

Supp: An Agentic Vision–Language Pipeline for Interactive Nutritional Estimation from Food Images

Gayatri Bhatambarekar
Virginia Tech
Blacksburg, VA, USA
gayatrimilind@vt.edu

Abhijit Sarkar
Virginia Tech
Blacksburg, VA, USA
asarkar1@vt.edu

Abstract

This document is a supplemental document to the CVPRW submission: An Agentic Vision–Language Pipeline for Interactive Nutritional Estimation from Food Images. Note that the main results are discussed in the main paper. The reader should consult the main paper for primary findings and contributions.

1. Introduction

In our proposed pipeline, the nutritional estimation process is grounded in a structured reasoning framework. This framework is implemented via a Chain-of-Thought (CoT) prompting strategy for static analysis and an Agentic mode where the model actively seeks missing information.

While the main paper details the quantitative performance and demonstrates how the model autonomously generates clarifying questions in the Agentic mode to resolve visual ambiguities, this supplemental document outlines the foundational logic behind our approach. Specifically, to ensure full reproducibility, we provide the exact CoT prompt templates used for our evaluations, along with the detailed reasoning for choosing the six core questions that drive both reasoning modes. These questions were deliberately selected to mirror the diagnostic process of human nutritionists:

1. **What foods are present in the image?** Identifying the foods or dishes correctly is foundational. Accurate recognition critically impacts all subsequent nutritional estimations [1]. Misidentifying a food item drastically skews the calorie estimates; hence, explicit verification at this stage is essential.
2. **What ingredients or components does each food contain?** Knowing exact ingredients is crucial, as caloric and macronutrient content directly correlate with specific components and preparation methods [1]. Accurate ingredient identification significantly improves calo-

rie prediction accuracy, especially in mixed dishes.

3. **How much of each food is there (portion size)?** Portion size estimation is the largest source of calorie estimation error [2, 3, 5]. Without proper volume or portion size assessment, errors compound rapidly. Explicitly prompting for portion sizes aligns closely with how nutrition professionals reduce estimation uncertainty.
4. **How was the food prepared, and are there any high-calorie add-ons (oils, sauces, etc.)?** Preparation methods and hidden ingredients drastically affect caloric content. For example, fried foods significantly exceed the caloric content of their grilled or boiled counterparts [6]. Sauces, dressings, or condiments often represent hidden calories not readily visible in images but critical for accurate estimates.
5. **What are the typical calories and macronutrients per unit for these foods?** Leveraging nutritional databases for known caloric densities ensures grounded and accurate predictions. This step directly mirrors standard nutrition expert workflows, retrieving known values (from USDA or similar databases) and multiplying by portion sizes identified earlier [4].
6. **Given all of the above, what is the total estimated calorie count and macronutrient breakdown for the entire meal?** Finally, the model integrates all previous reasoning, explicitly calculating and summing individual contributions to generate a comprehensive and transparent nutritional estimate. This final aggregation ensures systematic reasoning and facilitates error tracing if discrepancies arise.

2. Ingredient level error analysis

The analysis reveals that errors arise from different factors depending on dataset characteristics. In the Nutrition5k dataset, which contains visually complex mixed plates, the largest errors occur with calorie-dense but visually hidden ingredients such as oils, butter, and salad dressings. These components often melt or absorb into surrounding foods,

making them difficult to detect visually and leading to systematic calorie underestimation. The models also struggle with low-density foods such as cereals or croutons that appear visually bulky but contain little mass, causing incorrect density assumptions.

In contrast, the MetaFood3D dataset exhibits errors primarily due to incomplete ingredient annotations. The models frequently overestimate calories for foods such as fried chicken, processed meats, and seafood dishes because the visual appearance suggests a prepared, fried item while the ground-truth labels often include only the base ingredient (e.g., chicken breast) and omit preparation details such as batter or oil. These observations highlight two key multi-modal failure modes: hidden caloric components in complex real-world meals and incomplete textual annotations describing food preparation. Figures 1 and 2 quantitatively illustrate these failure modes, breaking down the specific ingredient categories where estimation discrepancies most frequently occur across the Nutrition5k and MetaFood3D datasets, respectively.




