

FoodLMM: A Versatile Food Assistant Using Large Multi-Modal Model

Yuehao Yin[✉], Huiyan Qi[✉], Bin Zhu[✉], Jingjing Chen[✉], Member, IEEE, Yu-Gang Jiang[✉], Fellow, IEEE, and Chong-Wah Ngo[✉], Senior Member, IEEE

Abstract—Large Multi-modal Models (LMMs) have made impressive progress in many vision-language tasks. Nevertheless, the performance of general LMMs in specific domains is still far from satisfactory. This paper proposes FoodLMM, a versatile food assistant based on LMMs with various capabilities, including food recognition, ingredient recognition, recipe generation, nutrition estimation, food segmentation, and multi-round conversation. To facilitate FoodLMM in dealing with tasks beyond pure text output, we introduce a series of novel task-specific tokens and heads, enabling the model to predict food nutritional values and multiple segmentation masks. We adopt a two-stage training strategy. In the first stage, we utilize multiple public food benchmarks for multi-task learning by leveraging the instruct-following paradigm. In the second stage, we construct a multi-round conversation dataset and a reasoning segmentation dataset to fine-tune the model, enabling it to conduct professional dialogues and generate segmentation masks based on complex reasoning in the food domain. Our fine-tuned FoodLMM achieves state-of-the-art results across several food benchmarks.

Index Terms—Food assistant, large multi-modal model.

I. INTRODUCTION

BENEFITING from the remarkable language understanding capabilities of Large Language Models (LLMs) [1], [2], [3], [4], [5] and image features obtained from vision-language models [6], [7], [8], [9], Large Multi-modal Models (LMMs) [8], [10], [11], [12], [13], [14], [15], [16] have exhibited outstanding performance in various vision-language tasks, such as image captioning, visual question answering and complex visual reasoning. Furthermore, LMMs are able to interact with humans with natural language, paving the way to build versatile conversational assistants in different vertical domains [17].

Received 27 May 2024; revised 20 September 2024; accepted 15 January 2025. Date of publication 30 July 2025; date of current version 21 October 2025. This work was supported by the National Natural Science Foundation of China Project under Grant 62072116. The associate editor coordinating the review of this article and approving it for publication was Dr. Li Cheng. (*Corresponding author: Jingjing Chen*.)

Yuehao Yin, Jingjing Chen, and Yu-Gang Jiang are with the School of Computer Science and Technology, Fudan University, Shanghai 200082, China (e-mail: yhyin21@m.fudan.edu.cn; chenjingjing@fudan.edu.cn; ygj@fudan.edu.cn).

Huiyan Qi, Bin Zhu, and Chong-Wah Ngo are with the School of Computing and Information Systems, Singapore Management University, Singapore 178902 (e-mail: huiyanqi@smu.edu.sg; binzhu@smu.edu.sg; cwngo@smu.edu.sg).

Our code, models, and datasets are available at <https://github.com/YuehaoYin/FoodLMM>.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TMM.2025.3590924>, provided by the authors.

Digital Object Identifier 10.1109/TMM.2025.3590924

Although previous LMMs perform well on general images and questions, due to a lack of domain-specific expertise, they often cannot provide reliable assistance in vertical domains and even produce incorrect responses or hallucinations [17], [18]. This is particularly evident in the food realm. For instance, when asked about nutrition in a food image, general LMMs typically can only answer what nutritional elements it contains but fail to provide specific quantities and precise nutritional content of the food. In the food domain, the previous efforts have been devoted to several tasks, ranging from food and ingredient recognition [19], [20], [21], [22], [23], [24], [25], recipe generation [26], [27], [28], [29], [30], [31], nutrition estimation [32], [33] to food segmentation [34], [35], [36], [37], [38], [39], [40], [41]. Though the existing works achieve promising results in each individual task, they are incapable of dealing with other tasks using a single model.

In this paper, we address the limitations by proposing FoodLMM, a versatile food assistant using a Large Multi-modal model tailored for various food-related tasks, as shown in Fig. 1. To the best of our knowledge, FoodLMM is the first unified and multi-task LMM in the food domain. FoodLMM is capable of handling a variety of food-related tasks, including: **Food Classification**, **Ingredient Recognition**, **Recipe Generation**, **Nutrition Estimation**, **Referring Segmentation**, and **Reasoning Segmentation**, achieving state-of-the-art (SOTA) performance in each task. The architecture of FoodLMM extends from LISA [42]. FoodLMM consists of a base multi-modal large language model LLaVA [11], a segmentation model SAM [43], and a series of novel task-specific head tokens. Our FoodLMM produces text outputs as other LMMs [17], [42] for food classification, ingredient recognition, and recipe generation tasks in an instruct-following fashion. We introduce a set of special segmentation tokens into the model’s vocabulary, and their hidden states are fed into the segmentation decoder to generate one or multiple masks. This mask generation mechanism, combined with our carefully designed instructions, enables FoodLMM to effectively address the complex one-to-many and one-to-zero challenges [44] in the **Referring Segmentation** task. Furthermore, we introduce nutritional task-specific tokens for nutrition estimation, where the hidden states of these tokens are processed by regression heads to provide accurate nutritional values for various food items.

The training process of FoodLMM is divided into two stages. In the first stage, we utilize multiple public datasets [20], [26], [32], [35], [36] to build large-scale instruction-tuning data to

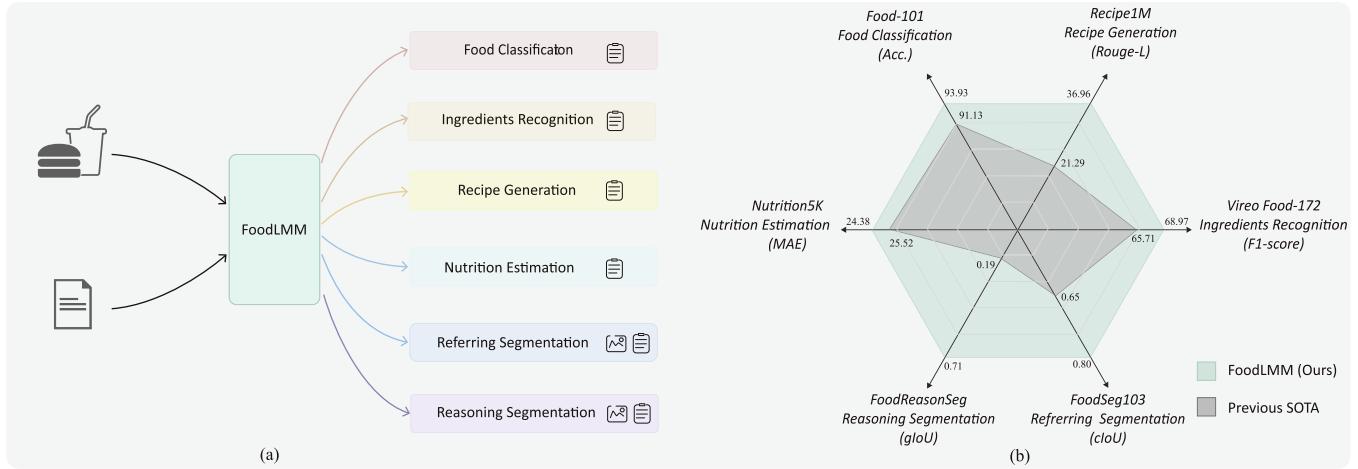


Fig. 1. (a) FoodLMM is a versatile large multi-modal model for various food-related tasks. (b) Performance comparison between our FoodLMM and previous state-of-the-art (SOTA) methods across various tasks.

conduct multi-task learning, aiming at injecting substantial basic food domain knowledge into FoodLMM. Specifically, we design a rich set of instructions and answer query templates for each task, enabling the model with the ability to handle different food tasks. This stage is crucial for building a general understanding of the food domain and further extending model capabilities for food knowledge perception, reasoning, and conversation. The second stage aims to endow the model with multi-round conversational interaction capabilities in free-form prompts. For this purpose, we create a food multi-round conversation dataset **FoodDialogues** and a reasoning segmentation dataset **FoodReasonSeg** with GPT-4 [45]. The FoodDialogues is constructed based on the Nutriton5k food nutrition dataset, where detailed nutritional information is fed to GPT-4, prompting it to generate multi-round dialogues on various food-related topics, such as calorie calculation, dietary planning, and metabolism. Similarly, FoodReasonSeg is derived from the FoodSeg103 food segmentation dataset, utilizing GPT-4 to create multi-round complex reasoning dialogues, where the segmentation masks of the mentioned ingredients are used as reasoning labels. In the second training stage, we fine-tune FoodLMM using these two datasets. As a result, FoodLMM acquires the ability to interact in multi-round conversations with users and provide reasoning segmentation masks for complex queries. This two-stage strategy ensures that the model first learns comprehensive domain knowledge before specializing in multi-turn interaction capabilities. Our main contributions can be summarized as follows:

- We propose **FoodLMM**, a versatile large multi-modal model in the food domain. We design specific instructions for different tasks and unify a variety of food tasks using LMM, achieving SOTA performance across multiple benchmarks. To the best of our knowledge, our FoodLMM is the first unified approach for multiple food tasks. FoodLMM provides practical guidance for the construction of LMM in other vertical domains.
- FoodLMM generates segmentation masks and nutritional predictions through a series of novel task-specific tokens. This paradigm easily solves the most challenging one-to-many and one-to-zero problems in referring segmentation

and enables the model to estimate the nutritional value of the entire dish or any ingredient in the food image.

