Decomposing Food Images for Better Nutrition Analysis: A Nutritionist-Inspired Two-Step Multimodal LLM Approach

Pitikorn Khlaisamniang[1,2,†] **Kun Kerdthaisong**[1,2,3,†] **Supasate Vorathammathorn**[1,2,†]
Nutchanon Yongsatianchot[3] **Hirunkul Phimsiri**[1,4] **Amrest Chinkamol**[1,5]
Teermade Thitseesaeng[6] **Kanyakorn Veerakanjana**[1,7,8] **Kaisorn Kachai**[1]
Piyalitt Ittichaiwong[1,7,9,*] **Tossaporn Saengja**[1,7,*]

[1]PreceptorAI Team, CARIVA Thailand
[2]Artificial Intelligence Association of Thailand
[3]Thammasat School of Engineering, Thammasat University
[4]Computer Engineering and Digital Technology Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University
[5]Vidyasirimedhi Institute of Science and Technology
[6]National Health Security Office (NHSO)
[7]Siriraj Informatics and Data Innovation Center (SIData+), Faculty of Medicine, Siriraj Hospital, Mahidol University
[8]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London
[9]School of Biomedical Engineering & Imaging Sciences, King's College London
[†]Equal first contributions
[*]Corresponding authors

Abstract

Accurate estimation of nutritional information from food images remains a challenging problem. Most existing approaches rely on deep image models fine-tuned with extensive food annotations or require detailed user inputs (e.g., portion size, cooking method), both of which are prone to error. Motivated by the workflow of nutrition experts, we propose a ***two-step prompting*** *framework leveraging* off-the-shelf *Multimodal Large Language Models (MLLMs). The first step deconstructs the dish into its components listing major ingredients, portion sizes, and cooking details while the second step computes total calories and macronutrients. This approach alleviates the need for heavy fine-tuning or large ingredient databases, by instead harnessing the compositional reasoning capabilities of general MLLMs. We evaluate the method on both a subset of the Nutrition5k dataset (****Nutrition320****) and real-world samples from the Gindee application (****Gindee121****), achieving more accurate estimates than one-step direct queries. Additional experiments with visual prompts (bounding boxes, segmentation masks) further demonstrate the robustness and adaptability of our approach. Notably, our findings reveal that guiding MLLMs through a structured two-step reasoning process—separating "what is on the plate" from "how it translates nutritionally"—substantially improves the reliability of image-based macronutrient estimation.*

1. Introduction

Maintaining a consistent, healthy diet has a profound impact on health outcomes; however, sustaining it remains a challenge. *Having balanced diet promotes health* by reducing the risks of malnutrition and chronic diseases [27]. However, *most people struggle to accurately estimate their daily caloric and macronutrient intake*, often due to the complexity of measuring portion sizes and identifying all ingredients of a meal. Such misestimation can lead to long-term health risks [36, 37]. Moreover, the training process for dieticians to develop these digital viewing skills can be extensive, often requiring several years of practice to achieve proficiency [10, 11]. To address this challenge, numerous calorie-tracking applications now help users log meals and compute macronutrient facts, such as **Foodvisor** [6], **Calzen.ai** [1] and **Gindee** [34]. Nevertheless, challenges persist in providing reliable macronutrient estimates especially when users must manually enter easily to misjudge details such as portion size and cooking method.

A prior study shows that inaccurate portion size estimation can lead to almost three times greater error in calorie predictions [35]. Ingredients, preparation methods, and even inedible parts (e.g., bones or peels) all affect nutri-

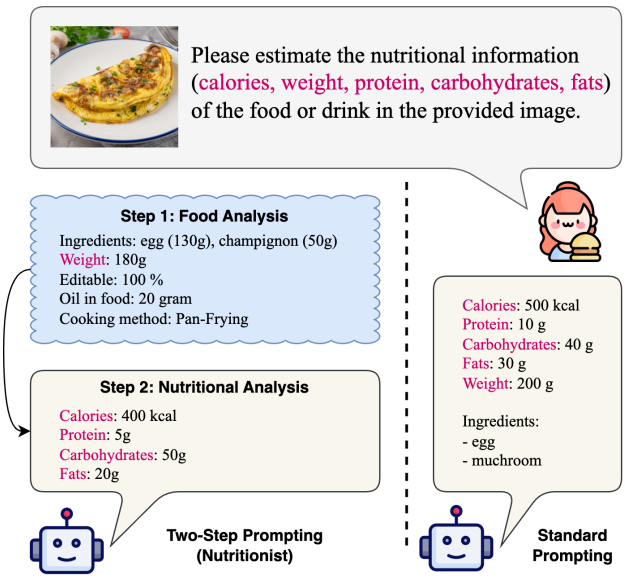

Figure 1. Comparison of **Two-Step Prompting** and **Standard Prompting** to estimate nutrition from a food photo. The left panel uses a two-step ingredient analysis, while the right panel offers a single-step estimate.

tional content yet are difficult to infer from visuals alone. Traditional calorie-tracking applications rely on manual user input of food type and weight, a process that is both time-consuming and prone to error [40].

