nutrients

MDPI

Article

Reasoning-Driven Food Energy Estimation via Multimodal Large Language Models †

Hikaru Tanabe and Keiji Yanai *

Department of Informatics, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu 182-8585, Tokyo, Japan; tanabe-h@mm.inf.uec.ac.jp

* Correspondence: yanai@cs.uec.ac.jp

† This paper is an integrated and extended version of our papers published in Tanabe, H.; Yanai, K. CalorieLLaVA: Image-based Calorie Estimation with Multimodal Large Language Models. In Proceedings of ICPR Workshop on Multimedia Assisted Dietary Management, Kolkata, India, 1 December 2024, and Tanabe, H.; Yanai, K. CalorieVoL: Integrating Volumetric Context Into Multimodal Large Language Models for Image-Based Calorie Estimation. In Proceedings of the International Conference on MultiMedia Modeling, Nara, Japan, 8–10 January 2025.

Abstract: Background/Objectives: Image-based food energy estimation is essential for user-friendly food tracking applications, enabling individuals to monitor their dietary intake through smartphones or AR devices. However, existing deep learning approaches struggle to recognize a wide variety of food items, due to the labor-intensive nature of data annotation. Multimodal Large Language Models (MLLMs) possess extensive knowledge and human-like reasoning abilities, making them a promising approach for image-based food energy estimation. Nevertheless, their ability to accurately estimate food energy is hindered by limitations in recognizing food size, a critical factor in energy content assessment. **Methods:** To address this challenge, we propose two approaches: fine-tuning, and volume-aware reasoning with fine-grained estimation prompting. **Results:** Experimental results on the Nutrition5k dataset demonstrated the effectiveness of these approaches in improving estimation accuracy. We also validated the effectiveness of adapting LoRA to enhance food energy estimation performance. **Conclusions:** These findings highlight the potential of MLLMs for image-based dietary assessment and emphasize the importance of integrating volume-awareness into food energy estimation models.

Keywords: image-based food energy estimation; multimodal large language models; volume injection; daily food intake tracking

Academic Editors: Mikołaj Kamiński and Damian Skrypnik

Received: 31 January 2025
Revised: 16 March 2025
Accepted: 18 March 2025
Published: 24 March 2025

Citation: Tanabe, H.; Yanai, K. Reasoning-Driven Food Energy Estimation via Multimodal Large Language Models. *Nutrients* **2025**, *17*, 1128. https://doi.org/10.3390/nu17071128

1. Introduction

Monitoring daily food intake is a critical step towards achieving health-related goals such as dieting and bodybuilding. Accurate food energy estimation is particularly important for supporting these efforts, yet traditional methods like Food Diaries, 24 h Dietary Recalls (24HR), and Food Frequency Questionnaires (FFQ) rely on participants to self-report their intake, making the data prone to various errors and biases [1,2]. These self-reporting approaches suffer from limitations such as recall bias, misreporting due to social desirability, and the burden of detailed tracking, often leading to inaccurate energy intake estimations. Furthermore, the requirement for extensive manual input discourages long-term adherence, limiting their effectiveness in real-world dietary monitoring.

This challenge highlights the need for more efficient and user-friendly solutions that reduce the user burden, while maintaining high accuracy. With the increasing availability of smartphones and AR-enabled devices, capturing food images has become a practical

alternative for dietary monitoring. By leveraging computer vision methods, image-based food energy estimation automates dietary tracking, minimizing user effort and reducing reliance on memory-based reporting [3–7]. This approach offers an intuitive and less intrusive means of monitoring food intake, making it particularly advantageous for individuals who struggle with traditional self-reporting methods.

However, despite its potential, image-based food energy estimation remains challenging due to the diversity in food types and portion sizes, which significantly impact energy estimation accuracy. Current methods often fail to adapt to such variations effectively [5].

Multimodal Large Language Models (MLLMs) offer a promising solution to these challenges. MLLMs combine visual and textual reasoning capabilities, enabling them to identify diverse food items and contextualize visual information for tasks like food energy estimation. Recent advancements have demonstrated their potential in solving food-related tasks by leveraging extensive knowledge base and reasoning capabilities [8,9]. However, existing MLLMs struggle with accurately recognizing food volume, which is crucial for determining energy content [10].

In this study, we propose an approach that builds upon MLLMs to improve image-based food energy estimation. By introducing a fine-tuning strategy and volume injection with fine-grained estimation prompting, our approach addresses the limitations of current methods in handling food diversity and volume estimation. Experimental evaluation on the Nutrition5k dataset [6] demonstrates the effectiveness of these strategies in enhancing the quality of energy estimation.

We review the methodology from our previous works [11,12] and incorporate a fine-grained prompting strategy to improve food energy estimation. Additionally, we conduct several ablation studies, including an evaluation of the effectiveness of adapting LoRA for food energy estimation performance.

The main contributions of this study are summarized as follows:

- We introduce an approach that leverages fine-tuning and volume injection to enhance the reasoning and recognition capabilities of MLLMs for image-based food energy estimation.
- We propose a fine-grained estimation prompting method to address the challenges of food volume recognition in MLLMs.
- We evaluated the proposed approach on the Nutrition5k dataset, showing significant improvements over baseline methods and discussing its strengths and limitations.

2. Related Work

2.1. Image-Based Food Energy Estimation

Estimating food energy content from images has been widely studied, due to its potential applications in health and nutrition [7]. Two primary approaches exist for this task: size-based methods, and direct estimation methods.

Size-based methods typically involve multiple steps, starting with segmenting food regions in an image, estimating the food category, and then calculating the volume or mass of the food regions. The energy content is derived from these intermediate estimations. This approach enables precise consideration of food quantity, a critical factor for food energy estimation accuracy.

Determining the actual size of food in images has been approached through various techniques. One common strategy involves using everyday objects as reference points, such as credit cards or wallets [3], chopsticks [13], or even rice grains [14]. Augmented reality (AR) has also been leveraged, where virtual anchors enable users to estimate food dimensions interactively [4]. Beyond these object-based methods, advancements in depth estimation have further refined food size and volume calculations. DepthCalorieCam [5]

improved caloric estimation by integrating depth cameras with segmentation models, while implicit surface reconstruction techniques enabled the creation of detailed 3D food meshes, capturing both the food and dish with high fidelity [15].

