CalorieVoL: Integrating Volumetric Context into Multimodal Large Language Models for Image-based Calorie Estimation

Hikaru Tanabe[0009-0003-5752-9175] and Keiji Yanai[0000-0002-0431-183X]

The University of Electro-Communications, Chofu, Tokyo, Japan
{tanabe-h, yanai}@mm.inf.uec.ac.jp

Abstract. Multimodal Large Language Models (MLLMs) can perform various food-related tasks with high quality. Notably, high-performance MLLMs, such as GPT-4V, can even estimate caloric content from food images. However, these MLLMs often struggle to accurately recognize volume information, which often leads to errors in calorie estimation. To address this issue, we propose a new MLLM framework called CalorieVoL, designed to enhance the recognition of volume information in food items. By integrating this framework into MLLMs like GPT-4V, we achieved higher scores in terms of MAE and correlation coefficients on Nutrition5k compared to simple MLLMs. Our experiments also showed that the volume-aware recognition improved responses in scenarios where accurate volume estimation is critical.

Keywords: Multimodal Large Language Models · Image-based Calorie Estimation · Volume Estimation.

1 Introduction

By recording daily food intake, we can obtain valuable information that helps achieve health-related goals such as dieting and bodybuilding. For this purpose, manual dietary survey frameworks like food diary method, 24-hour recall method, and food frequency questionnaire are widely used among nutrition experts. However, these methods are time-consuming and require participants to weigh or recall their food intake, which poses a challenge to them. Moreover, caloric content of daily food is a critical metric for maintaining a healthy lifestyle. Therefore, quickly and easily estimating the caloric content of food can have a significant impact on the healthcare field.

Taking pictures of food using smartphones or camera-equipped AR devices is an easier approach for people to keep a food diary. However, daily food varies widely in type and quantity, which leads to variance in the caloric content. Thus, it is important to build models that can correctly recognize the type and quantity of food from images for calorie estimation.

Large Language Models (LLMs) have acquired a wide range of commonsense knowledge about the world. Mainly, LLMs that have undergone recent instruction tuning can reason tasks based on the commonsense they possess when provided with prompts designed for the tasks [1]. Multimodal Large Language Models (MLLMs) can solve image recognition tasks while retaining the reasoning capabilities of LLMs [2]. Some models have acquired specific knowledge in fields such as bio-medicine, achieving human-level performance [3]. Some advanced models can also perform reasoning food-related tasks by recognizing various types of food [4]. However, these models cannot recognize the volumetric amount of food accurately.

In this study, we propose a framework called CalorieVoL that utilizes food recognition capabilities of MLLMs for image-based calorie estimation. While leveraging MLLMs to cover the diversity of food that conventional calorie estimation methods could not achieve, we introduce a novel volume estimator to complement the volume estimation capabilities that MLLMs struggle with. This approach allows us to estimate the caloric content of food images with high quality, even for food images that were not explicitly trained for calorie estimation tasks.

The main contributions of this study are as follows:

- We introduce CalorieVoL, a framework that enables volume-aware recognition for image-based calorie estimation using MLLMs.
- We introduce a new plug-in volume estimator by utilizing off-the-shelf SOTA models to integrate volume information into MLLMs.
- We evaluate the performance of CalorieVoL on the Nutrition5k [5] dataset and discuss the effectiveness and challenges of this method.

2 Related Work

2.1 Image-based Calorie Estimation

Estimating the caloric content of food items shown in images has been attempted through various methods due to its applicability [6]. There are primary approaches called size-based methods, where a pipeline is constructed that combines multiple image recognition modules to estimate caloric content. The basic procedure involves first segmenting the food regions from the meal image, then estimating the food category, followed by estimating the volume or mass of the food region. Subsequently, the caloric content is estimated based on these results. By taking these steps before estimating the caloric content, these methods can particularly consider the quantity of food.

To determine the metric size of the food region, some methods estimate the actual size of objects included in the food image. Okamoto *et al.* [7] used a credit card or a long wallet as a reference, Akpa *et al.* [8] used chopsticks as a reference, and Ege *et al.* [9] used rice grains as a reference. Furthermore, Tanno *et al.* [10] employed a method using anchors placed in an AR space, obtaining the actual size through interaction with the user. DepthCalorieCam [11] significantly reduced the error in calorie estimation by estimating the food volume using a depth camera and a segmentation model. Naritomi *et al.* [12] reconstructed high-quality 3D meshes of the dish and food using an implicit function representation.

