FoodAgent: A Multi-modal Mixture of Experts Reasoning Agent for Divide-and-Conquer Food Nutrition Estimation

Pengfei Zhang, Yutong Song, Chenhan Lyu, Ziyu Wang, Amir M. Rahmani
University of California, Irvine, USA
{pengfz5, yutons12, clyu4, ziyuw31, amirr1}@uci.edu

***Abstract*—Estimating nutrition from food images remains a challenging task, particularly for complex, multi-component dishes. While computer vision methods are effective at recognizing food elements, they typically treat entire meals as monolithic inputs, lacking the ability to decompose visual scenes into individual components. Large language models (LLMs), in contrast, offer strong identification and qualitative reasoning capabilities but struggle with quantitative estimation, especially for assessing volume and mass of individual elements. In this work, we propose FoodAgent, a multi-modal Mixture-of-Experts (MoE) reasoning framework that improves nutrition estimation through a divide-and-conquer strategy. By decomposing dishes into distinct food components, FoodAgent dynamically routes each element to one of three specialized expert modules: (1) monocular volume estimation for nutritionally important and visually clear elements, (2) Retrieval-Augmented Generation (RAG) for important but not clear elements, and (3) direct LLM inference for minor components. This conditional expert selection aligns estimation strategies with the visual and semantic characteristics of each food element, significantly reducing cumulative errors. Experiments show that our element-wise, MoE-driven approach outperforms holistic methods, especially in real-world dietary scenarios involving diverse and complex meals.**

***Index Terms*—Nutrition estimation, Large Language Model, Reasoning Agent, Mixture of Expert, Retrieval Augmented Generation**

I. Introduction

Nutrition estimation from food images has become a critical challenge in digital health [1]–[4], driven by the rise of diet-related health conditions and the growing demand for personalized nutrition management. Traditional manual food logging is time-consuming and prone to low user compliance, creating a need for intelligent systems that can automatically analyze food images and deliver accurate nutritional information.

Existing approaches to automated nutrition estimation typically fall into two paradigms: computer vision-based methods [1]–[3], [5], which use RGB, depth, or volumetric data to detect food elements and estimate nutritional content; and large language model (LLM)-based methods [4], [6], [7], which infer nutrition by leveraging internal visual understanding and external knowledge sources. While both have demonstrated potential, they face key limitations in practical, real-world settings involving complex meals. Computer vision methods often treat the entire dish as a single visual unit, producing a holistic nutritional estimate without decomposing the image into distinct food components. On the other hand, LLM-based approaches like ChatGPT [8], LLAVA [9], and Qwen [10] perform well in food recognition and qualitative reasoning but struggle with quantitative estimation, particularly in measuring the volume or mass of individual components in mixed or ambiguous visual contexts.

This mismatch between food complexity and estimation capability reveals the limitations of applying a uniform strategy to all inputs. In reality, different food elements demand different estimation techniques—a grilled chicken breast may require precise volumetric analysis, while a mixed salad with dressing may benefit more from ingredient-level retrieval or prior knowledge [6], [11]. Treating these heterogeneous elements with a single model introduces compounded errors and undermines estimation reliability.

To address these challenges, we propose FoodAgent, a multi-modal Mixture-of-Experts (MoE) reasoning framework that reimagines nutrition estimation through modularity and conditional routing. Rather than relying on a single monolithic model, FoodAgent decomposes complex dishes into individual food elements and intelligently routes each element to one of several specialized expert modules. Specifically, the agent classifies components into three categories: (1) important and visually clear elements processed through monocular volume estimation; (2) important but not clear elements estimated via Retrieval-Augmented Generation (RAG) [12], [13] from a nutrition database; and (3) minor components handled by direct LLM inference. This selective, expert-driven processing aligns the estimation strategy with the characteristics of each food element, minimizing cumulative errors and improving overall accuracy.

Our contributions are threefold: 1) We introduce a Mixture-of-Experts nutrition estimation framework that overcomes the limitations of holistic approaches in analyzing multi-element dishes. 2) We design a modular reasoning agent that dynamically selects the most suitable expert module for each food element based on its visual and semantic attributes. 3) We empirically demonstrate that this element-wise, MoE-based strategy significantly improves estimation accuracy over traditional methods, especially in real-world dietary scenarios.

