CalorieLLaVA: Image-based Calorie Estimation with Multimodal Large Language Models

Hikaru Tanabe[1], Keiji Yanai[1]

[1]The University of Electro-Communications

Dec. 2, 2024

Image-based Calorie Estimation

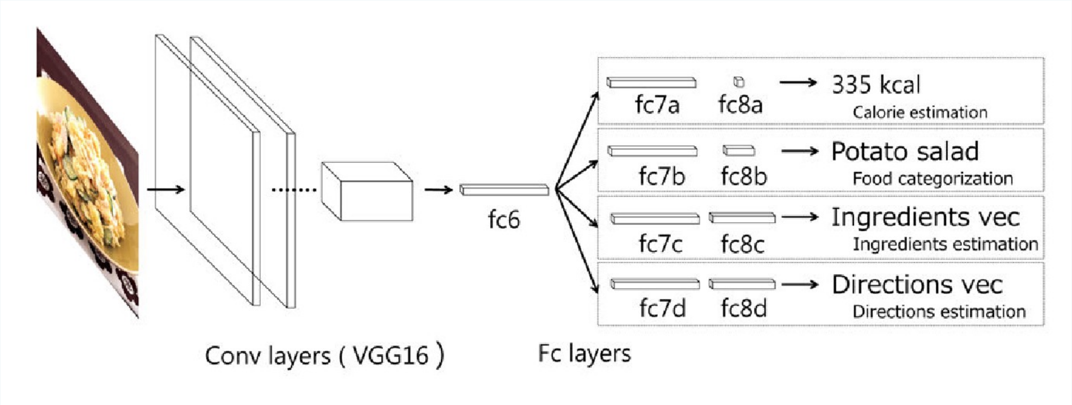
Food Image

Estimate

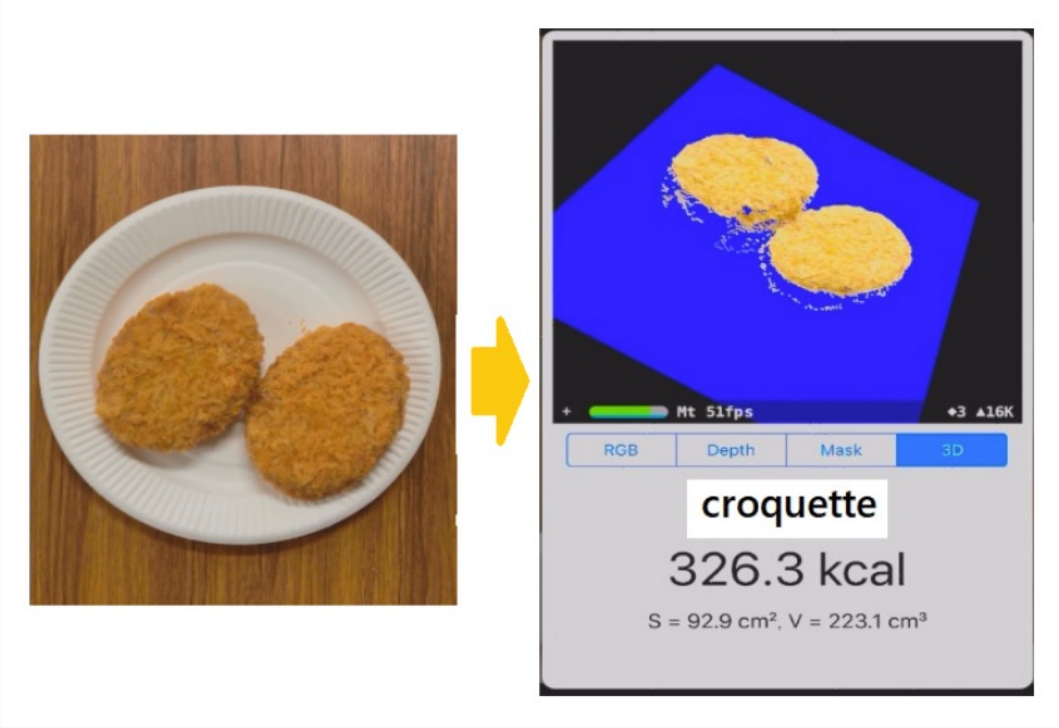
750 kcal

Energy Content
(Calorie)

Previous Methods for Calorie Estimation



End-to-end Methods [1]