However, these size-based methods lack variety in food. For example, DepthCalorieCam is limited to estimating only three categories of food, which causes the lack of applicability.

In this study, inspired by size-based methods, we create a food volume estimator by combining an open-set segmentation model, a promptable segmentation model, and a monocular depth estimation model. We aim to achieve zero-shot calorie estimation with high quality, without the need for training on the target dataset.

2.2 Multimodal Large Language Models

In recent years, Large Language Models (LLMs), which are language models trained under large-scale conditions with a substantial number of model parameters, data, and computational resources, have achieved high performance across various language tasks. These models exhibit a power-law improvement in performance as the scale of learning conditions increases [13]. they also demonstrate emergent abilities where their performance improves dramatically at a certain stage as the learning conditions are scaled up [14]. These new aspects, which were not observed in conventional language models, are attracting significant attention.

Multimodal Large Language Models (MLLMs) are constructed by extending the ability of LLMs to other modalities. Flamingo [15] acquired the ability to answer various vision-language questions by fusing the visual features with text features using gated cross-attention dense blocks. LLaVA [2] employed a linear or MLP layer to transform the visual features into the shape of the language tokens. Additionally, it adopted a training framework called Visual Instruction Tuning, which resulted in high-quality instruction-following ability for vision-language tasks. MiniGPT-4 [16] and InstructBLIP [17] also acquired abilities to solve a wide range of tasks by using Q-Former [18] as the vision-language connector and applied training framework similar to Visual Instruction Tuning.

In food domain, FoodLMM [19] achieved high performance in various food-related tasks, including image-based calorie estimation. we particularly focuses on improving the performance of calorie estimation from various food images without the need to train.

3 Methods

The overview of CalorieVoL is shown in Fig. 1. CalorieVoL consists of two main components: a part that uses MLLMs as a calorie estimator and a part that estimates the volume of food (Section 3.1). By combining these components, we construct CalorieVoL (Section 3.2).

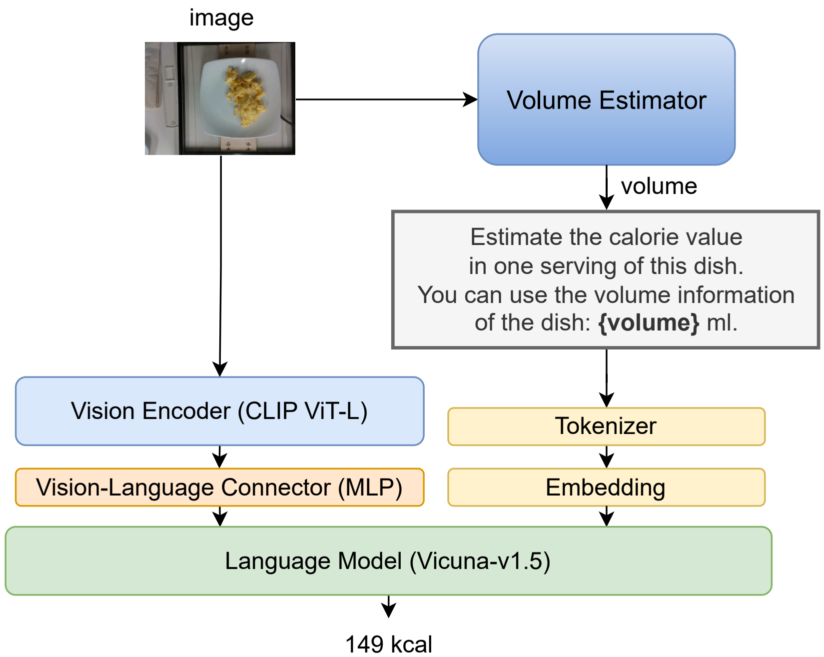

Fig. 1. Overview of CalorieVoL (the case of LLaVA-v1.5)

3.1 Volume Estimator

We construct a model that recognizes the food portion from a meal image and estimates its volume. The process from inputting the image to outputting the volume value is shown in Fig. 2.