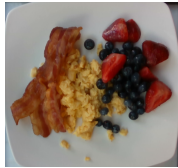

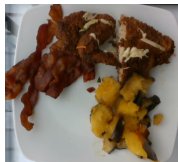

2.1. Qualitative Examples of Agentic Reasoning

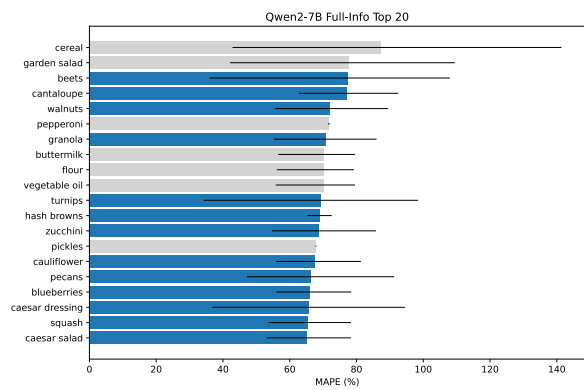
To further illustrate the model’s interactive reasoning capabilities, Table 1 provides qualitative examples of the Agentic Mode in action. For each dish, the model autonomously generates a series of clarifying questions aimed at reducing uncertainty regarding ingredient composition, portion sizes, and hidden preparation methods. By actively seeking these specific visual and contextual cues, the agent is able to ground its nutritional estimates more accurately, as demonstrated by the predicted calorie ranges aligning with the ground truth values.

References

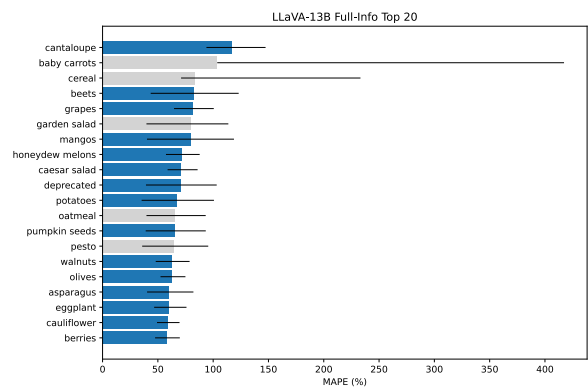
- [1] Takumi Ege and Keiji Yanai. Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 367–375. ACM, 2017. 1
- [2] Sharon I Kirkpatrick, Richard P Troiano, Brian Barrett, Christopher Cunningham, Amy F Subar, Yikyung Park, Heather R Bowles, Laurence S Freedman, Victor Kipnis, Eric B Rimm, et al. Measurement error affecting web-and paper-based dietary assessment instruments: insights from the multi-cohort eating and activity study for understanding reporting error. *American journal of epidemiology*, 191(6): 1125–1139, 2022. 1
- [3] D.A. Lucassen, R.F. Willemsen, A. Geelen, E.M. Brouwer-Brolsma, and E.J.M. Feskens. The accuracy of portion size estimation using food images and textual descriptions of portion sizes: an evaluation study. *Journal of Human Nutrition and Dietetics*, 34(6):945–952, 2021. 1
- [4] Anwar T Merchant and Mahshid Dehghan. Food composition database development for between country comparisons. *Nutrition journal*, 5(1):2, 2006. 1
- [5] Shally Sharma, Vidisha Sharma, Anjani Bakshi, and Ravinder Chadha. Food portion size estimation skill among post graduate dietetics students. *Current Research in Nutrition and Food Science Journal*, 14(1), 2025. 1
- [6] Aman Mohammad Ziaifar, Nawel Achir, Francis Courtois, Isabelle Trezzani, and Gilles Trystram. Review of mechanisms, conditions, and factors involved in the oil uptake phenomenon during the deep-fat frying process. *International Journal of Food Science and Technology*, 43(8):1410–1423, 2008. 1

Table 1. Examples of Agentic Mode questions with predictions and ground truth values.