Despite their potential, size-based methods are often limited in their ability to handle diverse food types. For instance, DepthCalorieCam is restricted to estimating the caloric content of only three food categories [5], significantly limiting its applicability.

Direct estimation methods bypass intermediate steps and use deep learning models trained end-to-end to directly estimate energy content. For example, Ege et al. [16] applied a multi-task learning framework based on VGG16 [17] to simultaneously estimate food category, ingredients, cooking methods, and energy content. While this approach simplifies the pipeline, it struggles to account for variations in food quantity, leading to potential inaccuracies when portions differ. Furthermore, these methods require extensive labeled datasets for training, imposing significant annotation costs.

In this study, we address the limitations of both approaches by leveraging Multimodal Large Language Models (MLLMs) and integrating volume estimation capabilities. Our method combines a promptable segmentation model, an open-set segmentation model, and monocular depth estimation to achieve high-quality zero-shot food energy estimation, reducing the reliance on annotated training data.

2.2. Multimodal Large Language Models (MLLMs)

Recent advancements in Large Language Models (LLMs) have demonstrated their ability to achieve remarkable performance across a wide range of language tasks by scaling model parameters, data, and computational resources [18]. These models exhibit emergent capabilities, where performance improves significantly at certain scaling thresholds [19]. Extending these models to handle visual information has led to the development of Multimodal Large Language Models (MLLMs), which integrate vision and language for enhanced reasoning.

Several MLLMs have been developed with superior performance across vision–language tasks. Flamingo [20] integrates visual and textual features using gated cross-attention, enabling it to handle diverse image and video tasks. LLaVA [8] employs linear layers to map visual features into a format compatible with LLMs and applies visual instruction tuning for task-specific improvements. Models like BLIP-2 [21], MiniGPT-4 [22], and InstructBLIP [23] further refine this approach, incorporating specialized modules such as Q-Former and leveraging instruction-following data for robust generalization.

In the food domain, FoodLMM [9] has demonstrated strong performance across various tasks, including food energy estimation. However, its accuracy is limited by insufficient volume recognition capabilities. In this study, we build on the reasoning and generalization capabilities of MLLMs and introduce enhancements to improve their ability to estimate caloric content from food images by incorporating explicit volume estimation and fine-tuning strategies.

3. Methods

In this study, we implement two approaches that enhance the food energy estimation capability of MLLMs: fine-tuning (Section 3.1), and volume injection with fine-grained prompting (Section 3.2).

3.1. Fine-Tuning MLLMs

We train MLLMs based on LLaVA-1.5 [24], to adapt their reasoning capabilities for image-based food energy estimation. The MLLM architecture consists of three main components (Figure 1):

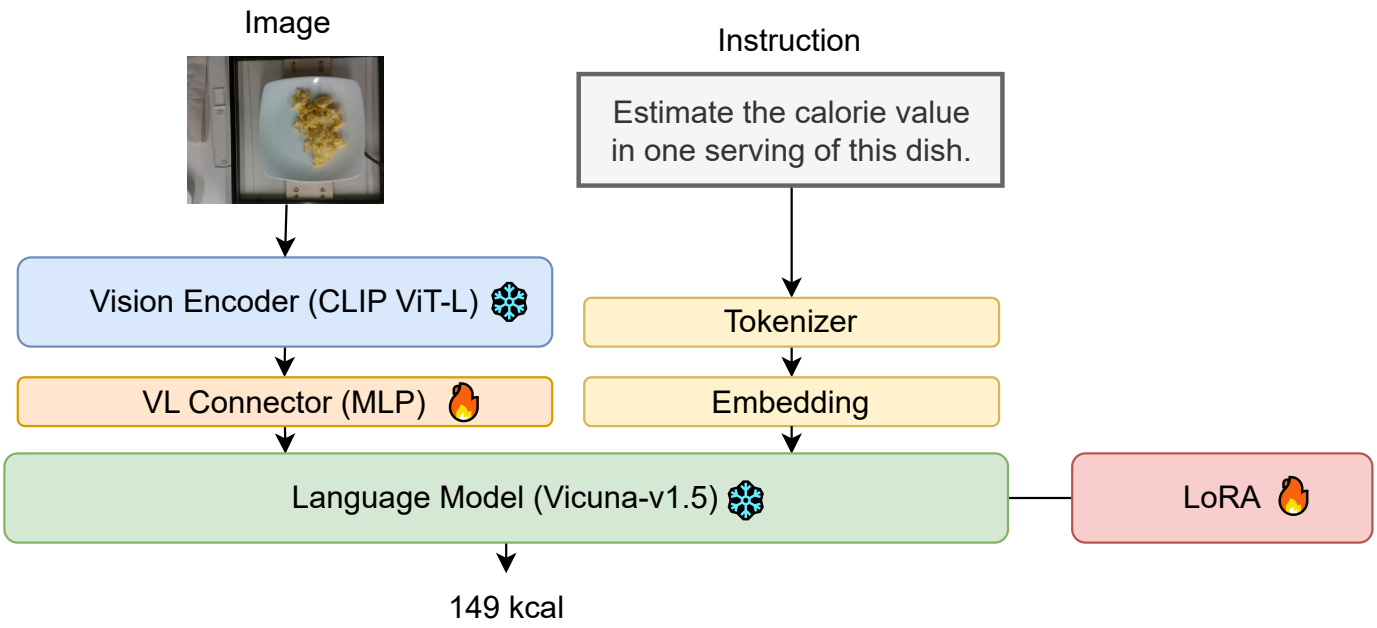

Figure 1. The architecture of MLLM

Visual Encoder: The input image is encoded into visual features using the OpenAI CLIP ViT-L encoder [25]. Vision–Language Connector: A two-layer MLP transforms the visual features to match the dimensions of text token embeddings. Language Model: The transformed visual features and text token embeddings are processed by Vicuna-v1.5 [26], a fine-tuned LLM, which outputs the estimated energy value.

To train the MLLMs, we used the Nutrition5k dataset [6], which includes paired food images and energy annotations. The dataset includes 3265 annotated food images with energy information. The dataset was converted into instruction-following format, where the instruction prompts the model to estimate energy values, and the response provides the energy value in a structured format (e.g., `[[300]] calories`). This consistent format facilitates extraction of energy content using regular expressions. Fine-tuning is performed using LoRA [27] for efficient adaptation with reduced computational overhead.

