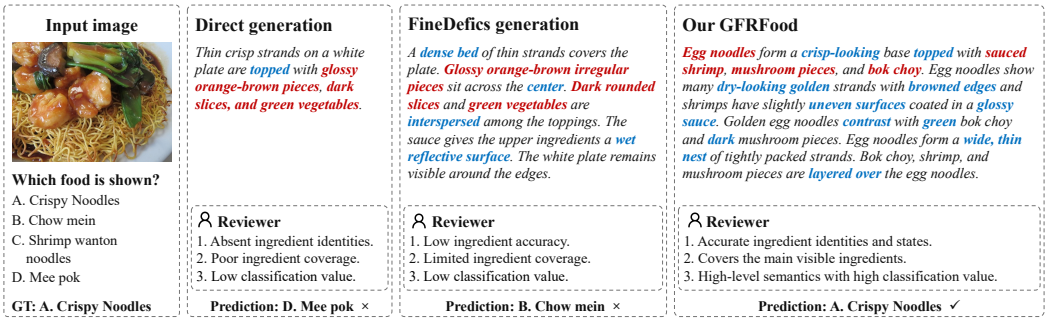

Figure 4: Visualization of existing attribute generation methods and our CIPSR. Red bold text marks ingredient references; blue bold text marks visual and relational semantics.

Qualitative Comparison of Food Attributes. To assess the advantages of CIPSR attributes, we qualitatively compare the generated descriptions in Figure 4. Direct generation produces generic ingredient descriptions, while FineDefics misses shrimp. CIPSR captures the ingredients with their shapes, colors, and textures, and identifies shrimp, mushrooms, and bok choy as toppings over a crisp egg-noodle base. Improvements in ingredient accuracy, ingredient coverage, and classification value enhance the MLLM's FGFR capability. Additional visualizations are provided in Appendix A.5.

5 Conclusion

In this paper, our objective is to address the three food-specific challenges faced by MLLMs in FGFR. Drawing on human food cognition, we decompose food understanding into *glancing*, *focusing*, and *reasoning*, and propose GFRFood, a cognition-inspired MLLM framework with two complementary modules. CIPSR constructs discriminative food attributes through progressive multi-agent reasoning over ingredient-level evidence and high-level semantics. TC-S3M models taxonomy semantics through selective state updates to refine category understanding from coarse to fine. Experiments on FoodFG demonstrate that GFRFood outperforms the evaluated MLLMs, while ablation studies verify the effectiveness of both modules. Future work will explore multi-agent attribute generation that more closely mirrors human cognition and extend TC-S3M to open-world food taxonomies.

AI USE STATEMENT

We used generative AI tools to edit parts of the manuscript, assist with translation, and revise scientific figures. We also used GPT-5 within CIPSR to generate synthetic annotations for existing food images, including ingredient annotations and visual and semantic attribute descriptions. The generated attributes were used as supervision for model training. The annotation generation and verification procedures are described in Section 3.1. We also used GPT-based comments to illustrate qualitative comparisons of generated attributes. We have not used generative AI tools to formulate the mathematical claims or proofs, design the method or experiments, implement the method, or interpret the results. All AI-assisted edits were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

M. Bar, K. S. Kassam, A. S. Ghuman, J. Boshyan, A. M. Schmid, A. M. Dale, M. S. Hämäläinen, K. Marinkovic, D. L. Schacter, B. R. Rosen, and E. Halgren. Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, 103(2):449–454, 2006. doi: 10.1073/pnas.0507062103.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV 2014)*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29. URL https://link.springer.com/chapter/10.1007/978-3-319-10599-4_29.

Jingjing Chen and Chong-Wah Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*, pp. 32–41. ACM, 2016. doi: 10.1145/2964284.2964315. URL https://dl.acm.org/doi/10.1145/2964284.2964315.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198. IEEE, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.html.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00276. URL https://openaccess.thecvf.com/content/CVPR2023/html/Cherti_Reproducible_Scaling_Laws_for_Contrastive_Language-Image_Learning_CVPR_2023_paper.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Damion M. Dooley, Emma J. Griffiths, Gurinder S. Gosal, Pier L. Buttigieg, Robert Hoehndorf, Matthew C. Lange, Lynn M. Schriml, Fiona S. L. Brinkman, and William W. L. Hsiao. FoodOn: A harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2(1):23, 2018. doi: 10.1038/s41538-018-0032-6. URL https://www.nature.com/articles/s41538-018-0032-6.

Jonatan Fridolfsson, Emma Sjöberg, Meri Thiwång, and Stefan Pettersson. Performance evaluation of 3 large language models for nutritional content estimation from food images. *Current Developments in Nutrition*, 9(10):107556, 2025. doi: 10.1016/j.cdnut.2025.107556. URL https://www.sciencedirect.com/science/article/pii/S2475299125030185.

