

Are Vision-Language Models Ready for Dietary Assessment? Exploring the Next Frontier in AI-Powered Food Image Recognition

Sergio Romero-Tapiador¹

Ruben Tolosana¹

Blanca Lacruz-Pleguezuelos²

Laura Judith Marcos-Zambrano²

Guadalupe X. Bazán²

Isabel Espinosa-Salinas²

Julian Fierrez¹

Javier Ortega-Garcia¹

Enrique Carrillo de Santa Pau²

Aythami Morales¹

¹Biometrics and Data Pattern Analytics Lab, Universidad Autonoma de Madrid, Madrid, Spain

²IMDEA Food, CEI UAM+CSIC, Madrid, Spain

Abstract

Automatic dietary assessment based on food images remains a challenge, requiring precise food detection, segmentation, and classification. Vision-Language Models (VLMs) offer new possibilities by integrating visual and textual reasoning. In this study, we evaluate six state-of-the-art VLMs (ChatGPT, Gemini, Claude, Moondream, DeepSeek, and LLaVA), analyzing their capabilities in food recognition at different levels. For the experimental framework, we introduce the FoodNExTDB, a unique food image database that contains 9,263 expert-labeled images across 10 categories (e.g., “protein source”), 62 subcategories (e.g., “poultry”), and 9 cooking styles (e.g., “grilled”). In total, FoodNExTDB includes 50k nutritional labels generated by seven experts who manually annotated all images in the database. Also, we propose a novel evaluation metric, Expert-Weighted Recall (EWR), that accounts for the inter-annotator variability. Results show that closed-source models outperform open-source ones, achieving over 90% EWR in recognizing food products in images containing a single product. Despite their potential, current VLMs face challenges in fine-grained food recognition, particularly in distinguishing subtle differences in cooking styles and visually similar food items, which limits their reliability for automatic dietary assessment. The FoodNExTDB database is publicly available at <https://github.com/AI4Food/FoodNExTDB>.

1. Introduction

Food is a fundamental source of energy for human life and plays a critical role in preventing chronic diseases. Maintaining a healthy diet has become increasingly challenging due to multiple factors, including food quality, dietary diversity, cooking styles, and nutrient absorption. An imbalance in these aspects can lead to suboptimal nutrition,

negatively impacting health. In recent decades, diets have shifted toward increased consumption of processed, high-calorie foods and reduced intake of fruits and vegetables, contributing to the rise of diet-related diseases [7, 13].

To counteract these trends, dietary guidelines such as the Mediterranean and Japanese diets have been widely promoted for their health benefits and balanced nutritional profiles [45, 52]. However, these guidelines provide general recommendations and require individualized monitoring to assess effectiveness. Traditional dietary assessment methods, such as 24-hour recall and food frequency questionnaires, often fail due to reliance on self-reporting, which can be tedious and prone to inaccuracies. Advancements in dietary assessment suggest that automatic and personalized nutrition may offer more effective solutions [20, 49].

Food computing has emerged as a rapidly growing field, applying computational techniques to various food-related domains [29]. The increasing digitalization of daily life has led to the availability of vast amounts of online food data, enabling the development of large-scale food datasets [37]. Moreover, advancements in artificial intelligence (AI) have facilitated the evaluation of nutritional content and the understanding of individual dietary habits, contributing to both health maintenance and disease prevention [40, 44]. machine learning (ML), particularly deep learning (DL), has demonstrated remarkable capabilities in food image recognition and classification tasks [23, 41]. However, these methods primarily focus on visual classification and struggle to extract contextual information, such as ingredient composition, preparation techniques, etc.

To address these limitations, foundation models have demonstrated impressive performance across a wide range of multimodal tasks [1, 3, 11]. Models such as ChatGPT and Gemini have achieved state-of-the-art results in applications such as food tracking, dietary recommendation, and food science research [21, 24, 34, 35, 49, 51]. More recently, Vision-Language Models (VLMs) have emerged as

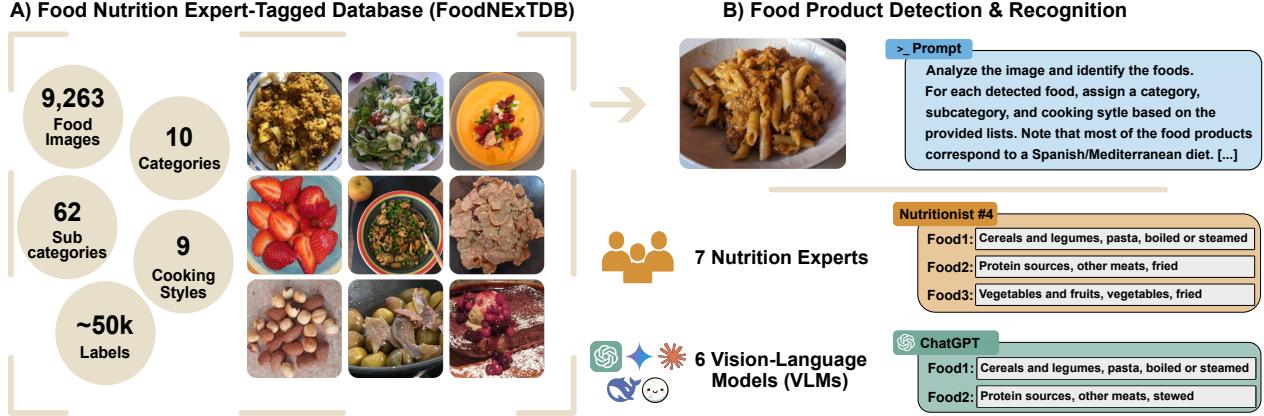


Figure 1. Overview of the proposed framework. (A) The FoodNExTDB consists of 9,263 food images labeled by nutrition experts across 10 food categories, 62 subcategories, and 9 cooking styles, with approximately 50,000 assigned labels. (B) The experimental setup, where six Vision-Language Models (VLMs) process food images using a structured prompt and generate predictions. These predictions are then compared against expert annotations to assess model performance in food product detection and recognition.

