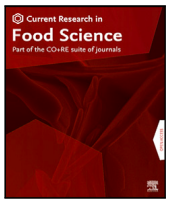
Current Research in Food Science 12 (2026) 101405

Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science



Research paper

A comparative study of vision–language models for food ingredient recognition and nutrient estimation

Shenglong Wang [a], Guorui Sheng [a], Hongfei Yan [a], Weiqing Min [b,c,*], Shuqiang Jiang [b,c]

[a] School of Computer Science and Artificial Intelligence, Ludong University, Yantai, 264025, China
[b] State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[c] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, 100190, China

ARTICLE INFO

Handling editor Georgios Leontidis

Keywords:
Vision–Language Models
Nutritional assessment
Food ingredient recognition
Nutrient estimation

ABSTRACT

The accurate assessment of food composition is essential to understanding its nutritional and sensory properties. Traditional dietary assessment methods are often constrained by subjective input and low reproducibility. This study explores the use of Vision–Language Models (VLMs) for automated food composition analysis, focusing on two key tasks: food ingredient recognition and nutrient estimation. We evaluated state-of-the-art VLMs using the Nutrition5K dataset, which contains real-world food images with ingredient-level annotations. To improve model sensitivity to complex food structures, we introduce a progressive multi-view image recognition approach that enhances ingredient recognition. We also propose a prompting strategy using ingredient labels to guide nutrient estimation. Results show that while most VLMs effectively identify primary food components, challenges persist in quantifying nutrient contents, particularly for composite or visually ambiguous dishes. Our findings highlight the promise and limitations of AI-assisted food composition analysis and offer insights for future methods integrating chemical, visual, and computational perspectives.

1. Introduction

Rising public health awareness and increasing efforts to prevent chronic diseases have intensified interest in improving both individual and population health through evidence-based dietary regulation. This trend is driven by the growing acceptance of the "food as medicine" paradigm and a substantial body of evidence linking dietary patterns to major health concerns such as obesity, diabetes, and cardiovascular disease. However, the effectiveness of dietary interventions critically depends on reliable and accessible methods for nutritional assessment. Traditional approaches — including 24-Hour Recalls (24HR), Dietary Records (DR), and Food Frequency Questionnaires (FFQ) (Shim et al., 2014) — typically rely on manual logging or dietitian evaluation, making them labor-intensive, time-consuming, and prone to inaccuracies.

The emergence of artificial intelligence has brought transformative advancements to this field (Almoselhy and Usmani, 2024; Theodore Armand et al., 2024; Deng et al., 2021). Specifically, the domain of Food Computing leverages tailored AI models to address nutrition-related tasks, enabling more efficient and accurate dietary analysis (Bagler and Goel, 2024). Large-scale datasets such as Food-101 (101 categories, 101,000 images) and Food2K (2000 categories, 1 million images) (Bossard et al., 2014; Min et al., 2023) have facilitated progress in food recognition, detection, and segmentation. However, the use of closed-set categories, high annotation costs, and practical deployment issues continue to hinder their effectiveness in real-world nutritional assessment. In response, recent developments have introduced advanced VLMs, which offer significant advantages over traditional deep learning architectures. These models achieve cross-domain generalization through pre-training on massive multimodal datasets, reducing the need for task-specific fine-tuning. Importantly, VLMs can support real-time inference on portable devices such as smartphones, making them promising tools for scalable, user-friendly nutritional monitoring. This study primarily investigates the capabilities of contemporary VLMs, such as ChatGPT and Qwen, in the context of nutritional assessment. Accordingly, this section first reviews prior research on deep learning applications in nutrition-related tasks, and then examines recent efforts involving VLMs in this domain. With their growing capabilities, VLMs are emerging as promising alternatives to traditional deep learning models in nutrition assessment.

Early studies largely relied on conventional deep learning approaches for tasks such as food ingredient recognition and nutrient estimation by integrating food imagery with multimodal data sources (Hu

* Corresponding author at: State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China.
E-mail addresses: wangshenglong@m.ldu.edu.cn (S. Wang), shengguorui@ldu.edu.cn (G. Sheng), hongfei.yan@m.ldu.edu.cn (H. Yan), minweiqing@ict.ac.cn (W. Min), sqjiang@ict.ac.cn (S. Jiang).

https://doi.org/10.1016/j.crfs.2026.101405
Received 9 November 2025; Received in revised form 7 February 2026; Accepted 8 April 2026
Available online 10 April 2026

et al., 2023; Bodnar et al., 2020; Keller et al., 2024; Parinayok et al., 2023; Sahoo et al., 2019; Ma et al., 2025; Shao et al., 2023; Wang et al., 2022). For example, Li et al. proposed the OptmWave deep learning framework, which leveraged near-infrared hyperspectral imaging for nutrient prediction and achieved notable performance in estimating protein content (Li et al., 2023). However, these approaches face two significant limitations: (1) they require precise alignment across multiple data modalities, and (2) their cross-domain adaptability is restricted by handcrafted feature fusion mechanisms. To overcome these challenges, recent studies have increasingly explored the application of Large Language Models (LLMs) and VLMs in the nutrition domain (Khamesian et al., 2025; Kopitar et al., 2024; Yin et al., 2023; Alkhalaf et al., 2024; Bergling et al., 2025; Papastratis et al., 2024). Yang et al. introduced the ChatDiet framework, which harnesses the reasoning capabilities of LLMs to deliver personalized, interpretable, and interactive food recommendations, offering a novel solution for dietary guidance and health management (Yang et al., 2024). In parallel, several works have proposed the development of domain-specific foundation models tailored for food computing tasks (Qi et al., 2023; Zhou et al., 2024). A growing body of research has also conducted performance evaluations of large-scale models, such as GPT-4 and similar VLMs, in various nutrition-related applications (Azimi et al., 2025; Adilmetova et al., 2025; Lo et al., 2024; Shi et al., 2023; Niszczota and Rybicka, 2023; Haman et al., 2024; Wang et al., 2024). These studies typically employ pre-defined prompts and task instructions without additional fine-tuning, thereby reflecting the models' out-of-the-box capabilities in real-world scenarios. Notably, Annalisa Szymanski et al. in collaboration with registered dietitians, assessed LLMs' effectiveness in delivering nutritional information. Their findings informed the development of design principles for optimizing GPT-4-based systems and culminated in a prototype nutrition assistant. This work not only offers practical insights but also provides a theoretical foundation for enhancing the role of LLMs in nutrition-related applications (Szymanski et al., 2024). Khlaisamniang et al. proposed a two-step approach for food nutrient analysis by separating ingredient recognition from nutrient calculation (Khlaisamniang et al., 2025). Unlike their approach, our study introduces multi-view fusion techniques and incorporates verified ingredient labels, highlighting their importance in enhancing assessment accuracy.

Accurately estimating nutritional content from food images remains a challenging task for traditional deep learning models. This process imposes strict performance requirements on multiple interdependent sub-tasks, including food recognition, portion size estimation, and ingredient recognition. The complexity is further exacerbated by visual factors such as shadows, occlusions, ingredient overlap, and image blur, which commonly occur in real-world dietary scenarios. Moreover, constructing high-quality annotated datasets for these tasks incurs substantial cost and labor. Given these limitations, it is both timely and necessary to investigate the performance of VLMs, which exhibit enhanced generalization and reasoning capabilities compared to conventional approaches.

Despite their impressive performance across various cross-domain tasks, VLMs encounter significant challenges when applied to food computing (Ma et al., 2024). First, food ingredient recognition scenarios are inherently complex: food items are frequently stacked, partially occluded, or presented under uneven lighting conditions. Moreover, regional cuisines and cultural diversity introduce substantial intra-class variation, requiring advanced fine-grained recognition capabilities. In the context of nutrient estimation, the reliability of VLMs presents several critical concerns: (1) visual information alone is insufficient for precise nutrient estimation; (2) substantial nutritional variability exists within food categories, and VLMs lack integration with authoritative nutrition databases; (3) model outputs are prone to hallucinations, occasionally misclassifying high-calorie items as healthier alternatives. For example, in preliminary tests, a leading VLM misclassified cream as milk, resulting in a 380 kcal error in caloric estimation. These limitations underscore the necessity of systematic model validation. To address these challenges (Vasiloglou et al., 2023), we systematically investigate the potential of VLMs for nutritional assessment, focusing on two core tasks: food ingredient recognition and nutrient estimation. This study provides practical insights into the comparative performance of different VLMs in nutritional assessment and explores strategies to enhance their practical utility.

The following section summarizes the key experimental findings derived from our evaluation of VLMs in the context of food ingredient recognition and nutrient estimation:

Food ingredient recognition. Most VLMs demonstrated acceptable performance, achieving an average precision of approximately 60%. They maintained relatively robust recognition even under challenging conditions such as shadowing, mild blurring, or partial occlusion, reflecting real-world scenarios. Progressive multi-view input testing revealed a trade-off between precision and recall. While precision slightly declined from the second to the fourth views, recall improved significantly. The largest recall improvement appeared when a second view was added, generally within a broad range of a few percentage points (around 2%–9%), while further gains became smaller (about 1%–3%) as more views were included. This pattern suggests that VLMs capture more food components from multiple perspectives but may misclassify newly visible items, resulting in the observed precision–recall trade-off. As an optimal strategy, providing two distinct food views achieves the best balance between performance and efficiency for VLM-based food ingredient recognition.

Nutrient estimation. Overall, most VLMs exhibited poor performance in estimating nutritional content. Dialogue-based analysis revealed that these models typically identify ingredients first, followed by nutrient estimation. However, errors in the initial identification stage, such as misclassifying chicken as smoked fish due to lighting or occlusion led to significant inaccuracies in the final estimation. In a controlled experiment where true ingredient labels were provided as input, most models paradoxically exhibited increased estimation errors, despite having correct ingredient information. This finding suggests that VLMs struggle with visual–nutritional reasoning, even when given accurate ingredient-level data, revealing a fundamental limitation in their capacity for nutritional quantification.

2. Data and methods

2.1. Data preparation and preprocessing

This study aims to evaluate the performance of several popular large-scale models (as of April 2025) in ingredient recognition and nutrient estimation tasks based on food images. To align with the research objectives, we selected the Nutrition5K dataset (Thames et al., 2021), which contains over 5000 real-world meals. This dataset includes 360-degree multi-angle videos along with corresponding ingredient categories and detailed nutritional information.