- We construct diverse multi-round conversation food dataset **FoodDialogues** and food reasoning segmentation dataset **FoodReasonSeg**. We release both datasets, along with our codebase and model checkpoints to facilitate future research.

II. RELATED WORK

A. Large Language Model

The remarkable language understanding and reasoning capabilities of large language models (LLMs) lead researchers to explore their use in visual tasks. Recently, closed-source GPT-4 [45] receives significant recognition for introducing the ability to process multimodal inputs. Additionally, many open source multimodal LLMs, such as Minigpt4 [12], Vicuna [46], LLaVA [11], MultiModal-GPT [13], and OpenFlamingo [47], have demonstrated their effectiveness across a wide range of tasks. In addition, there have been successful applications of LLMs in specific domains, such as biomedicine [48], [49], [50], finance [51], and law [52]. However, these LLMs are only limited to text modality. LLaVA-Med [17] introduces a multimodal LMM in the medical domain, which employs biomedical conversations generated by GPT-4 to instruct the model in learning biomedical knowledge. Inspired by LLaVA-Med, we aim to develop multimodal LLMs in the food domain.

B. Food Analysis

With the release of food-related datasets, such as Food-101 [19], VIREO Food-172 and 251 [20], [53], Recipe1M [26], Nutrition5K [32], UEC Food256 [54], FoodSeg103 [36], Food2K [55], the past efforts in food domain have been devoted to various tasks such as food classification [56], [57], [58], [59], [60], [61], [62], ingredient recognition [20], [21], [22], [53], [63], cross-modal recipe retrieval [26], [64], [65], [66], [67], [68], [69], recipe generation [27], [28], [29], food segmentation [36], [70], [71], [72], food recommendation [73], [74], [75],

nutrition estimation [32], [76], [77], [78] and food logging [79], [80], [81]. These works focus on a specific food task. In contrast, this paper introduces a unified model, FoodLMM, to deal with various tasks.

C. Referring Segmentation

Referring segmentation [82] tasks aim to use instructions to guide the segmentation of specific objects mentioned in the query text. Transformer-based backbones are dominant in referring segmentation [83], [84]. SAM [43] attracts widespread attention in the community due to its precise segmentation and powerful zero-shot capabilities, but its performance on natural language prompts is suboptimal. Additionally, most existing referring segmentation works assume a single object in the query text, ignoring scenarios without object or with multiple objects. Considering this issue, [44] proposes One-to-Many and One-to-Zero segmentation. LISA [42] introduces a novel reasoning segmentation task that demands complex query text. However, LISA rigidly provides a binary mask for each query. In contrast, FoodLMM accurately conducts reasoning segmentation in the food domain, offering multiple masks or refraining from providing masks for non-existent objects.

III. FOOD VISUAL INSTRUCTION-FOLLOWING DATA

Inspired by LLaVA [11], we adopt a two-stage strategy to train FoodLMM, which equips the model with capabilities for food knowledge perception, reasoning, and conversation. In the first stage, we utilize existing public food datasets to build large-scale instruction-tuning data, aiming at injecting extensive general knowledge into FoodLMM. In the second stage, the model is trained with a multi-turn food dialogue dataset, FoodDialogues, and a reasoning segmentation dataset, FoodReasonseg, further equipping FoodLMM with multi-turn conversational interaction capability in a free-form prompt. We follow LLaVA to adopt visual instruction tuning to train our FoodLMM. Denote \mathbf{X}_i as a food image, \mathbf{X}_q^m and \mathbf{X}_a^m as the textual query and answer of the m -th conversation turn respectively. The format of the instruction-following data is:

```

 $\mathbf{X}_{\text{system-message}} \langle \text{STOP} \rangle \backslash n$ 
USER:  $\mathbf{X}_i \mathbf{X}_q^1 \langle \text{STOP} \rangle \backslash n$  ASSISTANT:  $\mathbf{X}_a^1 \langle \text{STOP} \rangle \backslash n$ 
USER:  $\mathbf{X}_q^2 \langle \text{STOP} \rangle \backslash n$  ASSISTANT:  $\mathbf{X}_a^2 \langle \text{STOP} \rangle \backslash n \dots$ 

```

Only green sequence/tokens are used to compute the autoregressive loss.

A. Stage 1: Public Food Datasets

We construct the instruction-following data based on instruction templates from 5 most critical tasks in the food domain: Food Classification, Ingredient Recognition, Recipe Generation, Nutrition Estimation and Food Segmentation. The public datasets used for each task are listed in Table I, and we introduce them individually as follows.

Food VQA: We collectively refer to the three pure text output tasks: food classification, ingredient recognition and recipe generation as Food VQA and design instruction templates for

TABLE I
DATASET STATISTICS OF PUBLIC FOOD DATASETS TO CONSTRUCT VISUAL INSTRUCTION-FOLLOWING DATA IN STAGE 1

Dataset	Images	Class	Ingredients	Recipes	Annotations
VIREO Food-172 [20]	100k	172	353	—	Category, ingredient
Recipe1M [26]	626k	—	1,488	361k	Recipe
Nutrition5k [32]	125k	—	555	—	Ingredient, nutrition
FoodSeg103 [36]	7k	—	103	—	Ingredient, mask
UEC-FoodPIX [35]	10k	—	102	—	Ingredient, mask

each of them. We select two widely used benchmarks for Food VQA, VIREO Food-172 [20] and Recipe1M [26]. VIREO Food-172 [20] dataset contains around 100 k food images from 172 categories and includes 353 ingredients. Recipe1M [26] dataset includes about 922 k recipes accompanied by 819 k images, with some recipes corresponding to several images and others to none. This collection contains about 16 k food ingredients. We only use recipes from the Recipe1M dataset that have corresponding images as cross-modal recipe retrieval [20], [66], [85]. We elaborately design 10, 8, and 11 instruction templates for food classification, ingredient recognition, and recipe generation tasks, respectively. Please refer to the supplementary materials for more details.

Nutrition Estimation: Nutrition5k [32] provides fine-grained nutritional element values, food calories, and quality annotations, including RGB images and multi-angle videos for about 5,000 dishes. Furthermore, this dataset also provides detailed nutritional information for each dish, covering total mass, total calories, and total macro-nutrient (fat, carbohydrate, protein) contents, along with specific data on the weight, calorie, and macro-nutrient content for every ingredient in each dish. We select RGB images and video frames with good viewing angles. Note that we do not use depth images, as they are difficult for users to acquire.

We design a rich collection of instruction templates for nutrient estimation tasks that contain 7 types of instructions, including queries about overall calories or nutritional values, inquiries referring to one or more ingredients, questions about the primary nutrient, etc. There are 64 query templates and 68 answer templates in total, listed in the supplementary material.

Food Segmentation: The FoodSeg103 [36] and UEC-FoodPIX Complete [35] datasets are the two most commonly used benchmarks for food segmentation. Of these, the FoodSeg103 [36] dataset, which consists of 7,118 food images with fine-grained ingredient-level segmentation mask annotations for 103 ingredient types and a total of 42,097 ingredient mask annotations, provides a more challenging benchmark for food image segmentation [40]. UECFoodPixComplete [35] provides food category-level segmentation mask annotations for 10,000 images, containing a total of 102 food categories. We design 10 query templates and 9 answer templates for the food segmentation task. All the instruction templates can be found in the supplementary material.

B. Stage 2: GPT-4 Generated Conversation Datasets

Stage 2 aims to empower our FoodLMM with multi-round conversational ability on various topics based on food images and provide segmentation masks for queries requiring