Gregor Geigle, Radu Timofte, and Goran Glavaš. African or European swallow? benchmarking large vision-language models for fine-grained object classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2653–2669. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.154. URL https://aclanthology.org/2024.emnlp-main.154/.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *Proceedings of the International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=p3NKpom1VL.

Hulingxiao He, Zijun Geng, and Yuxin Peng. Fine-R1: Make multi-modal LLMs excel in fine-grained visual recognition by chain-of-thought reasoning. In *Proceedings of the International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=kyzHM557gE.

Hulingxiao He, Zhi Tan, and Yuxin Peng. Taxonomy-aware representation alignment for hierarchical visual recognition with large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31124–31134. IEEE, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/He_Taxonomy-Aware_Representation_Alignment_for_Hierarchical_Visual_Recognition_with_Large_Multimodal_CVPR_2026_paper.html.

Saihui Hou, Yushan Feng, and Zilei Wang. VegFru: A domain-specific dataset for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 541–549. IEEE, 2017. doi: 10.1109/ICCV.2017.66. URL https://openaccess.thecvf.com/content_iccv_2017/html/Hou_VegFru_A_Domain-Specific_ICCV_2017_paper.html.

Yue Hu and Guohang Zhuang. MultiFoodhat: A potential new paradigm for intelligent food quality inspection. *arXiv preprint arXiv:2510.13889*, 2025. URL https://arxiv.org/abs/2510.13889.

J. D. Ireland and A. Møller. LanguaL food description: A learning process. *European Journal of Clinical Nutrition*, 64(Suppl. 3):S44–S48, 2010. doi: 10.1038/ejcn.2010.209. URL https://www.nature.com/articles/ejcn2010209.

Pengkun Jiao, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. From holistic to localized: Local enhanced adapters for efficient visual instruction fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2728–2737. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00262. URL https://openaccess.thecvf.com/content/ICCV2025/html/Jiao_From_Holistic_to_Localized_Local_Enhanced_Adapters_for_Efficient_Visual_ICCV_2025_paper.html.

Pengkun Jiao, Xinlan Wu, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. RoDE: Linear rectified mixture of diverse experts for food large multi-modal models. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 2457–2466. ACM, 2026. doi:

10.1145/3805622.3810616. URL https://dl.acm.org/doi/10.1145/3805622.3810616.

Yoshiyuki Kawano and Keiji Yanai. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *Proceedings of the European Conference on Computer Vision Workshops (ECCV 2014)*, pp. 3–17. Springer, 2015. doi: 10.1007/978-3-319-16199-0_1. URL https://link.springer.com/chapter/10.1007/978-3-319-16199-0_1.

Jeonghwan Kim and Heng Ji. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6187–6207. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.356. URL https://aclanthology.org/2024.emnlp-main.356/.

Junho Kim, Yeonju Kim, and Yong Man Ro. What if...?: Thinking counterfactual keywords helps to mitigate hallucination in large multi-modal models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10672–10689. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.626. URL https://aclanthology.org/2024.findings-emnlp.626/.

Woojin Lee, Pranav Mekkoth, Ye Tian, Onat Gungor, and Tajana Rosing. FoodCHA: Multi-modal LLM agent for fine-grained food analysis. *arXiv preprint arXiv:2605.05499*, 2026. URL https://arxiv.org/abs/2605.05499.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. DyFo: A training-free dynamic focus visual search for enhancing LMMs in fine-grained visual understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9098–9108. IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_DyFo_A_Training-Free_Dynamic_Focus_Visual_Search_for_Enhancing_LMMs_CVPR_2025_paper.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV 2024)*, pp. 216–233. Springer, 2025. doi: 10.1007/978-3-031-72658-3_13. URL https://link.springer.com/chapter/10.1007/978-3-031-72658-3_13.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu. Visual aware hierarchy based food recognition. In *Proceedings of the International Conference on Pattern Recognition Workshops and Challenges*, pp. 571–598. Springer, 2021. doi: 10.1007/978-3-030-68821-9_47. URL https://link.springer.com/chapter/10.1007/978-3-030-68821-9_47.

Weiqing Min, Linhu Liu, Zhengdong Luo, and Shuqiang Jiang. Ingredient-guided cascaded multi-attention network for food recognition. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1331–1339. ACM, 2019. doi: 10.1145/3343031.3350948. URL https://dl.acm.org/doi/10.1145/3343031.3350948.

Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. ISIA Food-500: A dataset for large-scale food recognition via stacked global-local attention network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 393–401. ACM, 2020. doi: 10.1145/3394171.3414031. URL https://dl.acm.org/doi/10.1145/3394171.3414031.

Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9932–9949, 2023. doi: 10.1109/TPAMI.2023.3237871. URL https://ieeexplore.ieee.org/document/10019590.

Fnu Mohbat and Mohammed J. Zaki. LLaVA-Chef: A multi-modal generative model for food recipes. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1711–1721. ACM, 2024. doi: 10.1145/3627673.3679562. URL https://dl.acm.org/doi/10.1145/3627673.3679562.