Recent advances in *Vision-Language Models (VLMs)* like **OpenAI's GPT-4** [24] offer a promising path for automated image analysis and text generation, demonstrating remarkable performance in general image. However, directly asking a multimodal large language model (MLLM) to "estimate the calories in this image" often yields accuracy comparable to non-experts. Although fine-tuning MLLMs on nutrition data can potentially improve accuracy, the rapid evolution of these models and the associated training cost makes this approach impractical [18]. Instead, we propose a solution that can adapt to any off-the-shelf MLLM without retraining, guided by carefully designed prompts.

In this work, we propose a *two-step prompting* to guide the MLLMs through a reasoning process akin to the *nutritionist and dietician* examining a meal [11]. In our method, the automated macronutrients estimation from the food image task is decomposed into two sequential prompts to analyze the meal components: **Step 1 (Food Analysis)**, to make model examines the image and is prompted to output detailed attributes of the meal: estimate portion sizes and note other relevant factors such as the cooking method and inedible parts. **Step 2 (Nutritional Analysis)** the model incorporates the Step 1 output, and is then prompted to estimate the total calories and macronutrients (protein, fat, and carbohydrates). This method mimics how a nutrition expert would logically separate the assessment of "what and how much is on the plate" from "what does that translate into nutritionally" By breaking down the problem, we empower standard MLLMs to provide more accurate estimations [11].

We validate our approach on both standard and real-world datasets, across multiple MLLMs architectures. In summary, our contributions are:

- **Two-Step Prompting for Nutrition Estimation:** A novel prompting strategy that significantly improves the accuracy of calorie and macronutrient estimates from food images. By breaking the task into an *analysis step* and an *estimation step*, we enable general MLLMs to handle this complex regression problem *without fine-tuning*.
- **Evaluation on Benchmark and Real-world Data:** We demonstrate that our two-step approach outperforms a direct (one-step) baseline query baseline on the **Nutrition5k benchmark dataset** [35] , including a sampled subset Nutrition5k(**Nutrition320**), and on a new in-the-wild dataset, **Gindee121** collected from Gindee application [34]. Improvements are consistent across different MLLMs (including **Gemini-2.0-Flash** [33], **GPT-4o** [25], **Qwen2.5-VL-72B** [2], and **Llama3.2-90B-Vision** [21]), highlighting the robustness and generality of the method.
- **Visual Prompting:** we also examine multiple visual prompting techniques for macronutrients estimation, including segmentation masks from (segmentation masks from **SAM2.1** [29] and **FoodSAM** [16], as well as bounding-box annotations from **YOLO-World** [3] with human alignment.

2. Related Work

2.1. Automated Dietary Assessment

Early approaches to automated dietary assessment primarily relied on convolutional neural networks (CNNs) to identify food items and heuristic approaches to estimate portion sizes [39]. Despite being innovative at the time, these strategies encountered challenges due to the high variability of food presentation and limited scalability. A major milestone appeared with the introduction of the Nutrition5k dataset [35], which included real-world food images, depth data, and extensive nutritional annotations. Nutrition5k's comprehensive scope enabled the development of multi-task learning models capable of predicting calories, macronutrients, and portion sizes simultaneously, thereby establishing a critical benchmark for future research in nutrition analysis.

2.2. Vision Transformers, Visual Prompting, and Depth-Based Extensions

As deep learning techniques matured, attention mechanisms led to the advent of Vision Transformers (ViTs) [5], celebrated for their ability to capture global context. The concept of Learning Visual Prompts to guide the Attention of Vision Transformers [31] emerged as a self-supervised method to direct the focus of a model on key image regions, particularly significant in food analysis, where small details such as garnishes and textures can affect nutrition estimates. Concurrently, depth information became indispensable for improving food volume assessments. DPF-Nutrition [9] pioneered RGB-D fusion by integrating depth data with conventional RGB inputs, thus improving the calorie and portion size predictions through a more accurate three-dimensional view of food items.

2.3. Multimodal with Vision-Language Models

Recent developments extend dietary analysis beyond visual inputs alone, incorporating language signals to capture cooking methods and hidden ingredients. CaLoRAify [38] exemplifies this direction by employing visual-text pairings with low-rank adaptation (LoRA) [12] to align ingredient detection more closely with calorie estimates. Other systems such as CoCot [15] and CalorieMe [20] reinforce the importance of combining visual and textual clues for robust nutritional predictions. Summaries and surveys, like Advancements in Using AI for Dietary Assessment Based on Food Images: Scoping Review [4] and a variety of vision-language model surveys [7, 8, 19], underline the fast pace of progress in multimodal AI. They also mention emerging techniques, including mixture-of-experts frameworks (e.g., RoDE [13]) and high-throughput methods, that enhance the synergy between visual recognition and language understanding.

3. Methodology

3.1. Sample Selection on Nutrition5k

We selected a subset of samples from the Nutrition5K [35] dataset for our experiments to improve resource usage efficiency and decrease computational time.

To ensure data quality and remove outliers, we restricted the dataset by including only dishes that had overhead images, nonzero calorie values, and no more than 1,000 kcal. To maintain the original distribution shape, 10% of the samples were selected from each bin.

To select 10% of each bin and ensure diversity among the selected samplessample diversity, we embedded the images using **DINOv2** [26] to obtain image features. We then combined these features with Nutrition5K attributes, such as ingredient lists and nutritional information to form a unified food feature representation. Next, we applied K-means clustering within each bin, grouping dishes into clusters [14, 32] representing 10% of the total samples for each bin as Figure 2. We refer to this final subset as **Nutrition320**.