the next generation of explainable computer vision systems, integrating textual and visual data for a more comprehensive understanding of food-related information [18, 19, 43]. Despite their potential, several challenges still remain such as the reliability, interpretability, and accuracy of these models for food-related tasks [5]. This raises fundamental questions about their practical application in real-world nutritional assessments: *Are VLMs ready to assist nutrition experts in critical tasks such as supervised dietary quality assessment? Can they accurately recognize food products based on subtle factors like cooking styles, which can significantly impact nutritional values? Can current AI-based tools reliably evaluate an individual's dietary behavior solely through image uploads?*

Advancing knowledge around these questions, this paper has the following contributions: *i*) we introduce the Food Nutrition Expert-Tagged Database (FoodNExTDB), a unique food image database that contains 9,263 expert-labeled images, containing real dietary records from individuals who participated in a weight loss intervention [39], *ii*) we propose an experimental framework to assess state-of-the-art VLMs, analyzing their capabilities in food recognition at different levels and prompt comprehension, and *iii*) we design a novel evaluation metric, Expert-Weighted Recall (EWR), that accounts for inter-annotator variability. Differences in nutritional paradigms, dietary guidelines, and individual professional experiences can lead to inconsistencies in labeling the same food item. Also, cultural and regional differences further contribute to variations in annotation. The FoodNExTDB includes annotations from seven different nutrition experts (50,000 labels), representing a valuable contribution for the research community. Figure 1 illustrates the overall framework of the present study.

The remainder of this paper is structured as follows: Sec.

2 reviews key food image databases and recognition systems. Sec. 3 details the proposed FoodNExTDB. Sec. 4 presents the VLM models used in this study, while Sec. 5 outlines the experimental protocol and proposed metric. Sec. 6 analyzes the experimental results, followed by discussions and conclusions in Sec. 7 and 8, respectively.

2. Related Works

2.1. Food Image Databases

Food image databases are essential in training and evaluating automatic recognition models. These databases are typically acquired through three main methods: *i*) *self-collected*, where food images are manually captured in controlled or semi-controlled environments, *ii*) *web-scraped*, where large-scale databases are gathered from online platforms, and *iii*) *combination*, where multiple existing databases are merged to enhance diversity and coverage.

First, self-collected databases provide high-quality images but are limited in size due to manual acquisition. For instance, the UNIMIB2016 database includes around 20K images of various food products [10], while Nutrition5k offers 5K unique dishes with videos and nutritional data [47].

In contrast, web scraping allows for the collection of large-scale food image databases from social media and online sources. Examples include Food-101 [6] with 101,000 images spanning 101 food categories, VireFood-172, with 172 categories [8], and ISIA Food-500 [30], which covers 500 different food types.

Combination databases merge multiple sources to create more diverse and comprehensive databases. For example, Food-11 [46] integrates data from Food-101, UECFOOD-100 [27], and UECFOOD-256 [16], grouping food items into 11 main categories. MAFood-121 [2] focuses on global

cuisines, featuring 121 food products across 21,000 images. AI4FoodNutritionDB introduces a structured nutritional taxonomy, consisting of 553,000 images categorized into different nutritional levels, main categories, subcategories, and final food products [41]. Finally, Food-500 Cap enhances traditional food databases by incorporating 24,700 images with detailed captions that describe fine-grained visual attributes [26].

2.2. Food Image Recognition Models

Food image recognition has evolved significantly, transitioning from traditional classification techniques to DL approaches. Early methods struggled with complex and diverse food datasets, leading to the adoption of DL architectures such as convolutional neural networks (CNNs). These models integrated feature extraction and classification, achieving high accuracy rates exceeding 80% [28, 31]. Outstanding models include Squeeze-and-Excitation Networks (SENet) [14], Stacked Global-Local Attention Networks (SGLANet), and Progressive Region Enhancement Networks (PRENet), which demonstrated strong generalization across multiple food databases [30, 31]. Transfer learning and ensemble techniques further enhanced food classification performance, while object detection frameworks such as LOng-tailed FIne-Grained Network (LOFI) and YOLO improved food recognition [7, 23, 38].

The introduction of vision transformers (ViTs) marked a shift in food recognition by capturing global dependencies within images, enhancing classification accuracy [44]. However, ViTs alone required significant computational resources and struggled with fine-grained classification due to high intra-class variations in food appearance. As a result, hybrid architectures combining CNNs and ViTs emerged, leveraging the spatial awareness of CNNs and the contextual understanding of transformers [33].

In the present study, we explore the application of recent VLMs. Models such as CLIP align images and text in a shared multimodal space, enabling zero-shot classification and improved fine-grained food differentiation [25, 36]. Large multimodal models like FoodLLM and Large Language and Vision Assistants (LLaVA), specifically LLaVA-Chef, extend this capability by incorporating domain-specific knowledge about food ingredients, preparation methods, and cultural contexts. These models leverage multimodal prompting, combining textual and visual inputs to enhance classification accuracy, making them effective at extracting contextual information [17, 32, 50].

3. FoodNExTDB

The FoodNExTDB is a food image database derived from the AI4FoodDB, a comprehensive multimodal database acquired from a one-month randomized controlled trial (RCT) with 100 overweight and obese participants un-

dergoing a nutritional intervention [39]. Figure 1A) summarizes its key features, including food images collected over 14 days per participant. With many food products reflecting characteristics of Spanish cuisine, this database provides a valuable resource for studying food intake within a dietary pattern. The database is publicly available on GitHub¹.

3.1. Database Construction

Participants were instructed to capture images of all consumed foods and beverages using their smartphones. A total of 10,739 images were collected, of which approximately 14% were discarded during post-processing (e.g., due to non-food images, blurred images, etc.), resulting in a final database of 9,263 food images. Notably, around 88% of the images contain a valid timestamp, a key parameter for analyzing participants' eating behaviors.

Food images were annotated during the post-processing stage by a team of seven nutrition experts, ensuring that each image was reviewed by at least three annotators. This approach was implemented to enhance labeling reliability, given the varying complexity of food product recognition. While some food products are straightforward to classify, others require expert judgment due to their ambiguity.

The annotation process involved identifying the food items present in each image and categorizing them according to a predefined nutritional taxonomy. Each food product was assigned a *category*, *subcategory*, and *cooking style*. The taxonomy comprises 10 main food categories (e.g., “cereals and legumes”, “protein sources”, etc.), 62 subcategories (e.g., “alcoholic beverages”, “fruits”, etc.), and 9 cooking styles (e.g., “fried”, “boiled or stewed”, etc.). Additionally, free-text fields were provided to accommodate food products that did not fit into the predefined classifications. The complete list of *category*, *subcategory*, and *cooking style* classes is available in the supplementary material.

