

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

Hikaru Tanabe¹, Keiji Yanai¹

¹The University of Electro-Communications



Introduction

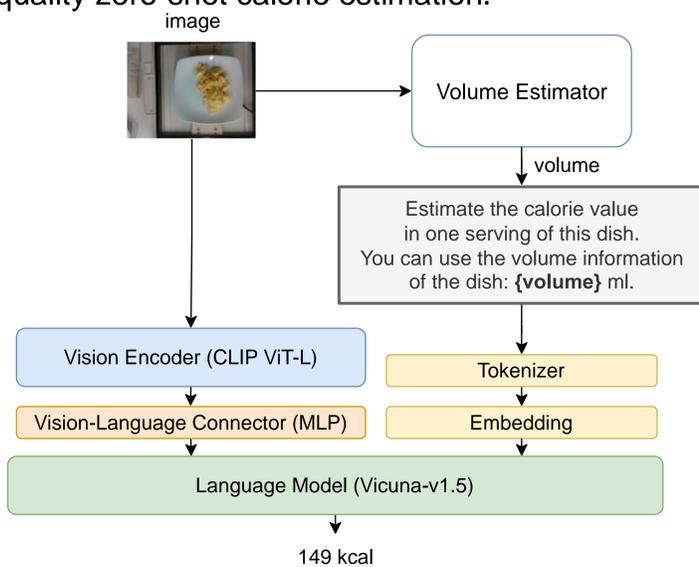
- **Image-based Calorie Estimation** plays a vital role in developing practical nutrition management applications for daily use.
- Existing calorie estimation methods suffer from poor scalability in the variety of subject foods [1].
- **Multimodal Large Language Models (MLLMs)** [2] are capable of handling various tasks, including food-related ones, but they face challenges in recognizing food size.

Can we enhance zero-shot calorie estimation by leveraging volume-aware MLLMs?

Methodology

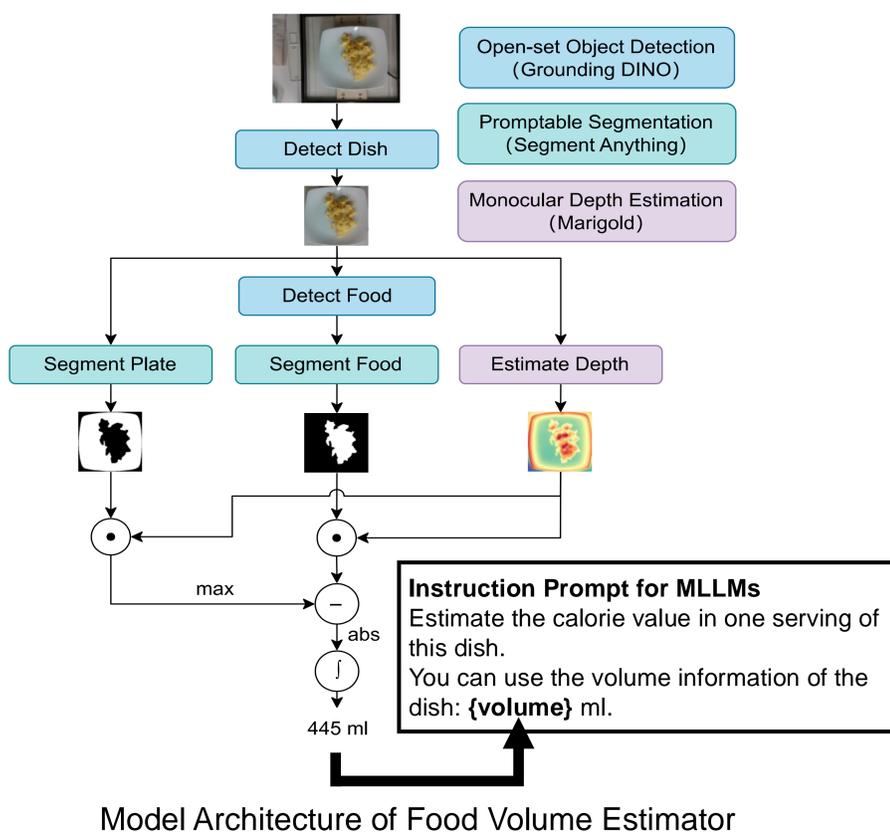
Model Overview

We integrate volume estimation into MLLMs to achieve high-quality zero-shot calorie estimation.