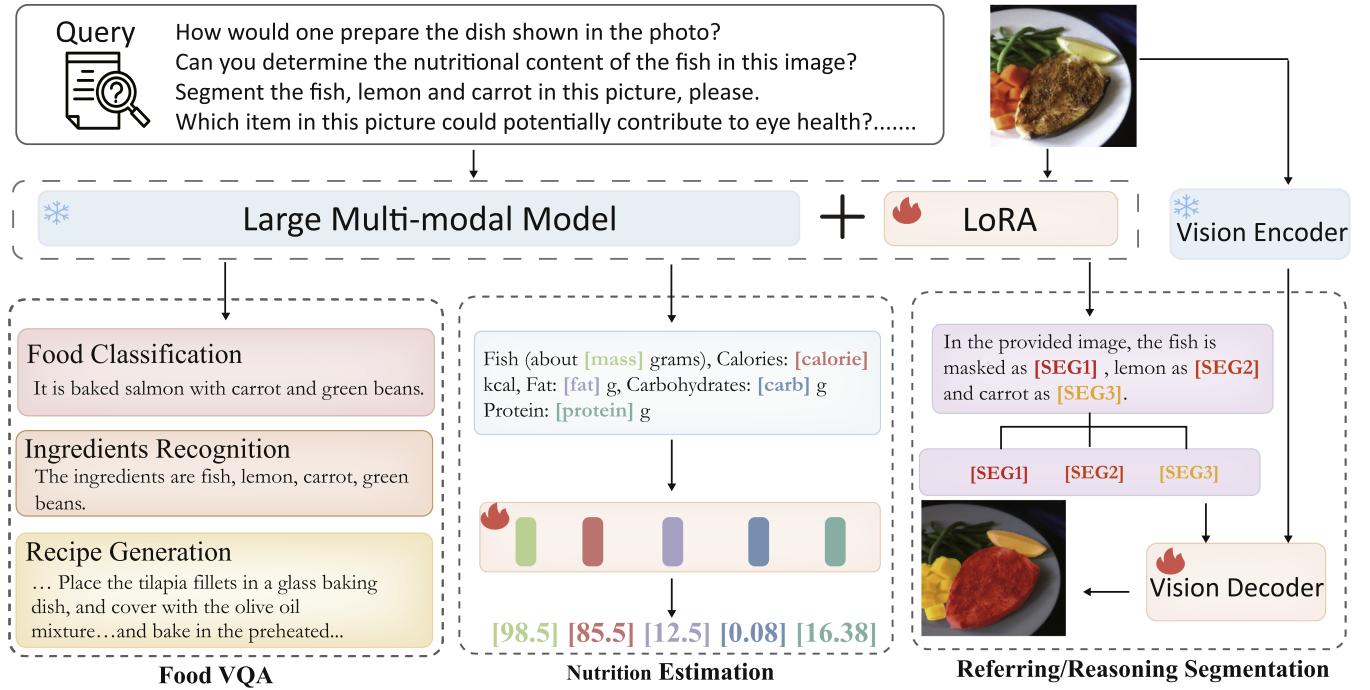


Fig. 2. Architecture Overview of the proposed FoodLMM.

TABLE II
DATASET STATISTICS OF FOODDIALOGUES

Split	Ingredients	Topics	Rounds	Samples
Train	1~3	1	3	1,480
	4~10	3	4	4,130
	11~19	7	5	818
Test	1~3	1	3	298
	4~10	3	4	647
	11~17	7	5	149

TABLE III
DATASET STATISTICS OF FOODREASONSEG

Ingredients	Rounds	Train Samples	Test Samples
3	3	1,535	664
4	4	1,224	450
5	4	646	323
6	4	304	108
7	4	140	52
8	4	39	18
9	4	12	4
10	4	6	1
11	5	1	0

complex reasoning. Nevertheless, the food domain lacks comprehensive instruction-following datasets and reasoning segmentation datasets suitable for LMMs. To bridge this gap, as shown in Table IV, we construct the first food multi-round conversation dataset **FoodDialogues** and food reasoning segmentation dataset **FoodReasonSeg**.

FoodDialogues is built from the Nutrition5k dataset, which contains ingredient labels and precise nutrition information, making it unique and suitable for various conversational topics. Specifically, we follow the training and testing splits of the original dataset and select an overhead RGB image and a well-angled (angle A or D) video frame for each sample. For each sample, we send the ingredient list and detailed nutritional information to GPT-4 in the form of plain text, requesting it to generate multiple rounds of conversations on various topics, ranging from nutrition, calorie calculation, health and diseases, metabolism, dietary planning, allergies, food pairing to substitution. The number of question-answer (QA) pairs in a multi-round conversation sample is set between 3 to 5 rounds. Taking into account the varying richness of ingredients in food images, we make the samples with more ingredients to cover more topics and include more rounds of QA, as detailed in Table II.

The prompt for GPT-4 is carefully designed based on expert opinions, where answers are required to be professional and explanatory. The two generated examples are shown in Fig. 3 and Fig. 4, originating from the FoodDialogues dataset and the FoodReasonSeg dataset, respectively.

FoodReasonSeg is constructed based on the food segmentation dataset FoodSeg103. Similarly, we send the ingredients list to GPT-4 and ask it to generate multi-round conversations where the questions should require complex reasoning. The corresponding masks provided in the original dataset for the ingredients mentioned in answers are used as segmentation labels. The more ingredients the food image contains, the greater the number of dialogue rounds. To generate QA pairs with complex reasoning, we filter out food images with less than three ingredients. The statistics are shown in Table III.

Fig. 5 showcases the word cloud visualizations for FoodDialogues and FoodReasonSeg. In these visual representations, the prominence of each word is directly proportional to its frequency within the respective datasets.

Moreover, we conduct a human evaluation to measure the quality of the generated **FoodDialogues** and **FoodReasonSeg**

A chat between a curious human and an artificial intelligence assistant specialized in food and nutrition topics. The assistant gives helpful, detailed, and polite answers to the human's questions.

USER: <im_start><image><im_end>

Based on the food items present in this meal, would this be appropriate for someone aiming to increase their metabolic rate?

ASSISTANT: The meal you mentioned contains diverse food items that can help to boost metabolism.

Protein-rich foods like sausage with [PRO1] grams and [CAL1] kcal can increase the thermic effect of food, which is the energy required for digestion, absorption, and disposal of nutrients. Meanwhile, mixed greens with [CARB1] grams are high in fiber, which can increase feelings of fullness and boost metabolism as well. However, high-fat and high-carbohydrate foods like hash browns consuming [CAL2] kcal with [FAT1] grams and [CARB2] grams might lead to weight gain if not consumed in moderation. Therefore, it's important to balance the intake of these foods.</s>

USER: May foods rich in antioxidants, like the berries here, have an effect on metabolism?

ASSISTANT: Yes, foods rich in antioxidants can potentially affect metabolism. Antioxidants present in berries such as blackberries and raspberries contributing [CAL1] and [CAL2] kcal respectively help in reducing oxidative stress in your body. Oxidative stress can slow down the metabolic rate as it hampers the body's ability to manage stress. By combating oxidative stress, antioxidants help maintain a healthy metabolic rate.</s>

USER: Does a person's metabolism increase upon eating a meal like this because it contains a lot of calories?

ASSISTANT: Eating increases your metabolism for a while. This phenomenon is known as the thermic effect of food (TEF) and it represents the energy required for digestion, absorption, and disposal of nutrients. More calorically dense foods, such as the hash browns and sausage in this meal which collectively provide [CAL1] kcal, increase TEF more so than less calorically dense foods. Nevertheless, it's crucial to remember that while consuming high-calorie foods will temporarily increase your metabolism due to TEF, over time, consistent overeating can lead to weight gain and potentially slow the basal metabolic rate.</s>

USER: How can my metabolism influence how I digest the different types of nutrients in this meal?

ASSISTANT: Metabolism significantly affects how your body digests and utilizes different nutrients in this meal. As a rule, protein-rich meals like the sausage here which has [PRO1] grams can increase metabolic rate because they require more energy to digest compared to fats or carbohydrates. Carbohydrates like those found in hash browns with [CARB1] grams and granola with [CARB2] grams are broken down into sugars, which are either used for immediate energy or stored for later use. Fat from food like hash browns and vinaigrette providing [FAT1] grams total is either stored or used to facilitate absorption of fat-soluble vitamins. Therefore, the metabolic rate can determine the rate at which these nutrients are processed and used.</s>



(a) Overhead RGB image.



(b) Frame extracted from the video.

Fig. 3. An example from FoodDialogues.

A chat between a curious human and an artificial intelligence assistant specialized in food and nutrition topics. The assistant gives helpful, detailed, and polite answers to the human's questions.

USER: <im_start><image><im_end>

What ingredients present in this meal are known for being high in protein? Please output segmentation mask.

ASSISTANT: Steak [SEG1] is a rich source of protein which is essential for body functions like muscle building and tissue repair.</s>

USER: What part of this meal may help lower blood pressure? Please output segmentation mask.

ASSISTANT: The Cucumber [SEG1] in this dish can assist in lowering blood pressure because it contains high amounts of potassium, an essential nutrient known for its blood pressure-lowering abilities.</s>

USER: Can you tell me what in this dish could enhance digestion? Please output segmentation mask.

ASSISTANT: The dietary fiber found in Bread [SEG1] aids digestion by adding bulk to the diet, preventing constipation, and promoting regular bowel movements.</s>

USER: Which ingredient in this meal supports a healthy immune system? Please output segmentation mask.

ASSISTANT: The Onion [SEG1] here is known to boost the immune system. It contains Vitamin C and phytochemicals that assist in improving immunity.</s>



(a) Input RGB image.



(b) Mask Annotation.



(c) Masked Image.

Fig. 4. An example from FoodReasonSeg.

TABLE IV
DATASET STATISTICS OF FOODDIALOGUES AND FOODREASONSEG

Dataset	Split	Images	Dialogues	QA Pairs
FoodDialogues	Train	8,094	6,428	25,821
	Test	1,418	1,094	4,325
FoodReasonSeg	Train	3,907	3,997	13,817
	Test	1,620	1,703	5,831

datasets. Specifically, users are asked to rate the quality of GPT-4 generated content on a five-point scale (5 being the best). The average score is 4.2 for FoodDialogues and 4.4 for FoodReasonSeg, with more than half of the participants choosing 5, which showcases the remarkable reliability and high quality of the datasets. More details are provided in the supplementary material.