OpenAI. GPT-5 model. OpenAI API Documentation, 2025. URL https://developers.openai.com/api/docs/models/gpt-5.

Philippe G. Schyns and Aude Oliva. From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, 5(4):195–200, 1994. doi: 10.1111/j.1467-9280.1994.tb00500.x.

Yosua Setyawan Soekamto, Andreas Lim, Leonard Christopher Limanjaya, Yoshua Kaleb Purwanto, Suk-Ho Lee, and Dae-Ki Kang. Pic2Plate: A vision-language and retrieval-augmented framework for personalized recipe recommendations. *Sensors*, 25(2):449, 2025. doi: 10.3390/s25020449. URL https://www.mdpi.com/1424-8220/25/2/449.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00914. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tong_Eyes_Wide_Shut_Exploring_the_Visual_Shortcomings_of_Multimodal_LLMs_CVPR_2024_paper.html.

Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786, 2006. doi: 10.1037/0033-295X.113.4.766.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL https://arxiv.org/abs/2409.12191.

Runze Yan, Hanqi Luo, Jiaying Lu, Darren Liu, Hannah Posluszny, Mehak Preet Dhaliwal, Janice MacLeod, Yao Qin, Carl Yang, Terry J. Hartman, and Xiao Hu. DietAI24 as a framework for comprehensive nutrition estimation using multimodal large language models. *Communications Medicine*, 5(1):458, 2025. doi: 10.1038/s43856-025-01159-0. URL https://www.nature.com/articles/s43856-025-01159-0.

Zhongqi Yang, Elahe Khatibi, Nitish Nagesh, Mahyar Abbasian, Iman Azimi, Ramesh Jain, and Amir M. Rahmani. ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. *Smart Health*, 32:100465, 2024. doi: 10.1016/j.smhl.2024.100465. URL https://www.sciencedirect.com/science/article/pii/S2352648324000217.

Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. FoodLMM: A versatile food assistant using large multi-modal model. *IEEE Transactions on Multimedia*, 27:6949–6961, 2025. doi: 10.1109/TMM.2025.3590924. URL https://ieeexplore.ieee.org/document/11105019.

A APPENDIX

A.1 CONSTRUCTION OF FOODFG

Food recognition benchmarks such as Food-101 (Bossard et al., 2014), VireoFood-172 (Chen & Ngo, 2016), ISIA Food-200 (Min et al., 2019), Food2K (Min et al., 2023), ISIA Food-500 (Min et al., 2020), UECFOOD256 (Kawano & Yanai, 2015), and VegFru (Hou et al., 2017) exhibit three limitations for fine-grained food recognition (FGFR): noise in images and class labels, inconsistent category names arising from synonyms and translation variants, and inconsistent category granularity when broad food categories coexist with specific subclasses. Accordingly, we construct FoodFG from these resources with reduced noise, standardized category names, and consistent category granularity.

We standardize category names and fine-grained class boundaries, assigning each fine-grained category a taxonomy path through L1 food clusters, L2 ingredient families, and L3 processing form. Images are screened for potential class-label errors by combining OpenCLIP image–text similarities (Cherti et al., 2023), class prototypes, and neighborhood evidence. The screened images undergo global duplicate and near-duplicate removal using perceptual hashing and feature embeddings before dataset splitting. From the remaining images, quality-guided k-center diversity sampling selects representative samples that capture the appearance variation within each category.

Table S1: Analysis of FoodFG data curation.

Curation stage	Images	Categories	Removed
Multi-source image collection from food recognition benchmarks	1,670,787	2,997	—
Standardize category names and fine-grained class boundaries	1,036,213	1,638	634,574
OpenCLIP image–text similarities, class prototypes, and neighborhood evidence	387,294	1,638	648,919
Perceptual hashing, feature embeddings, and quality-guided k-center sampling	287,149	1,634	100,145
Manual review	284,935	1,622	2,214
Manual audit pass rate: 94.7%			

In Table S1, standardizing category names and fine-grained class boundaries reduces the category inventory from $2{,}997$ to $1{,}638$. Annotation screening retains 37.38% of the input images while preserving all $1{,}638$ categories. Global deduplication and diversity sampling further reduce the image pool by 25.86%. Manual review removes another $2{,}214$ images, yielding $284{,}935$ images across $1{,}622$ fine-grained categories. The retained images constitute 17.05% of the initial image pool. Using seed 42, we randomly sampled $1{,}000$ images from the final benchmark. An author manually assessed image quality and image–label consistency, with 94.7% of the samples passing both criteria.

For supervised fine-tuning, we split the images into training, validation, and test sets in a $6{:}1{:}3$ ratio. Representative image sets R_y, category-level ingredient priors P_y, and class prototypes for recognition are constructed from training images only. For zero-shot evaluation, we split the categories into disjoint training and test sets in a $7{:}3$ ratio. All images, CIPSR attributes, and category-level ingredient priors used in attribute alignment fine-tuning and instruction tuning come exclusively from the training categories.