Integrating Volumetric Context into MLLMs

Extract the food region from the input image and estimate its volume using depth information.



Experiments

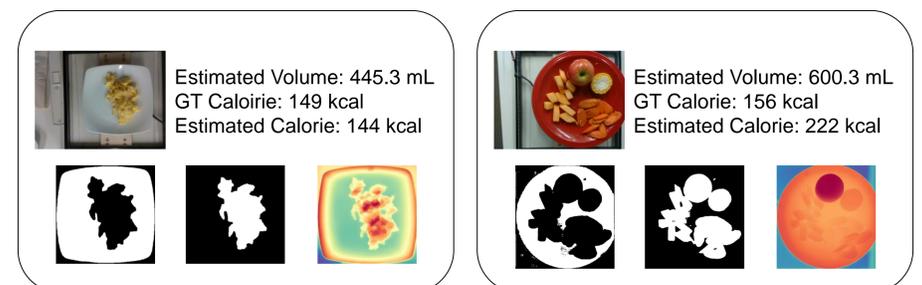
Zero-shot Calorie Estimation on Nutrition5k [3]

- Volume-aware reasoning performed higher-quality estimation
- We observed a tendency for producing outliers when the reasoning performance of MLLM was insufficient (such as LLaVA-1.5-13B)

Method	MAE / kcal ↓	MAPE / % ↓	r ↑
LLaVA-1.5-13B	109.6	92.8	0.656
GPT-4V	106.6	54.8	0.688
GPT-4o	82.7	46.7	0.817
LLaVA-1.5-13B + vol	6122.7	6591.4	-0.041
GPT-4V + vol	83.8	54.1	0.816
GPT-4o + vol	78.8	43.4	0.846

Volume Estimation Process with GPT-4V + Volume Injection

Segmentation masks of the plate and food, and depth capture detailed features of dish.



Reasoning Procedure with GPT-4V + Volume

Volume-awareness improves calorie estimation, especially in cases where food size is important.



Answer: **687 kcal**



A typical serving size of almonds is about **1 ounce** (28 grams). However, without a reference for the exact amount in this bowl, it's difficult to provide an accurate calorie count. The calorie estimate would be around **[[160]]** calories.

GPT-4V+Vol (Ours)

A typical serving size of almonds is about 1 ounce (28 grams). This serving size generally contains about **160 to 170 calories**. Given the volume information of the dish (**527 ml**), it could potentially hold around **4 to 5 ounces** of almonds. For the purpose of providing a single calorie value, I will estimate the calorie content of the bowl of almonds as follows: **[[700]]** calories

✘ Failure Case of Reasoning with LLaVA-1.5-13B + Volume

MLLM with insufficient reasoning performance makes mistakes in understanding units and multiplication.



Answer: **149 kcal**

LLaVA+Vol (Ours)

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**.

Conclusion and Future Work

We proposed CalorieVoL, a framework to enhance volume-awareness of MLLMs for more accurate calorie estimation

Future work:

- Improving reasoning procedure of calorie estimation with MLLMs
- Evaluation on volume estimation and more various foods

[1] Yoshikazu Ando, Takumi Ege, Jaehyeong Cho, and Keiji Yanai. DepthCalorieCam: A mobile application for volume-based food calorie estimation using depth cameras. In Proc. of the 5th International Workshop on Multimedia Assisted Dietary Management, p. 76–81, 2019.
 [2] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744, 2023.
 [3] 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, pp. 8903–8911, 2021.