A custom Windows GUI was developed to streamline labeling while ensuring consistency. The database uniquely combines food categorization with culinary techniques, offering technical annotations and nutritional insights.

3.2. Database Statistics and Characteristics

On average, each participant captured approximately 96 food images, with a standard deviation of 58 images. Notably, ~20% of participants took fewer than 50 images, while ~15% captured more than 150. Regarding temporal distribution, most images (~79%) were taken during Spain's main daily meals, 1,836 images (~20%) at breakfast, 2,540 (~27%) at lunch, and 2,988 (~32%) at dinner.

In total, nutrition experts assigned over 50,000 labels. The three most frequently assigned food categories are “vegetables and fruits” (~28%), “cereals and legumes” (~17%), and “beverages” (~16%). At the subcategory

¹<https://github.com/AI4Food/FoodNExTDB>

level, they are “*vegetables*” (~13%), “*fruits*” (~13%), and “*bread*” (~8%). Finally, for cooking styles they are “*none*” (~31%), “*fresh*” (~28%), and “*boiled or steamed*” (~10%), remarking the prevalence of raw foods.

3.3. Problems Encountered During Labeling

The analysis of the annotations from different nutrition experts revealed inconsistencies in the classification of certain food products and cooking styles, highlighting a lack of consensus in specific cases. This variability can significantly impact dietary assessments, as cooking styles influence the nutritional quality of a meal.

Additionally, in many cases, the cooking style could not be reliably determined from the image alone, nor could certain food components like sauces or oils, which are often visually indistinguishable. Moreover, complex dishes with multiple food categories and subcategories posed a significant challenge, making labeling tedious and requiring expert judgment for consistency.

4. Proposed Methods

4.1. Selected Vision-Language Models

In this study, six state-of-the-art VLMs are explored: ChatGPT (GPT-4o), Gemini 2.0 Flash, Claude 3.5 Sonnet, Moondream, DeepSeek Janus-Pro, and LlaVA. The primary objective is to compare VLMs that are accessible through API on limited-resource environments.

ChatGPT, Gemini, and Claude are closed-source models with dedicated APIs, while DeepSeek, LlaVA, and Moondream are open-source models, which in this study are deployed using Replicate², a platform offering efficient AI infrastructure to open-source VLMs through an API. To ensure unbiased results, each model instance is initialized independently for each image.

Several parameters can be adjusted by the user in models that support customization, including *temperature*, *top-p*, *max tokens*, and *seed*. The *temperature* parameter controls the randomness of model outputs, where lower values result in more deterministic responses. The *top-p* parameter, also known as nucleus sampling, defines the probability mass for selecting the next token, balancing diversity and coherence. *Max tokens* determines the maximum length of the generated output, while *seed* ensures reproducibility across different runs. For models that allow parameter tuning, we experimentally set the following values: *temperature* = 0.2, *top-p* = 0.95, *max tokens* = 200, and *seed* = 42.

We describe next the key details of the selected VLMs:

- **ChatGPT (GPT-4o):** this is a multimodal model from OpenAI, capable of processing both text and image inputs while generating text-based outputs. It has been pre-trained on a diverse dataset, including web content, pro-

prietary data, and multimodal sources [15]. For this study, GPT-4o is accessed via OpenAI’s API using the `openai` Python library. It supports a 128,000-token context window and a maximum output of 16,384 tokens³.

- **Gemini 2.0 Flash:** developed by Google, this is a high-speed multimodal model designed for diverse tasks, including text and image processing. It features next-generation capabilities such as native tool use and multimodal generation. The model has been pre-trained on a diverse dataset, enabling strong vision-language reasoning [3, 12]. For this study, Gemini is accessed via Google’s API using the `google` Python library. It supports a context window of around 1M tokens and a maximum output of 8,192 tokens⁴.
- **Claude 3.5 Sonnet:** developed by Anthropic, this is a state-of-the-art model with multilingual, vision, and reasoning capabilities. It features a 200,000-token context window and a maximum output of 8,192 tokens [4]. Trained on a proprietary dataset combining publicly available web data, third-party sources, and internally generated content, Claude employs a hybrid reasoning approach with strong performance in code generation, computational tasks, and extended context processing. For this study, it is accessed via Anthropic’s API⁵.
- **Moondream:** this is an open-source, lightweight VLM designed for efficient image analysis, object detection, visual reasoning, and scene comprehension. Moondream-2B, in particular, is optimized for visual understanding tasks with a minimal computational footprint⁶. For this study, we use Moondream-2B via Replicate’s API⁷.
- **DeepSeek Janus-Pro:** this is one of the first open-source models incorporating a vision module. It is built upon DeepSeek Janus [48], an autoregressive transformer framework designed for both multimodal understanding and generation. The core innovation lies in decoupling visual encoding, enhancing the model’s ability to process and generate text from visual inputs⁸ [9]. For this study, DeepSeek Janus-Pro is accessed via Replicate’s API⁹.
- **LlaVA:** this is an end-to-end trained multimodal model that employs a fully-connected vision-language connector¹⁰ [22]. For this study, we use the LlaVA-v1.6-mistral-7b model via Replicate’s API¹¹. Leveraging Mistral-7B, LlaVA enhances multimodal text generation and image-based reasoning while balancing performance and computational cost.

³<https://chatgpt.com/>

⁴<https://deepmind.google/>

⁵<https://claude.ai/>

⁶<https://moondream.ai/>

⁷<https://replicate.com/lucataco/moondream2>

⁸<https://chat.deepseek.com/>

⁹<https://replicate.com/deepseek-ai/janus-pro-1b>

¹⁰<https://LlaVA-v1.github.io/>

¹¹<https://replicate.com/yorickvp/LlaVA-v1.6-mistral-7b>

²<https://replicate.com/>

4.2. Prompt Design

To optimize the prompt for obtaining accurate and structured responses, several key aspects were considered. The primary goal was explicitly defined: analyzing the image and identifying the foods present. Since many of the food products align with Mediterranean and Spanish dietary patterns, the final prompt enforces a standardized format, requiring each detected food item to be categorized into a pre-defined *category*, *subcategory*, and *cooking style*. To maintain consistency, it restricts responses to the given taxonomy, prevents additional explanations or assumptions, and ensures outputs follow a strict, structured format.