II. Method

This section presents FoodAgent, our Mixture-of-Experts (MoE) framework for nutrition estimation from user-input

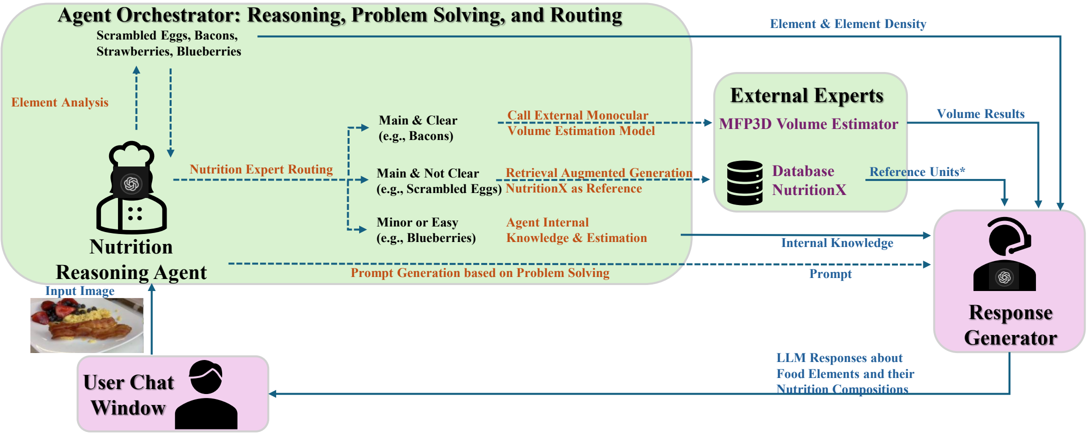


Fig. 1. Framework of FoodAgent. It consists of three core components: (1) external expert modules for specialized processing, (2) a nutrition reasoning agent that performs food decomposition and expert routing, and (3) a response generator that synthesizes the outputs into comprehensive nutrition analysis.

food images. Unlike traditional monolithic approaches, FoodAgent decomposes complex dishes into individual components and selectively routes each to a specialized expert based on its visual and nutritional characteristics. Figure 1 illustrates the overall architecture.

A. External Experts

As part of the Mixture-of-Experts architecture, our system incorporates multiple external expert modules that serve as specialized tools for handling distinct categories of food elements. These modules operate beyond the capacity of large language models alone, bringing in precise estimators and structured knowledge bases to enhance both quantitative and semantic accuracy.

1) MFP3D Volume Estimator: The MFP3D (Monocular Food Portion 3D) [5] volume estimator serves as a dedicated expert for food elements that are both nutritionally important and visually clear. It estimates portion volume from single RGB images by combining depth estimation and segmentation. Specifically, we first use Segment Anything Model (SAM) [14], [15] for identifying food regions of specific element, and then use ZoeDepth [16] for zero-shot depth estimation. The image is masked to isolate the target food element, based on SAM-generated bounding boxes, and paired with a sampled point cloud to form the input to the MFP3D model. The output is a precise volume measurement in cubic centimeters.

2) NutritionX Database for Retrieval Augmented Generation (RAG): The NutritionX database [17] functions as an external knowledge base for RAG, offering detailed nutritional information expressed in practical units (e.g., cups, tablespoons) along with per-unit nutritional profiles, and annotated images of a wide range of food elements and prepared dishes.

B. Nutrition Reasoning Agent

At the core of FoodAgent lies the nutrition reasoning agent, a controller that orchestrates expert selection through automated element-wise analysis and routing. Powered by a large language model (LLM) with strong visual and contextual understanding, Qwen2.5-Omni-7B [10], the agent performs conditional expert assignment—the defining feature of the Mixture-of-Experts framework—by selecting the most appropriate estimation pathway for each identified food element based on its visibility, complexity, and nutritional relevance.

1) Element Analysis and Decomposition: The reasoning process begins with comprehensive image-level analysis. The agent uses its multimodal understanding capabilities to identify all visible food elements within the input image and localize them spatially. It generates detailed per-element descriptions that include appearance, position, rough size, expected density (internal knowledge), and preliminary nutritional significance.