To ensure uniform sample class distribution and comprehensively evaluate the models' capabilities in food ingredient recognition and nutrient estimation, we selected 3466 samples from the Nutrition5K dataset for experimental testing. These samples cover 285 distinct food categories. From the provided four-viewpoint video for each food sample, we extracted the tenth frame of each video sequence as the corresponding input image for that viewpoint. The distribution of samples is illustrated in Fig. 1.

The majority of selected sample images contain multiple ingredients, providing a robust test of VLM recognition capabilities in complex scenarios. However, the diversity of ingredients increases interference in nutrient estimation experiments. Specifically, prediction errors for the nutrient content of one or a few ingredients may be substantial, while predictions for other ingredients remain accurate. To mitigate this interference, we additionally selected a subset of sample images

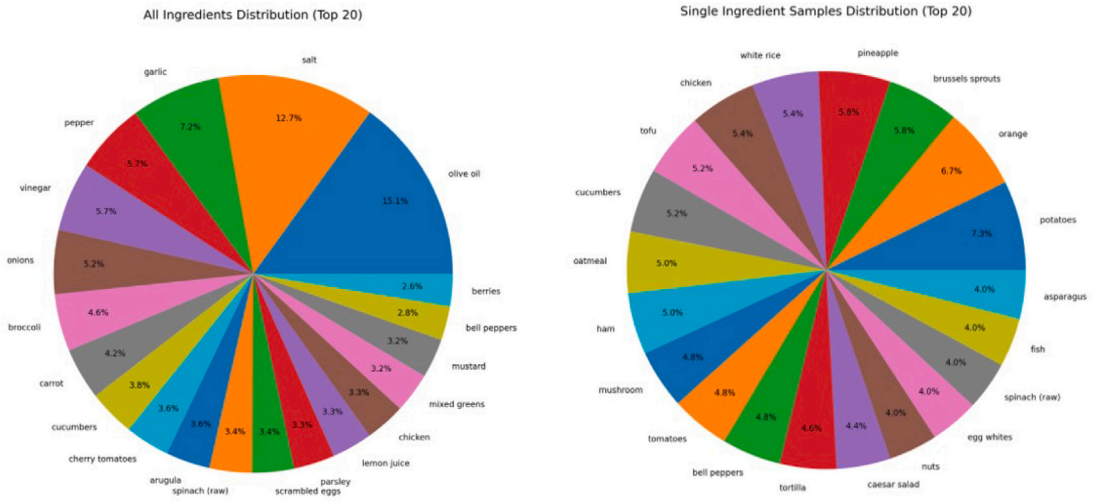


Fig. 1. Class distribution overview of the curated Nutrition5K dataset. (a) Proportions of the top 20 ingredient categories in the overall dataset. (b) Proportions of the top 20 ingredient categories in images containing only a single ingredient category.

Table 1
Summary of key notations used across recognition and estimation modules.

Symbol	Description
T	Ground truth ingredient set
P	Predicted ingredient set
$\mathrm{Sim}(a, b)$	Similarity function combining string matching and semantic analysis
y_k	nutrient value for sample k
$\hat{y}_k$	Predicted nutrient value for sample k
N	Total number of samples
V_i, V_j	View combinations (e.g., "A", "A+B", "A+B+C")
$\mathrm{LCS}(a, b)$	Longest Common Subsequence between strings a and b
$\mathrm{Var}(x)$	Synonym list for ingredient x (from ingredient categories)
$\lvert S\rvert$	Cardinality of set S

containing only a single ingredient. Although recognition is less challenging for these single-ingredient samples, they enable effective testing of VLMs' nutrient estimation capabilities by predicting the nutrient content of varying portion sizes of the same ingredient.

2.2. Experimental methods

To comprehensively and systematically evaluate the models' nutritional assessment capabilities, we designed two categories of experiments: food ingredient recognition and nutrient estimation. For each VLM, we provided carefully designed prompts to identify the categories or estimate the nutrient content of one or more ingredients per viewpoint within a single sample. In the following, we describe the specifics of these two task categories.

2.2.1. Food ingredient recognition

Current supervised learning-based food ingredient recognition methods exhibit two major limitations: firstly, model performance is constrained by the predefined taxonomy of the training data, making it difficult to handle the diversity of ingredients in real-world scenarios; secondly, recognition performance degrades when deployed on mobile devices due to limited computational resources. In contrast, self-supervised learning-based VLMs, such as ChatGPT and Qwen, pre-trained on massive multimodal datasets, not only adapt to open-domain ingredient recognition demands but also demonstrate superior performance on mobile devices.

In this experiment, the VLMs were tasked with identifying the categories of one or more ingredients from the four-viewpoint images of selected samples. The recognition results were compared with the ground truth for similarity assessment.

Specifically, if the prediction matched the ground truth, it was considered a correct prediction. Recognition performance was evaluated using Precision, Recall, and F1 score to quantify both the correctness and coverage of ingredient identification. The specific calculation formulas for these metrics are given in (1), (2) and (3). Explanations of all symbols used in the formulas are provided in Table 1.

$$\text{Precision} = \frac{\sum_{p\in P} \max_{t\in T} \mathrm{Sim}(p, t)}{|P|} \tag{1}$$

Recall is employed to evaluate the model's capability in identifying multiple ingredients within an image. A higher Recall value indicates more comprehensive ingredient recognition by the model. The calculation formula for *Recall* is given in (2):

$$\text{Recall} = \frac{\sum_{t\in T} \max_{p\in P} \mathrm{Sim}(t, p)}{|T|} \tag{2}$$

$F1$ is a composite metric combining *Precision* and *Recall*, providing a more balanced measure of the model's recognition accuracy and completeness. The calculation formula is given in (3):

$$\mathrm{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

The computation method for Sim in Eqs. (1) and (2) is defined by Eqs. (4), (5), and (6), where (5) is based on the Longest Common Subsequence (LCS), which identifies the longest sequence of characters shared between strings a and b in the same order:

$$\mathrm{Sim}(a, b) = \max\left(\underbrace{\mathrm{StrMatch}(a, b)}_{\text{Character-level}}, \underbrace{\mathrm{SemMatch}(a, b)}_{\text{Semantic}}\right) \tag{4}$$

Specifically, StrMatch(a, b) measures similarity at the character level using LCS, while SemMatch(a, b) evaluates semantic relatedness between two ingredients. The SemMatch(a, b) function is defined as follows:

$$\mathrm{StrMatch}(a, b) = \frac{2 \times |\mathrm{LCS}(a, b)|}{|a| + |b|} \tag{5}$$

$$\mathrm{SemMatch}(a, b) = \begin{cases} 1.0 & \text{if } a \in \mathrm{Var}(b) \text{ or } b \in \mathrm{Var}(a) \\ 0.0 & \text{otherwise} \end{cases} \tag{6}$$

This design allows the model to tolerate minor variations in ingredient naming and improves the robustness of food ingredient recognition by considering both string-level similarity and semantic equivalence. Here, Var(X) denotes a synonym list that contains alternative expressions or semantically related terms for ingredient X. Due to the diversity

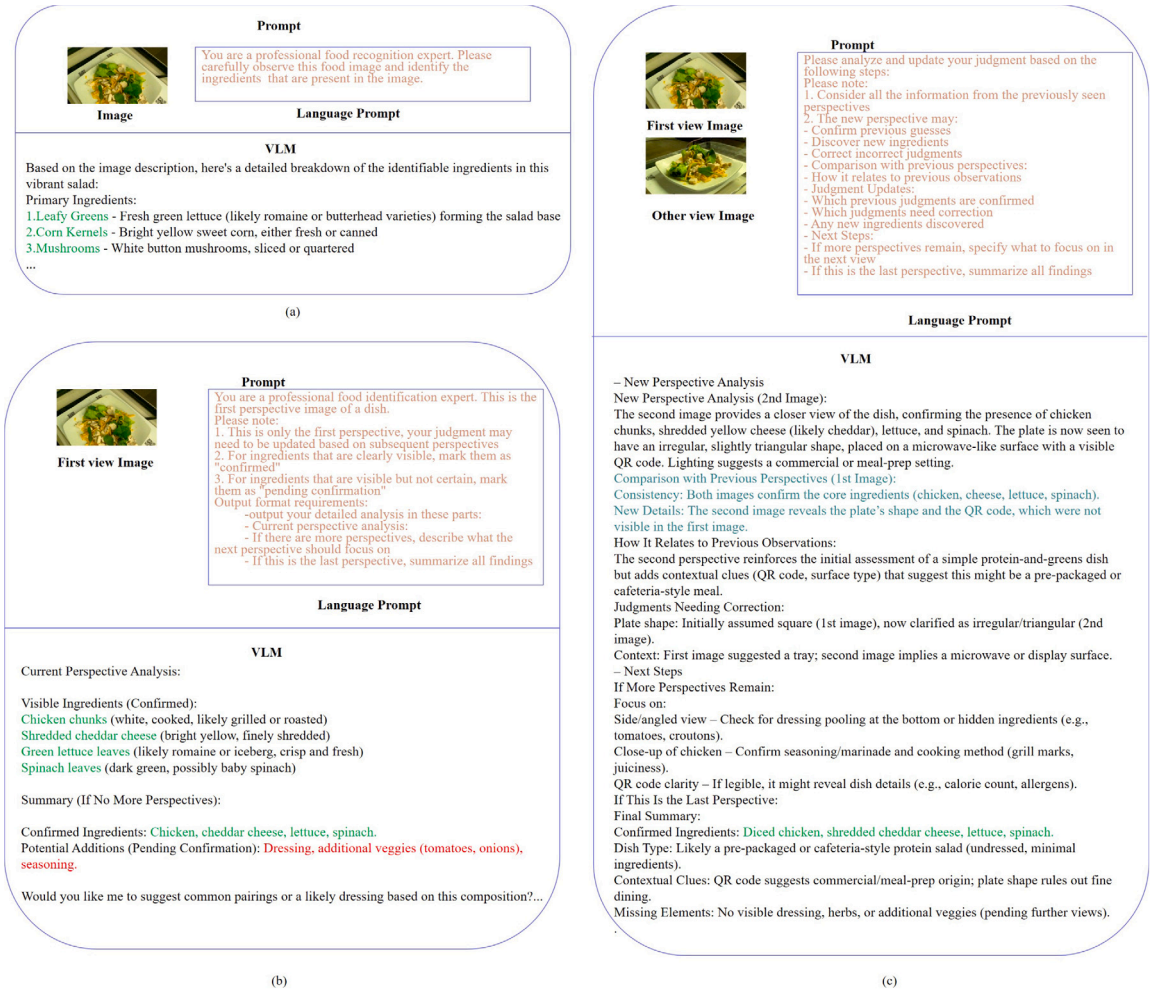


Fig. 2. Representative examples of VLMs applied to food ingredient recognition tasks. (a) Single-view recognition: ingredients are identified from a single food image. (b) First interaction in progressive multi-view recognition: the model identifies confirmed ingredients and lists uncertain ones based on the initial image. (c) Subsequent interaction with an additional image: new ingredients are identified, and information from both views is integrated to resolve previously uncertain categories.

of VLM outputs, ingredient predictions may include semantically correct but lexically different expressions compared to the ground-truth annotations. For example, a model may predict a general category such as "leafy greens", while the ground truth specifies individual ingredients such as "spinach", "kale", or "arugula". Although these predictions differ at the textual level, they are semantically consistent from a food ingredient recognition perspective. To account for such cases and avoid penalizing semantically valid predictions, we manually constructed a synonym list based on common naming variations observed in the dataset and the models' initial outputs. This synonym list was curated iteratively during early-stage evaluation and applied uniformly across all evaluated VLMs. Importantly, the same synonym mapping was used for all models, ensuring a fair and consistent comparison of model performance.