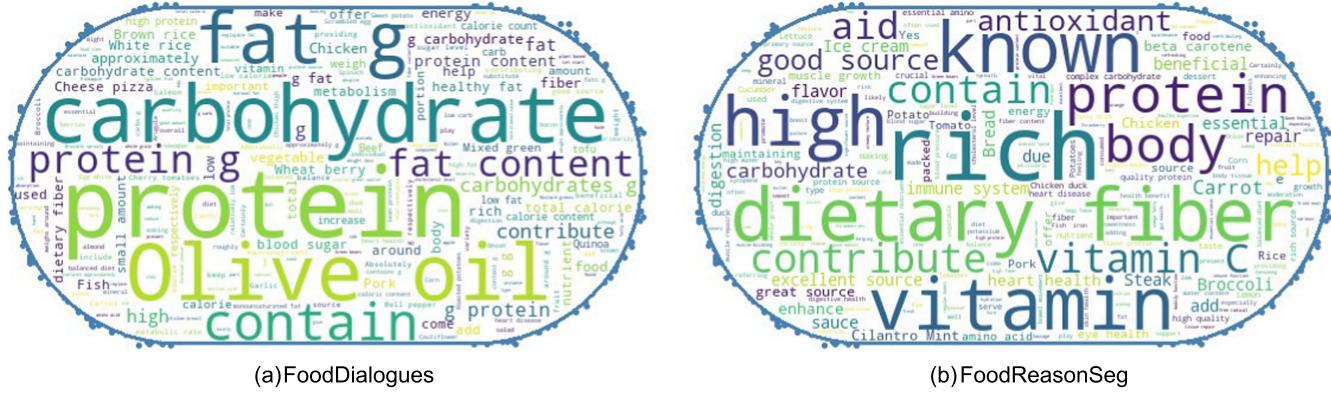


Fig. 5. Word cloud of FoodDialogues and FoodReasonSeg datasets.

IV. METHOD

The network architecture of FoodLMM is illustrated in Fig. 2. Our FoodLMM takes image and text prompts as input and is capable of handling various food tasks in a unified model. The architecture of FoodLMM is built upon Large Multi-modal Model (LMM) LISA [42], which is a LMM with powerful language generation and image segmentation ability. As we focus on not only language generation tasks, such as ingredient recognition and recipe generation but also food segmentation, we naturally choose LISA instead of other LMMs, such as LLava [17], to build our FoodLMM.

We adopt a two-stage training strategy to specialize the general LMM in the food domain. The first training stage aims to inject sufficient food-related basic knowledge into the LMM by using instruction-following data constructed in Section III-A based on five tasks in the food domain. As the output of the three tasks of food classification, ingredient recognition, and recipe generation are pure text, these tasks could be learned through pure language auto-regressive, similar to LLaVa. The output of food segmentation is multi-modal, with both masks and text. Nutrition estimation aims to produce nutrition values, e.g., mass and calories. We propose task-specific tokens and heads for food segmentation and nutrition estimation to address this issue. For the nutrition estimation task, different types of nutritional values are predicted by a series of nutrition tokens and regression heads. Specifically, we set 10 task-specific tokens corresponding to 10 nutritional regression heads, and then divide these tokens into two categories: ingredient-level and dish-level. The ingredient-level tokens include $\langle \text{mass} \rangle$, $\langle \text{cal} \rangle$, $\langle \text{carb} \rangle$, $\langle \text{fat} \rangle$, $\langle \text{pro} \rangle$ for regressing the mass, calories, carbohydrates, fats, and proteins of any ingredient respectively. And the dish-level tokens include $\langle \text{total_mass} \rangle$, $\langle \text{total_cal} \rangle$, $\langle \text{total_carb} \rangle$, $\langle \text{total_fat} \rangle$, $\langle \text{total_pro} \rangle$ for predicting the overall nutritional element values of the input food image. Note that in the implementation, we assign numbers to the ingredient-level tokens, such as $\langle \text{mass}_1 \rangle$, $\langle \text{mass}_2 \rangle$, ..., $\langle \text{mass}_n \rangle$ to differentiate the nutritional elements of different ingredients mentioned in the answers. The LMM base (LLaVA) generates only these tokens without any specific value in response. We then replace the corresponding tokens with the nutritional values obtained from the regression heads to produce the final output answer.

For the Referring/Reasoning Segmentation task, the segmentation masks are generated through multiple additional segmentation tokens and the segmentation encoder-decoder. We set task-specific token $\langle \text{seg} \rangle$ for referring and reasoning segmentation tasks. The LMM base (LLaVA) learns to assign $\langle \text{seg} \rangle$ tokens to the ingredients mentioned in the answer, and maps its corresponding hidden state to hidden embedding through the MLP. Hidden embedding serves as the segmentation prompt. The decoder is trained to generate a segmentation mask according to a $\langle \text{seg} \rangle$ hidden embedding. Consistent with ingredient-level nutrition estimation tokens, $\langle \text{seg} \rangle$ tokens also use numerical numbers to distinguish different ingredients, i.e., $\langle \text{seg}_1 \rangle, \langle \text{seg}_2 \rangle, \dots, \langle \text{seg}_n \rangle$. We then combine all generated masks to obtain the final segmented image containing multiple masks. The LMM is fine-tuned using LoRA [86] to learn the generation of task-specific tokens and the answers for various tasks. The second training stage focuses on endowing FoodLMM with the capability to conduct multiple rounds of highly professional conversations and provide detailed explanations with segmentation masks for queries requiring complex reasoning.

A. Stage 1: Multi-Task Learning

We adopt a multi-task learning approach in this stage, in order to enable LMM to handle different basic food tasks. The tasks include **Food VQA** (food classification, ingredient recognition, recipe generation), **Nutrition Estimation** and **Referring Segmentation**.

Food VQA: We design different question-and-answer templates for food classification, ingredient recognition and recipe generation tasks. Similar to LLaVA, FoodLMM is trained by language autoregression. The loss can be formalized as follows:

$$\mathcal{L}_{txt} = \text{CE}(\hat{\mathbf{y}}_{txt}, \mathbf{y}_{txt}), \quad (1)$$

where $\hat{\mathbf{y}}_{txt}$ stands for the generated sentences (tokens), \mathbf{y}_{txt} represents for the ground-truth.

Nutrition Estimation: We introduce nutritional tokens in FoodLMM for nutrition estimation. Specifically, a total of ten nutritional tokens are added into the vocabulary: $\langle mass \rangle$, $\langle cal \rangle$, $\langle carb \rangle$, $\langle fat \rangle$, $\langle pro \rangle$ are used for regressing the mass, calories, carbohydrates, fats, and proteins of any ingredient respectively, and $\langle total_mass \rangle$, $\langle total_cal \rangle$, $\langle total_carb \rangle$, $\langle total_fat \rangle$,

$\langle total_pro \rangle$ are used for predicting the overall nutritional element values of the input food picture.

The hidden states of these nutritional tokens are processed into embeddings through a Multilayer Perceptron (MLP). These embeddings are fed into the corresponding regression heads to obtain the predicted values. Finally, the predicted values replace the task-specific tokens in the original text output to obtain the answer with nutritional value.

We employ the Mean Absolute Error (MAE) and the Mean Squared Error Loss (MSE) as follows:

$$\mathcal{L}_{nutrition} = \lambda_{MAE} \left(\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \right) + \lambda_{MSE} \left(\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right), \quad (2)$$

where \hat{y}_i stands for the predictions, y_i for the ground-truth values, i for different regression heads, λ_{MAE} and λ_{MSE} are the corresponding weights of the MAE and MSE losses.

Referring Segmentation: To embed ingredient segmentation into our FoodLMM, we introduce segmentation tokens, denoted as $\langle seg_i \rangle$, with i representing the ingredient's index, into the primary vocabulary. When the model receives a query and its associated image, it generates a text response containing tokens for the ingredients to be segmented. The hidden states of these tokens are converted through an MLP into embeddings that capture the relevant ingredient information. These embeddings, combined with the visual representation from SAM's encoder, are processed by the SAM decoder to create the ingredient masks.

The segmentation loss is denoted as \mathcal{L}_{mask} , integrating the per-pixel binary cross-entropy (BCE) loss alongside the Dice coefficient loss (DICE) [87] as follows:

$$\begin{aligned} \mathcal{L}_{mask} &= \lambda_{bce} \frac{1}{n} \sum_{i=1}^n BCE(\hat{M}_i, M_i) \\ &\quad + \lambda_{dice} \frac{1}{n} \sum_{i=1}^n DICE(\hat{M}_i, M_i) \end{aligned} \quad (3)$$

where \hat{M}_i and M_i stand for the predicted mask and the ground-truth of token $\langle seg_i \rangle$ respectively. Every projected embedding of $\langle seg_i \rangle$ token is fed into the decoder successively to generate the corresponding \hat{M}_i .