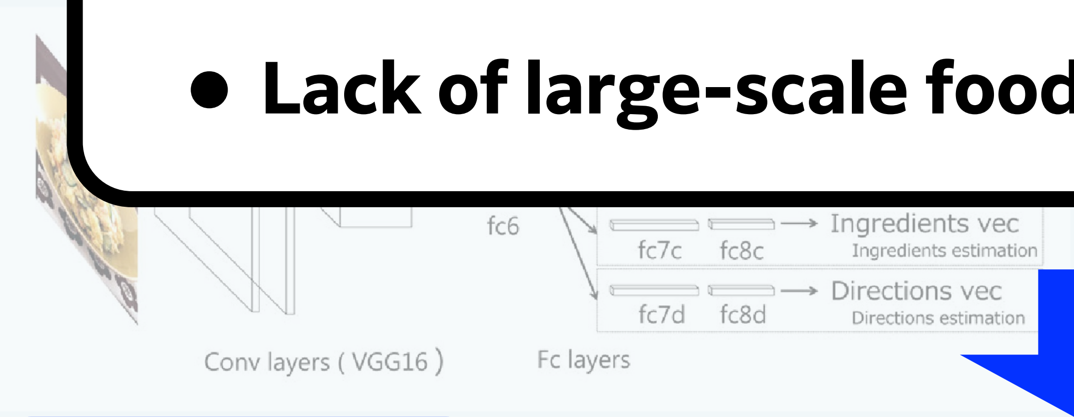


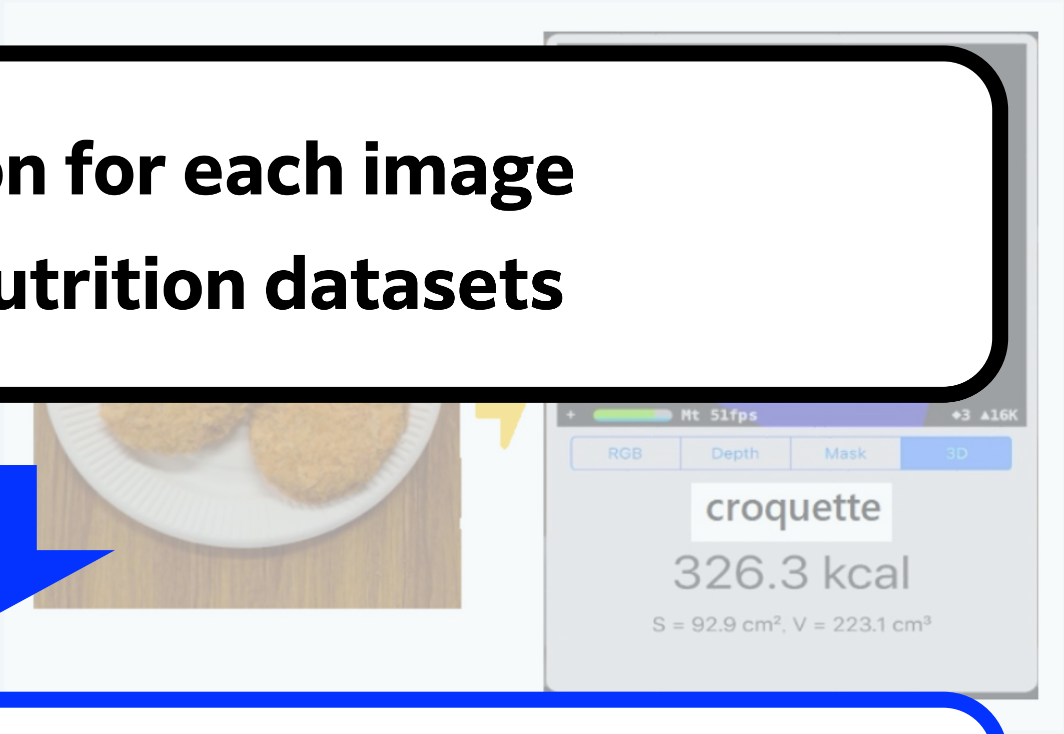
Size-based Methods [2]

[1] Ege, T., Yanai, K.: Image-Based Food Calorie Estimation Using Knowledge on Food Categories, Ingredients and Cooking Directions. In: Proc. of the on Thematic Workshops of ACM Multimedia 2017. pp. 367–375 (2017)

[2] Ando, Y., Ege, T., Cho, J., Yanai, K.:DepthCalorieCam: A Mobile Application for Volume-Based FoodCalorie Estimation using Depth Cameras. In: Proc. of the 5th, International Workshop on Multimedia Assisted Dietary Management. p. 76–81 (2019)

Previous Methods for Calorie Estimation





Previous Methods

- **Labor-intensive annotation for each image**
- **Lack of large-scale food nutrition datasets**