To further explore model capabilities, we designed an enhancement experiment as illustrated in Fig. 2. Specifically, when providing only a single viewpoint image, recognition inaccuracy often occurs due to occlusions between ingredients or shadows. To mitigate these environmental errors, we enabled the model to progressively observe images of different viewpoints for the same sample. We then calculated the improvement rate after each additional viewpoint compared to the previous state. The corresponding calculation formulas are provided in (7), (8), and (9):

$$\Delta\text{Precision}_{V_i \to V_j} = \frac{P_{V_j} - P_{V_i}}{P_{V_i}} \times 100\% \tag{7}$$

$$\Delta\text{Recall}_{V_i \to V_j} = \frac{R_{V_j} - R_{V_i}}{R_{V_i}} \times 100\% \tag{8}$$

$$\Delta\text{F1}_{V_i \to V_j} = \frac{\text{F1}_{V_j} - \text{F1}_{V_i}}{\text{F1}_{V_i}} \times 100\% \tag{9}$$

This progressive image input approach enables the quantification of interference caused by environmental factors. Simultaneously, it offers a straightforward approach for users to mitigate such interference during daily model interactions.

2.2.2. Nutrient estimation

Current methods primarily rely on supervised learning. These approaches lack support from large-scale nutritional databases, confining them to information within their training datasets. Furthermore, their dependence on food ingredient recognition, detection, and segmentation tasks results in excessively large models, hindering deployment on performance-constrained portable devices. In contrast, self-supervised learning-based VLMs do not require on-device deployment. Their powerful capabilities can be readily accessed via Application Programming Interface (API) calls.

In this task, the models were required to calculate the calories (kcal), total weight (g), fat content (g), carbohydrate content (g), and protein content (g) for each sample. The performance of each model in estimating the corresponding nutrients was evaluated by calculating the

Mean Absolute Error (MAE) and Relative Error (RE) for each nutrient. The specific calculation formulas are provided in (10) and (11). The MAE quantifies the average deviation between predicted and true values, while the RE characterizes the error magnitude relative to the true value. From a food science and nutrition perspective, the selected evaluation targets, including total weight, calories, fat, carbohydrates, and protein, represent the essential components of dietary assessment and are widely applied in both clinical nutrition and everyday dietary monitoring (Kumar et al., 2017). Total food weight reflects portion size, which is a major source of uncertainty in visual nutrient estimation and directly influences subsequent calculations of energy and macronutrient intake. Caloric content indicates the overall energy supplied by a meal and is central to maintaining energy balance, since carbohydrates and fats function as the primary substrates that support metabolic activity. The macronutrients themselves perform distinct physiological roles. Carbohydrates serve as the principal energy source and help spare protein from being diverted toward energy production. Proteins provide indispensable amino acids required for tissue synthesis, immune defense, enzymatic reactions, and the maintenance of body structure. Dietary fats contribute concentrated energy, facilitate the absorption of fat-soluble vitamins, and supply essential fatty acids that participate in cellular regulation and metabolic processes. Because imbalances in these nutrients are closely associated with health outcomes such as cardiovascular risk, glycemic control, metabolic disorders, and muscle preservation, accurate estimation of these nutritional components is fundamental for reliable dietary assessment and for understanding the health implications of food intake.

$$\mathrm{MAE} = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k| \tag{10}$$

$$\mathrm{RelErr} = \frac{1}{N}\sum_{k=1}^{N}\frac{|y_k - \hat{y}_k|}{|y_k|} \times 100\% \tag{11}$$

As shown in (12) and (13), we also calculated the mean of the estimated nutrient content values to assess the model's overall performance:

$$\mathrm{AvgMAE} = \frac{1}{5}(\mathrm{MAE}_{\mathrm{cal}} + \mathrm{MAE}_{\mathrm{mass}} + \mathrm{MAE}_{\mathrm{fat}} + \mathrm{MAE}_{\mathrm{carb}} + \mathrm{MAE}_{\mathrm{prot}}) \tag{12}$$

$$\mathrm{AvgRelErr} = \frac{1}{5}(\mathrm{RelErr}_{\mathrm{cal}} + \mathrm{RelEr}_{\mathrm{mass}} + \mathrm{RelErr}_{\mathrm{fat}} + \mathrm{RelErr}_{\mathrm{carb}} + \mathrm{RelErr}_{\mathrm{prot}}) \tag{13}$$

We further designed a comparative experiment to evaluate the impact of providing the model with ground truth ingredients, as illustrated in Fig. 3. Specifically, to investigate the relationship between the performance of ingredient recognition and the performance of nutrient estimation, we conducted two testing scenarios: (1) with explicit ingredient information provided to the model, and (2) without any ingredient information disclosure. This experimental design offers practical guidance for real-world applications of vision-based models in nutrient estimation.

2.2.3. Model selection

This section introduces the vision–language models (VLMs) selected for our study and compares their performance in nutritional assessment tasks under constrained computational resources via API calls. All models were accessed through the OpenAI API, with each client instance independently initialized for every sample. The temperature parameter, commonly used to control output randomness, was set to 0.2 to ensure more deterministic results. In addition, the maximum token limit set to 4096 tokens, was used to constrain output length. The specific model selection process is detailed in the following section.

GPT-4.1 mini: A lightweight multimodal language model released by OpenAI in April 2025, designed to significantly reduce latency and cost while maintaining high performance. This model supports image inputs with capabilities in image understanding and visual reasoning, enabling the processing of complex vision tasks (OpenAI, 2025). GPT-4.1 mini represents a substantial advancement in small-model performance, surpassing GPT-4o on many benchmarks. It matches or exceeds GPT-4o in intelligent evaluation tasks while reducing latency by 50% and lowering operational costs by 83%. Supporting context windows up to 1 million tokens, this architecture is particularly suitable for applications requiring large-scale context processing.

Gemini 2.5 Flash: A lightweight multimodal reasoning model released by Google in April 2025, featuring substantially enhanced reasoning capabilities while maintaining advantages in low latency, high throughput, and cost efficiency. This marks Google's first hybrid reasoning architecture (Google Cloud, 2025). The model accepts text, images, audio, and video inputs, enabling users to dynamically adjust the reasoning budget for processing complex information across different data modalities. It supports context windows up to 1 million tokens with maximum output capacity of approximately 65,000 tokens. Visual input accommodates multiple image formats with individual file size limits of 7MB per image, allowing processing of up to 3000 images or documents (e.g., PDFs) per prompt. Video and audio inputs support multiple formats with defined maximum duration and file quantity restrictions. During experiments, we permitted the model to automatically control its reasoning process, generating up to 8192 tokens.

Grok-2 Vision: The multimodal variant of Grok-2 released by xAI in December 2024 (xAI, 2025). This model demonstrates enhanced visual processing and comprehension capabilities, specifically designed to handle integrated text-image inputs. Grok-2 Vision exhibits exceptional performance in visual tasks, particularly excelling at processing diverse visual information types including documents, charts, screenshots, and photographs. It supports context windows up to 32,768 tokens. For image input, the model accepts JPG/JPEG or PNG formats with individual file size limits of 10MiB per image, with no explicit limit on image quantity per request.

Qwen2.5-VL (Qwen-VL-Max, Qwen2.5-VL-72B, Qwen2.5-VL-32B): A series of advanced VLMs successively released by Alibaba Cloud from 2024 to 2025 (Qwen Team, 2025; Bai et al., 2025). Designed to deliver superior multimodal comprehension capabilities, this architecture features substantially enhanced visual reasoning and instruction-following capabilities, demonstrating exceptional performance in complex vision tasks. The model exhibits comprehensive visual processing strengths, including image question answering, multilingual optical character recognition (OCR), mathematical problem solving, video analysis (capable of processing extended videos for summarization or event localization), and object detection. It supports context windows up to 128,000 tokens. In our experiments, we selected three variants from the Qwen-VL series: the continuously updated closed-source flagship model Qwen-VL-Max, along with two open-source variants Qwen2.5-VL-72B and Qwen2.5-VL-32B.

Doubao-1.5-vision-pro: A multimodal large model developed by Bytedance, initially released in January 2025 with subsequent optimization upgrades in March 2025 (ByteDance, 2025). This architecture implements comprehensive technical enhancements in multimodal data synthesis, dynamic resolution adjustment, multimodal alignment, and hybrid training methodologies. It supports image processing at arbitrary resolutions and extreme aspect ratios, significantly improving visual reasoning, document recognition, fine-grained information comprehension, and instruction-following capabilities. The model features a maximum context length of 128k tokens with configurable output capacity up to 16k tokens. Input image processing accommodates JPEG, PNG, GIF and other standard image formats, making it suitable for multi-image interleaved conversations, visual question answering, and complex tasks in specialized domains.