A.2 RATIONALE FOR THE FOOD TAXONOMY

The food taxonomy draws on FoodOn and LanguaL, which organize foods by source, product type, and processing characteristics (Dooley et al., 2018; Ireland & Møller, 2010). L1 food clusters reflect overall food composition and consumption form. L2 ingredient families distinguish the principal ingredient or product family within each cluster, and L3 processing form distinguishes processing methods and product forms. These levels provide coarse-to-fine category context for TC-S3M to model semantic relations among fine-grained food categories, as illustrated in Figure S1.

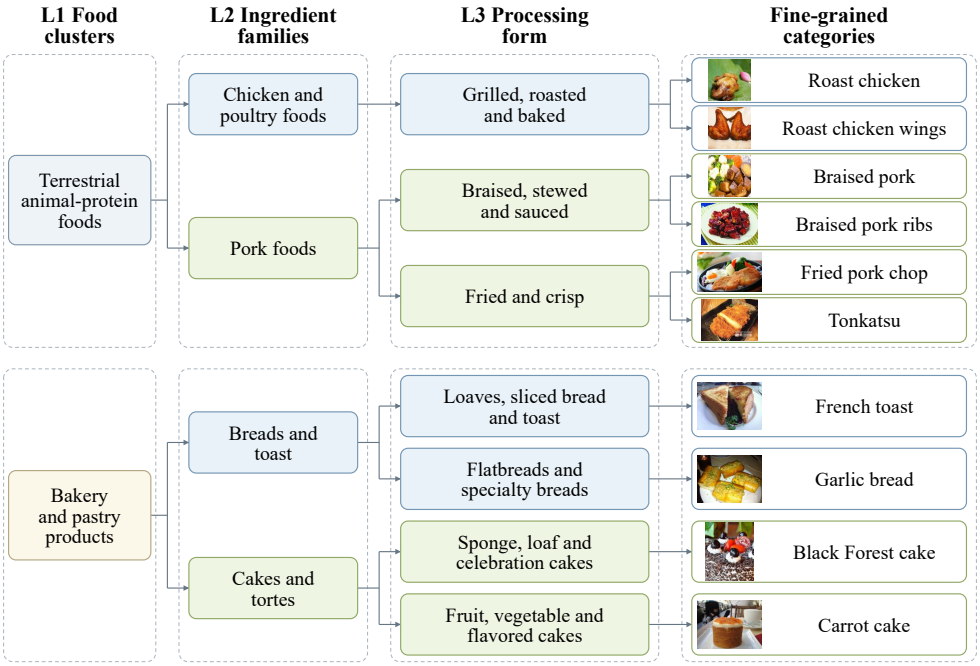

Figure S1: Representative branches of the food taxonomy for terrestrial animal-protein foods and bakery and pastry products.

A.3 Evaluation Metrics

For text-based MCQs and prototype-based MCQs, we evaluate each question under four cyclic option orders while keeping its candidate categories fixed. We measure recognition accuracy and robustness to option positions as:

$$\text{R-Acc.} = \frac{100}{N_e}\sum_{i=1}^{N_e}\delta_{i,0}, \tag{S1}$$

$$\text{Rot-Acc.} = \frac{100}{N_e}\sum_{i=1}^{N_e}\prod_{k=0}^{3}\delta_{i,k}, \tag{S2}$$

$$\text{Rot-Hit} = \frac{100}{4N_e}\sum_{i=1}^{N_e}\sum_{k=0}^{3}\delta_{i,k}, \tag{S3}$$

where N_e denotes the number of evaluated images and $\delta_{i,k} \in \{0,1\}$ indicates whether the answer for image i under cyclic order $k \in \{0,1,2,3\}$ selects the correct category. The original option order corresponds to $k = 0$. R-Acc. measures accuracy under the original order. Rot-Acc. measures the proportion of images answered correctly under all four orders, whereas Rot-Hit averages accuracy across the four orders.

A.4 More Ablation Studies

A.4.1 Effects of Key Agents in CIPSR

Gold-standard annotations. We construct a gold-standard dataset of 4,866 images, with three images from each fine-grained category. Experts annotate ingredient identities, visible states, and bounding boxes. The annotations undergo spot checks to ensure accuracy. Precision, recall, Macro F1, and Micro F1 assess multi-label ingredient-name classification, without scoring bounding boxes or physical states. Precision and recall are computed per image and averaged across images. Macro F1 and Micro F1 use class-wise and global aggregation, respectively.