3.2. Volume Injection

To overcome the limitation of MLLMs in recognizing food volume, we apply a volume injection approach. This involves a dedicated volume estimation module and the integration of volume information into the MLLM.

3.2.1. Overall Architecture

The integration of the volume estimation module with the MLLM is shown in Figure 2. The estimated food volume is injected into the MLLM via instruction. Specifically, the instruction includes a placeholder `{{volume}}`, which is replaced with the calculated volume value. This approach allows the MLLM to reason food energy content with explicit consideration of food volume, improving its estimation accuracy. Figure 3 is the instruction prompt used for GPT-4V and 4o. To self-refine the accuracy of the estimated energy value, the value is asked in two-step format. The estimated energy value is output as JSON format, defined in the JSON Schema. For LLaVA-1.5, we remove the specification of the JSON format and put the following sentence:

```
Return single calorie value in the following format: "[[x]] calories.''
```

due to a lack of capability to follow the JSON format.

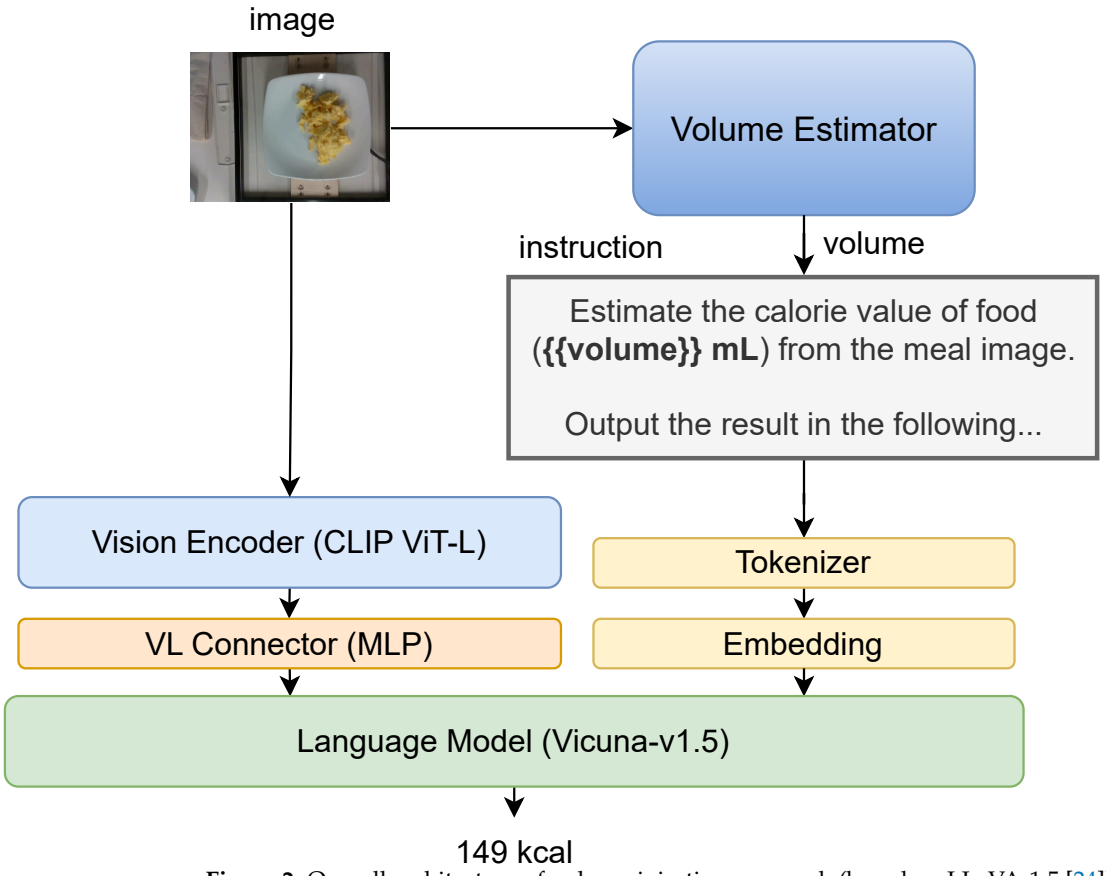

Figure 2. Overall architecture of volume injection approach (based on LLaVA-1.5 [24]).

```
You are a professional nutritionist who estimates the calorie value of meals from images.
Analyze images of meals provided by users to accurately estimate the calorie content of each
dish.
The output should focus only on calorie content.

Estimate the calorie value of food ({{volume}} ml) from the meal image.
Output the result in the following JSON format.

## JSON Schema
{
    ''type'': ''object'',
    ''properties'': {
        ''rough_calorie'': {''type'': ''integer'', ''description'': ''Calorie Value (kcal)''},
        ''calorie'': {''type'': ''float'', ''description'': ''Finer-Grained Calorie Value (kcal)''}
    },
    ''required'': [''calorie'']
}
```

Figure 3. Instruction prompt with fine-grained estimation prompting.

3.2.2. Volume Estimation Module

The volume estimation module processes food images to estimate the volume through the following steps (Figure 4):