2) Expert Assignment via Three-Category Classification: Based on the decomposition results, each food element is assigned to one of three expert modules in the MoE system. Each expert is designed to handle a specific class of food elements, reflecting different estimation challenges:

Category 1: Important and Clear Elements - Volume Estimator These elements are visually distinct, nutritionally significant, and geometrically regular—e.g., whole meats, bread, fruits. The reasoning agent routes them to the MFP3D volume estimator expert, which uses monocular depth prediction and image segmentation to produce volumetric estimates. These are later converted to mass and nutritional values using known density data.

Category 2: Important but Not Clear Elements - RAG This category includes elements with high nutritional impact but poor visual separability, such as scrambled eggs, mixed pasta, or dressed salads. These are routed to the RAG expert, which queries the NutritionX database to retrieve visually and semantically similar dishes. By grounding its estimation in reference data, this expert compensates for the ambiguity that makes direct measurement unreliable.

Category 3: Minor or Easily Estimated Elements - LLM Inference These are elements with minimal nutritional influence or low complexity, such as herbs, garnishes, or small condiments. The reasoning agent delegates them to a lightweight LLM-based inference expert, relying on the model's internal knowledge to produce approximate estimates with acceptable error margins. This avoids overprocessing negligible elements, preserving computational efficiency.

This dynamic expert routing forms the backbone of FoodAgent's MoE structure, where estimation is not handled uniformly, but selectively tailored to each element's attribute.

3) Output Structuring and Communication with the Response Generator: Once all elements have been routed through their respective experts, the nutrition reasoning agent formats the outputs for downstream processing. For each food element routed to: Category 1 outputs include the volume estimated by the MFP3D expert; Category 2 outputs provide matched NutritionX references, including estimated portion units (e.g., cups, tablespoons) and per-unit nutritional profiles. Category 3 outputs contain coarse nutrition estimates generated directly by the LLM.

Additionally, the agent compiles a unified structured prompt, summarizing all elements, expert routes, and intermediate estimations. This prompt forms the input to the response generator, which synthesizes the final nutritional analysis.

C. Response Generator

The response generator is the final module in FoodAgent. It aggregates outputs from all expert pathways and computes a unified nutritional profile for the entire input dish. Each expert's result is post-processed through standardized conversion and inference rules to ensure consistency and accuracy.

Category 1 (Important and Clear Elements): For food elements with precise volume measurements from the MFP3D estimator and density information from agent knowledge, the response generator calculates mass as:

$$\text{mass} = \text{volume} \times \text{density}, \tag{1}$$

and then computes total nutrition by multiplying the corresponding unit nutritional values.

Category 2 (Important but Not Clear Elements): The response generator estimates the quantity or amount of units present in the image compared to the reference unit, then calculates the total nutrition based on reference unit.

Category 3 (Minor or Easy): For elements with minimal nutritional contribution or low visual complexity, the response generator directly accepts the estimated nutrition values passed from the reasoning agent, which are inferred using internal LLM knowledge.

Finally, the response generator aggregates the nutrition estimates from all three categories to compute the total nutritional profile of the input dish image. This includes caloric content and macronutrient breakdown (carbohydrates, proteins, fats).

III. Experiments

Dataset. We evaluate our FoodAgent framework on the Nutrition5k [1] dataset, which consists of over 5,000 unique real-world dishes captured in campus cafeterias, each annotated with ground-truth mass, calories, and macronutrient composition (fat, carbohydrates, protein) based on per-ingredient measurements and standardized nutritional tables. Each dish includes overhead RGB images, depth data, and segmentation-level ingredient breakdowns, while we only leverage RGB images for training and inference. Dishes range from single-element plates to complex meals with over 30 ingredients, providing a rigorous benchmark for general nutrition estimation. We use the standard data split from Nutrition5k, with 90% of dishes in the training set and 10% in the test set.

TABLE I
Quantitative results on Nutrition5k comparing RGB-based methods, LLM-based models, and our FoodAgent.