Almost all the referring segmentation models are limited to producing a single binary mask for different queries [82], [88], [89], [90]. To address this issue, we aim to enable FoodLMM to go *Beyond One-to-One* [44], i.e., to produce different numbers of masks based on the queries. To achieve this, we design three different types of referring instructions: (1) to segment certain specified ingredients (one-to-one/one-to-many), e.g., Query: Segment the fish and lemon in this picture. Answer: The fish is masked as $\langle seg_1 \rangle$ and the lemon as $\langle seg_2 \rangle$. (2) to segment all visible ingredients (one-to-many), e.g., Query: Segment all the ingredients in this photo. Answer: The fish is masked as $\langle seg_1 \rangle$, lemon as $\langle seg_2 \rangle$, carrot as $\langle seg_3 \rangle$ and green beans as $\langle seg_4 \rangle$. (3) to segment objects that are not present in the image (one-to-zero), e.g., Query: Segment the watermelon in this picture. Answer: The watermelon is not found in this picture. Through these three

TABLE V
TRAINING HYPERPARAMETERS

Model	Training	LoRA
$\lambda_{txt}=1.0$	precision=bf16	r=8
$\lambda_{nutrition}=0.1$	optimizer=AdamW	alpha=16
$\lambda_{mask}=1.0$	learning_rate=3e-4	dropouts=5e-2
$\lambda_{MSE}=1.0$	weight_decay=0.0	
$\lambda_{MAE}=1e-3$	warmup_type=linear	
$\lambda_{bce}=2.0$	warmup_num_steps=100	
$\lambda_{dice}=0.5$	gradient_clipping=1.0	

instructions, the challenge of referring segmentation is dexterously addressed, realizing *One-to-Any* segmentation.

Our FoodLMM is trained by combining all the tasks and data with multi-task learning. Thus, the overall loss \mathcal{L} consists of \mathcal{L}_{txt} , \mathcal{L}_{mask} and $\mathcal{L}_{nutrition}$, weighted by λ_{txt} , $\lambda_{nutrition}$ and λ_{mask} , as formulated below:

$$\mathcal{L} = \lambda_{txt} \mathcal{L}_{txt} + \lambda_{nutrition} \mathcal{L}_{nutrition} + \lambda_{mask} \mathcal{L}_{mask}. \quad (4)$$

B. Stage 2: Fine-Tuning for a Versatile Food Assistant.

The second stage focuses on training FoodLMM's **Multi-round Conversation** and **Reasoning Segmentation** capabilities. Relying on the generated datasets **FoodDialogues** and **FoodReasonSeg**, a versatile food conversational assistant can be achieved.

Multi-round Conversation: We adjust the generated dialogues to the format of instruction-following data (defined in Section III) to train FoodLMM's multi-round conversation capability. In one conversation, the image is input only before the first round of questioning. In subsequent rounds, each new question is appended to the existing historical context, guiding the LMM to produce a subsequent response. The associated loss function is the autoregression loss \mathcal{L}_{txt} (1).

Reasoning Segmentation: The data format is consistent with multi-round conversation. Every answer contains segmentation tokens $\langle seg_i \rangle$, and segmentation loss \mathcal{L}_{mask} (3) is used to enhance the segmentation capability.

V. EXPERIMENT

A. Experimental Setup

Training Details: We train our FoodLMM from LISA-7B-v1-explanatory model [42], using four NVIDIA 40 G A100 GPUs. AdamW optimizer combined with the WarmupDecayLR is used as a learning rate scheduler. We set the initial learning rate to 0.0003, the weight decay to 0, and use 100 warm-up iterations. Following LISA, we set the weights λ_{BCE} and λ_{Dice} to 2.0 and 0.5 respectively. For the nutrition estimation task, we empirically adjust the weights λ_{MAE} and λ_{MSE} to 0.1 and 0.0001 respectively, to balance their magnitudes with other losses. Throughout the training, we employ LoRA to fine-tune LLaVa. In addition, the parameters for SAM's decoder, the MLP, and the heads dedicated to nutritional value estimation are all trainable.

In stage 1, the model is trained with a batch size of 4 over 10000 training steps, taking 2 days on 4 NVIDIA A100 (40 G) GPUs. The training hyperparameter settings and the sampling

TABLE VI
SAMPLING RATIO OF DIFFERENT DATASETS

Dataset	Stage 1	Stage 2
VIREO Food-172 [20]	5	5
Recipe1M [26]	15	10
Nutrition5k [32]	10	10
FoodSeg103 [36]	6	6
UECFoodPixComplete [35]	4	4
FoodDialogues	—	45
FoodReasonSeg	—	30

ratios of different datasets are detailed in Tables V and VI (Stage 1), respectively.

We denote the model trained from Stage 1 as **FoodLMM S1**. In Stage 2, the batch size is reduced to 2, and training takes 28 hours on 4 NVIDIA A100 (40 G) GPUs over 8000 training steps. The training hyperparameters are consistent with Stage 1 (Tab. V), and the sampling ratio is shown in Table VI (Stage 2). The model trained from Stage 2 is denoted as **FoodLMM Chat**. To better evaluate the ability of FoodLMM on specific tasks, we fine-tune FoodLMM on each task, recorded as **FoodLMM FT**, and the batch size during fine-tuning is set to 4.

B. Evaluation Metrics.

For **food classification**, top-1 accuracy is adopted as the evaluation metric. While for **ingredient recognition**, the Intersection over Union (IoU) and F1-score are used to evaluate the performances. For **recipe generation**, we follow [27], [91] and apply SacreBLEU, Rouge-L metrics to quantify the quality of the generated recipes. For **nutrition estimation**, following [32], we use Mean Absolute Error (MAE) and the percent of MAE to the respective mean for that field to measure regression accuracy for calories, mass, and individual macro-nutrient mass. Caloric MAE is measured in kilocalories, and all other amounts are measured in grams. For **referring segmentation**, in line with LISA [42], we employ the Complete Intersection over Union (cIoU) metric to evaluate the performance. Meanwhile, to evaluate the risk of our model incorrectly rejecting segmenting existing ingredients, we also report the probability of accurate response. For one-to-zero referring segmentation, we report the accuracy of our model successfully rejecting to segment non-existent items. For **reasoning segmentation**, following LISA [42], Generalized Intersection over Union (gIoU) along with cIoU are adopted for the evaluation of Reasoning Segmentation.

C. FoodVQA Results

Our FoodVQA consists of three tasks, i.e., food classification, ingredient recognition, and recipe generation. For these tasks, we compare the performance of our model with baseline models that achieve state-of-the-art performances. Note that other LMMs (e.g., LISA [42]) are general-purpose with limited knowledge in the food domain. Their outputs of the FoodVQA tasks are uncontrollable and difficult to be quantitatively evaluated. Empirically, the results are also not promising. Consequently, we do not include the results of other LMMs as baselines for the FoodVQA tasks. Table VII summarizes the results. As shown in

the table, for all these three tasks, our FoodLMM outperforms the SOTA methods and achieves the best results. Specifically, for food classification, by fine-tuning the Food-101 dataset, our model attains an accuracy of 93.93%, which is 3.07% higher than the SOTA method [55]. For ingredient recognition, our model outperforms CACLNet [92], a SOTA method on the VIREO Food-172 dataset, for 3.2% in terms of F1. For recipe generation, without any extra information, our FoodLMM significantly outperforms FIRE [91], the SOTA method on the Recipe1M dataset that leverages ground-truth ingredient labels for recipe generation. Fig. 6 further shows qualitative examples for foodVQA. As can be seen, our FoodLMM can generate accurate answers for food classification, ingredient recognition and recipe generation.

D. Nutrition Estimation Results

We further evaluate the performance of our model in nutrition estimation, the results are shown in Table VII. After training Stage 1, FoodLMM has been able to accurately estimate the overall nutritional value based on food images, decreasing the prediction error by 4.5% on average compared to the previous SOTA [32]. Examples of using FoodLMM to estimate the total nutritional value of food are shown in Fig. 6. Moreover, FoodLMM is also able to predict the nutritional value of specific ingredients in dishes, which is named Referring Nutrition Estimation. It is worthwhile mentioning that previous methods lack the ability of Referring Nutrition Estimation. For evaluating the performance of Referring Nutrition Estimation, we select the top three ingredients with the largest mass as referred ingredients. Table VII summarizes the results. In the table, refer@kst denotes referring the ingredient with the k th highest mass. The results show that although further fine-tuning does not enhance the performance of overall nutrition estimation, it significantly improves the accuracy of nutrition estimation for specific ingredients.

E. Referring Segmentation Results

We compare our FoodLMM with LISA [42] on FoodSeg103 and UECFoodPIX datasets, the results are shown in Table VII. For standard referring segmentation (one-to-one), FoodLMM's cIoU scores significantly exceed those of LISA, achieving 20% improvement (0.65 to 0.78) on FoodSeg103 and 25% (0.72 to 0.90) on UECFoodPIX datasets. Since we deliberately train FoodLMM's one-to-zero ability, the model may inevitably give a zero mask to the correct referring, i.e., mistakenly identifying the queried ingredient as non-existent. We examine the probability of accurate response (ACC) to measure this impact. Notably, the decrease in FoodLMM's accuracy is minimal (99.16% on FoodSEG103 and 98.69% on UECFoodPIX) despite significant improvements in cIoU.