Some *categories* are “cereals and legumes”, “vegetables and fruits”, “protein sources”, etc. Furthermore, *subcategories* such as “fruits”, “infusions”, and “pizza”, and *cooking styles*, including “fresh”, “fried”, “boiled” or “steamed”, and “oven-baked”, ensure a detailed and standardized representation of food items.

To ensure consistency, foods not in the predefined list were labeled as “*Others*”, while partially visible or distant items were excluded. Finally, for models that failed to generate structured responses (i.e., DeepSeek, LLaVA, and Moondream), we implemented a secondary post-processing step using ChatGPT-4o. This step refined outputs by correcting formatting errors, separating merged elements, and classifying food products into predefined *category*, *subcategory*, and *cooking style*. The complete prompts are available in the supplementary material.

5. Experimental Framework

5.1. Experimental Protocol

In order to systematically evaluate the performance of the six selected VLMs with our proposed FoodNExTDB, we design an experimental protocol consisting of three tasks. As indicated in Sec. 4.2, we insert to each VLM a structured prompt to analyze each image, identifying individual food products, and generating an appropriately formatted output. To ensure consistency, manual post-processing was applied to fewer than 1% of the generated labels, addressing cases where models produced similar but non-exact matches (e.g., generating food categories or cooking styles not included in the predefined taxonomy). The total cost of all experiments was \$434.04, covering all API executions. We describe next the three tasks analyzed in our experimental study.

5.1.1. Task 1: Food Image Recognition

This task evaluates the ability of VLMs to recognize food products from images by classifying them into predefined categories. The evaluation is conducted at three consecutive levels, from simpler to more complex classifications:

- *Category*: Recognizing general food groups, such as “*vegetables and fruits*” or “*protein sources*”.

- *Category + Subcategory*: Providing finer-grained classification within each category, for example, distinguishing between “*vegetables*” and “*fruits*” subcategories.
- *Category + Subcategory + Cooking Style*: Identifying both the food *category* and *subcategory* and its preparation method, such as differentiating between “*grilled*” fish and “*fried*” fish.

5.1.2. Task 2: Fine-Grained Food Recognition

The second task focuses on the model’s ability to recognize specific food products correctly. Performance is analyzed across different categorization levels to determine which food products are easier or more challenging to recognize. This task helps to assess whether VLMs can reliably differentiate between foods that look similar and correctly assign them to their respective *categories* and *subcategories*.

5.1.3. Task 3: Image Complexity Performance

The third task examines the impact of image complexity on model performance. Two scenarios are evaluated, with images containing only one identifiable food item (*single-product* images), and images featuring multiple food products in a single image (*multi-product* images).

5.2. Proposed Evaluation Metric: EWR

Our metric quantifies how well the VLMs match the annotations considering possible differences between annotators. In image i , the annotators identify p_j^i food products, with $j = 1, \dots, M^i$, where M^i is the total number of food products detected by the annotators. Let L^i be the total number of product labels l_k^i , with $k = 1, \dots, L^i$, assigned by the annotators (note that a single product can originate multiple labels from multiple experts). We assign a different weight for each marked label l_k^i based on the level of label agreement between nutrition experts $W_{\text{Nutritionists}}(i, l_k^i) = n_k^i/N^i$, where n_k^i is the number of nutritionists who marked the given label l_k^i for the same product p_j^i (out of a total of N^i annotators). Note that the number of products and annotators can vary between images, and some annotators may not label a particular product if they are not sure about it.

For VLMs predictions we consider the following weights: *i*) if a predicted product label appears in the annotations of experts, it is assigned the corresponding weight (i.e., $W_{\text{VLM}}(i, l_k^i) = n_k^i/N^i$), otherwise *ii*) if a prediction is not in the annotated list, then $W_{\text{VLM}}(i, l_k^i) = 0$. The propose Expert-Weighted Recall (EWR) metric is as follows:

$$\text{EWR}^i = \frac{\sum_{k=1}^{L^i} W_{\text{VLM}}(i, l_k^i)}{\sum_{k=1}^{L^i} W_{\text{Nutritionists}}(i, l_k^i)} \quad (1)$$

The final EWR score is obtained by averaging the individual EWR^i values from all images. The EWR metric ensures that higher-agreement predictions contribute more to the final score, allowing flexibility for partial agreement.