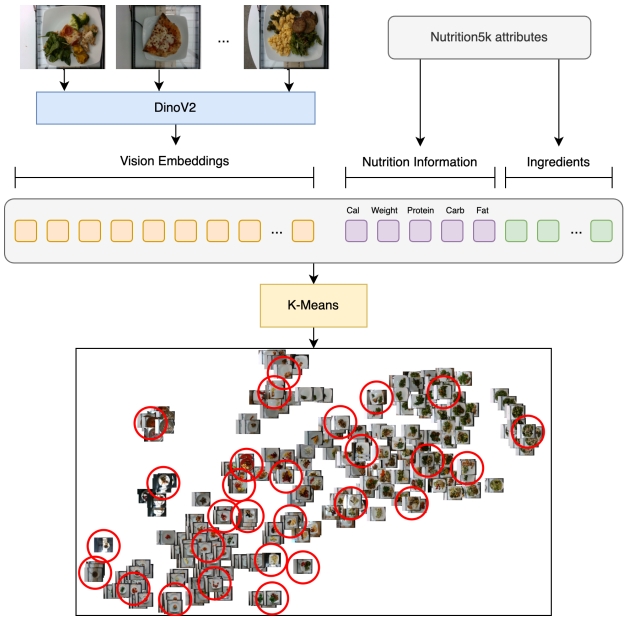

Figure 2. Dish images are embedded with DinoV2, combined with Nutrition5k attributes, and then clustered via K-Means. The bottom panel shows a 2D projection of the resulting clusters.

3.2. Gindee121

In **Gindee** platform, users contributed over two million food images to the app and we collected 140,855 user-edited entries (images plus attributes). and applied several filters: edited calorie values are below 20,000 edited carbohydrate values are under 2,000 and edited protein values are at least 0. Additionally, we also excluded instances where the app's original predicted carbohydrate exceeded 2,000 grams.

After filtering the dataset, we use **Gemini-1.5-Pro** with a "nutshell prompt" (Figure 3) to cluster the data into 121 groups. From each cluster, we selected the single entry that best represented its group, ensuring a diverse and well-distributed dataset for further analysis.

Finally, a trained nutritionist manually verified each chosen entry to correct any remaining inaccuracies. This expert review helped maintain the dataset's quality.

3.3. Visual Prompting Techniques

In this work, we also investigated the effectiveness of two visual prompting techniques, box prompting and mask prompting.

Figure 3. Gindee121 data collection pipeline: raw images are filtered and clustered, then one image per cluster is labeled by name, preparation method, and cultural origin.

3.3.1. Box Prompting

Bounding boxes are used to delineate objects or regions within an image, enabling MLLMs to extract visual features [17]. We annotated bounding boxes using **YOLO-World** [3], followed by human alignment and approval, on Nutrition320 and Gindee121 datasets at two levels: *food annotation*, which identifies the overall food item, and *ingredient annotation*, which captures a more fine-grained representation of individual ingredients.

3.3.2. Mask Prompting

Segmentation models, including **SAM2.1** [29] and **FoodSAM** [16], are employed to delineate and identify specific regions, objects, or structures within images, allowing models to focus more accurately on relevant visual information [17].

For our experiments, we employed the following approaches: SAM2.1 for segmenting entire food images, SAM2.1 segmentation using bounding boxes as prompts, filtering SAM2.1-segmented food images to retain only regions within the bounding box while also plotting the bounding box, and filtering SAM2.1-segmented food images without plotting the bounding box.

To enhance food image analysis, we also utilized semantic and panoptic segmentation from FoodSAM. Semantic segmentation was applied to generate refined food masks by combining coarse semantic masks with SAM-generated masks, improving segmentation quality. Furthermore, we leveraged FoodSAM's panoptic segmentation, which incorporates an object detector to capture food within the image. By integrating these segmentation techniques.

3.4. Two-Step Prompting

Nutrition estimation often involves breaking meals into ingredient portions, as in INMUCAL [28] and CaLoRAify [38]. However, these methods rely on fixed ingredient databases, causing coverage gaps and mapping inconsistencies.

To address these limitations, we introduce a two-step prompting method, inspired by Compositional Chain-of-Thought (CCoT) [22]. First, the model performs a "food analysis", extracting detailed on portion sizes, ingredients, and preparation methods. Second, it uses this structured representation to estimate key nutritional values such as calories, protein, carbohydrates, and fat. By separating the reasoning process into two logical steps, our approach enhances compositional understanding while remaining independent of extensive ingredient databases or fine-tuning. The pipeline and complete prompts, demonstrating this structured approach, are presented in Figure 4.

3.4.1. Step 1: Food Analysis

Our first step is to generate a structured representation of food data, eliminating the need for manually annotated food datasets. The food analysis prompt instructs the model to systematically extract key properties of a given meal, including ingredients, portion sizes, weight, edible percentage, oil used in food, and cooking method. Furthermore, we standardize the output format in JSON, enabling consistent interpretation by the model.

3.4.2. Step 2: Nutrition Analysis

The structured food representation from Step 1 serves as an intermediate reasoning step. The model estimates macronutrients by utilizing the extracted ingredients, portion sizes, weight, edible percentage, oil usage, cooking methods, and raw image.