Llama 4 Maverick: A multimodal large model released by Meta in April 2025, built upon the Mixture-of-Experts (MoE) architecture with 17 billion active parameters (17B) and 1.28 trillion expert parameters (128E) (Meta AI, 2025b). This model implements early fusion technology to achieve seamless integration of text and visual tokens, enabling joint reasoning over high-resolution images, complex charts,

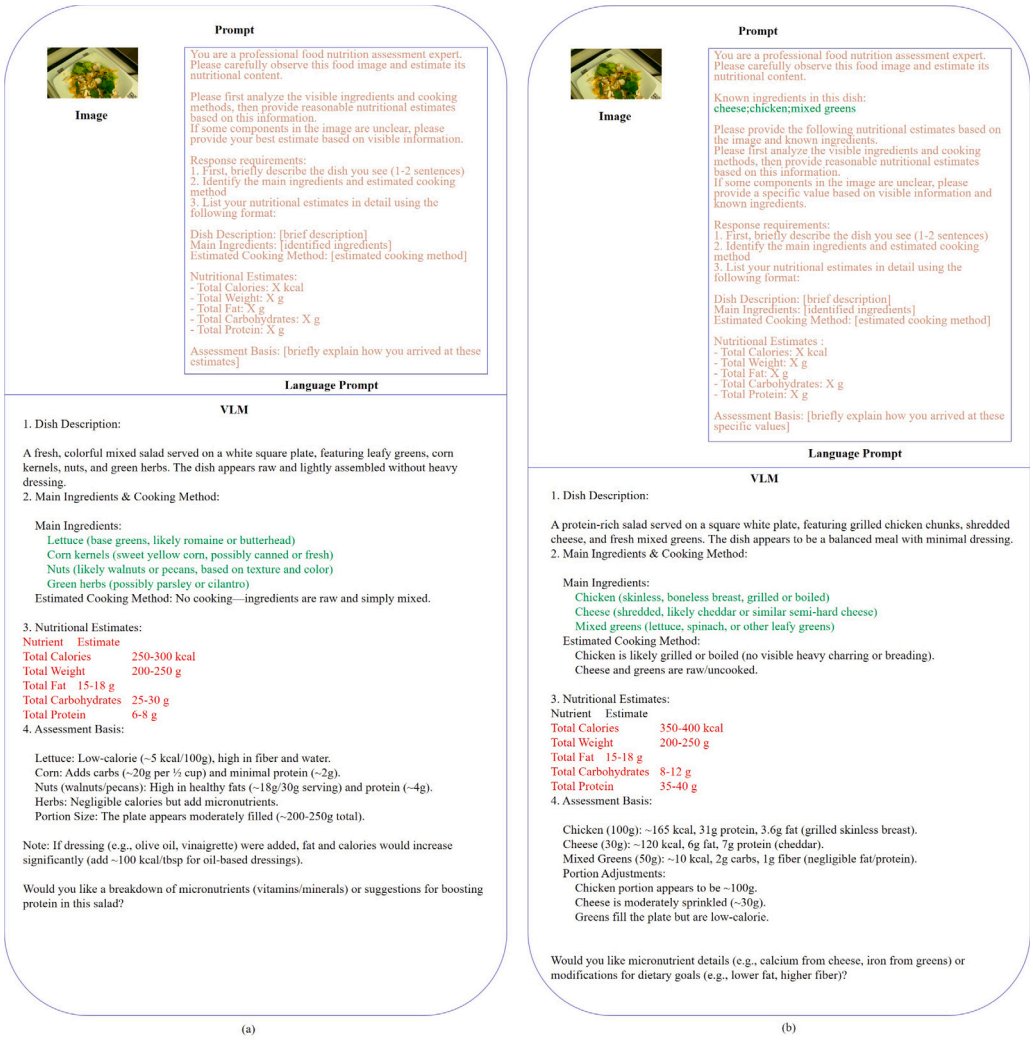


Fig. 3. Representative examples of VLMs applied to nutrient estimation tasks. (a) Estimation based solely on food images, without access to ground truth ingredient information. (b) Estimation with explicit ground truth ingredient information provided alongside the images.

and multimodal inputs. It significantly improves accuracy in vision-question answering and cross-modal retrieval tasks. The architecture supports image inputs at arbitrary resolutions and aspect ratios, with optimizations for long-context scenarios (specific length undisclosed). Model outputs are delivered in textual format, making it suitable for dynamic interactive multimodal applications. As a lightweight variant within the Llama 4 series, it is differentiated from the larger-scale Behemoth model through architectural specialization.

Llama 3.2-Vision (Llama 3.2-Vision 11B, Llama 3.2-Vision 90B): A multimodal large language model (MLLM) released by Meta in September 2024, featuring two-scale variants with 11 billion parameters (11B) and 90 billion parameters (90B) (Meta AI, 2025a). This architecture is built on an optimized transformer framework with autoregressive language modeling capabilities, enabling context-aware text generation through next-token prediction. Instruction-following capabilities are enhanced via fine-tuning techniques. The model supports integrated text-image inputs with textual output generation, demonstrating superior performance in processing high-resolution images, complex charts, and multimodal content. It significantly improves execution accuracy in visual question answering, cross-modal retrieval tasks, and other vision–language applications.

LLaVA-1.5-13B: MLLM introduced by the Microsoft Research team in October 2023, evolved from the original LLaVA architecture (Liu et al., 2024). The model employs CLIP-ViT-L-336px as its vision encoder and incorporates a two-layer fully connected network (MLP) to replace the conventional linear projection layer, significantly enhancing vision–language alignment capabilities. Its core architecture consists of three components: (1) a vision encoder for image feature extraction, (2) a LLM based on the Vicuna/LLaMA architecture, and (3) a vision–language connector enabling cross-modal interaction. LLaVA-1.5 achieved state-of-the-art (SOTA) performance on 11 benchmarks, requiring only 1.2 million publicly available training samples to match GPT-4V's multimodal comprehension capabilities. Supporting high-resolution image inputs at 336px, it demonstrates superior effectiveness in complex tasks such as visual question answering (VQA), cross-modal retrieval, and other vision–language applications.

Gemma 3-27B: A MLLM released by Google in March 2025, developed based on the Gemini 2.0 architecture with parameter scales ranging from 1B to 27B. The 27B variant achieves efficient execution on a single GPU (Team et al., 2025). This model incorporates a custom SigLIP vision encoder (supporting 896 × 896 resolution image inputs) combined with early fusion technology for text-visual token integration, enabling multimodal reasoning across text, images, and short videos. It significantly enhances accuracy in visual question answering (VQA), cross-modal retrieval tasks, and other vision–language applications. Supporting context windows up to 128,000 tokens, the architecture accommodates arbitrary resolutions and aspect ratios for image processing, making it suitable for multi-turn dialogues and complex task scenarios. Additionally, Gemma 3-27B achieves state-of-the-art performance on benchmarks like MMLU-Pro (67.5 score) and

Table 2
Model performance comparison based on average evaluation metrics.

Model	Single View Camera A			Single View Camera B		
	Precision	Recall	F1	Precision	Recall	F1
GPT-4.1 mini	0.6413	0.5930	0.6037	0.6289	0.5790	0.5906
Gemini 2.5 Flash	0.6630	**0.6421**	**0.6398**	0.6513	**0.6286**	**0.6267**
Grok-2 Vision	0.6624	0.5935	0.6144	0.6391	0.5637	0.5875
Qwen-VL-Max	0.6360	0.5792	0.5945	0.6187	0.5534	0.5724
Qwen2.5-VL-72B	0.6420	0.5764	0.5963	0.6250	0.5507	0.5740
Qwen2.5-VL-32B	0.5728	0.5936	0.5700	0.5409	0.5693	0.5409
Doubao-1.5-vision-pro	0.6731	0.5967	0.6182	0.6603	0.5811	0.6044
Llama 4 Maverick	0.6460	0.5999	0.6094	0.6267	0.5756	0.5876
Llama 3.2-vision 11B	0.6061	0.5693	0.5700	0.5912	0.5519	0.5539
Llama 3.2-vision 90B	**0.6829**	0.5984	0.6245	**0.6791**	0.5883	0.6173
LLaVA-1.5 13B	0.4116	0.4371	0.4070	0.3996	0.4180	0.3920
Gemma 3 27B	0.5531	0.6066	0.5607	0.5153	0.5723	0.5252
Phi-4-multimodal	0.4726	0.3801	0.4084	0.4468	0.3480	0.3790
Mistral Small 3.1	0.5277	0.5013	0.5034	0.4928	0.4699	0.4698
Pixtral 12B	0.6063	0.5307	0.5543	0.5704	0.4873	0.5136
DeepSeek-VL2	0.4830	0.4469	0.4536	0.4677	0.4247	0.4349
InternVL3-14B	0.5956	0.5479	0.5587	0.5868	0.5344	0.5474
Model	Single View Camera C			Single View Camera D		
	Precision	Recall	F1	Precision	Recall	F1
GPT-4.1 mini	0.6293	0.5807	0.5915	0.6449	0.5959	0.6063
Gemini 2.5 Flash	0.6423	**0.6232**	**0.6194**	0.6681	**0.6450**	**0.6434**
Grok-2 Vision	0.6307	0.5626	0.5833	0.6639	0.5931	0.6143
Qwen-VL-Max	0.6186	0.5584	0.5751	0.6364	0.5751	0.5924
Qwen2.5-VL-72B	0.6296	0.5586	0.5805	0.6461	0.5765	0.5978
Qwen2.5-VL-32B	0.5439	0.5768	0.5461	0.5649	0.5878	0.5632
Doubao-1.5-vision-pro	0.6536	0.5774	0.5990	0.6727	0.5921	0.6159
Llama 4 Maverick	0.6195	0.5732	0.5830	0.6513	0.5977	0.6105
Llama 3.2-vision 11B	0.5867	0.5464	0.5502	0.6055	0.5635	0.5683
Llama 3.2-vision 90B	**0.6626**	0.5774	0.6034	**0.6967**	0.6061	0.6347
LLaVA-1.5 13B	0.4056	0.4344	0.4025	0.4098	0.4282	0.4011
Gemma 3 27B	0.5137	0.5786	0.5271	0.5411	0.6020	0.5523
Phi-4-multimodal	0.4374	0.3440	0.3731	0.4673	0.3699	0.4003
Mistral Small 3.1	0.4925	0.4756	0.4729	0.5184	0.4945	0.4951
Pixtral 12B	0.5700	0.4947	0.5179	0.6091	0.5268	0.5529
DeepSeek-VL2	0.4590	0.4210	0.4285	0.4747	0.4346	0.4429
InternVL3-14B	0.5838	0.5322	0.5454	0.6021	0.5498	0.5622

Average performance across all views

Model	Avg Precision	Avg Recall	Avg F1
GPT-4.1 mini	0.6361	0.5872	0.5980
Gemini 2.5 Flash	0.6562	**0.6347**	**0.6323**
Grok-2 Vision	0.6490	0.5782	0.5999
Qwen-VL-Max	0.6274	0.5665	0.5836
Qwen2.5-VL-72B	0.6357	0.5655	0.5872
Qwen2.5-VL-32B	0.5556	0.5819	0.5551
Doubao-1.5-vision-pro	0.6649	0.5868	0.6094
Llama 4 Maverick	0.6359	0.5866	0.5976
Llama 3.2-vision 11B	0.5974	0.5578	0.5606
Llama 3.2-vision 90B	**0.6803**	0.5926	0.6200
LLaVA-1.5 13B	0.4067	0.4294	0.4007
Gemma 3 27B	0.5308	0.5899	0.5413
Phi-4-multimodal	0.4560	0.3605	0.3902
Mistral Small 3.1	0.5078	0.4853	0.4853
Pixtral 12B	0.5889	0.5099	0.5347
DeepSeek-VL2	0.4711	0.4318	0.4400
InternVL3-14B	0.5921	0.5411	0.5534

features lightweight design optimizations for localized deployment on consumer-grade hardware such as smartphones and laptops.