Agent contributions. In Table S2, prior-guided **Extractor** achieves 73.56% in precision and 91.27% in recall by incorporating category-level ingredient priors from Glancing, outperforming **Extractor** without Glancing in these metrics by 9.04% and 46.52%, respectively. **Critic** uses counterfactual verification to revise ingredient identities or states when alternatives have stronger local visual support and to remove unsupported candidates. Adding **Critic** to prior-guided **Extractor** improves precision, Macro F1, and Micro F1 by 10.27%, 5.67%, and 5.38%, respectively, but reduces recall by 1.68%. Counterfactual verification can nevertheless favor incorrect alternative explanations and reject valid candidates. **Refiner** performs result-oriented reassessment of rejected candidates, blinded to whether they originate from **Extractor** or **Critic**. Adding **Refiner** improves recall, Macro F1, and Micro F1 by 3.67%, 6.18%, and 7.45%, respectively. **Judge** conducts evidence-based adjudication of the rejected candidates and **Refiner**'s reassessments, blinded to reviewer identities. Adding **Judge** improves precision and recall by 4.48% and 1.49%, respectively. Focusing achieves 93.40% in Macro F1 and 93.32% in Micro F1 through multi-agent verification and self-correction.

Table S2: Effects of key CIPSR agents (%) on the gold-standard dataset.

Cumulative stage	Precision	Recall	Macro F1	Micro F1
w/o Glancing **Extractor**	64.52	44.75	57.51	55.00
Prior-guided **Extractor**	73.56	91.27	79.11	78.24
+ **Critic**	83.83	89.59	84.78	83.62
+ **Refiner**	87.94	93.26	90.96	91.07
+ **Judge**	**92.42**	**94.75**	**93.40**	**93.32**

Reasoning effectiveness. To evaluate Reasoning in CIPSR, we concatenate the ingredient names identified by Focusing into text as the baseline attributes. In Table S3, adding discriminative cues to the concatenated ingredient text improves R-Acc., Rot-Acc., and Rot-Hit by 2.23%, 2.06%, and 2.09%, respectively. Using full six-dimensional attributes further improves these metrics by 3.56%, 4.05%, and 3.55%, respectively. These attributes encompass ingredient identities and states while integrating low-level visual semantics, such as texture, shape, and size, with high-level semantics of cooking methods and ingredient relations, enhancing the MLLM's FGFR capability.

Table S3: Effects of Reasoning in CIPSR on FGFR (%).

Variant	R-Acc.	Rot-Acc.	Rot-Hit
Ingredient concatenation	72.52	70.54	72.68
+ discriminative cues	74.75	72.60	74.77
Full six-dimensional attributes	**78.31**	**76.65**	**78.32**

A.4.2 Effects of TC-S3M Components

We verify the effectiveness of key components in TC-S3M. In Table S4, removing the depth and parent-node embeddings reduces L1, L2, and L3 accuracy by 5.98%, 9.71%, and 13.78%, respectively. The larger drops at deeper levels highlight the need to condition state updates on category granularity and ancestor taxonomy semantics. Removing the residual shortcut reduces L1, L2, and L3 accuracy by 2.67%, 4.30%, and 6.08%, respectively. The residual shortcut enables successive S3M layers to refine taxonomy semantics while retaining visual cues that distinguish closely related categories.

A.4.3 Comparison of Taxonomy Modeling Methods

TC-S3M uses two stacked S3M layers across three taxonomy levels and is optimized with $L_{\text{path}} + L_{\text{fine}}$. To evaluate its taxonomy modeling, we replace the S3M layers with a three-layer MLP or a three-layer Transformer. The resulting taxonomy features are combined with visual features for fine-grained classification before late logit fusion. All three modules use a hidden dimension of 512 and are combined with the same fine-tuned MLLM.

In Table S5, TC-S3M outperforms the three-layer MLP in R-Acc., Rot-Acc., and Rot-Hit by 5.83%, 7.14%, and 5.98%, respectively, while outperforming the three-layer Transformer by 3.64%, 5.58%, and 3.50% on the same metrics.

Table S4: Effects of TC-S3M components on taxonomy accuracy (%).

Variant	L1 Acc.	L2 Acc.	L3 Acc.
w/o depth & parent embeddings	87.50	80.40	72.08
w/o residual shortcut	90.81	85.81	79.78
TC-S3M (ours)	**93.48**	**90.11**	**85.86**

Table S5: Effects of taxonomy semantic modeling methods on FGFR (%).

Method	R-Acc.	Rot-Acc.	Rot-Hit
3-layer MLP	84.41	82.58	84.25
3-layer Transformer	86.60	84.14	86.73
TC-S3M	**90.24**	**89.72**	**90.23**

TC-S3M gains an advantage over the MLP and Transformer alternatives in FGFR by selectively propagating ancestor taxonomy semantics.

A.4.4 Performance on Existing Food Benchmarks

We evaluate GFRFood on Food-101 (Bossard et al., 2014) and VireoFood-172 (Chen & Ngo, 2016) using text-based MCQs with two candidate pools: Within-benchmark selects hard negatives from the evaluated benchmark, whereas Within-FoodFG selects hard negatives from FoodFG. We follow each benchmark's original train/test split and exclude test images included in the FoodFG training set.