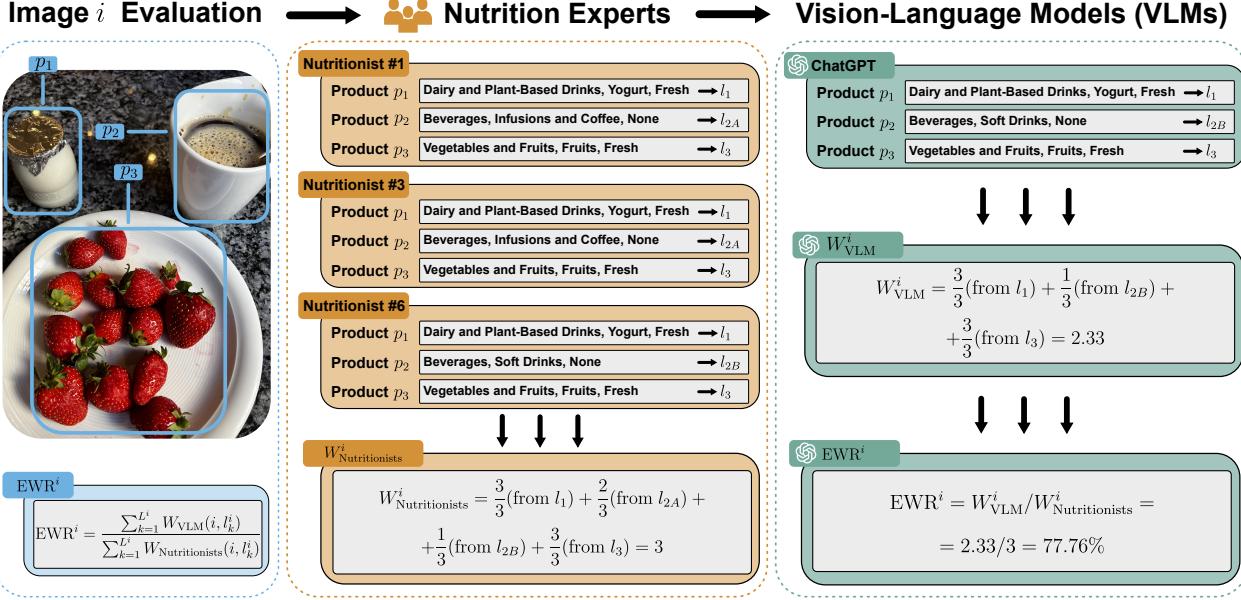


Figure 2. Illustration of the proposed Expert-Weighted Recall (EWR) computation for a food image i (left). This graphical example compares annotations (e.g. label $l_1 = \text{Yogurt}$ for product p_1) from three nutrition experts (middle) with the predictions made by a VLM (right). The proposed EWR metric reflects how well the VLM aligns with expert consensus while accounting for annotation variability.

	Model Name	Category	+Subcat.	+Cooking Style	Average
Closed Source	ChatGPT	80.67	69.87	42.41	64.32
	Gemini	85.79	74.69	50.00	70.16
	Claude	82.60	69.88	45.09	65.86
Open Source	Moondream	76.94	63.28	23.91	54.71
	DeepSeek	49.39	37.42	15.30	34.04
	LLaVA	63.67	48.84	28.48	47.00

Table 1. Performance comparison in terms of Expert-Weighted Recall (EWR) of the selected VLMs in food image recognition. Results are reported across three categorization levels, from simpler to more complex ones: *i*) Category, *ii*) Category + Subcategory, and *iii*) Category + Subcategory + Cooking Style.

Fig. 2 shows the EWR computation by comparing the annotations (middle) with the VLM predictions (right). In the food image i (left), three food products are detected: p_1^i , p_2^i , and p_3^i (in Fig. 2 we remove the upper i notation for simplicity). All experts agreed on p_1 (label $l_1 = \text{"yogurt"}$), while p_2 was labeled differently by two experts. All identified p_3 with the label $l_3 = \text{"fruits"}$. ChatGPT correctly predicted l_1 , matched l_{2B} , and identified l_3 . The 77.76% EWR achieved reflects how well ChatGPT aligns with the experts while accounting for annotation variability.

6. Results

This section evaluates the VLM performance in food image recognition. Sec. 6.1 analyzes classification across different granularity levels, Sec. 6.2 explores the most and least chal-

lenging food products, and Sec. 6.3 evaluates the impact of image complexity on recognition accuracy.

6.1. Task 1: Food Image Recognition

Table 1 presents the performance of each VLM in terms of EWR across three classification levels: *i*) category, *ii*) category + subcategory, and *iii*) category + subcategory + cooking style. Overall, closed-source models (ChatGPT, Gemini, and Claude) consistently outperform open-source ones (Moondream, DeepSeek, and LLaVA) at all classification levels. Gemini achieves the highest average EWR across all levels (70.16%), indicating a broader detection range, while ChatGPT (64.32%) and Claude (65.86%) maintain strong performance.

As classification complexity increases, all models experience performance drops. For instance, Gemini's EWR declines from 85.79% (category) to 74.69% (category + subcategory) and 50.00% (category + subcategory + cooking style), with a similar trend observed for ChatGPT (80.67%, 69.87% and 42.41%, respectively). These results highlight the challenge of fine-grained recognition.

Among open-source models, Moondream outperforms DeepSeek and LLaVA (avg. 54.71% vs. 34.04% and 47.00%, respectively), excelling in category and category + subcategory recognition. DeepSeek, with the lowest average EWR (34.04%), struggles due to limited exposure to food datasets. Figure 3 presents challenging cases. In Figure 3A), a multi-component dish is inconsistently labeled by experts, with VLMs only recognizing individual ingre-

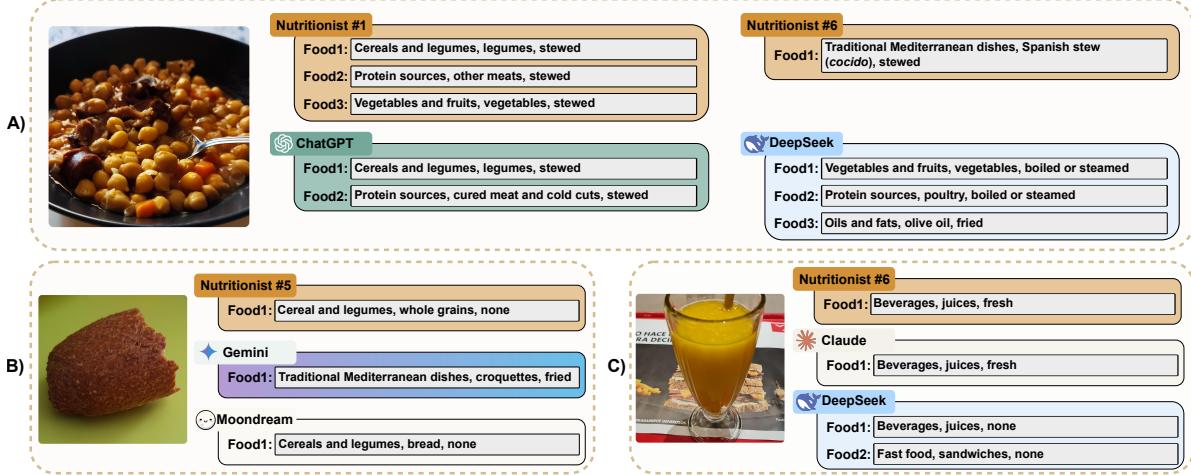


Figure 3. Examples of VLMs predictions compared to nutritionist’s annotations. (A) A multi-component dish where some experts identify individual ingredients (“legumes”, “meat”, “vegetables”), while others classify it as “Spanish stew (*cocido*)” (B) A whole grain bread misclassified by Gemini as “croquettes” due to shape similarity. (C) An orange juice image where some VLMs such as DeepSeek incorrectly identifies a sandwich from a background image on the paper tray liner.

dients. Figure 3B) shows whole-grain bread misclassified by Gemini as croquettes due to shape similarity. In Figure 3C), DeepSeek mistakenly identifies a sandwich from a background image on a paper tray liner, highlighting VLM susceptibility to background distractions.