4. Experimental Results

4.1. Experiment Setup

Experiments were conducted using various visual prompts, including bounding boxes, segmentation, and their combination, along with two-step prompting, as shown in Table 1. For visual prompts that require image segmentation, we also concatenated the original image to retain complete information. The multimodal models used for testing include GPT-4o (2024-11-20), Gemini-2.0-Flash, Qwen2.5-VL[2], and Llama3.2-Vision[21]. The experiments were conducted on the Nutrition320 and Gindee121 datasets. Mean Absolute Error (MAE) was used as the evaluation metric to assess the accuracy of macronutrient estimation across different prompting methods.

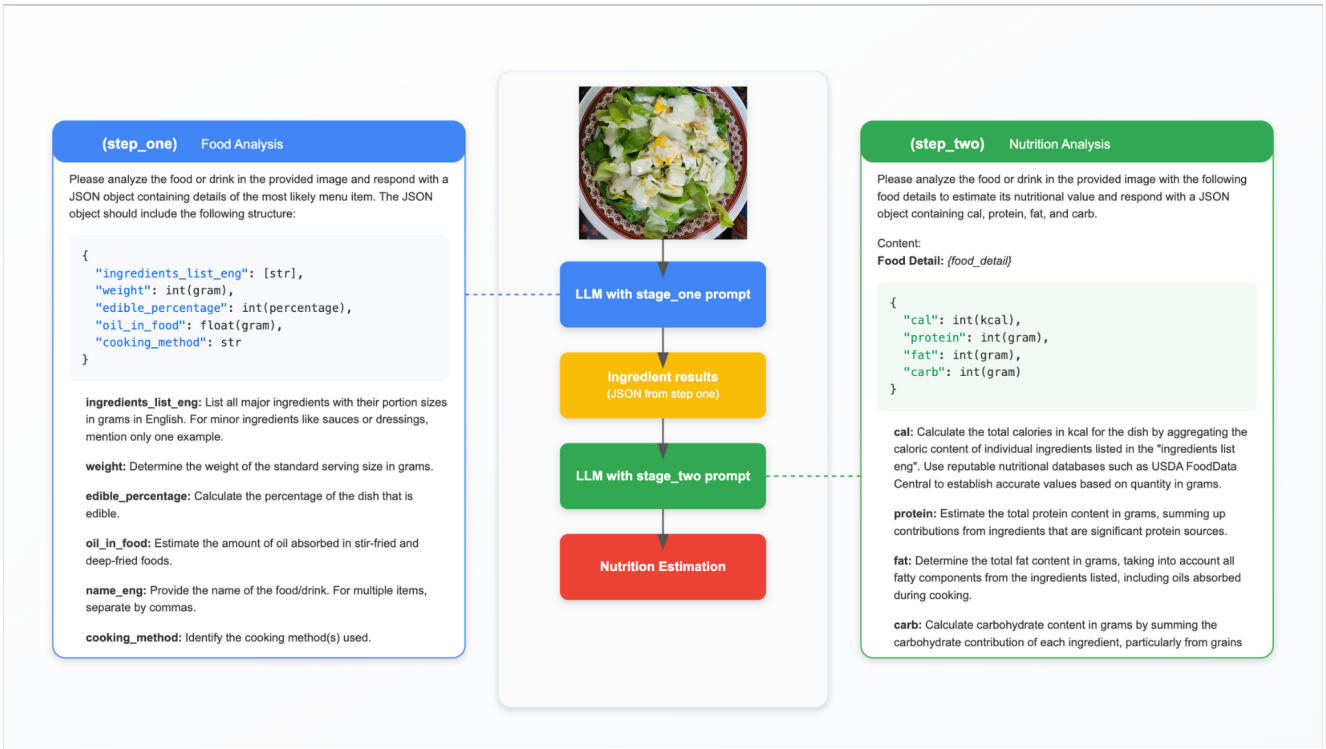

Figure 4. A two-step framework for food and nutrition analysis. The first step identifies key ingredients, portion sizes, and cooking Two-step framework for food nutrition analysis. Step 1 extracts ingredients and portion details; Step 2 uses that information to estimate calories and macronutrients **food_detail**