Table S6: Comparison of Qwen3-VL and GFRFood (ours) on Food-101 and VireoFood-172 (%). † denotes instruction tuning.

	Food-101			VireoFood-172		
Metric	Qwen3-VL	Qwen3-VL†	**Ours**	Qwen3-VL	Qwen3-VL†	**Ours**
Within-benchmark						
R-Acc.	90.93	93.50	**96.15**	73.20	86.50	**93.65**
Rot-Acc.	88.05	91.08	**95.47**	67.20	81.75	**91.75**
Rot-Hit	90.59	93.47	**96.04**	73.50	86.40	**93.34**
Within-FoodFG						
R-Acc.	67.25	74.75	**79.25**	57.35	69.75	**78.35**
Rot-Acc.	63.67	71.68	**77.94**	48.60	62.00	**75.25**
Rot-Hit	67.41	74.61	**79.46**	56.84	69.64	**78.03**

In Table S6, GFRFood outperforms Qwen3-VL with instruction tuning in R-Acc. by 2.65% on Food-101 and 7.15% on VireoFood-172 with within-benchmark hard negatives. With hard negatives from FoodFG, GFRFood outperforms Qwen3-VL with instruction tuning in R-Acc. and Rot-Acc. by 4.50% and 6.26% on Food-101 and by 8.60% and 13.25% on VireoFood-172, respectively. Hard negatives from FoodFG reduce all three methods' scores, indicating more challenging category distinctions. GFRFood maintains the highest scores under both candidate pools, demonstrating effective transfer of food knowledge from FoodFG to other benchmarks such as Food-101 and VireoFood-172.

A.5 More Qualitative Results

Figures S2 and S3 compare the attribute generation quality of direct generation, FineDefics (He et al., 2025), and GFRFood. GPT-based comments qualitatively assess ingredient accuracy, coverage, and classification value. In these examples, direct generation and FineDefics often describe colors, shapes, and containers without identifying discriminative ingredients. Our CIPSR descriptions identify stuffed tofu and its filling in Hakka tofu, lotus root with black fungus, and pork belly with visible fatty portions, capturing distinctions absent from "pale rectangular pieces" and "browned slices."

CIPSR also captures high-level semantics, including cooking methods, ingredient co-occurrence, and spatial relations. The descriptions characterize pork belly with green and red chili peppers as a stir-fry, associate lotus root with black fungus, carrots, and green pepper, and place flat egg noodles beneath pork liver, pork slices, and minced pork. These relations provide category-specific cues beyond individual ingredient appearance.