Phi-4-Multimodal: A multimodal language model launched by Microsoft in February 2025, featuring 5.6 billion parameters (5.6B) and built upon the Phi-4-Mini language model (380 million parameters). The architecture employs a 32-layer Transformer framework and integrates text, vision, and speech/audio modality adapters through LoRA-hybrid (low-rank adaptation) technology, enabling cross-modal interaction (Abouelenin et al., 2025). This model supports multimodal inputs including text, images, and speech/audio, generating textual outputs for complex tasks such as visual question answering (VQA), image analysis, automatic speech recognition (ASR), translation, and summarization. It demonstrates exceptional performance particularly in ASR applications, with vision capabilities spanning high-resolution images, charts, and screenshots. Notably, it achieves multimodal joint processing without relying on standalone models or complex pipelines. As an edge-optimized small language model (SLM), Phi-4-Multimodal balances performance efficiency with compact design, making it ideal for resource-constrained multimodal applications.

Mistral Small 3.1: MLLM released by Mistral AI in March 2025 under an open-source license, featuring 24 billion parameters (24B) and built upon a Transformer-based Mixture of Experts (MoE) architecture. By dynamically activating sparse parameter subsets, the model enhances computational efficiency while supporting multilingual processing, long-context handling, and cross-modal reasoning with text-image inputs (Mistral AI, 2025). The architecture integrates a unified modality encoder and projection module, enabling advanced visual capabilities for processing high-resolution images, complex charts, and multimodal content. This results in significantly improved accuracy for visual question answering (VQA), cross-modal retrieval tasks, and other vision–language applications. Supporting context windows up to 128,000 tokens with output generation speed of 150 tokens per second, the model is optimized for deployment on consumer-grade hardware like single NVIDIA RTX 4090 GPUs. It demonstrates strong adaptability in edge computing, real-time interactions, and open-source research applications.

Pixtral 12B: The first MLLM introduced by Mistral AI in September 2024, featuring 12 billion parameters (12B) and built upon a 40-layer architecture with 14,336 hidden dimensions and 32 attention heads. This model integrates a specialized vision encoder for text-image fusion processing (Agrawal et al., 2024). It inherits the text-side capabilities from the Nemo 12B language model while supporting arbitrary-resolution image inputs (up to 1024 × 1024 pixels) through an advanced vision encoder. Capable of handling multimodal tasks involving multiple images and text, the model generates textual outputs. Its vision capabilities span natural image understanding and document parsing while preserving original resolution characteristics. Supporting interleaved multi-image inputs and complex vision–language reasoning, Pixtral 12B is applicable to scenarios such as visual question answering (VQA), cross-modal retrieval, and other vision–language applications.

DeepSeek-VL2: An open-source VLM with 4.5 billion parameters, released by DeepSeek in December 2024 (Wu et al., 2024). Built upon the Mixture-of-Experts (MoE) architecture, this model employs a dynamic chunked vision encoding strategy to improve computational efficiency while supporting arbitrary resolutions and extreme aspect ratios for image processing. It integrates the SigLIP-SO400M image encoder combined with local sub-image and global thumbnail segmentation strategies, significantly enhancing high-resolution image analysis capabilities. The architecture introduces novel functionalities including meme comprehension, visual localization, and visual storytelling generation, covering applications in visual question answering (VQA), optical character recognition (OCR), document/table/chart understanding, and other multimodal tasks.

InternVL3-14B: A 14-billion-parameter (14B) open-source MLLM released by the Shanghai Artificial Intelligence Laboratory in April 2025 (Zhu et al., 2025). As the 14B variant within the InternVL3 series, this architecture inherits the "ViT-MLP-LLM" paradigm from its predecessor, initializing with pre-trained ViT (vision encoder) and LLM components connected through randomly initialized MLP layers for cross-modal interaction, significantly reducing computational costs. The model demonstrates comprehensive visual capabilities spanning high-resolution images, complex charts, document scans, and video content processing, supporting arbitrary aspect ratio inputs. It achieves performance levels approaching those of InternVL2.5-78B and Gemini-2.5-Pro on multimodal reasoning benchmarks, particularly excelling in OCR tasks with superior processing capabilities for sensitive document scans.

Table 3
Incremental performance impact of multi-view inputs (Relative change compared to previous configuration).

Model	View Config	Precision	ΔP	Recall	ΔR	F1	ΔF1
GPT-4.1 mini	A	0.6636	-	0.5815	-	0.6066	-
	A+B	0.6592	-0.67%	0.6213	6.84%	0.6280	3.53%
	A+B+C	0.6546	-0.69%	0.6310	1.56%	0.6312	0.51%
	A+B+C+D	0.6508	-0.58%	0.6367	0.91%	0.6323	0.17%
Gemini 2.5 Flash	A	0.6824	-	0.5949	-	0.6180	-
	A+B	0.6778	-0.66%	0.6066	1.98%	0.6238	0.94%
	A+B+C	0.6700	-1.16%	0.6209	2.35%	0.6294	0.90%
	A+B+C+D	0.6564	-2.04%	0.6476	4.31%	0.6389	1.50%
Grok-2 Vision	A	0.6525	-	0.5516	-	0.5833	-
	A+B	0.5692	-12.76%	0.5016	-9.06%	0.5233	-10.29%
	A+B+C	0.5684	-0.14%	0.5110	1.89%	0.5289	1.07%
	A+B+C+D	0.5680	-0.07%	0.5220	2.14%	0.5355	1.26%
Qwen2.5-VL-72B	A	0.6276	-	0.5448	-	0.5712	-
	A+B	0.6268	-0.12%	0.5731	5.21%	0.5881	2.97%
	A+B+C	0.6229	-0.62%	0.5815	1.45%	0.5911	0.52%
	A+B+C+D	0.6224	-0.09%	0.5887	1.24%	0.5952	0.69%
Qwen-VL-Max	A	0.6264	-	0.5408	-	0.5683	-
	A+B	0.6388	1.99%	0.5692	5.26%	0.5902	3.85%
	A+B+C	0.6383	-0.08%	0.5820	2.25%	0.5975	1.24%
	A+B+C+D	0.6412	0.45%	0.5917	1.66%	0.6044	1.15%
Qwen2.5-VL-32B	A	0.5502	-	0.5105	-	0.5204	-
	A+B	0.3760	-31.66%	0.4548	-10.91%	0.3997	-23.20%
	A+B+C	0.4873	29.58%	0.5190	14.12%	0.4918	23.05%
	A+B+C+D	0.4898	0.53%	0.5236	0.88%	0.4957	0.79%
Doubao-1.5-vision-pro	A	0.6726	-	0.5638	-	0.5987	-
	A+B	0.6707	-0.28%	0.5961	5.72%	0.6189	3.39%
	A+B+C	0.6651	-0.83%	0.6080	2.00%	0.6236	0.75%
	A+B+C+D	0.6650	-0.02%	0.6190	1.80%	0.6297	0.99%
Llama 4 Maverick	A	0.6636	-	0.5563	-	0.5892	-
	A+B	0.6513	-1.86%	0.5904	6.13%	0.6068	2.99%
	A+B+C	0.6469	-0.67%	0.5984	1.35%	0.6098	0.49%
	A+B+C+D	0.6443	-0.41%	0.6042	0.97%	0.6115	0.29%
Phi-4-multimodal	A	0.4097	-	0.4122	-	0.3953	-
	A+B	0.3057	-25.37%	0.4254	3.20%	0.3399	-14.02%
	A+B+C	0.3346	9.46%	0.4172	-1.93%	0.3544	4.28%
	A+B+C+D	0.3522	5.25%	0.4165	-0.17%	0.3650	2.98%
Mistral Small 3.1	A	0.4500	-	0.4237	-	0.4234	-
	A+B	0.4496	-0.08%	0.4630	9.28%	0.4437	4.78%
	A+B+C	0.4545	1.08%	0.4782	3.28%	0.4544	2.41%
	A+B+C+D	0.4537	-0.17%	0.4869	1.83%	0.4575	0.69%
Pixtral 12B	A	0.5975	-	0.4959	-	0.5280	-
	A+B	0.5910	-1.09%	0.5348	7.84%	0.5491	3.98%
	A+B+C	0.5815	-1.61%	0.5354	0.11%	0.5447	-0.80%
	A+B+C+D	0.5435	-6.53%	0.5362	0.15%	0.5239	-3.81%
InternVL3-14B	A	0.6236	-	0.5238	-	0.5558	-
	A+B	0.6128	-1.74%	0.5464	4.32%	0.5663	1.88%
	A+B+C	0.6140	0.20%	0.5523	1.07%	0.5701	0.68%
	A+B+C+D	0.6138	-0.03%	0.5559	0.66%	0.5722	0.37%

3. Experimental results

This section evaluates the performance of VLMs in food image recognition and nutrient estimation tasks. It first analyzes classification performance under single-view settings while investigating how progressive multi-view (four food images) inputs enhance model capabilities. Subsequently, it assesses VLM performance in nutrient estimation tasks, specifically examining the impact of providing ground truth ingredient information on model performance.

3.1. Food ingredient recognition

Table 2 lists the precision, recall, and F1 scores of each VLM across four viewing angles, along with their average values. Overall, Llama-3.2-90B-Vision achieves the highest precision across all perspectives, ranging from 0.6626–0.6967, indicating its strongest precision in identifying specific ingredients. However, its relatively lower recall results in suboptimal F1 scores, suggesting that some ingredients in multi-class images were missed, thereby affecting its comprehensive performance.

Similarly, Doubao-1.5-Vision-Pro also demonstrates high precision but slightly lower recall. In contrast, Gemini-2.5-Flash exhibits marginally lower precision but achieves the highest recall and F1 scores across all viewing angles, reaching up to 0.6450 and 0.6434 respectively. This highlights its superior comprehensive ingredient coverage in multi-class, complex scenarios. Therefore, Llama-3.2-90B-Vision is better suited for scenarios with fewer ingredient categories where high precision is critical, while Gemini-2.5-Flash excels in real-world applications involving diverse ingredient varieties and demanding higher recall and overall recognition capabilities.