The process of volume estimation is as follows (Fig. 2). First, the bounding box of the dish is obtained using Grounding DINO [20], an open-set object detection model. Next, the process is divided into three parts using this region of interest. First, the dish's region mask is obtained by applying Segment Anything (SAM) [21] to the region of interest. Second, the bounding box of the food portion is obtained by applying Grounding DINO to the region of interest, followed by applying SAM to obtain the region mask. Third, the depth map is obtained by applying Marigold [22], a monocular depth estimation model, to the region of interest.

Based on these two obtained masks and the depth map, the actual volume is estimated. First, the element-wise Hadamard product between the depth map and each mask is taken to extract the regions of each mask. Next, the maximum value within the dish's region in the depth map is obtained and used as the depth

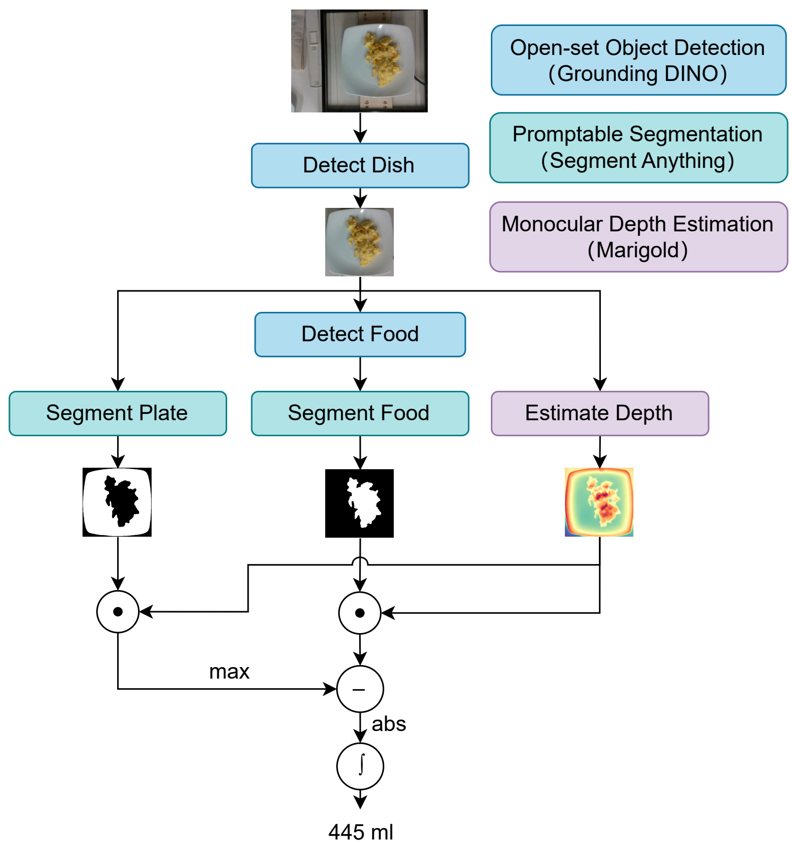

Fig. 2. Structure of the food volume estimator

of the dish's reference plane. Then, the difference is taken between each value in the depth map of the food region and the depth of the dish's top surface. This provides the ratio of the height from the dish's reference plane to the top surface of the food to the height from the shooting position. Furthermore, the actual height from the dish's reference plane to the top surface of the food for each pixel is obtained based on the actual height from the dish's reference plane to the shooting position and this ratio. Finally, the volume integration is performed for all pixels in the food region's depth map based on this height and the actual area per pixel. The volume integration value at this time is calculated using (1). Here, V represents the actual volume of the food, D_{ij} represents the actual height of the pixel at row i and column j, and A_{ij} represents the actual area of

Table 1. Results of zero-shot calorie estimation by MLLMs on the Nutrition5k dataset

Model	MAE / kcal ↓	MAPE / % ↓	r ↑
LLaVA-13B	109.6	92.8	0.656
GPT-4V	106.6	**54.8**	0.688
LLaVA-13B + CalorieVoL (Ours)	6122.7	6591.4	-0.041
GPT-4V + CalorieVoL (Ours)	**101.7**	56.8	**0.708**

the pixel at row i and column j.