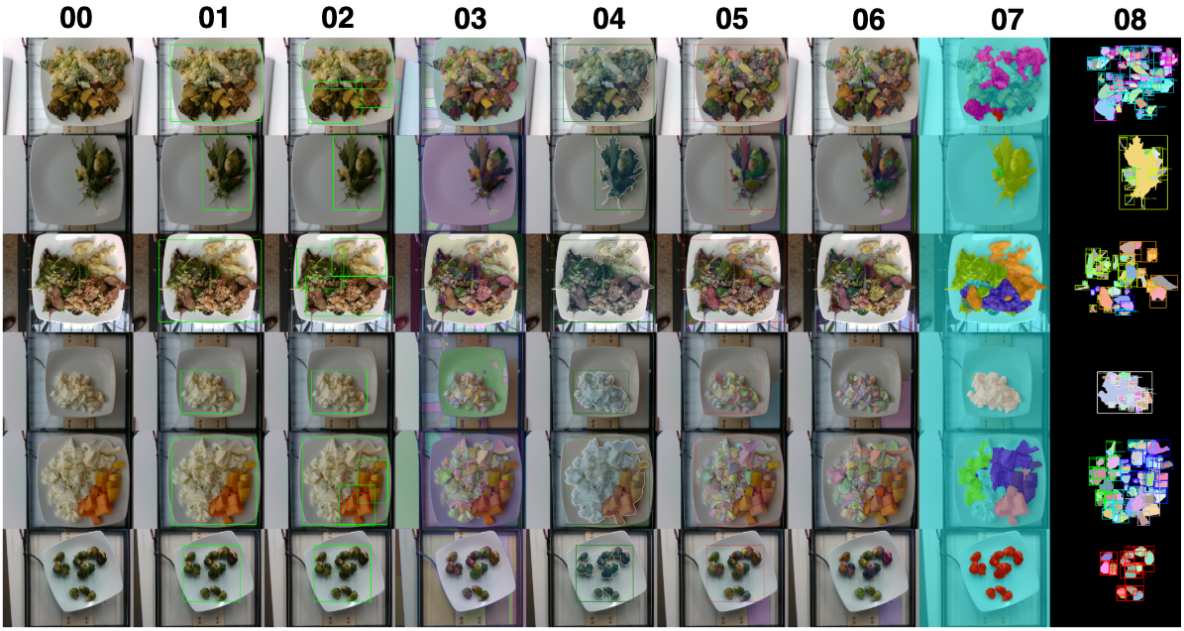

Figure 5. The different experiment setups (**EXP 00-08**) applied to the same dish image. Each column corresponds to a segmentation or prompting method from Table 1, illustrating how various approaches isolate and label ingredients.

Exp	Description
00	Standard Prompting as a baseline with raw image
01	Bounding box at food level
02	Bounding boxes at ingredient level
03	SAM2.1 segmented entire food image
04	SAM2.1 segmented using a bounding box as prompt
05	SAM2.1 segmented within a bounding box (w/ box)
06	SAM2.1 segmented within a bounding box (w/o box)
07	Semantic segmentation (FoodSAM)
08	Panoptic segmentation (FoodSAM)
09	Two-Step Prompting with raw image **(ours)**

Table 1. Overview of the experiment setups, each employing a unique segmentation/prompting method, culminating in our Two-Step Prompting (Exp 09).

4.2. Performance Comparison

We tested multiple prompting methods and multimodal models (Table 1) using the Nutrition320 and Gindee121 datasets.

- GPT-4o: Two-Step Prompting consistently excelled. On Nutrition320, it achieved the lowest calorie error (78.49 kcal, 12.11% better than baseline) and outperformed others in weight, carbs, and fat. On Gindee121, it improved calorie error by 10% (85.5 kcal), narrowly behind Qwen2.5-VL with bounding boxes, and ranked top 3 in all nutrients.
- Gemini-2.0-Flash: Two-Step Prompting improved calorie, protein, and carb estimation on Nutrition320, though bounding boxes gave better overall calorie results.
- Qwen2.5-VL: Bounding box prompting performed best for calorie estimation, exceeding GPT-4o Two-Step by 2% on Gindee121.
- Llama3.2-Vision: Performs poorly in all tests.

Overall, GPT-4o with Two-Step Prompting offers top-tier accuracy for calorie and macronutrient estimation across both datasets.

4.3. Ablation Study

Two-Step Prompting on Nutrition5k

We also compared Two-Step Prompting to a standard one-step baseline on the full Nutrition5k dataset. As shown in Table 2, Two-Step Prompting outperforms the standard prompting in calorie estimation and macronutrient breakdown for GPT-4o, Gemini-2.0-Flash, and Llama3.2-Vision, while Qwen2.5-VL performs slightly worse than the baseline. Among all models, GPT-4o achieves the best performance in calorie, weight, carbohydrate, and fat estimation.

Step 2: Text-Based Nutrition Analysis

After Step 1, using GPT-4o on the Nutrition5k dataset, Step 2 was conducted without an input image to assess the impact of image presence in Step 2. This resulted in a significantly higher MAE score for both calorie and macronutrient estimation, except for protein, as shown in Figure 6.

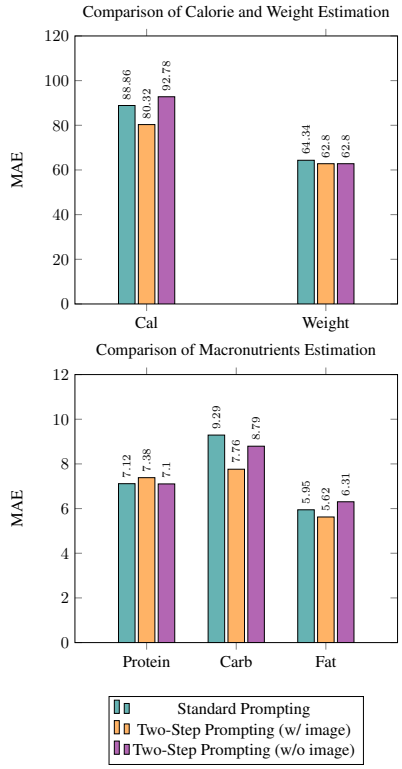

Figure 6. MAE comparison: Top—GPT-4o's calorie and weight estimates; Bottom—model comparisons for protein, carbohydrate, and fat on **Nutrition5k**.