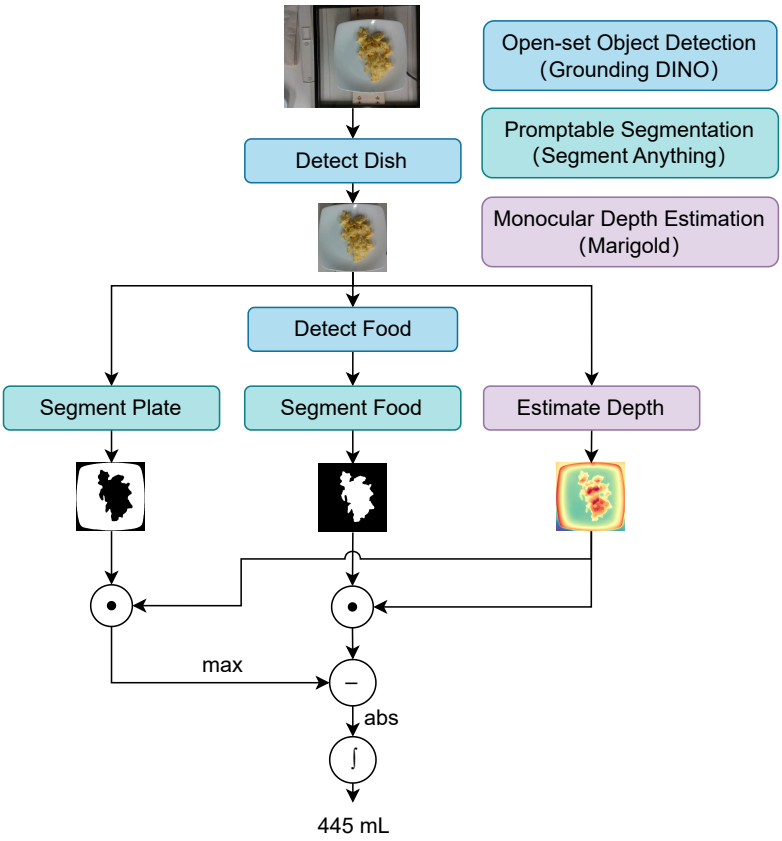

Figure 4. Architecture of the volume estimation module. This module processes a food image to estimate the food volume, which is then injected into the MLLM.

First, the dish's bounding box is detected using Grounding DINO [28], an open-set object detection model. Then, the Segment Anything Model (SAM) [29] generates masks for the food and the plate within the bounding box. A depth map is also estimated using Marigold [30], a monocular depth-estimation model.

The actual volume of the food is calculated using the extracted masks and depth map. The depth values corresponding to the food region are isolated by taking the Hadamard product of the depth map and the food region mask. The height of the food above the dish's reference plane is computed by subtracting the depth of the dish from the depth values of the food region. The per-pixel volume is calculated using the actual height and area of each pixel, and the total volume is obtained by summing the contributions of all pixels, as shown in Equation (1):

$$V = \sum_{i=1}^{n} \sum_{j=1}^{m} D_{ij} A_{ij} \tag{1}$$

here, D_{ij} and A_{ij} represent the height and area of each pixel at row i and column j. n and m indicate the height and width of the image, respectively.

4. Experimental Results

We conducted experiments to evaluate the proposed approaches using the Nutrition5k dataset [6], which is designed for nutritional analysis of food images. It comprises 3265 top-down food images, each paired with detailed nutritional information, including food energy content. For this study, we used 2759 samples for the training split, with the remaining 506 samples for the test split.

The evaluations included supervised food energy estimation, and zero-shot food energy estimation with volume injection.

4.1. Experiment Settings

We fine-tuned LLaVA-1.5-7B and 13B models for supervised setting using the training split of Nutrition5k, referring to the trained models as LLaVA-1.5-7B FT and 13B FT. The training was conducted using the AdamW optimizer with linear warmup and cosine decay, a peak learning rate of 2×10^{-4}, and a batch size of 64.

The models were evaluated on the test split of Nutrition5k. For text generation, the temperature value was set to 0 to ensure deterministic outputs. Failed extractions were retried with a temperature of 0.2 up to five times.

The volume injection approach was evaluated by combining the proposed food volume estimator with LLaVA-13B, GPT-4V, and 4o. Since these models were not fine-tuned on Nutrition5k, this evaluation focused on zero-shot performance. Similar extraction rules were applied in failed cases as in the fine-tuning evaluation.

4.2. Fine-Tuning Results

Table 1 presents the results of fine-tuning. The results were evaluated using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the correlation coefficient between the estimated values and the ground truth. The direction of the arrows indicates that a smaller value represents better performance for MAE and MAPE, while a larger value indicates better performance for the correlation coefficient. The LLaVA-1.5 FT models outperformed the baseline models, including GPT-4V and 4o for all metrics. In particular, LLaVA-1.5-13B FT also outperformed supervised expert models such as Google-nutrition-monocular [6] and FoodLMM [9], in terms of MAE. Figures 5 and 6 show the distributions of the energy values estimated by LLaVA-1.5-13B and 13B FT, respectively. They indicate that the fine-tuning approach was effective for improving food energy estimation.

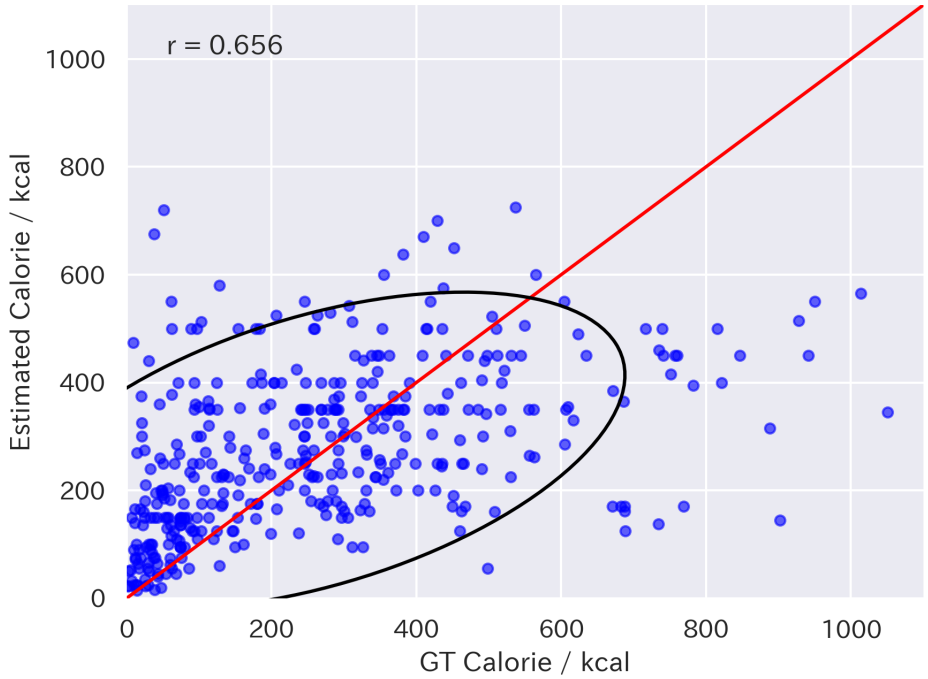

Figure 5. Scatter plot of energy values estimated by LLaVA-1.5-13B. The red line indicates equality between estimated and ground-truth values. The black line represents the 95% confidence ellipse.