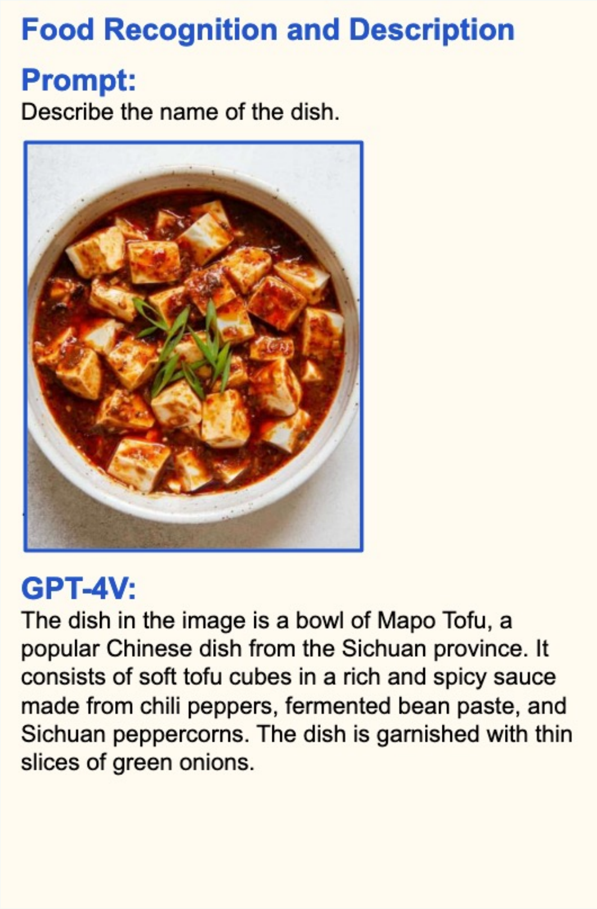
Suffer from

😢 **Poor scalability in the variety of food items**

[1] Ege, T., Yan e on Thematic Workshops of ACM Multimedia 2017. pp. 367–375 (2017)

[2] Ando, Y., Ege, T., Cho, J., Yanai, K.:DepthCalorieCam: A Mobile Application for Volume-Based FoodCalorie Estimation using Depth Cameras. In: Proc. of the 5th, International Workshop on Multimedia Assisted Dietary Management. p. 76–81 (2019)

Multimodal Large Language Models (MLLMs)