As shown in Table 3, under progressive image input settings, most VLMs exhibit a counterintuitive trend: while recognition precision slightly decreases with additional viewing angles, recall rates demonstrate significant improvement. For example, when evaluating a representative sample using Doubao-1.5-Vision-Pro, and the ground truth ingredients were cantaloupe, cherry tomatoes, cauliflower, and almonds, the model stably identified cauliflower and almonds under single-view (A) and dual-view (A+B) inputs, achieving an F1 score of 0.711, though with occasional omissions in cantaloupe detection.

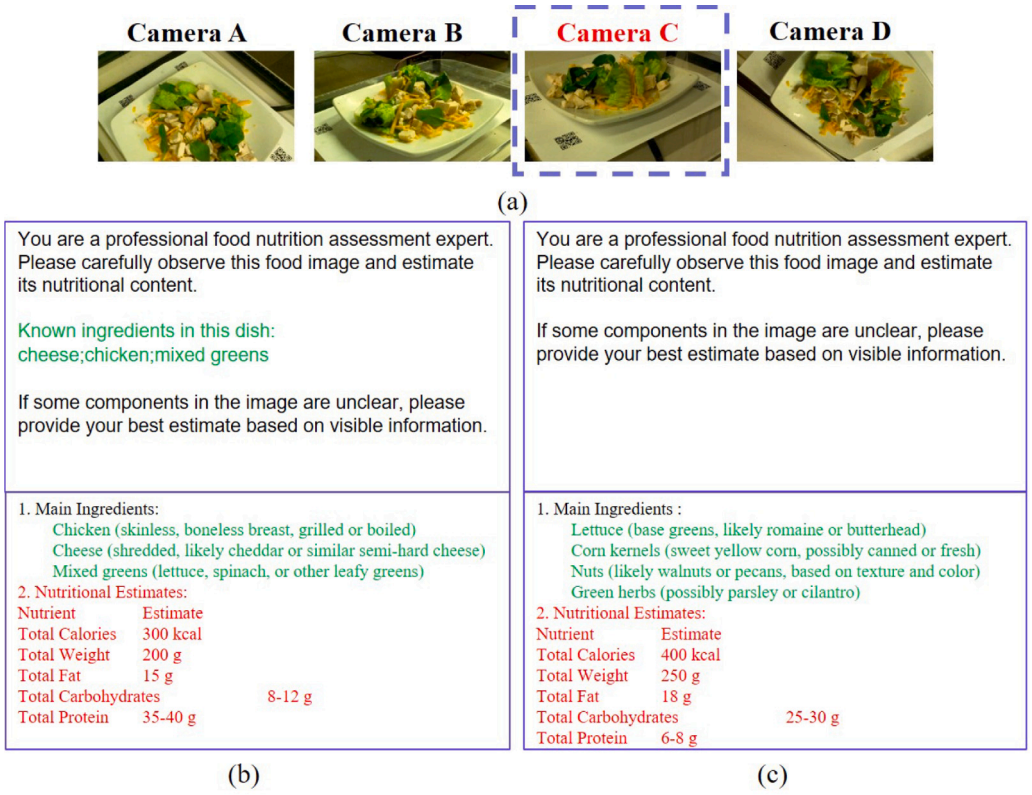


Fig. 4. Nutrient estimation results for a sample using View C images. (a) Multi-view sample images (View C used for estimation). (b) Estimation with ground-truth ingredient prompts. (c) Estimation without ingredient information.

As the number of input views increased (three views or more), recall improved marginally but introduced false positives for carrot identification, causing precision to decline slightly while maintaining stable F1 scores (0.698). This example aligns with the overall trend presented in the table, indicates that multi-view inputs enhance partial ingredient recall at the cost of introducing misclassifications, resulting in limited overall improvement in comprehensive performance (F1). The results reveal that most VLMs' ability to integrate multi-view information varies with their inherent capabilities: stronger models effectively leverage multi-view data for better recognition, whereas weaker models perceive additional views as noise that degrades performance.

3.2. Nutrient estimation

As shown in Table 4, most VLMs demonstrate improved performance when provided with ground truth ingredient labels compared to scenarios without such information, as exemplified by Qwen2.5-VL-32 and Gemma 3-27B. To validate the impact of ground truth ingredient assistance on nutrient estimation, comparative analysis was conducted using Doubao-1.5-Vision-Pro under two experimental settings: (1) with explicit ingredient labels and (2) without ingredient information. Specifically, in the informed condition, structured ground truth ingredient labels (e.g., "cheese; chicken; mixed greens") were explicitly included in model inputs. The generated reasoning text not only achieved comprehensive reconstruction of all input ingredients but also conducted systematic nutritional analysis based on these ingredients. For instance, the model output demonstrated in Fig. 4, the aforementioned reasoning content indicates that the model can fully "read" and utilize structured ground truth ingredient information, with its reasoning process heavily relying on these prior knowledge inputs. The itemized basis for nutrient estimation demonstrates high consistency with the provided ingredient data. In contrast, when ground truth ingredient information is withheld, the model solely depends on autonomous image-based ingredient identification, leading to potential omissions or misidentifications in the generated reasoning text. The model output was shown in Fig. 4, as observed, in the absence of real ingredient information, the model fails to correctly identify most ingredients, with nutritional estimation relying more on visual features and model-internal empirical reasoning. Results indicate that under informed conditions, the model's reasoning content achieves high consistency with ground truth labels, effectively leveraging prior knowledge for nutritional estimation. However, when integrating ground truth ingredient information, nutrient estimation errors did not significantly decrease and, in some cases, even increased, highlighting notable limitations in current mainstream VLMs' nutrient estimation capabilities.

As shown in Table 5, different VLMs exhibit complex differentiation patterns in nutrient estimation errors after integrating structured ground truth ingredient information. While models like Llama 4 Maverick demonstrate comprehensive optimization capabilities, achieving significant error reduction across all nutrient categories — most mainstream models show contrasting performance characteristics: Qwen2.5-VL-72B experiences error expansion in five metrics including mass (−6.21%) and protein (−10.41%); DeepSeek-VL2's caloric content (−68.94%) and protein (−136.58%) errors exhibit extreme deterioration, revealing that structured information exacerbates systematic bias. Some models display contradictory outcomes. For example, Doubao-1.5-Vision-Pro improves fat estimation (+46.71%) yet sees substantial error surges in mass (−27.23%) and carbohydrates (−15.51%).

From the nutrient dimension perspective, structured information exhibits distinct specificity in its impact. Fat estimation benefits most universally, with over half of the models including Gemini 2.5 Flash (+15.65%) and Qwen-VL-Max (+9.88%) achieving error reductions exceeding 5%, potentially attributable to strong correlations between fat content and ingredient categories. Protein estimation, however, demonstrates polarized outcomes: Grok-2 Vision (+10.98%) and InternVL3-14B (+11.24%) achieve optimization, while DeepSeek-VL2 (−136.58%)

Table 4
Nutrient estimation performance: Comparison between unlabeled and labeled ingredient inputs.

	MAE			RelErr (%)		
Model	w/o Ingre.	w/Ingre.	Δ	w/o Ingre.	w/Ingre.	Δ
GPT-4.1 mini	39.20	41.01	-4.41%	119.15	104.26	+14.28
Gemini 2.5 Flash	45.55	44.12	+3.24%	161.19	138.95	+16.00
Grok-2 Vision	47.91	46.97	+2.00%	159.77	140.42	+13.78
Qwen-VL-Max	55.84	57.34	-2.62%	179.39	165.23	+8.57
Qwen2.5-VL-72B	47.95	49.36	-2.85%	156.00	165.43	-5.70
Qwen2.5-VL-32B	54.88	53.19	+3.18%	251.56	163.29	+54.06
Doubao-1.5-vision-pro	38.01	41.98	-9.46%	99.42	94.25	+5.49
Llama 4 Maverick	57.60	52.11	+10.53%	221.13	176.18	+25.51
Llama 3.2-vision 11B	63.84	55.70	+14.61%	217.24	194.09	+11.93
Llama 3.2-vision 90B	59.35	51.09	+16.13%	127.40	121.97	+4.45
LLaVA-1.5 13B	82.88	74.87	+10.68%	475.23	637.78	-25.49
Gemma 3 27B	84.31	65.75	+28.23%	513.56	293.24	+75.13
Phi-4-multimodal	92.73	125.12	-25.90%	352.71	356.82	-1.15
Mistral Small 3.1	71.84	67.42	+6.55%	310.55	224.18	+38.52
Pixtral 12B	57.53	59.04	-2.56%	259.34	278.43	-6.86
DeepSeek-VL2	102.99	82.72	+24.53%	188.11	368.37	-48.93
InternVL3-14B	44.64	47.30	-5.62%	160.81	151.21	+6.35

and Phi-4 (−57.19%) experience catastrophic error escalation. In contrast, caloric content and carbohydrate estimation exhibit minimal error fluctuations, with InternVL3-14B showing only slight carbohydrate error adjustment (+4.52%). Mass estimation errors concentrate within moderate ranges, featuring both positive optimization in Llama 3.2-Vision 90B (+8.86%) and significant deterioration in Pixtral 12B (−12.85%). Overall, the impact of ground-truth ingredient information on nutrient estimation varies with model architecture and nutrient type. Superior models achieve comprehensive performance improvements by effectively leveraging additional ingredient prompts, while a minority of mainstream models expose significant integration bottlenecks in information processing.

Our analysis reveals a notable synchrony in the trends between nutrient estimation accuracy and the performance of ingredient recognition and mass estimation. By examining both the quantitative results and the qualitative experiments, we find that models with stronger recognition capabilities, such as GPT-4.1 mini and Doubao-1.5-vision-pro, generally achieve better nutrient estimation outcomes, whereas models with weaker recognition performance, including DeepSeek-VL2 and Phi-4-multimodal, do not exhibit meaningful improvement even when provided with ground-truth ingredient labels. Table 5 further indicates that supplying correct ingredient categories primarily benefits fat and protein estimation, suggesting that part of the performance gap originates from foods rich in these macronutrients. In addition, the error patterns in estimated food mass reveal a clear correlation between mass estimation accuracy and downstream nutrient predictions: models with smaller mass errors tend to produce more reliable calorie and macronutrient estimates, while those with larger mass errors, such as Phi-4 Multimodal and LLaVA-1.5-13B, show consistently higher errors across multiple nutrient dimensions. Taken together, the results suggest that VLMs with superior ingredient recognition and mass estimation accuracy generally demonstrate relative advantages in nutrient estimation. Evidence from multi-view experiments and ingredient-guided evaluations, together with qualitative observations, further implies that better-performing models are more effective at integrating heterogeneous information for nutritional reasoning.