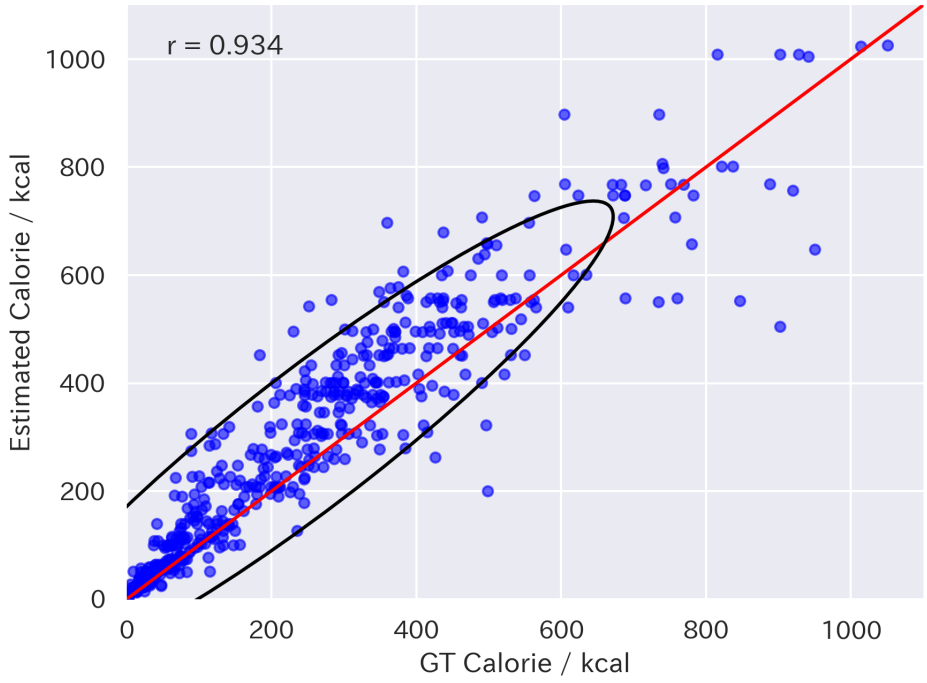

Figure 6. Scatter plot of energy values estimated by LLaVA-1.5-13B FT. The red line indicates equality between estimated and ground-truth values. The black line represents the 95% confidence ellipse.

Table 1. Results of food energy estimation on Nutrition5k (fine-tuning).

Method	MAE/kcal ↓	MAPE/% ↓	r ↑
Google-nutrition-monocular [6]	70.6	26.1	-
LLaVA-1.5-7B	178.8	129.5	0.637
LLaVA-1.5-13B	177.1	92.8	0.656
GPT-4V	80.7	55.7	0.833
GPT-4o	82.7	46.7	0.817
FoodLMM FT [9]	67.3	26.6	-
LLaVA-1.5-7B FT	74.2	41.5	0.927
LLaVA-1.5-13B FT	64.3	39.8	0.934

Table 2 compares the food energy estimation results under different training strategies. The model trained with LoRA outperformed the model using full-parameter tuning across all metrics, demonstrating its effectiveness for improving accuracy, while reducing training costs. Since LoRA requires fewer trainable parameters than full-parameter tuning, it helps mitigate overfitting, especially when using a relatively small dataset such as Nutrition5k.

Table 2. Comparison of full-parameter tuning and LoRA.

Method	MAE/kcal ↓	MAPE/% ↓	r ↑
LLaVA-1.5-13B FT (Full)	77.7	48.8	0.869
LLaVA-1.5-13B FT (LoRA)	64.3	39.8	0.934

4.3. Zero-Shot Results with Volume Injection

Table 3 shows the results of the zero-shot food energy estimation. In particular, combining the proposed volume injection with GPT-4o led to improvements in all metrics. Figures 7 and 8 are the distributions of estimated energy values by GPT-4o and 4o with volume injection, respectively. They illustrate that volume injection with fine-grained

estimation prompting improved the variance in estimated energy values, which resulted in more accurate estimation.

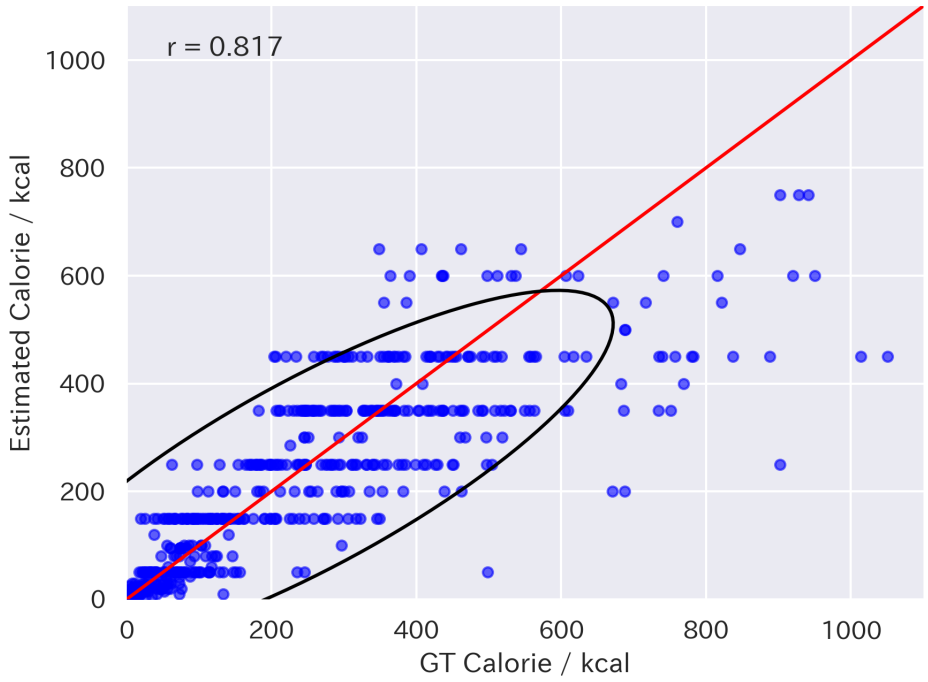

Figure 7. Scatter plot of energy values estimated by GPT-4o. The red line indicates equality between estimated and ground-truth values. The black line represents the 95% confidence ellipse.