	Method	MAE ↓	PMAE ↓
RGB-Based	Nutrition5k Baseline [1]	70.6	26.1%
	MultiLabel Food [3]	59.0	22.6%
	MonoFood [2]	40.1	15.8%
LLM-based	Qwen-omni-2.5 [10]	73.0	28.9%
	LLAVA [9]	82.4	31.7%
	FoodLMM [4]	67.3	26.6%
Ours	**FoodAgent**	**36.9**	**14.7%**
	FoodAgent(w/o MFP3D)	68.3	27.0%
	FoodAgent(w/o NutritionX)	57.5	22.4%

Metrics. We conduct two tasks on Nutrition5k. Nutrition Estimation, which predicts the nutritional values (calories and macronutrient) of the total dish; and Referring Nutrition Estimation, which predicts those values of specific ingredients in the dish. Following previous work, we use mean absolute error (MAE) and the percent of MAE (PMAE) to the respective mean for that field to measure regression accuracy for calories, mass, and individual macronutrient mass. Caloric MAE is measured in kilocalories, all others are measured in grams.

Baselines. We compare our methods to both RGB-based Deep Learning methods [1]–[3] and LLM-based methods [4], [9], [10]. For all the experiments, only RGB dish images serve as input, and the output would be numbers of estimated calories and macronutrient composition.

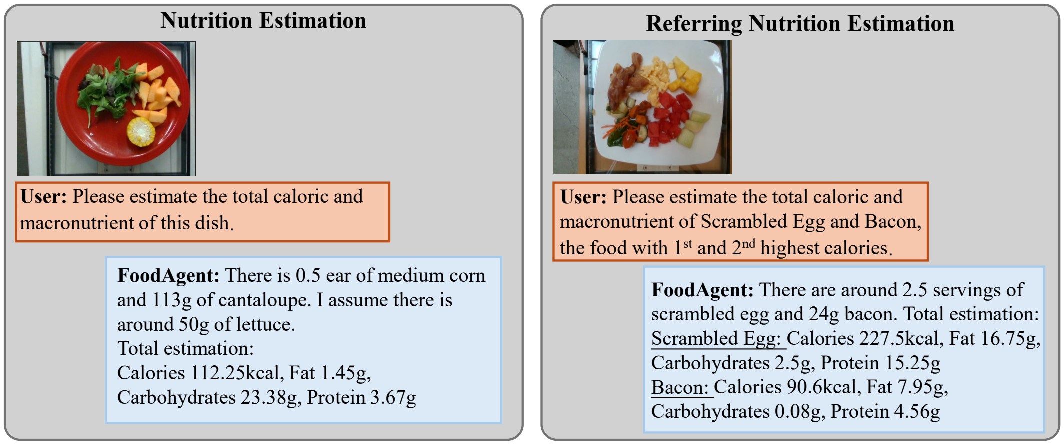
Results. Examples of the two tasks are shown in Figure 2. Table 1 reports the Nutrition Estimation task performance of our proposed FoodAgent model compared to both RGB-based and LLM-based baselines on the Nutrition5k dataset. FoodAgent achieves the best overall performance with a MAE of 36.9 and PMAE of 14.7%, significantly outperforming all baselines. Compared to the Nutrition5k baseline (70.6 MAE, 26.1% PMAE), our model reduces absolute error by more than half. Notably, LLM-based methods like LLAVA and Qwen-Omni-2.5 perform worse ($PMAE > 25\%$), demonstrating their limitations in quantitative tasks like nutrition estimation.

Compared to RGB-based traditional methods, LLM-based methods can execute more complicated tasks such as referring nutrition recognition. Table 2 shows the performance on predicting the nutrition of the first and second referenced ingredients in a dish. FoodAgent consistently outperforms all LLM-based baselines (Qwen-Omni-2.5, LLAVA, FoodLMM) across all nutrition metrics.

Ablation Study. We further validate the contribution of each expert module via ablation experiments, as shown on Table 1. Removing the monocular volume expert (MFP3D+SAM/ZoeDepth) results in a substantial performance drop to 27.0% PMAE, while removing the retrieval-based expert (NutritionX) degrades performance to 22.4% PMAE. These results underscore the importance of modular

TABLE II
REFERRING NUTRITION ESTIMATION RESULTS ON NUTRITION5K FOR THE 1ST AND 2ND CALORIC INGREDIENT IN THE DISH IMAGES.