Evaluating Model Understanding of Food ingredients

The models' comprehension of food composition was evaluated by extracting their predicted ingredients (text-based) and comparing them to the ground truth using Jaccard similarity. To standardize ingredient names, MiniLM-L6-v2 [30] was employed for name mapping, aligning the predicted ingredients with the USDA and Nutrition5k databases prior to similarity computation. The complete pipeline is depicted in Figure 7.

As shown in Table 3, Two-Step Prompting improves ingredient prediction accuracy across all models, consistently outperforming Standard Prompting in ingredient listing.

Comparison with Reasoning Models

We also conducted experiments on reasoning models, such as GPT-o1 [23]and Gemini-2.0-Flash-Thinking [33]. The results, shown in Figure 8 for the Nutrition320 dataset, indicate that reasoning models perform worse than standard models. However, for the Gindee121 dataset, GPT-o1 outperforms all other models, while Gemini-2.0-Flash-

Model	Exp	Nutrition320					Gindee121			
		Cal	Weight	Protein	Carb	Fat	Cal	Protein	Carb	Fat
GPT-4o	00	89.305†	79.122†	7.789	10.205	5.325†	97.554	6.389	13.958	**6.063**
	01	148.211	98.463	11.262	14.163	10.281	125.711	10.179	21.309	7.947
	02	123.133	92.122	10.525	13.580	7.965	130.731	10.287	19.408	8.022
	03	94.962	84.872	8.302	10.987	5.714	104.718	7.446	18.621	7.238
	04	92.433	81.681	8.334	9.939	5.597	103.655	6.249	12.934‡	6.999
	05	93.728	83.628	8.202	10.308	5.589	102.252	7.263	16.451	6.712
	06	97.867	86.459	8.433	11.204	5.745	96.669	7.469	16.642	6.662‡
	07	95.516	84.372	8.319	10.748	5.556‡	100.416‡	**6.064**	12.596†	6.818
	08	96.559	85.372	8.208	11.185	5.678	104.487	6.099†	15.667	7.009
	09	**78.489**	**73.959**	7.747	**8.220**	**5.035**	85.512†	6.111‡	**10.920**	6.287†
Gemini-2.0-Flash	00	108.480	83.762‡	7.636	10.231	6.974	116.732	6.800	13.086	10.609
	01	120.419	91.572	7.473†	10.951	7.688	112.836	6.825	13.551	10.397
	02	96.143	81.941	7.602‡	9.535‡	6.232	108.290	7.815	12.913	9.601
	03	124.436	96.934	8.310	11.475	7.980	124.324	6.733	14.639	10.327
	04	124.005	96.694	8.166	11.660	7.960	120.178	7.167	17.206	10.294
	05	134.803	104.431	8.323	12.648	8.422	122.663	7.235	16.846	11.065
	06	124.221	97.838	8.364	12.048	8.024	129.500	6.796	15.882	11.027
	07	128.319	94.644	8.115	11.896	8.258	113.396	8.635	17.020	10.292
	08	135.316	100.838	8.282	12.791	8.581	128.624	6.949	15.058	10.354
	09	102.663	90.131	**7.285**	9.006†	7.426	139.699	11.620	14.533	12.348
Qwen2.5-VL-72B	00	90.110	83.952	8.952	10.510	6.171	109.839	8.041	17.649	7.484
	01	103.690	98.014	9.272	10.770	6.517	**83.809**	7.856	15.126	8.172
	02	103.885	97.798	8.946	10.330	6.314	107.427	7.328	15.155	6.887
	03	98.722	91.154	8.788	11.601	6.327	114.123	8.546	17.541	7.690
	04	96.127	94.268	8.981	10.857	5.982	103.654	7.053	18.391	9.484
	05	104.431	94.914	9.406	11.637	6.628	102.974	7.705	19.144	7.077
	06	102.202	93.027	8.623	12.405	6.138	103.460	6.859	15.164	8.921
	07	107.332	114.306	10.071	10.725	6.611	127.758	9.314	15.854	8.419
	08	101.944	116.739	9.922	12.371	5.925	107.275	10.102	17.768	8.906
	09	89.641‡	88.381	8.186	12.590	7.214	113.670	9.237	23.688	9.028
Llama3.2-90B-Vision	00	104.585	99.211	10.351	16.201	8.293	142.809	12.964	32.959	12.637
	01	131.021	120.846	11.565	17.938	11.360	150.114	14.048	22.265	11.539
	02	124.265	122.960	11.005	17.107	8.110	135.760	11.702	27.713	11.260
	03	142.188	142.270	12.064	21.726	10.901	136.614	13.235	28.062	14.904
	04	128.041	129.071	11.834	16.254	9.723	143.439	14.421	30.098	12.629
	05	146.995	136.653	12.057	16.644	9.627	132.116	10.438	20.402	9.915
	06	132.434	133.731	12.375	18.582	11.028	150.096	13.835	21.381	11.003
	07	129.283	123.286	11.778	17.190	10.603	142.151	10.693	21.660	9.558
	08	134.030	134.792	12.819	24.739	12.180	156.583	11.313	23.940	12.075
	09	128.677	127.781	16.773	19.816	11.511	189.796	21.024	43.403	19.043

Table 2. **Results on Nutrition320 and Gindee121.** Calorie and macronutrient MAE (kcal) are reported, with lower values indicating better performance across both the Nutrition320 and Gindee121 datasets. Experiment numbers correspond to Table 1. Bold text indicates the best performance, while † and ‡ denote the second and third best results, respectively.