One-to-many: We then evaluate the performance of segmentation involving multiple ingredients referred in a query. Table VII summarizes the segmentation results of three (refer@3) and five (refer@5) referred ingredients. The results show that FoodLMM, after the first training stage (FoodLMM S1), achieves high cIoU scores (around 0.7) on the refer@3 task on two datasets, while also maintaining high accuracy. However,

TABLE VII
PERFORMANCE COMPARISON ON DIFFERENT FOOD BENCHMARKS

Food Classification in Food-101 [19]		Ingredient Recognition in VIREO Food-172 [20]			Recipe Generation in Recipe1M [26]		
Method	Acc.	Method	IoU	F1	Method	SacreBLEU	Rouge-L
PRENet [55]	91.13	CACLNet [92]	—	65.71	FIRE [91]	6.02	21.29
FoodLMM FT	93.93	FoodLMM FT	56.94	68.97	FoodLMM FT*	6.24	36.96
Dataset	Method	Total Caloric MAE	Total Mass MAE	Total Fat MAE	Total Carb MAE	Total Protein MAE	Average
Nutrition5k [32]	2D Direct Prediction [32]	70.6 / 26.1%	40.4 / 18.8%	5.0 / 34.2%	6.1 / 31.9%	5.5 / 29.5%	25.52 / 28.1%
	FoodLMM S1	67.3 / 26.6 %	38.8 / 20.2 %	5.5 / 40.4 %	6.1 / 31.9 %	4.2 / 26.2 %	24.38 / 29.1%
	FoodLMM FT	67.3 / 26.6 %	39.7 / 20.7 %	5.4 / 39.7 %	5.9 / 31.1 %	4.1 / 25.8 %	24.48 / 28.8%
Dataset	Refer ingredient	Referring Nutrition Estimation					
Nutrition5k [32]	refer@1st	Model	Caloric MAE	Mass MAE	Fat MAE	Carb MAE	Protein MAE
	FoodLMM S1	69.3 / 45.8 %	56.1 / 43.9 %	3.8 / 47.6 %	3.2 / 30.3 %	3.1 / 29.4 %	—
	FoodLMM FT	25.2 / 34.7 %	47.3 / 37.1 %	3.7 / 46.1 %	3.7 / 34.1 %	2.5 / 22.4 %	—
	refer@2nd	FoodLMM S1	34.8 / 45.5 %	21.2 / 34.1 %	1.1 / 29.4 %	2.0 / 29.8 %	1.7 / 33.4 %
	FoodLMM FT	21.0 / 27.4 %	16.9 / 27.1 %	1.1 / 29.0 %	2.1 / 30.5 %	1.4 / 27.4 %	—
	refer@3rd	FoodLMM S1	30.5 / 71.2 %	21.0 / 55.4 %	1.1 / 50.5 %	1.7 / 45.2 %	1.6 / 66.7 %
	FoodLMM FT	16.9 / 39.5 %	14.2 / 37.5 %	0.9 / 39.9 %	1.5 / 38.1 %	1.2 / 48.2 %	—
Dataset	Method	Referring Segmentation (one-to-one/one-to-many)			refer@3		
FoodSeg103 [36]	LISA-7B [42]	cIoU	Acc.	cIoU	Acc.	cIoU	Acc.
	FoodLMM S1	0.65	100.00%	—	—	—	—
	FoodLMM FT	0.80	99.16%	0.73	98.99%	0.58	98.34%
	LISA-7B [42]	0.72	98.69%	0.74	99.51%	0.62	99.29%
UEC-FoodPIX [35]	FoodLMM S1	0.90	100.00%	—	—	—	—
	FoodLMM FT	0.90	99.70%	0.72	97.51%	0.64	98.10%
	FoodLMM FT	0.90	99.60%	0.76	99.38%	0.68	99.05%
Dataset	Method	Referring Segmentation (one-to-zero)			refer@5		
FoodSeg103 [36]	refer@1	Acc.	refer@3	Dataset	Method	refer@1	Acc.
	LISA-7B	0.00%	—	UEC-FoodPIX [35]	LISA-7B	0.00%	—
	FoodLMM S1	73.07%	57.80%		FoodLMM S1	77.50%	52.30%
	FoodLMM FT	87.87%	88.99%		FoodLMM FT	83.70%	83.50%

"FoodLMM S1" denotes the model trained from the first stage, and "FoodLMM FT" represents the model finetuned on specific tasks.

Food VQA



USER: What is the dish in the picture made of?

FoodLMM: The ingredients are chicken, margarine, cracker, cheese, parsley.

Reasoning Segmentation



USER: Can you identify a dessert item in this image that I could indulge in to satisfy my sweet tooth? Please output segmentation mask.

FoodLMM: Sure, the dessert item you could indulge here is **Ice Cream [SEG1]**. It's typically high in sugar and fat, which may be calorie-dense and satisfying for those with a sweet tooth.

Nutrition Estimation



USER: Can you provide the total caloric, fat, carbohydrate, and protein content of this dish directly?

FoodLMM:

- Weight: 672.0 g
- Calories: 760.0 kcal
- Fat: 46.5 g
- Carbohydrates: 24.0 g
- Protein: 68.0 g

Referring Segmentation



USER: Please segment the beef, egg and tomatoes in this image.

FoodLMM: Sure, the beef is masked as **[SEG1]** egg as **[SEG2]** and tomato as **[SEG3]**.



USER: Please segment the watermelon in this image.

FoodLMM: The watermelon is not found in this picture.

Multi-round Conversation



USER: Could you please provide some nutritional information about the frozen yogurt in the image?

ASSISTANT: Absolutely! The Frozen yogurt weighs approximately 112.0 grams and contains 176.0 kcal. It consists of 6.66g of fat, 22.75g of carbohydrates, and a significant amount of carbohydrates, as it often contains a range of sugar and fruits.



USER: Is frozen yogurt healthier than traditional ice cream?

ASSISTANT: Among the two, frozen yogurt is generally healthier. It typically contains less fat and more protein compared to traditional ice cream. It's also often a source of calcium and probiotics, which are beneficial for bone health and digestive health respectively.

Fig. 6. Qualitative examples of various capabilities of FoodLMM.

TABLE VIII
THE HUMAN EVALUATION RESULTS OF FOODLMM CHAT

Score	1	2	3	4	5
Number	5	22	50	73	50

TABLE IX
PERFORMANCE COMPARISON IN REASONING SEGMENTATION

Method	gIoU	cIoU
LISA-7B	0.19	0.17
FoodLMM S1	0.35	0.34
FoodLMM Chat	0.71	0.72

as the number of referred ingredients increases the segmentation becomes more challenging. Nevertheless, the fine-tuned model (FoodLMM FT) attains higher cIoU scores across all scenarios and maintains a high probability of accurate responses (over 99%) even in complex segmentation tasks. Fig. 6 shows the examples of referred ingredient segmentation.

One-to-zero: To further examine the reasoning performance of our FoodLMM, we evaluate its performance under a challenging scenario, where the model is required to segment the ingredients that are not present in the image. For such a query, the model should reject to return the segmentation mask. The performances of one-to-zero segmentation are shown in Table VII. As LISA is not trained with one-to-zero samples, the accuracy in identifying absent ingredients on both FoodSeg103 and UECFoodPIX datasets is 0.0%. In contrast, our FoodLMM, attain an accuracy of 73.07% on FoodSeg103 and 77.50% on UECFoodPIX, when handling one single nonexistent referring (*refer@1*), after the first train stage. The task becomes more challenging when there are three absent referring (*refer@3*) in the query. Nevertheless, FoodLMM still maintains a high accuracy, surpassing 80% after fine-tuning. Fig. 6 presents an example where the user requests FoodLMM to segment watermelon from the seafood vegetable soup, but the model declines.

F. Performance on the Proposed Benchmarks

Multi-turn Conversation: We further evaluate the performance of our model on the Multi-turn Conversation dataset. The evaluation is performed by user study. Specifically, we invite 10 participants to manually evaluate the satisfaction score (1 to 5) of 20 sets of multi-turn dialogues (totalling 200 dialogue sets, all of the questions are randomly selected from the test set of our FoodDialogues). The results are summarized in Table VIII. The majority of the dialogues (173 out of 200) receive high scores (≥ 3), and 4 is the most frequent score. It indicates that our FoodLMM is capable of closely aligning with user preferences and generating high-quality answers that are both accurate and explanatory.

Reasoning Segmentation: We compare the performance of our FoodLMM with LISA in reasoning segmentation on our FoodReasonSeg benchmark. The results are shown in Table IX. From the results, LISA achieves a quite low gIoU and cIoU, which are 0.19 and 0.17, respectively. The results indicate that although LISA is excellent in the general domain's reasoning

TABLE X
PERFORMANCE COMPARISON BETWEEN MULTI-TASK LEARNING AND SINGLE-TASK LEARNING

Referring Segmentation One-to-One/Many (cIoU/Acc)		Referring Segmentation One-to-Zero (Acc)	
multi-task	0.75 / 99.25%	multi-task	86.02%
Single-task	0.76 / 98.20%	Single-task	89.07%
Nutrition Estimation (MAE)		Referring Nutrition Estimation (MAE)	
multi-task	24.48 / 28.8%	multi-task	10.64 / 34.60%
Single-task	25.6 / 30.5%	Single-task	16.40 / 43.62%
Ingredient Recognition (IoU/F1)		Recipe Generation (BLEU/Rouge-L)	
multi-task	56.94 / 68.97	multi-task	5.89 / 35.12
Single-task	56.49 / 68.72	Single-task	6.24 / 36.65

The average results are reported.

segmentation, due to a lack of specialized knowledge in the food domain, its performance in food reasoning segmentation is less impressive. In contrast, our FoodLMM S1 achieves much better performance by learning food knowledge from public datasets. After the Stage 2 fine-tuning, the performance of FoodLMM is significantly improved, with the gIoU score increasing from 0.35 to 0.71 and the cIoU score from 0.34 to 0.72 compared to FoodLMM S1, indicating that FoodLMM acquires more expertise and stronger reasoning ability from the constructed datasets.