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

3.2 CalorieVoL

We use GPT-4V [4] and LLaVA-v1.5 [2] as the MLLMs for the reasoning of calorie estimation. The text prompt is constructed to encourage the LLM to estimate the calorie value in one serving of the dish. Additionally, it includes a placeholder `{volume}`, where the volume value estimated from the volume estimator will be substituted. This allows the LLM to estimate the calorie value based on the volume of the food.

4 Experiments

4.1 Evaluation of Calorie Estimation

We conducted evaluations using the test split of the Nutrition5k dataset. It should be noted that none of the models were trained on Nutrition5k dataset, making this an evaluation of zero-shot calorie estimation. The temperature is set at 0 through the evaluation.

Table 1 shows the results of zero-shot calorie estimation. If the output text of a model did not include the calorie value, the result was obtained by repeating the same question up to five times with a temperature parameter set to 0.2. As a result, 79 data points were excluded from the evaluation for the model that combined GPT-4V with the food volume estimator. It can be observed that the model combining the proposed food volume estimator with the base model, GPT-4V, achieved better scores on MAE and correlation coefficient.

Fig. 3 and Fig. 4 show scatter plots of the estimated values and the ground truth for zero-shot calorie estimation. Although there is not a significant difference overall, the correlation coefficient is higher for the model combined with the food volume estimator.

Additionally, Fig. 5 and Fig. 6 present examples of the model responses in zero-shot calorie estimation. In Fig. 5, it can be seen that the estimation results are improved significantly as the volume estimation results from the proposed method are considered in the calorie estimation process. On the other hand, Fig. 6 shows an example of overestimation when LLaVA-v1.5 is used as the MLLM. Observing the reasoning process, it can be seen that the calorie value of the dish is initially estimated relatively accurately. However, in the latter part, the calorie value seems to have been incorrectly multiplied by the volume value, leading to the final overestimated result. This suggests that the overestimation occurred due to the MLLM's inability to properly recognize units and perform calculations accurately.

4.2 Evaluation of Volume Estimation

Fig. 7 shows the estimated volume values. It can be observed that the shape of the distribution resembles the distribution of the true calorie values (Fig. 3).

Additionally, Fig. 8 presents the results of object detection, segmentation, and depth estimation during the volume estimation process. For object detection and segmentation, appropriate regions for both the dish and food were successfully extracted, demonstrating overall high-quality estimation. For depth estimation, variations in the uneven surfaces within the image are well represented. Furthermore, in the image containing multiple types of food, the depth values in areas with different heights are noticeably distinct from the surrounding values.

5 Direction of Improvement in Volume Estimation

The food volume estimator proposed in this study has two main characteristics that may lead to an overestimation of volume (Fig. 9). First, the volume from the bottom of the food to the reference plane of the dish may be overestimated. Second, when the lowest part of the dish is covered by food, an incorrect reference plane of the dish may be selected.

In the research of DepthCalorieCam [11], after the volume is calculated using a method similar to the one in this study, the volume value is input into a food mass regression model. This model adjusts the estimate by accounting for the overestimated volume, leading to an accurate calorie estimation. Another approach, as suggested by Naritomi *et al.* [12], is to reconstruct high-quality 3D shapes of the dish and food, ensuring that the area between the bottom of the food and the reference plane of the dish does not affect the food volume.

However, creating models to implement these methods would require a large amount of food data, which would be burdensome to prepare. For the method that estimates the excess volume, a large amount of annotated data on volume would be needed to train the model. Additionally, for the method of reconstructing 3D shapes, the existing performance of 3D shape reconstruction methods is not sufficient, especially in the food domain, necessitating a large amount of 3D shape data for improvement.