Thinking still underperforms compared to Gemini-2.0-Flash.

5. Conclusion

We have introduced a two-step prompting technique that emulates the logical workflow of nutrition experts in ana-

Model	Experiment	Nutrition5k						
		Mean Absolute Error (MAE)					Jaccard Similarity	
		Cal	Weight	Protein	Carb	Fat	USDA	Nutrition5k
GPT-4o	Standard Prompting (00)	88.86	64.34	**7.12**	9.29	5.95	0.311983	0.329593
	Two-Step Prompting (09)	**80.32**↓	**62.80**↓	7.38↑	**7.76**↓	**5.62**↓	**0.338045**↑	0.337873↑
Gemini-2.0-Flash	Standard Prompting (00)	126.03	86.56	7.76	10.94	8.26	0.320652	0.344203
	Two-Step Prompting (09)	102.93↓	79.85↓	7.38↓	8.70↓	7.54↓	0.337509↑	**0.346119**↑
Qwen2.5-VL-72B	Standard Prompting (00)	96.95	75.96	8.34	10.21	6.62	0.300879	0.327524
	Two-Step Prompting (09)	100.22↑	74.37↓	8.60↑	12.10↑	7.19↑	0.330620↑	0.335915↑
Llama3.2-90B-Vision	Standard Prompting (00)	118.90	116.40	10.38	14.01	9.03	0.264221	0.280697
	Two-Step Prompting (09)	106.26↓	82.52↓	8.48↓	11.73↓	7.21↓	0.329180↑	0.322352↑

Table 3. Comparison of model performance on the **Nutrition5k** dataset under **Standard Prompting (Exp 00)** versus **Two-Step Prompting (Exp 09)**. Mean Absolute Errors (MAE) are reported for calories, weight, protein, carbohydrate, and fat, alongside Jaccard similarity measures (USDA and Nutrition5k). Arrows indicate improvement (green) or decline (red) relative to the Standard Prompting.

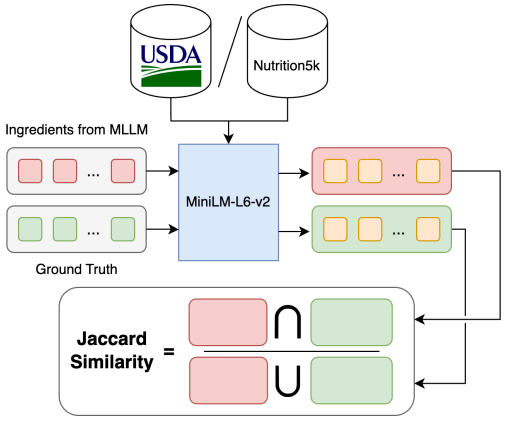

Figure 7. Pipeline for Evaluating Model Understanding of Food Ingredients.

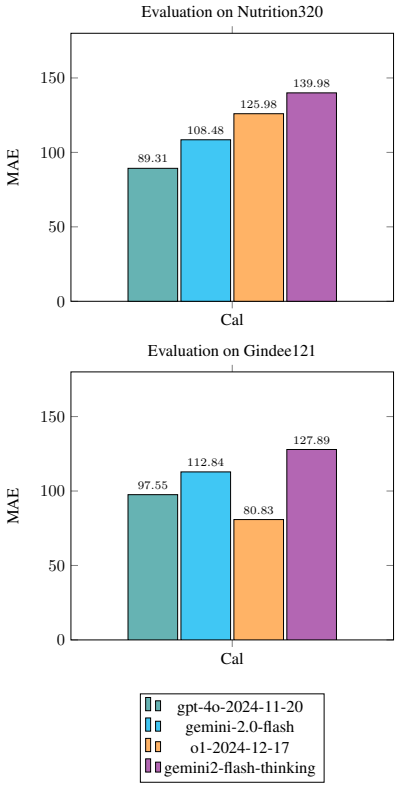

Figure 8. MAE comparison in calorie estimation across standard and reasoning models on Nutrition320 and Gindee121.

lyzing food items. By decomposing the task into image-based food analysis and subsequent nutrient calculation, our approach capitalizes on the compositional reasoning of off-the-shelf MLLMs without demanding additional fine-tuning. The experiments on both benchmark (Nutrition5k) and real-world (Gindee121) datasets demonstrate that our method consistently outperforms a conventional one-step prompt, yielding more accurate calorie and macronutrient estimates. Moreover, we explored various visual prompting strategies—bounding boxes and segmentation—to enhance the baseline performance, highlighting the adaptability and extensibility of the two-step framework. Future research directions include integrating domain-specific nutritional databases for refined ingredient analysis, leveraging depth information for improved portion sizing, and applying our framework to broader contexts such as meal planning and personalized dietary management. Ultimately, the proposed method offers a practical and scalable solution for automated dietary assessment, promising to reduce the user burden while increasing estimation accuracy.