6.2. Task 2: Fine-Grained Food Recognition

The ability of VLMs to recognize specific food products varies significantly across different categorization levels. Figure 4 presents radar charts illustrating the percentage of correctly recognized food products for each level, similar to Task 1. It is important to highlight that these radar charts include some examples of all possible classes considered in our proposed FoodNExTDB, i.e., 10 categories, 62 subcategories, and 9 cooking styles.

At the *category* level, VLMs perform best in recognizing food categories such as “cereals and legumes”, “protein sources”, and “dairy and plant-based drinks”. In these particular categories, ChatGPT consistently outperforms other models, achieving the highest accuracy across most categories. Conversely, “fast food” remains the most challenging category for all models. At the *subcategory* level, recognition declines, with “fruits” being more accurately detected than “vegetables” and “fish” outperforming “poultry” within their corresponding main categories. “Pasta” is also more frequently recognized than “rice”.

At the *cooking style* level, all models struggle significantly. “Fresh” is the most identifiable style, followed by “grilled”, while “fried” and “stewed” are the least accurately predicted. Notably, Gemini excels in detecting “fresh” foods, Moondream in “oven-baked” products, ChatGPT in “preserved” foods, and LLaVA in “stewed”

	Model Name	Single-Product Image	Multi-Product Image	Average
Closed Source	ChatGPT	90.71	76.51	83.61
	Gemini	94.52	82.18	88.35
	Claude	92.48	78.51	85.50
Open Source	Moondream	90.03	71.52	80.78
	DeepSeek	43.48	51.83	47.66
	LLaVA	68.29	61.76	65.03

Table 2. Performance comparison in terms of EWR of the selected VLMs in food recognition (*category* level) based on the image complexity (i.e., *single-product* and *multi-product* images).

dishes, despite overall lower performance in this category.

6.3. Task 3: Image Complexity Performance

Finally, we evaluate the impact of image complexity (i.e., the number of food products present in one image) on the VLMs’ performances. This analysis is carried out for *single-product* images (i.e., only one food product appears in the image) and *multi-product* images (i.e., multiple food products can appear in one image). Table 2 presents the EWR of each VLM across these scenarios. For all models except DeepSeek, performance is higher for the scenario of single-product images, with VLMs such as ChatGPT, Gemini, Claude, and Moondream are able to achieve EWRs above 90%. In contrast, EWR results decrease considerably for the scenario of multi-product images. Again, Gemini demonstrates the highest overall performance in both scenarios (88.35% EWR) whereas DeepSeek is the worst (47.66% EWR), showing a difference of over 20% EWR with the rest of the VLMs.

A more detailed analysis of our proposed database reveals that, for the scenario of single-product images, approximately 28% of images contain “fruits”, 10% include

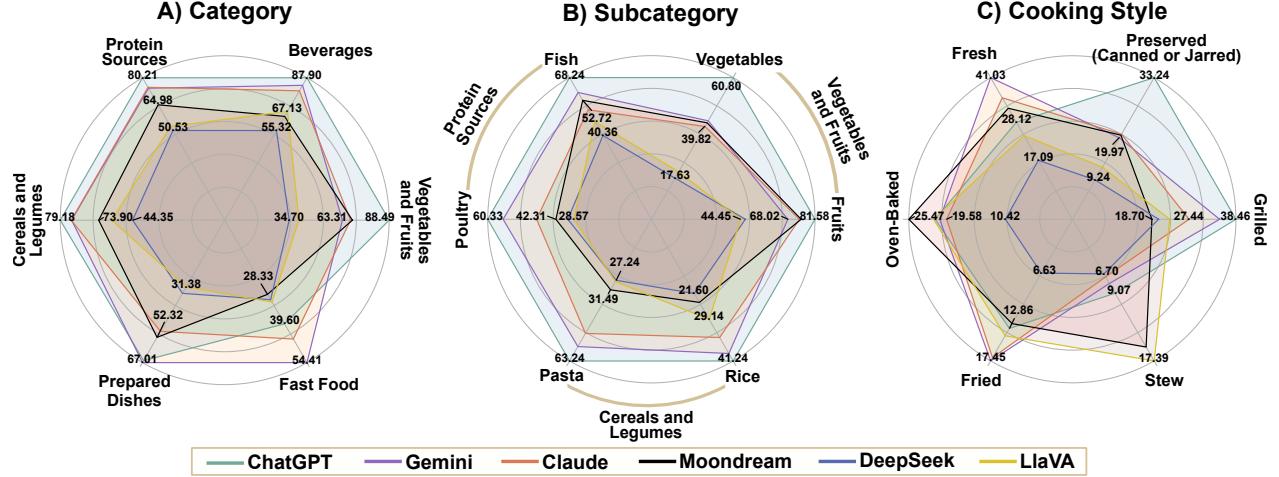


Figure 4. Radar charts illustrating VLM performance in fine-grained food recognition. We include some examples of all available classes considered in our proposed FoodNExTDB, i.e., 10 categories, 62 subcategories, and 9 cooking styles. (This figure is best viewed in color.)

“beverages” such as “*infusions or coffee*”, and 7% feature “*yogurt and fresh cheese*”. As observed in previous experiments, these are among the most accurately recognized food *categories* and *subcategories*, which explains the higher performance.

7. Discussion and Future Work

Diet analysis remains a major challenge in nutrition, requiring the consideration of multiple interrelated factors. While pure image recognition models have improved significantly in the task of food recognition, they still struggle with complex, multi-food images and fail to provide sufficient contextual understanding. Although these models bring automated nutritional assessment closer to reality, they remain insufficient for a comprehensive analysis.

VLMs present a promising alternative by integrating textual and visual reasoning, improving explainability in food recognition and dietary analysis. However, they still face difficulties with fine-grained tasks such as identifying cooking style, which require additional multi-modal data integration. More research is needed to benchmark their performance against transformer-based models in this area.

A key observation is the performance gap between open- and closed-source VLMs. Open-source models consistently underperform, often struggling with structured prompts and generating accurate responses. As a result, improving fine-tuning strategies, dataset diversity, and domain-specific training are crucial to bridge this gap.