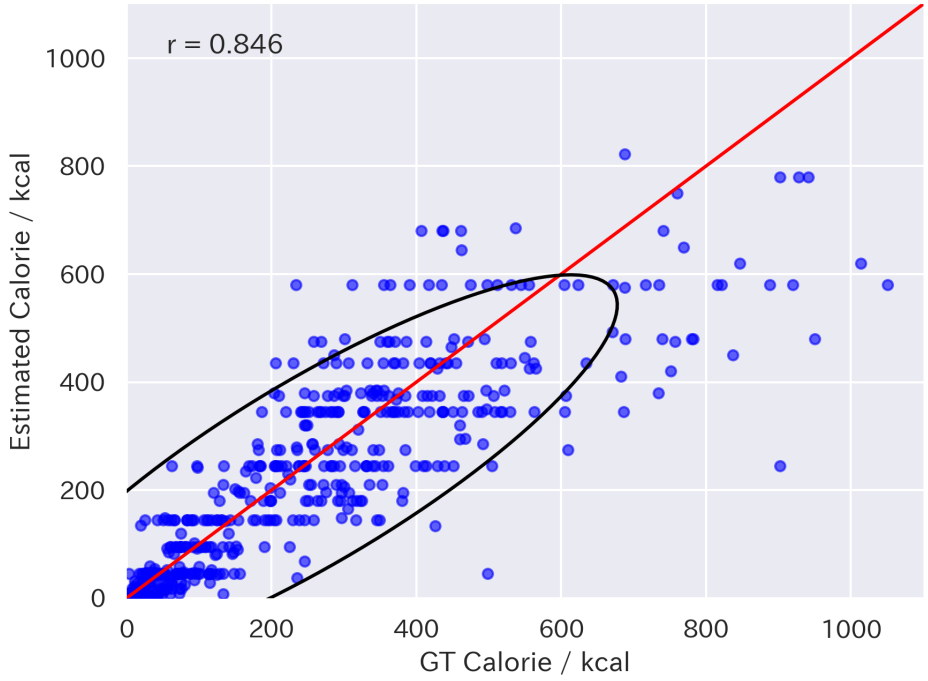

Figure 8. Scatter plot of energy values estimated by GPT-4o with volume injection. The red line indicates equality between estimated and ground-truth values. The black line represents the 95% confidence ellipse.

Table 3. Results of zero-shot food energy estimation by MLLMs on Nutrition5k (volume injection).

Model	MAE/kcal ↓	MAPE/% ↓	r ↑
LLaVA-1.5-13B	109.6	92.8	0.656
GPT-4V	80.7	55.7	0.833
GPT-4o	82.7	46.7	0.817
LLaVA-1.5-13B w/vol	6122.7	6591.4	−0.041
GPT-4V w/vol	83.8	54.1	0.816
GPT-4o w/vol	78.8	43.4	0.846

Additionally, Figure 9 demonstrates the segmentation and depth estimation outputs, highlighting the volume estimation module's ability to localize food regions and compute volumes effectively. For object detection and segmentation, both the dish and food regions were accurately extracted, indicating high-quality estimation. In depth estimation, variations in uneven surfaces within the image were well captured. Additionally, in images containing multiple food items, areas with differing heights exhibited distinct depth values compared to their surroundings. These results highlight the effectiveness of the volume estimation module in localizing food regions and accurately computing their volumes.

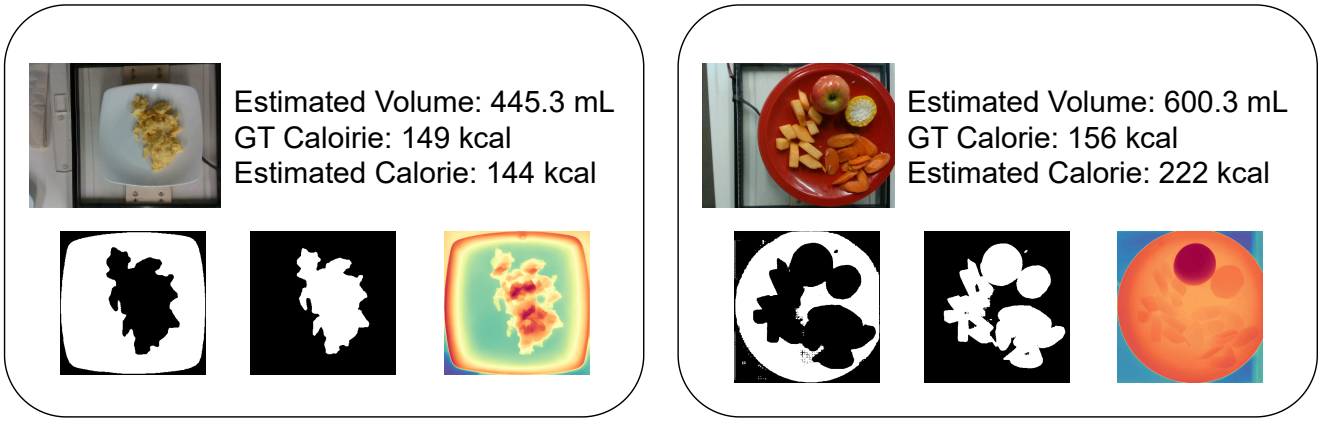

Figure 9. Outputs of the volume estimation module: object detection, segmentation, and depth estimation (the estimated energy values are the outputs by GPT-4V with volume injection).

5. Discussion

5.1. Challenges in Volume Estimation

The food volume estimation module had challenges in overestimation of food volume (Figure 10):

- The estimated volume between the bottom of the food and the reference plane of the dish could be over-calculated.
- If the lowest part of the dish is obscured by food, an incorrect reference plane for the dish may be selected.

To mitigate these issues, methods from prior research can serve as valuable references. For example, DepthCalorieCam [5] used a food mass regression model to adjust for overestimated volumes, resulting in more accurate food energy estimation. Similarly, Naritomi et al. [15] reconstructed high-quality 3D shapes of food and dishes to exclude the area beneath the food and the reference plane, ensuring accurate volume estimation.