References

[1] 42apps LLC. Calzen.ai, 2024. Accessed via Mobile App. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and et al. Qwen2.5-vl technical report, 2025. 2, 4

[3] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16901–16911, 2024. 2, 4

[4] P. Chotwanvirat, A. Prachansuwan, P. Sridonpai, and W. Kriengsinyos. Advancements in using ai for dietary assessment based on food images: Scoping review. *Journal of Medical Internet Research*, 26:e51432, 2024. 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3

[6] Foodvisor. Foodvisor, 2023. Accessed via Mobile App. 1

[7] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions, 2024. 3

[8] Tobias Groot and Matias Valdenegro Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, Mexico City, Mexico, 2024. Association for Computational Linguistics. 3

[9] Yuzhe Han, Qimin Cheng, Wenjin Wu, and Ziyang Huang. Dpf-nutrition: Food nutrition estimation via depth prediction and fusion. *Foods*, 12(23), 2023. 3

[10] D. K. N. Ho, S. H. Tseng, M. C. Wu, C. K. Shih, A. P. Atika, Y. C. Chen, and J. S. Chang. Validity of image-based dietary assessment methods: A systematic review and meta-analysis. *Clinical Nutrition*, 39(10):2945–2959, 2020. 1

[11] Dang Khanh Ngan Ho, Wan-Chun Chiu, Yu-Chieh Lee, Hsiu-Yueh Su, Chun-Chao Chang, and Chih-Yuan et al. Yao. Integration of an image-based dietary assessment paradigm into dietetic training improves food portion estimates by future dietitians. *Nutrients*, 13(1), 2021. 1, 2

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 3

[13] Pengkun Jiao, Xinlan Wu, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yugang Jiang. Rode: Linear rectified mixture of diverse experts for food large multi-modal models, 2024. 3

[14] Xin Jin and Jiawei Han. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA, 2010. 3

[15] Trevor Kann, Lujo Bauer, and Robert K. Cunningham. Cocot: Collaborative contact tracing. In *Proceedings of the Fourteenth ACM Conference on Data and Application Security and Privacy*, page 175–186, New York, NY, USA, 2024. Association for Computing Machinery. 3

[16] Xing Lan, Jiayi Lyu, Han Jiang, Kunkun Dong, Zehai Niu, Yi Zhang, and Jian Xue. Foodsam: Any food segmentation. *ArXiv*, abs/2308.05938, 2023. 2, 4

[17] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable MLLMs to comprehend what you want. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[18] Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. In *Advances in Neural Information Processing Systems*, pages 29615–29627. Curran Associates, Inc., 2023. 2

[19] Peihua Ma, Yixin Wu, Ning Yu, Yang Zhang, Michael Backes, Qin Wang, and Cheng-I Wei. Vision-language models boost food composition compilation. *CoRR*, abs/2306.01747, 2023. 3

[20] Bavlly Magid, Mohamed Ibrahim, Yomna A. Kawashti, Mazen Mohamed, Mohamed Sabry, Hanan Hindy, Mazen Khaled, and Waleed Mohamed. Calorieme: An image-based calorie estimator system. In *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 555–560, 2023. 3

[21] meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, 2024. Accessed: 2024-09-25. 2, 4

[22] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models, 2024. 4

[23] openai. Introducing openai o1, 2025. Accessed via Mobile App. 6

[24] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and et al. Gpt-4 technical report, 2023. 2

[25] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and et al. Gpt-4o system card, 2024. 2

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and et al. Dinov2: Learning robust visual features without supervision, 2023. 3

[27] World Health Organization. Healthy diet, 2020. WHO fact sheets (accessed 29 April 2020). 1

[28] Keeratichamroen Arisa Chammari Kantanit Ponprachanuvut Punnee, Srisangwan Nuttarat. inmucal. 4

[29] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, and et al. Sam 2: Segment anything in images and videos, 2024. 2, 4

[30] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of*

the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2019. 6

[31] Razieh Rezaei, Masoud Jalili Sabet, Jindong Gu, Daniel Rueckert, Philip Torr, and Ashkan Khakzar. Learning visual prompts for guiding the attention of vision transformers, 2025. 3

[32] Andreas Stephan, Lukas Miklautz, Kevin Sidak, Jan Philip Wahle, Bela Gipp, Claudia Plant, and Benjamin Roth. Text-guided image clustering. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2960–2976, St. Julian's, Malta, 2024. Association for Computational Linguistics. 3

[33] Koray Kavukcuoglu Sundat Pichar, Demis Hassabis. Introducing gemini 2.0: our new ai model for the agentic era, 2024. 2, 6

[34] Cariva (Thailand). Gindee food nutrition ai, 2024. Accessed via Mobile App. 1, 2

[35] Quin Thames et al. Nutrition5k: Towards automatic nutritional understanding of generic food. In *CVPR*, 2021. 1, 2, 3

[36] Psychology Today. Why we underestimate what we eat, 2021. 1

[37] Brian Wansink and Pierre Chandon. Meal size, not body size, influences calorie underestimation. *Annals of Internal Medicine*, 145(7), 2006. 1

[38] Dongyu Yao, Keling Yao, Junhong Zhou, and Yinghao Zhang. Caloraify: Calorie estimation with visual-text pairing and lora-driven visual language models, 2024. 3, 4

[39] Weishan Zhang, Dehai Zhao, Wenjuan Gong, Zhongwei Li, Qinghua Lu, and Su Yang. Food image recognition with convolutional neural networks. In *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, pages 690–693, 2015. 2

[40] Yaping Zhao, Ping Zhu, Yizhang Jiang, and Kaijian Xia. Visual nutrition analysis: leveraging segmentation and regression for food nutrient estimation. *Frontiers in Nutrition*, 11, 2024. 2