4. Discussion

VLMs show great promise in food ingredient recognition, especially in open-domain tasks without fine-tuning. Our study introduces a progressive multi-view fusion strategy, demonstrating that multiple food image perspectives improve ingredient recall, which helps with challenges like occlusion. We also explored how ground-truth ingredient labels affect nutrient estimation, an underexplored area. We found that despite VLMs accurately identifying ingredients, they struggle with precise nutrient estimation even with correct ingredient data. This highlights a core gap between their visual understanding and nutritional reasoning; even with accurate ingredient input, calorie/nutrient calculations are often erroneous. Another contributing factor may be the internal nutrition knowledge or databases implicitly relied upon by different models, which are not explicitly aligned with standardized nutritional references. As demonstrated in Section Experimental Results, some models exhibit increased estimation errors after receiving correct ingredient labels. This suggests that accurate ingredient identification alone is insufficient for reliable nutritional reasoning. Fig. 4 further illustrates this behavior, where the model's responses change dramatically after being provided with ground-truth ingredients, often shifting from coarse, visually driven estimates to overconfident yet quantitatively inaccurate predictions. These observations indicate that current VLMs lack a robust mechanism for integrating ingredient identity, portion size, and nutritional knowledge into a coherent estimation process, highlighting a fundamental limitation in their visual–nutritional reasoning capability.

In practical applications, the usability of VLMs depends not only on model performance but also on computational efficiency. As shown in Table 6, the tested models exhibit clear differences in inference latency and token consumption: lightweight models such as GPT-4.1 mini and Llama 4 Maverick achieve the fastest responses (around 5–6 s) with relatively low output token counts, whereas Qwen-VL-Max requires more than 15 s despite producing short outputs, and Gemini 2.5 Flash generates substantially longer responses that increase overall token cost. From a deployment perspective, models that combine low latency with small token footprints offer more advantageous cost–performance trade-offs, particularly for mobile or resource-limited dietary assessment applications.

It is also important to note that this study evaluates VLMs primarily using the Nutrition5K dataset, which predominantly reflects Western dietary patterns and ingredient distributions. As a result, the benchmark does not cover a wide range of regional cuisines, such as East Asian, South Asian, or Middle Eastern foods, which often involve different ingredient vocabularies, cooking styles, and compositional structures. Moreover, contemporary VLMs are trained on large-scale multimodal corpora with varying data sources and regional biases, which may lead to model-specific preferences toward certain food types or culinary contexts. For example, models developed and trained predominantly in East Asia, such as DeepSeek or Qwen, may implicitly encode richer knowledge of East Asian culinary traditions, ingredient naming conventions, and preparation styles, whereas Western-developed models such as GPT-4.1 mini may demonstrate stronger familiarity with Western dishes and ingredient combinations. These

Table 5
Comparison of relative errors across specific nutrients.

	Mass (%)			Calories (%)			Carb (%)			Fat (%)			Protein (%)		
Model	w/o Ingre.	w/Ingre.	Δ	w/o Ingre.	w/Ingre.	Δ	w/o Ingre.	w/Ingre.	Δ	w/o Ingre.	w/Ingre.	Δ	w/o Ingre.	w/Ingre.	Δ
GPT-4.1 mini	42.54	44.63	-4.68	77.00	82.19	-6.31	101.79	98.43	+3.41	287.95	224.09	+28.47	86.45	71.94	+20.17
Gemini 2.5 Flash	47.12	49.31	-4.45	92.57	89.04	+3.96	90.07	76.87	+17.17	482.32	417.05	+15.65	93.85	62.51	+50.14
Grok-2 Vision	71.14	68.14	+4.40	117.40	115.11	+1.99	152.43	149.12	+2.22	352.97	275.20	+28.29	104.93	94.53	+10.98
Qwen-VL-Max	77.85	76.96	+1.16	116.23	121.42	-4.27	176.80	151.52	+16.70	399.84	363.90	+9.88	126.25	112.36	+12.37
Qwen2.5-VL-72B	61.50	65.58	-6.21	100.33	113.32	-11.45	158.96	164.03	-3.09	346.93	358.86	-3.33	112.28	125.35	-10.41
Qwen2.5-VL-32B	67.12	68.01	-1.31	127.62	115.93	+10.06	175.01	137.83	+26.94	734.32	391.91	+87.36	153.74	102.76	+49.59
Doubao-1.5-vision-pro	44.01	60.46	-27.23	65.95	78.28	-15.75	90.42	107.02	-15.51	222.71	151.81	+46.71	74.02	73.68	+0.47
Llama 4 Maverick	99.93	89.97	+11.07	146.48	135.30	+8.27	222.78	175.08	+27.24	492.49	363.93	+35.30	143.95	116.60	+23.46
Llama 3.2-vision 11B	101.22	79.59	+27.17	149.84	140.49	+6.66	177.20	236.16	-24.97	489.28	319.32	+53.23	168.63	194.90	-13.48
Llama 3.2-vision 90B	72.91	66.97	+8.86	97.20	106.28	-8.54	116.85	127.79	-8.56	240.82	192.86	+24.87	109.22	115.96	-5.81
LLaVA-1.5 13B	119.19	117.93	+1.05	200.54	301.14	-50.16	419.37	659.76	-57.32	1261.70	1498.68	-18.78	375.36	611.41	-62.89
Gemma-3 27B	112.75	97.22	+13.76	287.14	203.24	+29.22	341.77	260.70	+23.72	1511.46	724.56	+52.06	294.69	179.24	+39.18
Phi-4 Multimodal	107.83	137.19	-27.25	265.43	279.20	-5.19	307.19	321.03	-4.50	860.18	696.24	+19.07	222.93	350.44	-57.19
Mistral Small 3.1	118.56	120.07	-1.27	173.22	168.61	+2.66	255.98	218.27	+14.74	758.39	447.90	+40.95	246.61	166.07	+32.65
Pixtral 12B	79.75	90.00	-12.85	149.64	223.92	-49.64	247.43	334.55	-35.23	605.66	506.70	+16.35	214.20	236.98	-10.63
DeepSeek-VL2	93.10	121.14	-30.14	193.07	326.22	-68.94	187.64	391.38	-108.56	331.19	682.37	-105.97	135.53	320.74	-136.58
InternVL3-14B	50.60	67.98	-25.57	101.65	117.85	-13.75	151.71	145.14	+4.52	374.94	312.58	+19.94	125.13	112.47	+11.24

Table 6
Comparison of average inference speed and cost of some models.

Model	Inference speed (s)	Prompt token cost	Output token cost
GPT-4.1 mini	5.835	10 180	28
Gemini 2.5 Flash	8.102	7288	2227
Qwen-VL-Max	15.778	5068	17
Qwen2.5-VL-72B	5.689	8793	24
Qwen2.5-VL-32B	6.362	8452	89
Llama 4 Maverick	5.533	5068	12

culturally shaped knowledge distributions can influence recognition behavior and nutrient estimation outcomes, contributing to performance differences that are not solely attributable to model architecture or prompting strategy. Consequently, the performance observed in this study may not fully generalize to food ingredient recognition and nutrient estimation tasks involving non-Western cuisines. This limitation suggests that the results should be interpreted as a baseline evaluation under Western-centric dietary distributions, and highlights the need for future studies to systematically assess VLM generalization across diverse cultural and regional food domains.

Furthermore, we focused on the overall performance of mainstream VLMs, not deeply analyzing the specific impact of model architectures or parameter sizes. While we tested various representative models (e.g., Transformer-based and MoE-based models), our selection does not cover the full diversity of VLM designs. Future work should include a broader range of models and systematically analyze how network structure, cross-modal interaction, and pre-training objectives affect model performance, offering insights for more robust systems.

From a food science and nutrition perspective, estimation errors in different nutritional components have distinct implications for dietary assessment and health interpretation. Large errors in total food weight primarily reflect unreliable portion size estimation, which is a major source of uncertainty in dietary intake assessment and can systematically bias downstream calorie and nutrient calculations. Inaccurate calorie estimation may lead to misjudgment of overall energy intake, which is particularly relevant for weight management and metabolic health monitoring. Errors in macronutrient estimation, such as fat, carbohydrate, or protein, can further distort the evaluation of dietary composition, potentially affecting assessments related to glycemic control, cardiovascular risk, or protein adequacy. Therefore, while current VLMs demonstrate promising capabilities in food ingredient recognition, their nutrient estimation outputs should be interpreted with caution and are more suitable for coarse-grained dietary monitoring rather than precise or clinical-grade nutritional assessment.

Although the experimental setup in this study partially reflects real-world usage scenarios, it is important to further discuss how VLM-based dietary assessment may be applied in everyday contexts. In practical settings such as daily food logging, users typically capture only a single image of their meal, and capturing multiple viewpoints is often inconvenient. Acquiring several images increases the time and effort required from users and also leads to higher computational and inference costs for VLMs, which may limit the practicality of multi-view approaches in mobile or time-sensitive applications. Recent advances by model providers have led to the deployment of lightweight and efficient VLM-based applications, enabling users to access these models directly on mobile devices such as smartphones without substantial deployment overhead. In such cases, users can simply activate the camera and capture a single image of their meal to obtain relevant dietary information or recommendations. This significantly lowers the barrier to use and highlights the potential of VLMs as practical assistive tools for real-world dietary assessment under single-view constraints. In addition, most of the tested VLMs are not open-source, and future progress will be greatly facilitated if more high-performance models become openly available for fine-tuning in nutrition-related applications, enabling more accurate, adaptable, and domain-specific nutrient estimation systems.

Despite these limitations, VLMs offer significant opportunities in automated dietary assessment, particularly for the demands of ease of use, real-time capability, and scalability in mobile health applications. Although most current models are still unstable in nutrient estimation, and high-performance models often incur high deployment costs, making them unsuitable for standalone use in clinical settings, they show promising application prospects as assistive tools in daily dietary tracking, personalized dietary recommendations, and health management.