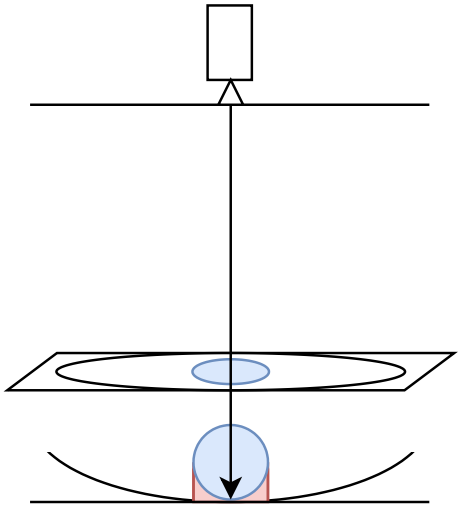

Figure 10. Overestimation of volume by the food volume estimator when the food is assumed to be spherical. Blue: food region, Red: excess region.

While these approaches are promising, they require extensive food data for training. Adjusting overestimated volumes demands large-scale annotated volume data, while improving 3D reconstruction methods necessitates high-quality 3D shape data, which remains a challenge in the food domain.

In particular, recent advancements in 3D reconstruction, such as Neural Radiance Fields (NeRF) [31] and 3D Gaussian Splatting [32], offer promising directions for addressing the limitations of volume estimation. These methods can generate detailed 3D representations of objects from 2D images, potentially overcoming the limitations of current 3D reconstruction models in the food domain. Integrating these techniques into the volume estimation process is expected to improve the accuracy of volume measurements.

5.2. Improving Commonsense Reasoning in MLLMs

Another potential enhancement involves leveraging the commonsense knowledge of MLLMs to revise and correct unreasonable energy estimations. For example, under zero-shot conditions, the language model could be prompted to reevaluate its reasoning process when it produces energy values that deviate significantly from expected norms. By incorporating such prompts, MLLMs could refine their estimations and provide outputs that align better with real-world expectations.

5.3. Future Directions

In future work, integrating advanced 3D reconstruction techniques and fine-tuning MLLMs to better capture spatial information will be crucial. Additionally, exploring ways to reduce the reliance on large-scale annotated datasets could lower the barrier to implementing these improvements. By addressing these challenges, the proposed framework could achieve more accurate and robust food energy estimation.

Beyond technical improvements, expanding the application of the proposed framework to interdisciplinary healthcare settings presents a promising direction. In hospital environments, this technology could assist clinicians in estimating caloric intake, particularly in cases of hospital malnutrition, where precise dietary monitoring is critical [33]. Integrating this model into multidisciplinary healthcare-technology projects, such as those involving nutritionists, nurses, and engineers, could facilitate real-time dietary assessment and personalized nutritional interventions [34,35].

Moreover, establishing standardized protocols for data collection and model evaluation across different clinical settings will be essential to ensure the reliability and gener-

alizability of AI-driven dietary assessment systems. Interdisciplinary collaboration will also be key in defining best practices for integrating these technologies into existing hospital workflows, minimizing disruption, while maximizing clinical utility. Furthermore, collaboration with nursing and engineering researchers could enable the development of integrated healthcare-technology systems that leverage food intake monitoring for patient management [36,37]. Applications in other medical-technological domains, such as metabolic disorder management and postoperative dietary tracking, also warrant exploration [38,39]. By extending the reach of this framework into healthcare and beyond, the impact of AI-driven dietary assessment can be maximized across various fields.

5.4. Discussion Summary

Despite the advancements achieved in this study, several limitations remain. First, the accuracy of volume estimation was affected by overestimation issues, due to the challenges in determining the reference plane of the dish and occlusions in food images. Future work could address these issues by incorporating improved 3D reconstruction techniques or refining volume estimation models with additional data. Second, while our approach enhances reasoning capabilities for food energy estimation, commonsense reasoning in MLLMs remains an open challenge, particularly in cases where the estimated values deviate significantly from expected norms. Incorporating self-refinement mechanisms or domain-specific constraints could further improve estimation accuracy. Lastly, our evaluation primarily relied on the Nutrition5k dataset, which, although diverse, may not fully represent real-world variations in food presentation and composition. Expanding evaluation to broader datasets and real-world scenarios will be necessary to ensure versatility. Addressing these limitations will be crucial for enhancing the robustness and applicability of MLLM-based food energy estimation models.

6. Conclusions

In this study, we applied MLLMs to food energy estimation, enhancing their capabilities by introducing fine-tuning and volume injection with fine-grained prompting. Through fine-tuning MLLMs and incorporating a food volume estimation module, our approach demonstrated improvements in food energy estimation accuracy. Evaluations on the Nutrition5k dataset showed that the proposed method outperformed baseline models and contemporary MLLMs, both in fine-tuned and zero-shot settings.

Despite these advancements, several challenges remain. The volume estimation module occasionally overestimated the food volume, due to inaccuracies in determining the reference plane of the dish or the height of the food. To address these limitations, future work could incorporate mass regression models and advanced 3D reconstruction techniques. These methods have the potential to refine the volume estimation process and improve model accuracy.

Additionally, more comprehensive evaluations of energy and volume estimation across a wider variety of food items are needed. Constructing a dataset that includes diverse food items annotated with volume and energy values would be a valuable contribution toward this goal.