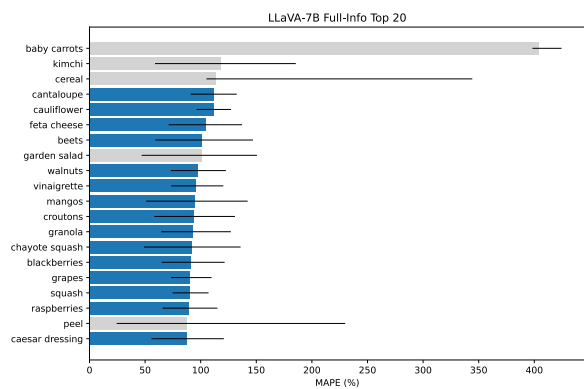
Dish Image	Agent Questions Asked	Predicted Calories	Ground Truth Calories
	<ol style="list-style-type: none"> 1. What type of berries are present on the plate? 2. How many pieces of each type of berry are visible? 3. Is there any additional food or liquid on the plate besides the berries? 4. What is the approximate size of the plate? 5. Are there any other ingredients or items in the background that might affect the calorie count? 6. Is the image taken under natural light or artificial lighting? 	20–30	38.10
	<ol style="list-style-type: none"> 1. What is the type of bacon shown in the image? 2. Is the bacon cooked or raw? 3. How many pieces of bacon are on the plate? 4. What is the size of the plate? 5. Are there any other ingredients or items visible in the image besides the bacon? 6. Is the image taken indoors or outdoors? 	50–70	70.33
	<ol style="list-style-type: none"> 1. What is the cooking method used for the cauliflower? 2. Are there any visible seasonings or spices on the dish? 3. How was the avocado prepared (e.g., sliced, diced)? 4. Is there any dressing or sauce on the salad? 5. Are there any other ingredients or components in the salad besides the greens? 6. What type of greens are used in the salad? 	180–250	204.88
	<ol style="list-style-type: none"> 1. How many strips of bacon are on the plate? 2. What is the exact amount of scrambled eggs visible in the image? 3. Are there any additional ingredients or condiments on the plate? 4. How many strawberries and blueberries are present? 5. Is the plate size standard or oversized? 6. Are there any other items in the background that might affect the calorie count? 	300–400	357.81
	<ol style="list-style-type: none"> 1. What type of tofu is used in the dish? 2. Are there any visible seasonings or spices on the plate? 3. What kind of mushrooms are present in the dish? 4. How are the vegetables (if any) prepared? 5. Is there any dressing or sauce accompanying the dish? 6. Approx how many pieces of each ingredient are on the plate? 	100–150	103.66
	<ol style="list-style-type: none"> 1. What is the main dish on the plate? 2. Are there any visible toppings or garnishes on the meat? 3. What appears to be the type of potatoes used in the side dish? 4. Is there any sauce or dressing accompanying the meal? 5. How many pieces of bacon are there on the plate? 6. Can you describe the texture and color of the meat? 	800–1100	851.21
	<ol style="list-style-type: none"> 1. What is the type of fruit shown in the image? 2. How many pieces of fruit are there on the plate? 3. Is the fruit cut into uniform pieces or irregularly shaped? 4. What is the size of each piece of fruit? 5. Are there any other ingredients or items visible in the image? 6. Is the plate being weighed or displayed as part of a food preparation process? 	30–50	32.30



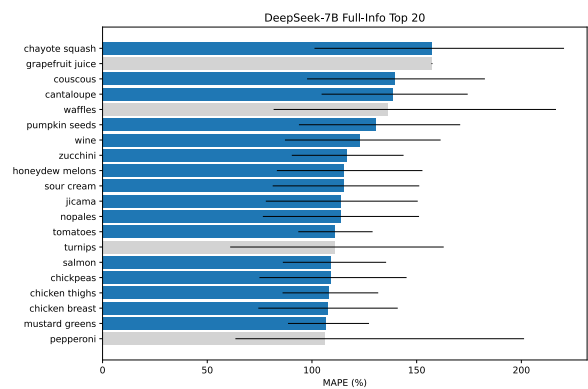
(a) Qwen 2-7B (Full-Info)