5. Conclusion

This study systematically evaluated the nutrition assessment capabilities of representative mainstream VLMs across two core tasks: food ingredient recognition and nutrient estimation, without any task-specific fine-tuning. Most VLMs exhibited promising performance in food ingredient recognition, with notable improvements observed when incorporating multi-view image inputs. However, despite generating seemingly plausible reasoning in nutrient estimation, their actual performance remains suboptimal. Even with the provision of ground truth ingredient information, performance improvements were modest, indicating that current VLMs are not yet reliable for standalone nutritional assessment based solely on visual–textual inputs. Therefore, outputs from these models should be interpreted as auxiliary references rather than definitive conclusions. Future research focuses on enhancing input modalities, such as incorporating depth information and structured recipes, and improving the models' multimodal reasoning and integration capabilities to achieve more accurate and trustworthy assessments.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al., 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. arXiv preprint arXiv:2503.01743.

Adilmetova, G., Nassyrov, R., Meyerbekova, A., Karabay, A., Varol, H.A., Chan, M.Y., 2025. Evaluating ChatGPT's multilingual performance in clinical nutrition advice using synthetic medical text: insights from central Asia. J. Nutr. 155 (3), 729–735.

Agrawal, P., Antoniak, S., Hanna, E.B., Bout, B., Chaplot, D., Chudnovsky, J., Costa, D., De Monicault, B., Garg, S., Gervet, T., et al., 2024. Pixtral 12b. arXiv preprint arXiv:2410.07073.

Alkhalaf, M., Shen, J., Chang, H.C., Deng, C., Yu, P., 2024. Fine-tuning large language models for effective nutrition support in residential aged care: a domain expertise approach. MedRxiv, 2024-2007.

Almoselhy, R.I., Usmani, A., 2024. AI in food science: Exploring core elements, challenges, and future directions. Challenges, and Future Directions (December 12, 2024).

Azimi, I., Qi, M., Wang, L., Rahmani, A.M., Li, Y., 2025. Evaluation of LLMs accuracy and consistency in the registered dietitian exam through prompt engineering and knowledge retrieval. Sci. Rep. 15 (1), 1506.

Bagler, G., Goel, M., 2024. Computational gastronomy: capturing culinary creativity by making food computable. NPJ Syst. Biology Appl. 10 (1), 72.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al., 2025. Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923.

Bergling, K., Wang, L.C., Shivakumar, O., Nandorine Ban, A., Moore, L.W., Ginsberg, N., Kooman, J., Duncan, N., Kotanko, P., Zhang, H., 2025. From bytes to bites: application of large language models to enhance nutritional recommendations. Clin. Kidney J. 18 (4), sfaf082.

Bodnar, L.M., Cartus, A.R., Kirkpatrick, S.I., Himes, K.P., Kennedy, E.H., Simhan, H.N., Grobman, W.A., Duffy, J.Y., Silver, R.M., Parry, S., et al., 2020. Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes. Am. J. Clin. Nutr. 111 (6), 1235–1243.

Bossard, L., Guillaumin, M., Van Gool, L., 2014. Food-101–mining discriminative components with random forests. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. Springer, pp. 446–461.

ByteDance, 2025. Doubao-1.5-pro. URL https://seed.bytedance.com/en/special/doubao_1_5_pro. (Accessed 04 April 2025).

Deng, X., Cao, S., Horn, A.L., 2021. Emerging applications of machine learning in food safety. Annu. Rev. Food Sci. Technol. 12 (1), 513–538.

Google Cloud, 2025. Gemini 2.5 flash: Generative AI on vertex AI. URL https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash. (Accessed 04 April 2025).

Haman, M., Školník, M., Lošťák, M., 2024. AI dietician: Unveiling the accuracy of ChatGPT's nutritional estimations. Nutrition 119, 112325.

Hu, G., Ahmed, M., L'Abbé, M.R., 2023. Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods. Am. J. Clin. Nutr. 117 (3), 553–563.

Keller, M., Tai, C.e.A., Chen, Y., Xi, P., Wong, A., 2024. NutritionVerse-direct: exploring deep neural networks for multitask nutrition prediction from food images. arXiv preprint arXiv:2405.07814.

Khamesian, S., Arefeen, A., Carpenter, S.M., Ghasemzadeh, H., 2025. NutriGen: Personalized meal plan generator leveraging large language models to enhance dietary and nutritional adherence. arXiv preprint arXiv:2502.20601.

Khlaisamniang, P., Kerdthaisong, K., Vorathammathorn, S., Yongsatianchot, N., Phimsiri, H., Chinkamol, A., Thitseesaeng, T., Veerakanjana, K., Kachai, K., Ittichaiwong, P., et al., 2025. Decomposing food images for better nutrition analysis: A nutritionist-inspired two-step multimodal LLM approach. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 482–491.

Kopitar, L., Bedrac, L., Strath, L.J., Bian, J., Stiglic, G., 2024. Identifying and decomposing compound ingredients in meal plans using large language models. arXiv preprint arXiv:2411.05892.

Kumar, V., Shukla, A.K., Sharma, P., Choudhury, B., Singh, P., Kumar, S., et al., 2017. Role of macronutrient in health. World J. Pharm. Res. 6 (3), 373–381.

Li, T., Wei, W., Xing, S., Min, W., Zhang, C., Jiang, S., 2023. Deep learning-based near-infrared hyperspectral imaging for food nutrition estimation. Foods 12 (17), 3145.

Liu, H., Li, C., Li, Y., Lee, Y.J., 2024. Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306.

Lo, F.P.W., Qiu, J., Wang, Z., Chen, J., Xiao, B., Yuan, W., Giannarou, S., Frost, G., Lo, B., 2024. Dietary assessment with multimodal ChatGPT: a systematic analysis. IEEE J. Biomed. Health Informatics.

Ma, P., Hong, H.C., Jia, X., Chi, C.J., Xiao, N., Fan, B., Wang, F., Wei, C.I., 2025. Structure from motion-convolutional neural network model (sfm-CNN) achieved accurate portable Chinese dietary chemical composition estimation for dietary recall. Food Chem. 144908.

Ma, P., Tsai, S., He, Y., Jia, X., Zhen, D., Yu, N., Wang, Q., Ahuja, J.K., Wei, C.I., 2024. Large language models in food science: Innovations, applications, and future. Trends Food Sci. Technol. 104488.

Meta AI, 2025a. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. (Accessed 05 April 2025).

Meta AI, 2025b. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. (Accessed 05 April 2025).

Min, W., Wang, Z., Liu, Y., Luo, M., Kang, L., Wei, X., Wei, X., Jiang, S., 2023. Large scale visual food recognition. IEEE Trans. Pattern Anal. Mach. Intell. 45 (8), 9932–9949.

Mistral AI, 2025. Mistral small 3.1. URL https://mistral.ai/news/mistral-small-3-1. (Accessed 18 March 2025).

Niszczota, P., Rybicka, I., 2023. The credibility of dietary advice formulated by ChatGPT: Robo-diets for people with food allergies. Nutrition 112, 112076.

OpenAI, 2025. Introducing GPT-4.1 in the API. URL https://openai.com/index/gpt-4-1/. (Accessed 04 April 2025).

Papastratis, I., Konstantinidis, D., Daras, P., Dimitropoulos, K., 2024. AI nutrition recommendation using a deep generative model and ChatGPT. Sci. Rep. 14 (1), 14620.

Parinayok, S., Yamakata, Y., Aizawa, K., 2023. Open-vocabulary segmentation approach for transformer-based food nutrient estimation. In: Proceedings of the 5th ACM International Conference on Multimedia in Asia. pp. 1–7.

Qi, Z., Yu, Y., Tu, M., Tan, J., Huang, Y., 2023. Foodgpt: A large language model in food testing domain with incremental pre-training and knowledge graph prompt. arXiv preprint arXiv:2308.10173.

Qwen Team, 2025. Introducing qwen-VL. URL https://qwenlm.github.io/zh/blog/qwen-vl/. (Accessed 04 April 2025).

Sahoo, D., Hao, W., Ke, S., Xiongwei, W., Le, H., Achananuparp, P., Lim, E.P., Hoi, S.C., 2019. FoodAI: Food image recognition via deep learning for smart food logging. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 2260–2268.

Shao, W., Min, W., Hou, S., Luo, M., Li, T., Zheng, Y., Jiang, S., 2023. Vision-based food nutrition estimation via RGB-d fusion network. Food Chem. 424, 136309.

Shi, Y., Ren, P., Wang, J., Han, B., ValizadehAslani, T., Agbavor, F., Zhang, Y., Hu, M., Zhao, L., Liang, H., 2023. Leveraging GPT-4 for food effect summarization to enhance product-specific guidance development via iterative prompting. J. Biomed. Informatics 148, 104533.

Shim, J.S., Oh, K., Kim, H.C., 2014. Dietary assessment methods in epidemiologic studies. Epidemiology Health 36, e2014009.

Szymanski, A., Wimer, B.L., Anuyah, O., Eicher-Miller, H.A., Metoyer, R.A., 2024. Integrating expertise in llms: crafting a customized nutrition assistant with refined template instructions. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. pp. 1–22.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al., 2025. Gemma 3 technical report. arXiv preprint arXiv:2503.19786.

Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., Sim, J., 2021. Nutrition5k: Towards automatic nutritional understanding of generic food [dataset]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8903–8911.

Theodore Armand, T.P., Nfor, K.A., Kim, J.I., Kim, H.C., 2024. Applications of artificial intelligence, machine learning, and deep learning in nutrition: a systematic review. Nutrients 16 (7), 1073.

Vasiloglou, M.F., Marcano, I., Lizama, S., Papathanail, I., Spanakis, E.K., Mougiakakou, S., 2023. Multimedia data-based mobile applications for dietary assessment. J. Diabetes Sci. Technol. 17 (4), 1056–1065.

Wang, W., Min, W., Li, T., Dong, X., Li, H., Jiang, S., 2022. A review on vision-based analysis for automatic dietary assessment. Trends Food Sci. Technol. 122, 223–237.

Wang, L.C., Zhang, H., Ginsberg, N., Ban, A.N., Kooman, J.P., Kotanko, P., 2024. Application of ChatGPT to support nutritional recommendations for dialysis patients–A qualitative and quantitative evaluation. J. Ren. Nutr. 34 (6), 477–481.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al., 2024. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302.

xAI, 2025. Grok-2 beta release. URL https://x.ai/news/grok-2. (Accessed 04 April 2025).

Yang, Z., Khatibi, E., Nagesh, N., Abbasian, M., Azimi, I., Jain, R., Rahmani, A.M., 2024. ChatDiet: Empowering personalized nutrition-oriented food recommender chatbots through an LLM-augmented framework. Smart Health 32, 100465.

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

Zhou, P., Min, W., Fu, C., Jin, Y., Huang, M., Li, X., Mei, S., Jiang, S., 2024. FoodSky: A food-oriented large language model that passes the chef and dietetic examination. arXiv preprint arXiv:2406.10261.

Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al., 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479.