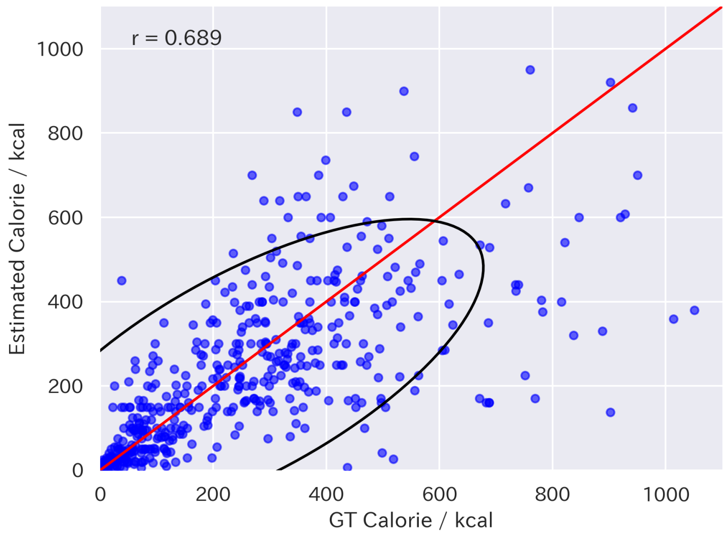

Fig. 3. Scatter plot of estimated calorie values by GPT-4V

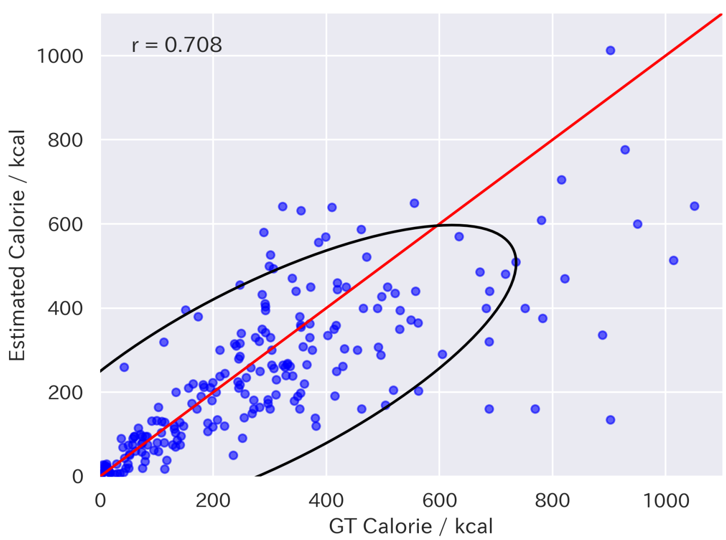

Fig. 4. Scatter plot of estimated calorie values by GPT-4V and the food volume estimator

Answer: **687 kcal**

User

Estimate the calorie value in one serving of the dish.
Return the single calorie value in the following format: "[[x]] calories".
This information is essential and must be provided in every response.

Example: "[[100]] calories"

GPT-4V

This is a bowl of almonds. A typical serving size of almonds is about 1 ounce (28 grams), which is roughly 23 almonds and contains approximately [[160]] calories. However, without a reference for the exact amount in this bowl, it's difficult to provide an accurate calorie count. If this bowl contains about one serving, then the calorie estimate would be around **[[160]] calories**. If it contains more, the calorie count would be higher accordingly.

User

Estimate the calorie value in one serving of the dish.
Return the single calorie value in the following format: "[[x]] calories".
This information is essential and must be provided in every response.

Example: "[[100]] calories"

You can use the volume information of the dish: 527 ml.

GPT-4V+Vol (Ours)

The image shows a bowl of almonds. Almonds are a nutrient-dense food, and a typical serving size of almonds is about 1 ounce (28 grams), which is approximately 23 almonds. This serving size generally contains about 160 to 170 calories. However, the bowl in the image appears to contain more than a single serving.

Without knowing the exact weight or number of almonds in the bowl, it's challenging to provide an accurate calorie count. If we assume the bowl contains around 3 servings (which is a rough estimate based on visual assessment), the calorie content would be approximately 480 to 510 calories.

Given the volume information of the dish (527 ml), and assuming the bowl is filled mostly to the top, it could potentially hold around 4 to 5 ounces of almonds (since almonds are not a liquid and their volume would not directly convert to weight). This would mean the bowl might contain around 640 to 850 calories, depending on the exact amount and whether the almonds are raw or roasted, which can slightly affect the calorie content.

For the purpose of providing a single calorie value, I will estimate the calorie content of the bowl of almonds as follows:

[[700]] calories

Please note that this is a rough estimate and the actual calorie content could be higher or lower depending on the precise amount of almonds in the bowl.

Fig. 5. Example of calorie estimation by the model combining GPT-4V and the food volume estimator

Answer: **149 kcal**

User

Estimate the calorie value in one serving of the dish.
Return the single calorie value in the following format: "[[x]] calories".
This information is essential and must be provided in every response.

Example: "[[100]] calories"

You can use the volume information of the dish: 445 ml.