G. Discussion About Multi-Task and Single-Task Learning

In this section, we compare the performance of our FoodLMM in single-task and multi-task learning. Multi-task model is capable of dealing with various food tasks, while single-task model focuses on a specific task. Table X lists the performance differences between multi-task and single-task models. For the Referring Segmentation task, we present average segmentation results for variable number of ingredients across FoodSeg103 [36] and UEC-FoodPIX Complete [35]. For Nutrition Estimation task, we report the average prediction results for five essential nutrients. We observe that the advantage of multi-task is the ability to leverage knowledge across different tasks, which benefits some basic recognition tasks such as ingredient recognition and nutrition estimation. For more complex tasks, multi-task is only able to boost the performance of referring nutrition estimation. In general, single-task learning may benefit complex tasks such as recipe generation, as model training can be more easily optimized. Nevertheless, multi-task learning still suffers from the problem of gradient conflicts between tasks [93], [94], [95], which remains an open issue and is worthy of further research.

VI. CONCLUSION

We have presented FoodLMM, a versatile large multi-modal model for the food domain. This model is proficient in understanding and responding to various food-related questions, including Food Classification, Ingredient Recognition, Recipe Generation, Nutrition Estimation, Referring Segmentation, and Reasoning Segmentation. We also establish two benchmarks specific to the food domain: one for multi-turn dialogues that involve complex reasoning, and another for food-related reasoning segmentation. These benchmarks are crucial for assessing the reasoning and dialogue capabilities. Fine-tuned via reasoning instructions, FoodLMM shows outstanding performance in complex reasoning tasks and multi-turn dialogues about food-related subjects.

REFERENCES

- [1] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 27730–27744.
- [2] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.
- [3] T. Wu et al., "A brief overview of chatGPT: The history, status quo and potential future development," *IEEE/CAA J. Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, May 2023.
- [4] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.
- [5] H. Touvron et al., "LLaMA 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.
- [6] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [7] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 12888–12900.
- [8] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.
- [9] Y. Fang et al., "Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19358–19369.
- [10] W. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," *Adv. Neural Inf. Process. Syst.*, vol. 36 pp. 49250–49267, 2023.
- [11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Adv. neural Inf. Process. Syst.*, vol. 36, pp. 34892–34916, 2023.
- [12] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*.
- [13] T. Gong et al., "MultiModal-GPT: A vision and language model for dialogue with humans," 2023, *arXiv:2305.04790*.
- [14] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "NEXT-GPT: Any-to-any multimodal LLM," in *Proc. 41th Int. Conf. Mach. Learn.*, 2024.
- [15] J. Chen et al., "MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning," 2023, *arXiv:2310.09478*.
- [16] X. Yang, F. Liu, and G. Lin, "Effective end-to-end vision language pre-training with semantic visual loss," *IEEE Trans. Multimedia*, vol. 25, pp. 8408–8417, 2023.
- [17] C. Li et al., "LLaVA-MED: Training a large language-and-vision assistant for biomedicine in one day," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 28541–28564, 2023.
- [18] S. Wei, C. Luo, Y. Luo, and J. Xu, "Privileged modality learning via multimodal hallucination," *IEEE Trans. Multimedia*, vol. 26, pp. 1516–1527, 2024.
- [19] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101-mining discriminative components with random forests," in *Proc. 13th Euro. Conf. Comput. Vis.*, 2014, pp. 446–461.
- [20] J. Chen and C. -W. Ngo, "Deep-based ingredient recognition for cooking recipe retrieval," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 32–41.
- [21] M. Bolaños, A. Ferrà, and P. Radeva, "Food ingredients recognition through multi-label learning," in *Proc. New Trends Image Anal. Process.* Sep. 2017, 2017, pp. 394–402.
- [22] J. Chen et al., "Zero-shot ingredient recognition by multi-relational graph convolutional network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 07, pp. 10542–10550.
- [23] M. Zhang, G. Tian, Y. Zhang, and H. Liu, "Sequential learning for ingredient recognition from images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 5, pp. 2162–2175, May 2023.
- [24] J. Gao, J. Chen, H. Fu, and Y.-G. Jiang, "Dynamic mixup for multi-label long-tailed food ingredient recognition," *IEEE Trans. Multimedia*, vol. 25, pp. 4764–4773, 2023.
- [25] B. Zhu, C. -W. Ngo, and W.-K. Chan, "Learning from web recipe-image pairs for food recognition: Problem, baselines and performance," *IEEE Trans. Multimedia*, vol. 24, pp. 1175–1185, 2021.
- [26] A. Salvador et al., "Learning cross-modal embeddings for cooking recipes and food images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3020–3028.
- [27] A. Salvador, M. Drozdzal, X. Giró-i Nieto, and A. Romero, "Inverse Cooking: Recipe generation from food images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10453–10462.
- [28] H. H. Lee et al., "RecipeGPT: Generative pre-training based cooking recipe generation and evaluation system," in *Proc. Companion Web Conf.* 2020, pp. 181–184.
- [29] H. Wang, G. Lin, S. C. Hoi, and C. Miao, "Structure-aware generation network for recipe generation from images," in *Proc. 16th Euro. Conf. Comput. Vis.*, 2020, pp. 359–374.
- [30] H. Wang, G. Lin, and S. C. Hoi, "Learning structural representations for recipe generation and food retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3363–3377, Mar. 2022.
- [31] M. Zhang, G. Tian, Y. Zhang, and P. Duan, "Reinforcement learning for logic recipe generation: Bridging gaps from images to plans," *IEEE Trans. Multimedia*, vol. 24, pp. 352–365, 2021.
- [32] Q. Thaines et al., "Nutrition5k: Towards automatic nutritional understanding of generic food," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8903–8911.
- [33] F. P. W. Lo, Y. Guo, Y. Sun, J. Qiu, and B. Lo, "An intelligent vision-based nutritional assessment method for handheld food items," *IEEE Trans. Multimedia*, vol. 25, pp. 5840–5851, 2023.
- [34] C. N. Freitas, F. R. Cordeiro, and V. Macario, "MyFood: A food segmentation and classification system to aid nutritional monitoring," in *Proc. 33rd Conf. Graph., Patterns Images*, 2020, pp. 234–239.
- [35] K. Okamoto and K. Yanai, "UEC-foodpix complete: A large-scale food image segmentation dataset," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 647–659.
- [36] X. Wu, "A large-scale benchmark for food image segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 506–515.
- [37] Z. Zhu and Y. Dai, "A new CNN-based single-ingredient classification model and its application in food image segmentation," *J. Imag.*, vol. 9, no. 10, 2023, Art. no. 205.
- [38] X. Dong, W. Wang, H. Li, and Q. Cai, "Windows attention based Pyramid network for food segmentation," in *Proc. IEEE 7th Int. Conf. Cloud Comput. Intell. Syst.*, 2021, pp. 213–217.
- [39] Y. Honbu and K. Yanai, "Unseen food segmentation," in *Proc. 2022 Int. Conf. Multimedia Retrieval*, 2022, pp. 19–23.
- [40] X. Lai et al., "FoodSAM: Any food segmentation," *IEEE Trans. Multimedia*, 2023.
- [41] G. Sinha et al., "Transferring knowledge for food image segmentation using transformers and convolutions," 2023, *arXiv:2306.09203*.
- [42] X. Lai et al., "LISA: Reasoning segmentation via large language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9579–9589.
- [43] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.
- [44] Y. Hu et al., "Beyond one-to-one: Rethinking the referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4067–4077.
- [45] A. Koubaa, "GPT-4 vs. GPT-3.5: A concise showdown," 2023.
- [46] W.-L. Chiang et al., "Vicuna: An open-source chatbot impressing GPT-4 with 90% chatgpt quality," 2023. Accessed: 14 Apr. 2023. [Online]. Available: <https://vicuna.lmsys.org>
- [47] A. Awadalla et al., "OpenFlamingo: An open-source framework for training large autoregressive vision-language models," 2023, *arXiv:2308.01390*.
- [48] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Pmc-llama: Further finetuning llama on medical papers," 2023, *arXiv:2304.14454*.
- [49] Y. Li, "ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge," *Cureus*, vol. 15, no. 6, 2023.
- [50] Y. Luo, "BiomedGPT: Open multimodal generative pre-trained transformer for biomedicine," 2023, *arXiv:2308.09442*.
- [51] S. Wu, "BloombergGPT: A large language model for finance," 2023, *arXiv:2303.17564*.
- [52] J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, "ChatLaw: Open-source legal large language model with integrated external knowledge bases," 2023, *arXiv:2306.16092*.
- [53] J. Chen, B. Zhu, C.-W. Ngo, T.-S. Chua, and Y.-G. Jiang, "A study of multi-task and region-wise deep learning for food ingredient recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 1514–1526, 2020.
- [54] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. Comput. Vis.* 2015, pp. 3–17.
- [55] W. Min, "Large scale visual food recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9932–9949, Aug. 2023.
- [56] C. Liu, "DeepFood: Deep learning-based food image recognition for computer-aided dietary assessment," in *Proc. Int. Conf. Smart Homes Health Telematics*, 2016, pp. 37–48.