Finally, integrating VLM with personalized nutrition strategies could improve dietary tracking and chronic disease prevention. Combining AI-driven food recognition with multimodal data from wearables, dietary questionnaires, and expert supervision may improve accuracy and adherence to automated dietary assessments [42].

8. Conclusion

This study presents FoodNExTDB, a food image database with 9,263 expert-labeled images, many reflecting Mediterranean and Spanish diets. A key strength of this database is that all images were annotated by seven nutrition experts, adding value by providing structured nutritional information, including the main *category*, *subcategory*, and *cooking style* for each detected food product.

We also propose an experimental framework to assess six state-of-the-art VLMs, analyzing their capabilities in food recognition and prompt comprehension. Due to inter-annotator variability, we designed a novel weighted evaluation metric named Expert-Weighted Recall (EWR). Our findings reveal a clear gap in model performance across classification levels. While in main categories such as “*protein sources*” and “*vegetables and fruits*” the top models reach around 80% EWR, *cooking style* recognition remains challenging, with top VLMs achieving only 50% EWR. This discrepancy highlights VLMs’ limits in capturing fine-grained food attributes, suggesting the need for advances in context-aware learning and multimodal integration.

9. Acknowledgments

This study is supported by projects: AI4FOOD-CM (Y2020/TCS6654), FACINGLCOVID-CM (PD2022- 004-REACT-EU), INTER-ACTION (PID2021-126521OB-I00 MICINN/FEDER), HumanCAIC (TED2021-131787BI00 MICINN), PowerAI+ (SI4/PJI/2024-00062), and Cátedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR TSI-100927-2023-2). We thank the nutrition experts for their valuable image annotations (B. Lacruz-Pleguezuelos, L.J. Marcos-Zambrano, S. Gallego Pozo, J. Haya, M. Izquierdo-Muñoz, J.A. Lemke Flores, G.X. Bazán).

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Eduardo Aguilar, Marc Bolaños, and Petia Radeva. Food Recognition using Fusion of Classifiers Based on CNNs. In *Proc. of the International Conference on Image Analysis and Processing*, 2017. 2
- [3] Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 4
- [4] AI Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. *Claude-3 Model Card*, 1:1, 2024. 4
- [5] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An Introduction to Vision-Language Modeling. *arXiv preprint arXiv:2405.17247*, 2024. 2
- [6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – Mining Discriminative Components with Random Forests. In *Proc. of the European Conference on Computer Vision*, 2014. 2
- [7] Le Bu, Caiping Hu, and Xiuliang Zhang. Recognition of Food Images Based on Transfer Learning and Ensemble Learning. *Plos One*, 19(1):e0296789, 2024. 1, 3
- [8] Jingjing Chen and Chong-Wah Ngo. Deep-based Ingredient Recognition for Cooking Recipe Retrieval. In *Proc. of the International Conference on Multimedia*, 2016. 2
- [9] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified Multimodal Understanding and Generation with Data and Model Scaling. *arXiv preprint arXiv:2501.17811*, 2025. 4
- [10] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Food Recognition: A New Dataset, Experiments, and Results. *IEEE Journal of Biomedical and Health Informatics*, 21(3):588–598, 2016. 2
- [11] Ivan Deandres-Tame, Ruben Tolosana, Ruben Vera-Rodriguez, Aythami Morales, Julian Fierrez, and Javier Ortega-Garcia. How Good is ChatGPT at Face Biometrics? a First Look into Recognition, Soft Biometrics, and Explainability. *IEEE Access*, 12:34390–34401, 2024. 1
- [12] Petko Georgiev, Ving Ian Lei, Ryan Burnell, et al. Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context. *arXiv preprint arXiv:2403.05530*, 2024. 4
- [13] Pan He, Zhu Liu, Giovanni Baiocchi, Dabo Guan, Yan Bai, and Klaus Hubacek. Health–environment Efficiency of Diets Shows Nonlinear Trends over 1990–2011. *Nature Food*, 5 (2):116–124, 2024. 1
- [14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2018. 3
- [15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welinda, Alan Hayes, Alec Radford, et al. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024. 4
- [16] Y. Kawano and K. Yanai. Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation. In *Proc. of the Workshop on Transferring and Adapting Source Knowledge in Computer Vision*, 2014. 2
- [17] Jun-Hwa Kim, Nam-Ho Kim, Donghyeon Jo, and Chee Sun Won. Multimodal Food Image Classification with Large Language Models. *Electronics*, 13(22), 2024. 3
- [18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-image Pre-training with Frozen Image Encoders and Large Language Models. In *Proc. of the International Conference on Machine Learning*, 2023. 2
- [19] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On Pre-training for Visual Language Models. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2024. 2
- [20] Jakob Linseisen, Britta Renner, Kurt Gedrich, Jan Wirsam, Christina Holzapfel, Stefan Lorkowski, Bernhard Watzl, Hannelore Daniel, Michael Leitzmann, et al. Perspective: Data in Personalized Nutrition: Bridging Biomedical, Psycho-behavioral, and Food Environment Approaches for Population-wide Impact. *Advances in Nutrition*, page 100377, 2025. 1
- [21] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 Technical Report. *arXiv preprint arXiv:2412.19437*, 2024. 1
- [22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2024. 4
- [23] Lihua Luo. Research on Food Image Recognition of Deep Learning Algorithms. In *Proc. of the International Conference on Computers, Information Processing and Advanced Education*, 2023. 1, 3
- [24] Peihua Ma, Shawn Tsai, Yiyang He, Xiaoxue Jia, Dongyang Zhen, Ning Yu, Qin Wang, Jaspreet K.C. Ahuja, and Cheng-I Wei. Large Language Models in Food Science: Innovations, Applications, and Future. *Trends in Food Science & Technology*, 148:104488, 2024. 1
- [25] Peihua Ma, Yixin Wu, Ning Yu, Xiaoxue Jia, Yiyang He, Yang Zhang, Michael Backes, Qin Wang, and Cheng-I Wei. Integrating Vision-Language Models for Accelerated High-Throughput Nutrition Screening. *Advanced Science*, 11(34):2403578, 2024. 3
- [26] Zheng Ma, Mianzhi Pan, Wenhan Wu, Kanzhi Cheng, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Food-500 Cap: A Fine-Grained Food Caption Benchmark for Evaluating Vision-Language Models. In *Proc. of the International Conference on Multimedia*, 2023. 3
- [27] Y. Matsuda, H. Hoashi, and K. Yanai. Recognition of Multiple-Food Images by Detecting Candidate Regions. In *Proc. of the International Conference on Multimedia and Expo*, 2012. 2