Author Contributions: Conceptualization, H.T.; methodology, H.T.; software, H.T.; validation, H.T.; formal analysis, H.T.; investigation, H.T.; resources, H.T.; data curation, H.T.; writing—original draft preparation, H.T.; writing—review and editing, H.T.; visualization, H.T. supervision, K.Y.; project administration, K.Y.; funding acquisition, K.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JSPS KAKENHI Grant Numbers, 22H00540 and 22H00548.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available on the project page of Nutrition5k [6].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Naska, A.; Lagiou, A.; Lagiou, P. Dietary assessment methods in epidemiological research: Current state of the art and future prospects. *F1000Research* **2017**, *6*, 926. [PubMed]
2. Bailey, R.L. Overview of dietary assessment methods for measuring intakes of foods, beverages, and dietary supplements in research studies. *Curr. Opin. Biotechnol.* **2021**, *70*, 91–96.
3. Okamoto, K.; Yanai, K. An automatic calorie estimation system of food images on a smartphone. In Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, Amsterdam, The Netherlands, 16 October 2016.
4. Tanno, R.; Ege, T.; Yanai, K. AR DeepCalorieCam V2: Food Calorie Estimation with CNN and AR-Based Actual Size Estimation. In Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology, Tokyo, Japan, 28 November–1 December 2018.
5. Ando, Y.; Ege, T.; Cho, J.; Yanai, K. DepthCalorieCam: A Mobile Application for Volume-Based FoodCalorie Estimation Using Depth Cameras. In Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, Nice, France, 21 October 2019; pp. 76–81.
6. Thames, Q.; Karpur, A.; Norris, W.; Xia, F.; Panait, L.; Weyand, T.; Sim, J. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 8903–8911.
7. Sultana, J.; Ahmed, B.M.; Masud, M.M.; Huq, A.K.O.; Ali, M.E.; Naznin, M. A Study on Food Value Estimation From Images: Taxonomies, Datasets, and Techniques. *IEEE Access* **2023**, *11*, 45910–45935. . [CrossRef]
8. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In Proceedings of the Advances in Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2023.
9. Yin, Y.; Qi, H.; Zhu, B.; Chen, J.; Jiang, Y.G.; Ngo, C.W. FoodLMM: A Versatile Food Assistant using Large Multi-modal Model. *arXiv* **2023**, arXiv:2312.14991.
10. Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.C.; Liu, Z.; Wang, L. The Dawn of LMMs: Preliminary Explorations with GPT-4V (ision). *arXiv* **2023**, arXiv:2309.17421.
11. Tanabe, H.; Yanai, K. CalorieLLaVA: Image-based Calorie Estimation with Multimodal Large Language Models. In Proceedings of the Proceedings of ICPR Workshop on Multimedia Assisted Dietary Management, Kolkata, India, 1 December 2024.
12. Tanabe, H.; Yanai, K. CalorieVoL: Integrating Volumetric Context Into Multimodal Large Language Models for Image-Based Calorie Estimation. In Proceedings of the International Conference on MultiMedia Modeling, Nara, Japan, 8–10 January 2025.
13. Akpa, E.A.H.; Suwa, H.; Arakawa, Y.; Yasumoto, K. Smartphone-Based Food Weight and Calorie Estimation Method for Effective Food Journaling. *SICE J. Control. Meas. Syst. Integr.* **2017**, *10*, 360–369.
14. Ege, T.; Shimoda, W.; Yanai, K. A New Large-scale Food Image Segmentation Dataset and Its Application to Food Calorie Estimation Based on Grains of Rice. In Proceedings of the ICPR Workshop on Multimedia Assisted Dietary Management, Nice, France, 21–25 October 2019.
15. Naritomi, S.; Yanai, K. Hungry Networks: 3D mesh reconstruction of a dish and a plate from a single dish image for estimating food volume. In Proceedings of the 2nd ACM International Conference on Multimedia in Asia, Singapore, 7 March 2021.
16. Ege, T.; Yanai, K. Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions. In Proceedings of the on Thematic Workshops of ACM Multimedia 2017, Mountain View, CA, USA, 23–27 October 2017; pp. 367–375.
17. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
18. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. *arXiv* **2020**, arXiv:2001.08361.
19. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.
20. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. In Proceedings of the Advances in Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 23716–23736.
21. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.

22. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
23. Dai, W.; Li, J.; Li, D.; Tiong, A.M.H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; Hoi, S. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv* **2023**, arXiv:2305.06500.
24. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. *arXiv* **2023**, arXiv:2310.03744.
25. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Shenzhen, China, 26 February–1 March 2021; pp. 8748–8763.
26. Chiang, W.L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J.E.; et al. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023. Available online: https://lmsys.org/blog/2023-03-30-vicuna/ (accessed on 17 March 2025).
27. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
28. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. *arXiv* **2023**, arXiv:2303.05499.
29. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
30. Ke, B.; Obukhov, A.; Huang, S.; Metzger, N.; Daudt, R.C.; Schindler, K. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. *arXiv* **2023**, arXiv:2312.02145.
31. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
32. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*, 139:1–139:14. [CrossRef]
33. Heighington-Wansbrough, A.J.; Gemming, L. Dietary intake in hospitals: A systematic literature review of the validity of the visual estimation method to assess food consumption and energy and protein intake. *Clin. Nutr. ESPEN* **2022**, *52*, 296–316. [CrossRef]
34. Roberts, S.; Hopper, Z.; Chaboyer, W.; Gonzalez, R.; Banks, M.; Desbrow, B.; Marshall, A.P. Engaging hospitalised patients in their nutrition care using technology: Development of the NUTRI-TEC intervention. *BMC Health Serv. Res.* **2020**, *20*, 148. [CrossRef]
35. Chaudhry, B.M.; Siek, K.A.; Connelly, K. The Usability and Feasibility of a Dietary Intake Self-Monitoring Application in a Population with Varying Literacy Levels. *J. Pers. Med.* **2024**, *14*, 1001. [CrossRef]
36. Pfisterer, K.J.; Boger, J.; Wong, A. Prototyping the automated food imaging and nutrient intake tracking system: Modified participatory iterative design sprint. *JMIR Hum. Factors* **2019**, *6*, e13017. [CrossRef]
37. Yinusa, G.; Scammell, J.; Murphy, J.; Ford, G.; Baron, S. Multidisciplinary provision of food and nutritional care to hospitalized adult in-patients: A scoping review. *J. Multidiscip. Healthc.* **2021**, *14*, 459–491. [PubMed]
38. Lo, F.P.W.; Qiu, J.; Jobarteh, M.L.; Sun, Y.; Wang, Z.; Jiang, S.; Baranowski, T.; Anderson, A.K.; McCrory, M.A.; Sazonov, E.; et al. AI-enabled wearable cameras for assisting dietary assessment in African populations. *NPJ Digit. Med.* **2024**, *7*, 356. [CrossRef] [PubMed]
39. Phalle, A.; Gokhale, D. Navigating next-gen nutrition care using artificial intelligence-assisted dietary assessment tools—A scoping review of potential applications. *Front. Nutr.* **2025**, *12*, 1518466.