LLaVA-13B+Vol
(Ours)

To estimate the calorie value in one serving of the dish, we need to know the ingredients and their respective calorie counts. Since the image shows a plate of scrambled eggs, we can assume that the dish is primarily composed of eggs.

The calorie value of one serving of scrambled eggs can vary depending on the number of eggs used and the cooking method. However, as a rough estimate, we can assume that one serving of scrambled eggs contains around 200-250 calories.

To calculate the calorie value for the entire dish, we need to know the volume of the dish, which is 445 ml. To convert the volume to weight, we can assume that the dish is made up of 445 grams of scrambled eggs.

So, the calorie value of the dish would be approximately **98,000 to 112,000 calories**. However, this is a rough estimate and may not be accurate, as the actual calorie count can vary depending on the specific ingredients and cooking method used.

Fig. 6. Example of overestimation by the model combining LLaVA-v1.5 and the food volume estimator

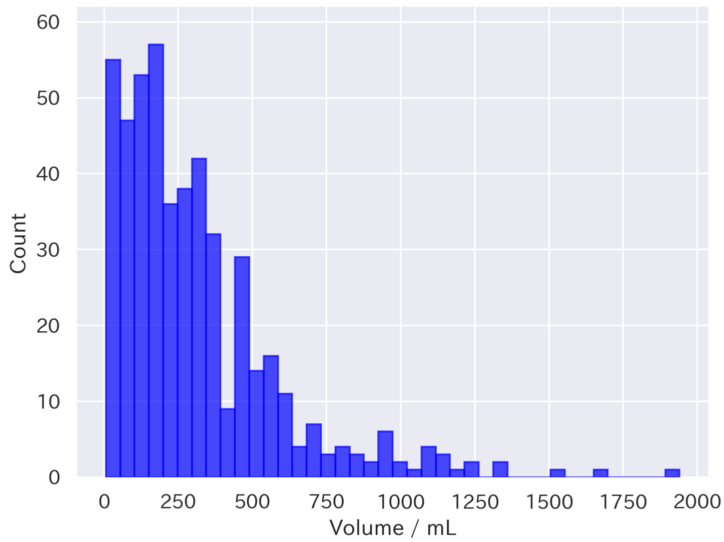

Fig. 7. Distribution of estimated food volumes

On the other hand, recent efforts have been made to achieve high-quality 3D shape reconstruction methods based on techniques such as NeRF [23] and 3D Gaussian Splatting [24]. By leveraging these approaches, it is expected that the challenges in volume estimation will be comprehensively addressed.

In terms of accurately capturing the spatial information of input images in MLLMs, methods such as SpatialVLM [25] exist. This method extracts various information from the input image using expert models like depth estimation and segmentation, integrates these results, and trains the MLLM based on the reconstructed 3D information. It is a promising direction to utilize off-the-shelf models to construct 3D information and train MLLMs to recognize food more spatially.

Furthermore, to prevent overestimation of calorie values under zero-shot conditions, it seems promising to prompt the language model to revise and correct the reasoning process when an output with an unreasonable calorie value is generated, based on the commonsense knowledge on food in MLLMs.

6 Conclusion

In this study, we proposed CalorieVoL, a novel framework that enhances the calorie estimation from food images by incorporating volumetric context into MLLMs. Our evaluations on the Nutrition5k dataset under zero-shot conditions demonstrate that CalorieVoL with GPT-4V improves the estimation accuracy, particularly in scenarios where the quantity of food plays a critical role.

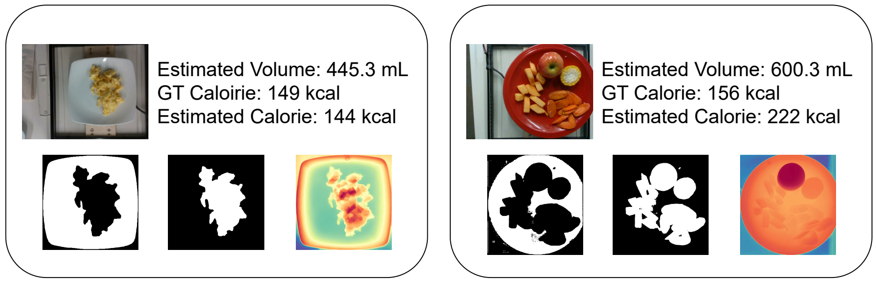

Fig. 8. Results of object detection, segmentation, and depth estimation. Top left: original image, bottom: dish region mask, food region mask, and depth map.