- [28] Patrick McAllister, Huiru Zheng, Raymond Bond, and Anne Moorhead. Combining Deep Residual Neural Network Features with Supervised Machine Learning Algorithms to Classify Diverse Food Image Datasets. *Computers in Biology and Medicine*, 95:217–233, 2018. 3
- [29] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A Survey on Food Computing. *ACM Computing Surveys*, 52(5):1–36, 2019. 1
- [30] Weiqing Min, Linhu Liu, Zhiling Wang, Zhengdong Luo, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. ISIA Food-500: A Dataset for Large-Scale Food Recognition via Stacked Global-Local Attention Network. In *Proc. of the International Conference on Multimedia*, 2020. 2, 3
- [31] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large Scale Visual Food Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9932–9949, 2023. 3
- [32] Fnu Mohbat and Mohammed J Zaki. Llava-chef: A Multi-modal Generative Model for Food Recipes. In *Proc. of the International Conference on Information and Knowledge Management*, 2024. 3
- [33] Kintoh Allen Nfor, Tagne Poupi Theodore Armand, Kenesbaeva Periyat Ismaylovna, Moon-Il Joo, and Hee-Cheol Kim. An Explainable CNN and Vision Transformer-Based Approach for Real-Time Food Recognition. *Nutrients*, 17(2):362, 2025. 3
- [34] Vasiliki Pitsilou, George Papadakis, and Dimitrios Skoutas. Using LLMs to Extract Food Entities from Cooking Recipes. In *Proc. of the International Conference on Data Engineering Workshops*, 2024. 1
- [35] Zhixiao Qi, Yijiong Yu, Meiqi Tu, Junyi Tan, and Yongfeng Huang. FoodGPT: A Large Language Model in Food Testing Domain with Incremental Pre-training and Knowledge Graph Prompt. *arXiv preprint arXiv:2308.10173*, 2023. 1
- [36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proc. of the International Conference on Machine Learning*, 2021. 3
- [37] Jesús M Rodríguez-de Vera, Pablo Villacorta, Imanol G Estepa, Marc Bolaños, Ignacio Sarasúa, Bhalaji Nagarajan, and Petia Radeva. Dining on Details: LLM-Guided Expert Networks for Fine-Grained Food Recognition. In *Proc. of the International Workshop on Multimedia Assisted Dietary Management*, 2023. 1
- [38] Jesús M Rodríguez-De-Vera, Imanol G Estepa, Marc Bolaños, Bhalaji Nagarajan, and Petia Radeva. LOFI: LOnge-tailed FIne-Grained Network for Food Recognition. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2024. 3
- [39] Sergio Romero-Tapiador, Blanca Lacruz-Pleguezuelos, Ruben Tolosana, et al. AI4FoodDB: A Database for Personalized e-Health Nutrition and Lifestyle through Wearable Devices and Artificial Intelligence. *Database*, 2023: baad049, 2023. 2, 3
- [40] Sergio Romero-Tapiador, Ruben Tolosana, Aythami Morales, et al. AI4Food-NutritionFW: A Novel Framework for the Automatic Synthesis and Analysis of Eating Behaviours. *IEEE Access*, 1:112199 – 112211, 2023. 1
- [41] Sergio Romero-Tapiador, Ruben Tolosana, Aythami Morales, et al. Leveraging Automatic Personalised Nutrition: Food Image Recognition Benchmark and Dataset Based on Nutrition Taxonomy. *Multimedia Tools and Applications*, 84:1945–1966, 2024. 1, 3
- [42] Sergio Romero-Tapiador, Ruben Tolosana, Aythami Morales, et al. Personalized Weight Loss Management through Wearable Devices and Artificial Intelligence. *arXiv preprint arXiv:2409.08700*, 2024. 8
- [43] Aditya Sharma, Michael Saxon, and William Yang Wang. Losing Visual Needles in Image Haystacks: Vision Language Models are Easily Distracted in Short and Long Contexts. *arXiv preprint arXiv:2406.16851*, 2024. 2
- [44] Guorui Sheng, Weiqing Min, Xiangyi Zhu, Liang Xu, Qingshuo Sun, Yancun Yang, Lili Wang, and Shuqiang Jiang. A Lightweight Hybrid Model with Location-preserving ViT for Efficient Food Recognition. *Nutrients*, 16(2):200, 2024. 1, 3
- [45] Ram B Singh, Jan Fedacko, Ghizal Fatima, Aminat Magomedova, Shaw Watanabe, and Galal Elkilany. Why and How the Indo-Mediterranean Diet May Be Superior to Other Diets: The Role of Antioxidants in the Diet. *Nutrients*, 14(4):898, 2022. 1
- [46] Ashutosh Singla, Lin Yuan, and Touradj Ebrahimi. Food/Non-Food Image Classification and Food Categorization Using Pre-Trained GoogLeNet Model. In *Proc. of the International Workshop on Multimedia Assisted Dietary Management*, 2016. 2
- [47] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In *Proc. of the Conference on Computer Vision and Pattern Recognition*, 2021. 2
- [48] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation. *arXiv preprint arXiv:2410.13848*, 2024. 4
- [49] Zhongqi Yang, Elahe Khatibi, Nitish Nagesh, Mahyar Abbasian, Iman Azimi, Ramesh Jain, and Amir M Rahmani. ChatDiet: Empowering Personalized Nutrition-oriented Food Recommender Chatbots through an LLM-Augmented Framework. *Smart Health*, 32:100465, 2024. 1
- [50] Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. FoodLMM: A Versatile Food Assistant Using Large Multi-modal Model. *arXiv preprint arXiv:2312.14991*, 2023. 3
- [51] Dongyu Zhang, Ruofan Hu, Dandan Tao, Hao Feng, and Elke Rundensteiner. LLM-based Hierarchical Label Annotation for Foodborne Illness Detection on Social Media. In *Proc. of the International Conference on Big Data*, 2024. 1
- [52] Ping Zhang. Influence of Foods and Nutrition on the Gut Microbiome and Implications for Intestinal Health. *International Journal of Molecular Sciences*, 23(17):9588, 2022. 1