- [57] F. Offi, Y. Aytar, I. Weber, R. Al Hammouri, and A. Torralba, "Is Saki# delicious? the food perception gap on Instagram and its relation to health," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 509–518.
- [58] C. Kiourt, G. Pavlidis, and S. Markantonatou, "Deep learning approaches in food recognition," *Mach. Learn. Paradigms: Adv. Deep Lear.-based Technol. Appl.*, pp. 83–108, 2020.
- [59] S. Mezgec and B. Koroušić Seljak, "NutriNet: A deep learning food and drink image recognition system for dietary assessment," *Nutrients*, vol. 9, no. 7, 2017, Art. no. 657.
- [60] X. Chen, Y. Zhu, H. Zhou, L. Diao, and D. Wang, "ChineseFoodNet: A large-scale image dataset for chinese food recognition," 2017, *arXiv:1705.02743*.
- [61] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *Proc. 2018 IEEE Winter Conf. Appl. Comput. Vis.* 2018, pp. 567–576.
- [62] G. Liu, Y. Jiao, J. Chen, B. Zhu, and Y. -G. Jiang, "From canteen food to daily meals: Generalizing food recognition to more practical scenarios," *IEEE Trans. Multimedia*, vol. 27, pp. 2724–2733, 2025.
- [63] C. Liu, Y. Liang, Y. Xue, X. Qian, and J. Fu, "Food and ingredient joint learning for fine-grained recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 27336, pp. 2480–2493, Jun. 2020.
- [64] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R2GAN: Cross-modal recipe retrieval with generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11477–11486.
- [65] J.-J. Chen, C.-W. Ngo, F.-L. Feng, and T.-S. Chua, "Deep understanding of cooking procedure for cross-modal recipe retrieval," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1020–1028.
- [66] J.-j. Chen, C.-W. Ngo, and T.-S. Chua, "Cross-modal recipe retrieval with rich food attributes," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1771–1779.
- [67] H. Wang et al., "Cross-modal food retrieval: Learning a joint embedding of food images and recipes with semantic consistency and attention mechanism," *IEEE Trans. Multimedia*, vol. 24, pp. 2515–2525, 2021.
- [68] B. Zhu, C.-W. Ngo, and J. -j. Chen, "Cross-domain cross-modal food transfer," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3762–3770.
- [69] F. Song, B. Zhu, Y. Hao, and S. Wang, "Enhancing recipe retrieval with foundation models: A data augmentation perspective," in *Proc. Eur. Conf. Comput. Vis.*, Cham: Springer Nature Switzerland, 2024, pp. 111–127.
- [70] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, "Food image analysis: Segmentation, identification and weight estimation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2013, pp. 1–6.
- [71] T. Ege, W. Shimoda, and K. Yanai, "A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice," in *Proc. 5th Int. Workshop Multimedia Assist. Dietary Manage.*, 2019, pp. 82–87.
- [72] L. Hollywood, G. Armstrong, and M. Durkin, "Using behavioural and motivational thinking in food segmentation," *Int. J. Retail Distrib. Manage.*, vol. 35, no. 9, pp. 691–702, 2007.
- [73] W. Min, S. Jiang, and R. Jain, "Food recommendation: Framework, existing solutions, and challenges," *IEEE Trans. Multimedia*, vol. 22, no. 10, pp. 2659–2671, Oct. 2019.
- [74] D. Elsweiler, C. Trattner, and M. Harvey, "Exploiting food choice biases for healthier recipe recommendation," in *Proc. 40th Int. acm sigir Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 575–584.
- [75] J. Freyne and S. Berkovsky, "Intelligent food planning: Personalized recipe recommendation," in *Proc. 15th Int. Conf. Intell. User Interfaces*, 2010, pp. 321–324.
- [76] K. H. Brown, S. E. Wuehler, and J. M. Peerson, "The importance of zinc in human nutrition and estimation of the global prevalence of zinc deficiency," *Food Nutr. Bull.*, vol. 22, no. 2, pp. 113–125, 2001.
- [77] L. B. Mirel et al., "National health and nutrition examination survey: Estimation procedures, 2007-2010," *Vital Health Statistics. Ser. 2 Data Eval. Methods Res.*, no. 159, pp. 1–17, 2013.
- [78] T.-C. Chen, J. Clark, M. K. Riddles, L. K. Mohajer, and T. H. Fakhouri, "National health and nutrition examination survey, 2015-2018: Sample design and estimation procedures," *Vital Health Statist. Ser. 2, Data Eval. Methods Res.*, no. 184, pp. 1–35, 2020.
- [79] K. Kitamura, T. Yamasaki, and K. Aizawa, "Food log by analyzing food images," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 999–1000.
- [80] D. Sahoo et al., "FoodAI: Food image recognition via deep learning for smart food logging," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 2260–2268.
- [81] J. Chen, W. Berkman, M. Bardouh, C. Y. K. Ng, and M. Allman-Farinelli, "The use of a food logging app in the naturalistic setting fails to provide accurate measurements of nutrients and poses usability challenges," *Nutrition*, vol. 57, pp. 208–216, 2019.
- [82] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Euro. Comp. Vision 14th Euro. Conf.*, 2016, pp. 108–124.
- [83] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16321–16330.
- [84] Z. Wang et al., "CRIS: CLIP-driven referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11686–11695.
- [85] X.-J. Zhang, Y.-F. Lu, and S.-H. Zhang, "Multi-task learning for food identification and analysis with deep convolutional neural networks," *J. Comput. Sci. Technol.*, vol. 31, no. 3, pp. 489–500, 2016.
- [86] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, vol. 1, no. 2, 2022, Art. no. 3.
- [87] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.* 2016, pp. 565–571.
- [88] R. Li et al., "Referring image segmentation via recurrent refinement networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5745–5753.
- [89] E. Margffoy-Tuay, J. C. Pérez, E. Botero, and P. Arbeláez, "Dynamic multimodal instance segmentation guided by natural language queries," in *Proc. Eur. Conf. Comput. Vis.* 2018, pp. 630–645.
- [90] Z. Yang et al., "LAVT: Language-aware vision transformer for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18155–18165.
- [91] P. Chhikara, D. Chaurasia, Y. Jiang, O. Masur, and F. Ilievski, "Fire: Food image to recipe generation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 8184–8194.
- [92] M. Luo, W. Min, Z. Wang, J. Song, and S. Jiang, "Ingredient prediction via context learning network with class-adaptive asymmetric loss," *IEEE Trans. Image Process.*, vol. 32, pp. 5509–5523, 2023.
- [93] S. Liu, Y. Liang, and A. Gitter, "Loss-balanced task weighting to reduce negative transfer in multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 9977–9978, 2019.
- [94] T. Yu, "Gradient surgery for multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 5824–5836.
- [95] M. Crawshaw, "Multi-task learning with deep neural networks: A survey," 2020, *arXiv:2009.09796*.



Yuehao Yin received the bachelor's degree in electronic and information engineering from Nanchang University, Nanchang, China, and the master's degree from the Department of Computer Science, Fudan University, Shanghai, China, in January 2024. His research primarily encompassed multi-modal learning, domain adaptation, and large multi-modal models.



Huiyan Qi received the B.E. degree and M.S. degree from Northwest Agriculture And Forestry University, Xianyang, China, in 2021 and Fudan University, Shanghai, China, in 2024, respectively. She is currently a Research Engineer with Singapore Management University, Singapore. Her research interests include transfer learning and cross-modal learning.



Bin Zhu received the B.E. degree from Southeast University, Nanjing, China, the M.E. degree from Zhejiang University, Hangzhou, China, and the Ph.D. degree from the City University of Hong Kong, Hong Kong, China. He is currently an Assistant Professor of Computer Science with the School of Computing and Information Systems, Singapore Management University (SMU). Prior to SMU, He was a Postdoctoral Researcher with the Department of Computer Science, University of Bristol, Bristol, U.K. His research interests mainly lie in cross-modal retrieval, LLMs, domain adaptation and egocentric video understanding. He is the co-organizer of EPIC KITCHENS100 CHALLENGES and Special Session on Multimedia on Cooking and Eating Activities hosted in ACM Multimedia Asia 2023.



Jingjing Chen (Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2018. She is currently an Associate Professor with the School of Computer Science, Fudan University, Shanghai, China. Before joining Fudan University, she was a Postdoc Research Fellow with the School of Computing, the National University of Singapore, Singapore. Her research interest lies in diet tracking and nutrition estimation based on multi-modal processing of food images, including food recognition, cross-modal recipe retrieval.



Yu-Gang Jiang (Fellow, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2009 and worked as a Postdoctoral Research Scientist with Columbia University, New York, NY, USA, during 2009–2011. He is currently a Professor of Computer Science with Fudan University, Shanghai, China. His research lies in the areas of multimedia, computer vision, and robust and trustworthy AI. He was the recipient of the inaugural ACM China Rising Star Award, 2015 ACM SIGMM Rising Star Award, Research Award for Excellent Young Scholars from NSF China, and Chang Jiang Distinguished Professorship appointed by Ministry of Education of China.



Chong-Wah Ngo (Senior Member, IEEE) is currently a Professor with the School of Computing and Information Systems, Singapore Management University, Singapore. His research interests include large-scale multimedia information retrieval, video computing, multimedia mining, and visualization. He was an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA from 2011 to 2014. He is the general chair of The Web Conference (WWW) 2024 and program chair of ACM Multimedia 2019.