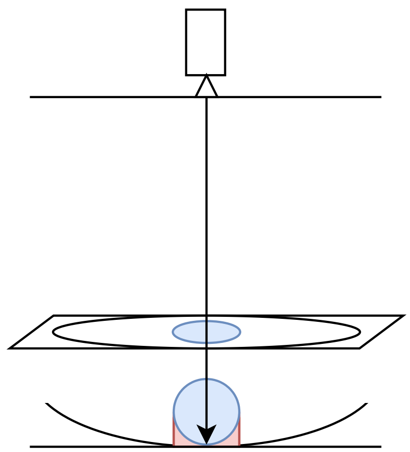

Fig. 9. Overestimation of volume by the food volume estimator when the food is assumed to be spherical. Blue: food region, Red: excess region.

We also identified challenges related to the volume estimation and the overestimation of food volume. We discussed potential approaches to address these issues, including the incorporation of mass regression models and advanced 3D shape reconstruction techniques. Looking forward, integrating the methods such as NeRF and 3D Gaussian Splatting, as well as adopting spatial-aware MLLMs training frameworks, could further enhance the accuracy and applicability of calorie estimation models.

Overall, CalorieVoL represents a promising step towards more accurate and reliable calorie estimation from food images, with potential applications in personalized nutrition and healthcare. Future work will focus on refining the volume estimation process and correcting the reasoning procedure with MLLMs.

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

References

1. Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *Proc. of International Conference on Learning Representations*, 2021.
2. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.
3. Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.
4. Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The Dawn of LMMs: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.
5. Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: Towards automatic nutritional understanding of generic food. In *Proc. of IEEE Computer Vision and Pattern Recognition*, pages 8903–8911, 2021.
6. Jamalia Sultana, Benzir Md. Ahmed, Mohammad Mehedy Masud, A. K. Obidul Huq, Mohammed Eunus Ali, and Mahmuda Naznin. A study on food value estimation from images: Taxonomies, datasets, and techniques. *IEEE Access*, 11:45910–45935, 2023.
7. Koichi Okamoto and Keiji Yanai. An automatic calorie estimation system of food images on a smartphone. *Proc. of the 2nd International Workshop on Multimedia Assisted Dietary Management*, 2016.
8. Yutaka Arakawa Elder Akpro Hippocrate Akpa, Hirohiko Suwa and Keiichi Yasumoto. Smartphone-based food weight and calorie estimation method for effective food journaling. *SICE Journal of Control, Measurement, and System Integration*, 10(5):360–369, 2017.
9. Takumi Ege, Wataru Shimoda, and Keiji Yanai. A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice. In *Proc. of ACMMM Workshop on Multimedia Assisted Dietary Management*, 2019.
10. Ryosuke Tanno, Takumi Ege, and Keiji Yanai. AR DeepCalorieCam V2: Food calorie estimation with cnn and ar-based actual size estimation. In *Proc. of the 24th ACM Symposium on Virtual Reality Software and Technology*, 2018.
11. Yoshikazu Ando, Takumi Ege, Jaehyeong Cho, and Keiji Yanai. DepthCalorieCam: A mobile application for volume-based foodcalorie estimation using depth cameras. In *Proc. of the 5th International Workshop on Multimedia Assisted Dietary Management*, page 76–81, 2019.
12. Shu Naritomi and Keiji Yanai. Hungry Networks: 3d mesh reconstruction of a dish and a plate from a single dish image for estimating food volume. In *Proc. of the 2nd ACM International Conference on Multimedia in Asia*, 2021.
13. Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
14. Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H.

Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
15. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736, 2022.
16. Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
17. Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
18. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. of International Conference on Machine Learning*, 2023.
19. Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. FoodLMM: A versatile food assistant using large multi-modal model. *arXiv preprint arXiv:2312.14991*, 2023.
20. Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
21. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.
22. Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. *arXiv preprint arXiv:2312.02145*, 2023.
23. Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. of European Conference on Computer Vision*, 2020.
24. Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for real-time radiance field rendering. In *Procs. of ACM Transactions on Graphics*, volume 42, 2023.
25. Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024.