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

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

I. Reviewer Zdsw

1. An evaluation of the individual components would have been useful. Although the ablation study gives an indication of the importance of two components (volume expert and retrieval-based expert), this is not apparent at all levels. For example, the segmentation/localization quality with SAM/ZoeDepth was not quantified.

Answer: An additional explanation has been added to the ablation study

2. The Nutrition Reasoning Agent is an essential component of the framework but is only explained and validated to a limited depth. In particular, the decision regarding which path should be taken for further calculation is very unclear. An evaluation would have been useful here to assess how appropriate the decision for a particular path was.

Answer: The agent's routing decision is based on heuristics combining visibility, separability, and nutritional importance. For example, geometrically regular and visually clear items (e.g., bread slices, fruits) are routed to MFP3D, while amorphous foods (e.g., scrambled eggs) are routed to NutritionX. This decision process is validated indirectly in the ablation study: when MFP3D or NutritionX are removed, performance drops substantially (Table I), indicating the routing choices are effective."

3. There is a fundamental lack of explanation of the limitations in this work. For example, there is no discussion of how robust and generalizable this method is. No other settings are tested (size and type of plates and environmental parameters such as lighting conditions), nor is it clear from the text or the example images in Figure 2 when this method fails. Please discuss failure cases like occlusions, sauces and extremely mixed foods. A more difficult example in Figure 2 might also be helpful in identifying the limitations.

Answer: add limitation on Method - Category 2, and Fig 2 can also indicates some errors. Also, add limitation discussions on conclusion.

4. There are a few spelling mistakes in Figure 2 and in the FoodAgent's response (e.g., letucce, Scrabled egg). Is this a mistake by the author, or is the agent making these mistakes? If it is the former, please correct it.

Answer: Fix all typos.

II. Reviewer cUen

The limitations of our work have been discussed at Section-Discussion, and some important references have been added.

III. Reviewer Hahp

1. Limited Dataset Evaluation: The evaluation is conducted solely on the Nutrition5k dataset. Given the diversity of global cuisines and food presentation styles, evaluation on additional datasets would strengthen the generalizability claims.

Answer: We focus on Nutrition5k as it is currently the largest and most standardized benchmark with paired RGB images and ground-truth nutrition labels. Other benchmarks are either too simple food so our agent framework is not necessary, or missing some labels. While evaluation on additional datasets (e.g., Recipe1M, Food2K) would be valuable for assessing cross-cultural generalization, this remains future work due to differences in available annotations.

2. Error Analysis Insufficient: While overall performance metrics are provided, the paper would benefit from deeper error analysis examining which types of dishes or food components contribute most to remaining errors, and under what conditions the system fails.

Answer: Add discussions on Method and Ablation Study.

3. User Study Absence: Given the practical application focus, the lack of user studies or real-world deployment evaluation limits understanding of practical utility and user acceptance.

Answer: As our focus is still on nutrition recognition precision, it's tough to ask user to evaluate the performance, as it's not pure generative work.

4. A major issue that is overlooked is how do we know if the QWen LLM (or any LLM) model does not have this dataset (and any other publicly available dataset) within its training? This would significantly bias the result of such a system's generalizability beyond internet available datasets.

Answer: Data contamination is hard to avoid during the LLM pretraining. However, based on QwenLLM official documents, we don't see any proof that Nutrition5K is utilized. Also, Nutrition5K doesn't fit the LLM pretraining requirements, as it doesn't provide sufficient text labels.