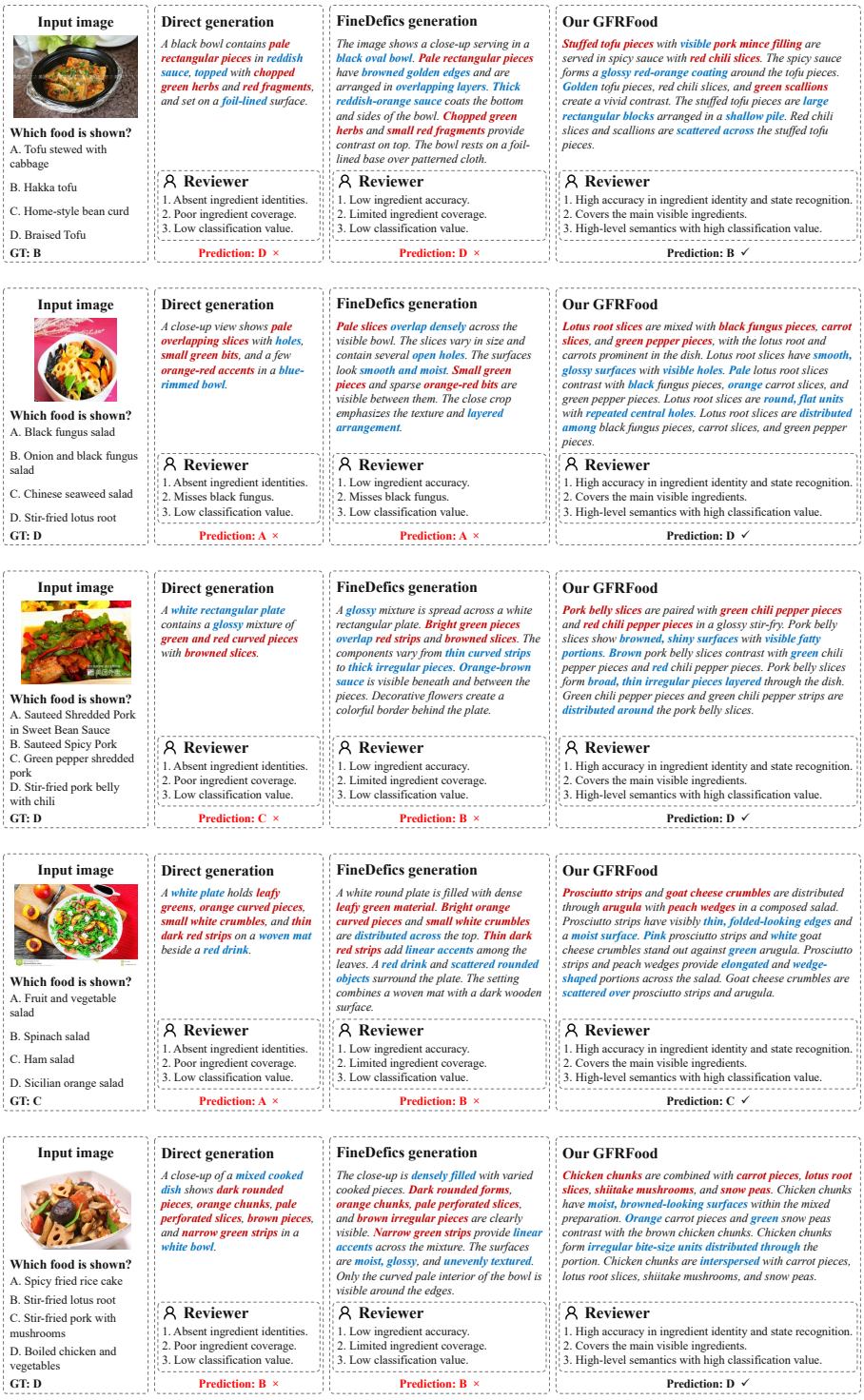

Figure S2: More qualitative comparisons of direct generation, FineDefics, and GFRFood. Each row presents an input image, four candidate categories, generated descriptions, qualitative assessments, and predicted options. Red text highlights ingredient-related expressions, and blue text highlights visual and relational semantics. GT denotes the ground-truth option. AI Review provides qualitative assessments of ingredient accuracy, ingredient coverage, and classification value: whether descriptions correctly identify visible ingredients, cover the visible ingredients, and distinguish candidate categories.