	Method	Caloric PMAE ↓	Fat PMAE ↓	Carb PMAE ↓	Protein PMAE ↓
Refer 1st Ingredient	Qwen-omni-2.5 [10]	39.1%	41.3%	26.9%	36.5%
	LLAVA [9]	45.8%	43.6%	30.2%	29.5%
	FoodLMM [4]	34.7%	46.1%	34.1%	22.4%
	FoodAgent	**15.1%**	**16.4%**	**12.9%**	**18.0%**
Refer 2nd Ingredient	Qwen-omni-2.5 [10]	35.6%	22.7%	27.9%	45.0%
	LLAVA [9]	45.5%	29.1%	29.8%	33.6%
	FoodLMM [4]	27.4%	29.0%	30.5%	27.4%
	FoodAgent	**16.2%**	**17.8%**	**14.0%**	**20.2%**



User: Please estimate the total caloric and macronutrient of this dish.

FoodAgent: There is 0.5 ear of medium corn and 113g of cantaloupe. I assume there is around 50g of lettuce.
Total estimation:
Calories 112.25kcal, Fat 1.45g, Carbohydrates 23.38g, Protein 3.67g

User: Please estimate the total caloric and macronutrient of Scrambled Egg and Bacon, the food with 1st and 2nd highest calories.

FoodAgent: There are around 2.5 servings of scrambled egg and 24g bacon. Total estimation:
Scrambled Egg: Calories 227.5kcal, Fat 16.75g, Carbohydrates 2.5g, Protein 15.25g
Bacon: Calories 90.6kcal, Fat 7.95g, Carbohydrates 0.08g, Protein 4.56g

Fig. 2. Qualitative Examples of FoodAgent for nutrition recognition.

expert routing: different food elements benefit significantly from different estimation strategies. Among all food elements, those in category 2 will result in most errors.

In summary, the superior performance of FoodAgent stems from its Mixture-of-Experts strategy, which leverages the complementary strengths of specialized expert modules to address the limitations of monolithic approaches. By intelligently routing each food element to the most suitable expert, the system achieves more accurate and interpretable estimations. Furthermore, our results demonstrate that combining prior knowledge from external sources, with LLM-based reasoning yields significant improvements.

IV. CONCLUSION.

In this work, we proposed FoodAgent, a novel multi-modal Mixture-of-Experts (MoE) framework for accurate nutrition estimation from food images. By decomposing complex dishes into individual components and routing each to a specialized expert module, our system effectively overcomes the limitations of holistic methods. One limitation is the latency due to multi-agent collaboration, rooted at the low speed of LLMs. The adaptability of our approach offer a promising direction for real-world dietary assessment tools.

REFERENCES

[1] Q. Thames, A. Karpur *et al.*, "Nutrition5k: Towards automatic nutritional understanding of generic food," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[2] Z. Shao, G. Vinod, J. He, and F. Zhu, "An end-to-end food portion estimation framework based on shape reconstruction from monocular image," 2023. [Online]. Available: https://arxiv.org/abs/2308.01810

[3] R. Ismail and Z. Yuan, "Food ingredients recognition through multi-label learning," 2022. [Online]. Available: https://arxiv.org/abs/2210.14147

[4] Y. Yin, H. Qi, B. Zhu, J. Chen, Y.-G. Jiang, and C.-W. Ngo, "Foodlmm: A versatile food assistant using large multi-modal model," *arXiv preprint arXiv:2312.14991*, 2023.

[5] J. Ma *et al.*, "Mfp3d: Monocular food portion estimation leveraging 3d point clouds," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 49–62.

[6] M. Abbasian, Z. Yang, E. Khatibi, P. Zhang, N. Nagesh, I. Azimi, R. Jain, and A. M. Rahmani, "Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients," *arXiv preprint arXiv:2402.10153*, 2024.

[7] F. P. W. Lo *et al.*, "Dietary assessment with multimodal chatgpt: A systematic analysis," 2023. [Online]. Available: https://arxiv.org/abs/2312.08592

[8] "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[10] J. Xu, Z. Guo *et al.*, "Qwen2.5-omni technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.20215

[11] J. Wu *et al.*, "Knowledge graph retrieval enhanced language models," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[12] P. Lewis, E. Perez *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[13] Y. Song, C. Lyu *et al.*, "Dementia-plan: An agent-based framework for multi-knowledge graph retrieval-augmented generation in dementia care," 2025. [Online]. Available: https://arxiv.org/abs/2503.20950

[14] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[15] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[16] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.

[17] "Nutritionx database," *https://www.nutritionix.com/*.