(b) LLaVA-13B (Full-Info)

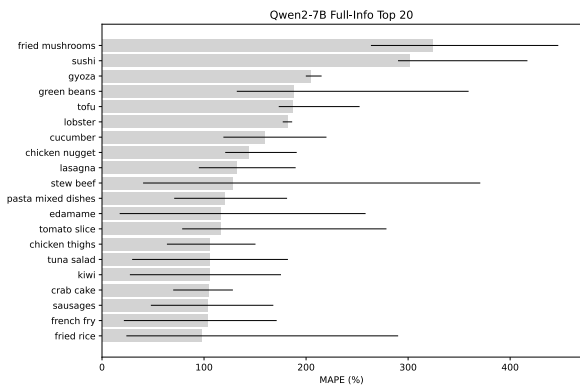


(c) LLaVA-7B (Full-Info)

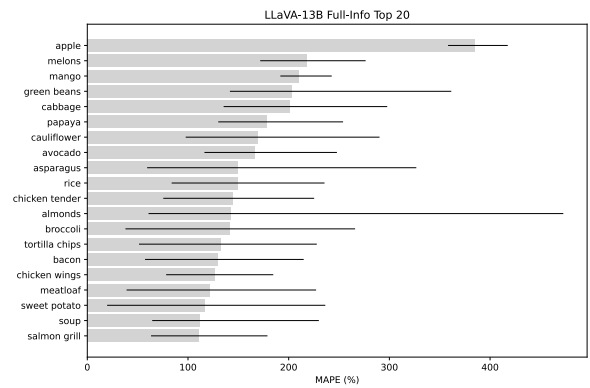


(d) DeepSeek-7B (Full-Info)

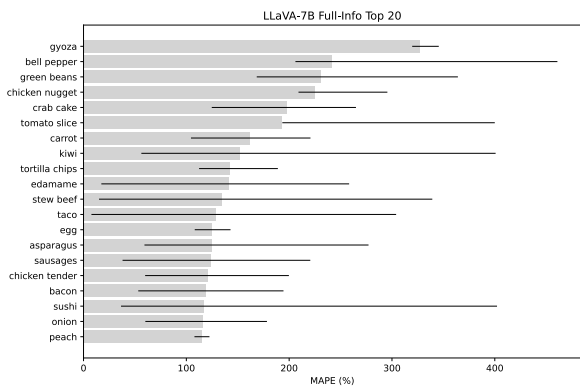
Figure 1. Ingredient-level top-20 errors under *Full-Info* for Nutrition-5k dataset.



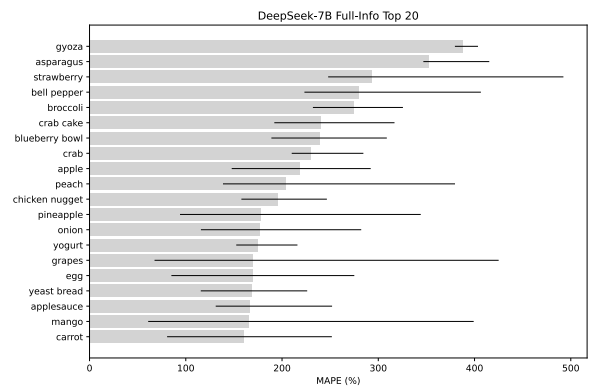
(a) Qwen 2-7B (Full-Info)



(b) LLaVA-13B (Full-Info)



(c) LLaVA-7B (Full-Info)



(d) DeepSeek-7B (Full-Info)

Figure 2. Ingredient-level top-20 errors under *Full-Info* for metafood3d dataset.