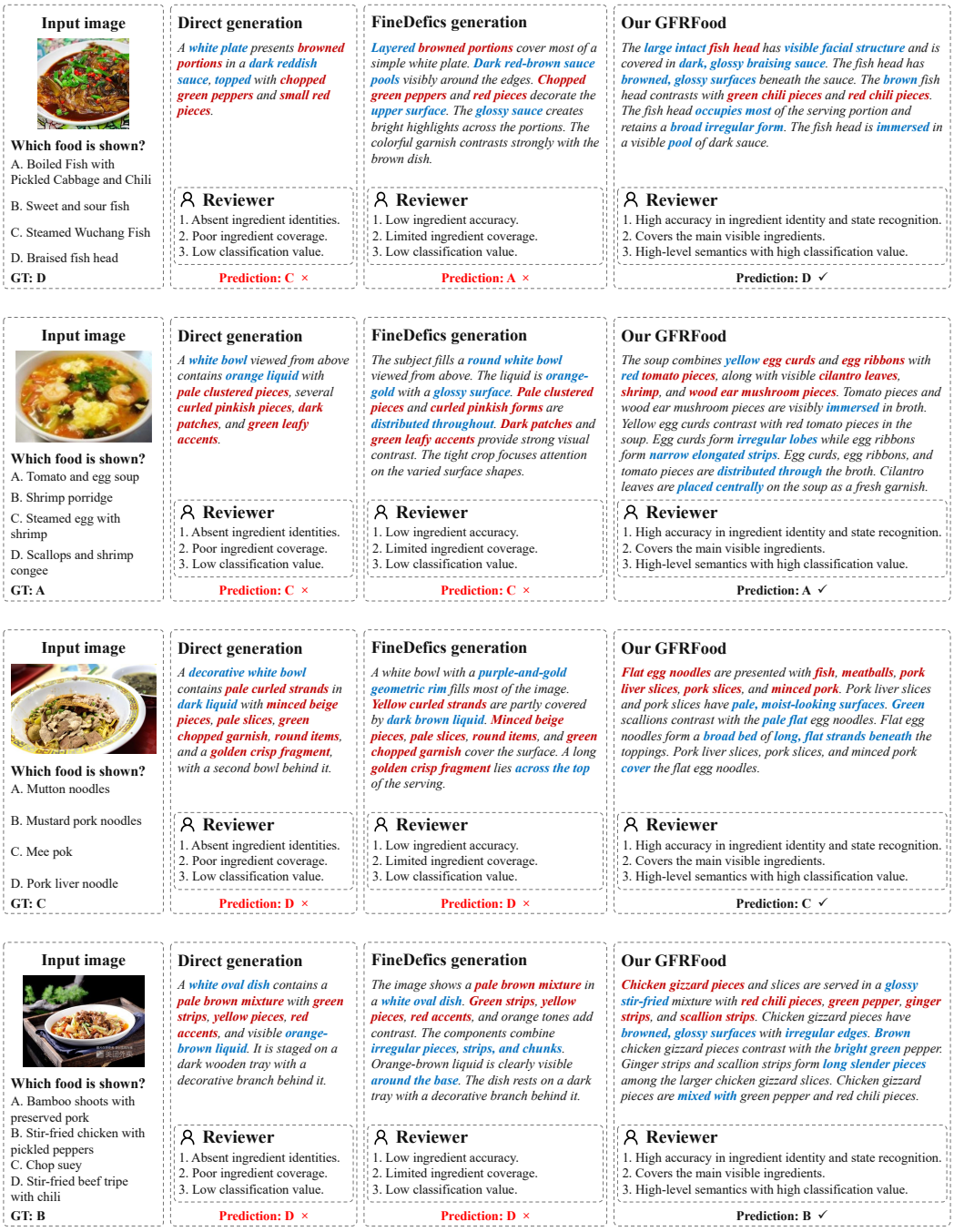

Figure S3: More qualitative comparisons of direct generation, FineDefics, and GFRFood. Colors and symbols follow Figure S2.

A.6 Prompts for Key Agents

We present the prompts for two key CIPSR agents, **Extractor** and **Critic**, which perform prior-guided ingredient recognition and counterfactual verification, respectively.

The prior-guided branch of **Extractor** uses category-level ingredient priors to identify candidate ingredients and their visible physical forms from local visual evidence (Prompt S1).

Prompt S1 — Extractor: Prior-guided ingredient recognition

Role and task.

You are **Extractor**, responsible for prior-guided, closed-set ingredient recognition. Examine every candidate for direct visual support of its identity and level of specificity. When a physical form is specified, verify that it matches the ingredient's visible form after food preparation.

Input.

A food image and candidate ingredients from the category-level ingredient priors. Food category names are not provided.

Instructions.

1. Ground each positive decision in a localizable food object and a distinguishing visual cue. Prioritize contours, cut surfaces, texture, and structural details over color alone. Do not infer ingredient presence from recipes or typical food composition.
2. Inspect the image twice. First, search for each candidate. Then scan the center, quadrants, image and plate edges, small toppings, occluded or low-contrast regions, and spatially separate instances to recover overlooked evidence.
3. Do not reject a visually supported candidate solely because it is small, cropped, occluded, low-contrast, or covered by sauce or broth. Fully blended ingredients without separable visual structure are not visually supported. A negative decision indicates insufficient visual evidence, not absence from the recipe.
4. Compare aliases, broader and finer labels, confusable identities, and alternative forms within the same region. Reject redundant interpretations clearly superseded by a more precise, visually supported candidate, while retaining any candidate with its own direct visual support. Distinct regions or separately visible forms may support multiple candidates.

Output.

Do not infer or output food category names. For each candidate, report whether it is visually supported. For each supported candidate, provide a tight bounding box and a brief description of the distinguishing visual evidence. Enclose only the visible extent of the ingredient, excluding plate, utensils, background, and inferred hidden portions. Keep the true extent of small objects without enlarging their boxes.

Critic compares each ingredient candidate with its strongest counterfactual alternative, revises the ingredient identity or state when the alternative has stronger local visual support, and rejects unsupported candidates (Prompt S2).

Prompt S2 — Critic: Counterfactual verification and correction

Role and task.

You are **Critic**, responsible for counterfactual verification and correction of ingredient candidates. Assume each candidate is incorrect and propose the strongest alternative interpretation of the same image region.

Input.

A food image, candidate ingredient identities and physical forms with bounding boxes, and enlarged views of the candidate regions. Food category names are not provided.

Instructions.

1. Compare each candidate with its strongest alternative using contours, cut surfaces, texture, layers, fibers, and repeated geometry. Distinguish visually similar ingredients and physical forms rather than relying on shared family resemblance.
2. Retain the supplied interpretation only when it has stronger visual support than its alternative. Revise the ingredient identity or visible physical form when the alternative has stronger local visual support. Reject unsupported claims. Evidence for a broader ingredient family alone does not establish a specific identity.
3. Recipe conventions, candidate confidence, color alone, and unrelated food within the bounding box do not establish identity. Do not infer concealed or fully blended ingredients without separable visual structure. Require visible structural evidence for small toppings, while retaining candidates supported by clear form and spatial cues despite low resolution.
4. Assess ingredient interpretation and localization separately. Inspect the candidate region in the full image and enlarged view, and check the rest of the image for a possible localization error. Correct an inaccurate bounding box without rejecting an otherwise supported interpretation. Merge redundant candidates only when they refer to the same visible identity, form, and instance.

Output.

Do not infer or output food category names. For each candidate, report whether its interpretation is retained, revised, or rejected, with the decisive visual evidence and strongest alternative. For any visible target, report the supported identity or broader family, its physical form when visually supported, and a tight bounding box enclosing only its visible extent.