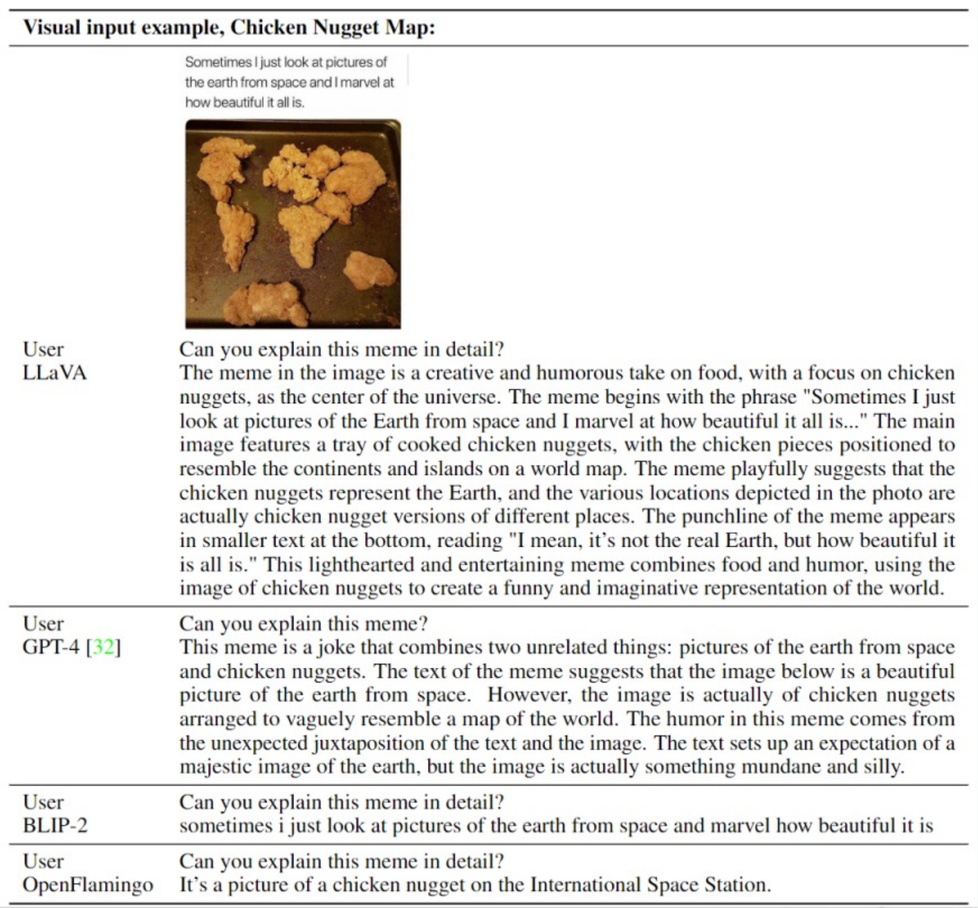


Food Recognition and Description

Prompt:
Describe the name of the dish.

GPT-4V:
The dish in the image is a bowl of Mapo Tofu, a popular Chinese dish from the Sichuan province. It consists of soft tofu cubes in a rich and spicy sauce made from chili peppers, fermented bean paste, and Sichuan peppercorns. The dish is garnished with thin slices of green onions.

GPT-4V [2, 3]



Visual input example, Chicken Nugget Map:

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.

User LLaVA	Can you explain this meme in detail? The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.
User GPT-4 [32]	Can you explain this meme? This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.
User BLIP-2	Can you explain this meme in detail? sometimes i just look at pictures of the earth from space and marvel how beautiful it is
User OpenFlamingo	Can you explain this meme in detail? It's a picture of a chicken nugget on the International Space Station.

LLaVA [4]

[2] OpenAI, GPT-4V(ision) System Card, 2023.

[3] 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.

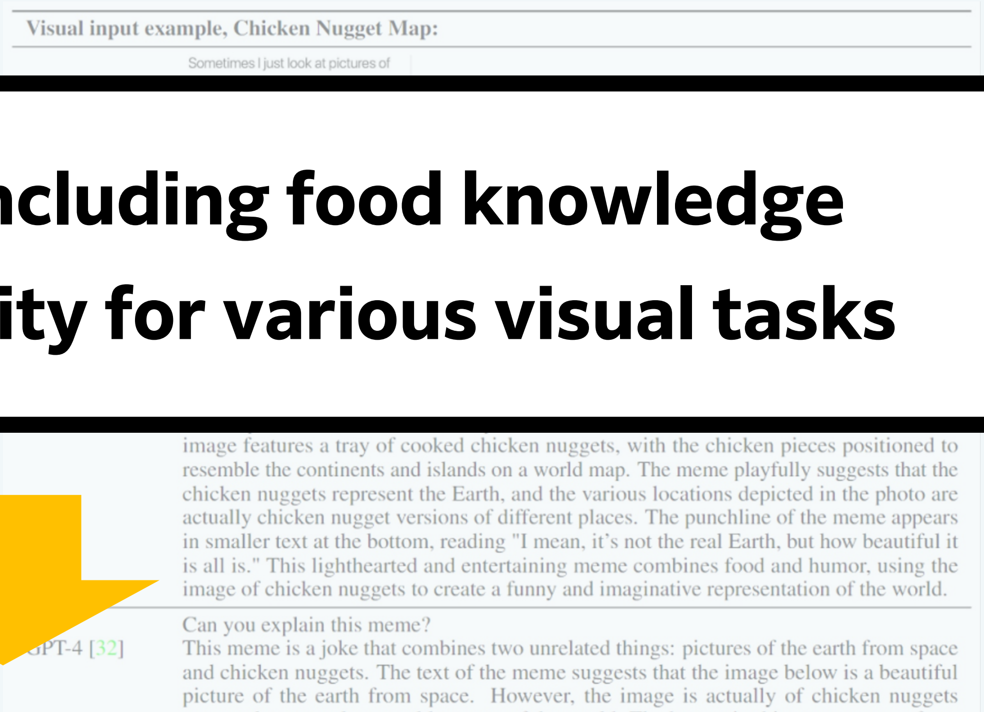
[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In Advances in Neural Information Processing Systems, 2023.

Multimodal Large Language Models (MLLMs)

Food Recognition and Description

GPT-4V:

The dish in the image is a bowl of Mapo Tofu, a popular Chinese dish from the Sichuan province. It consists of soft tofu cubes in a rich and spicy sauce made from chili peppers, fermented bean paste, and



Visual input example, Chicken Nugget Map:

Sometimes I just look at pictures of

image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.

PT-4 [32]

Can you explain this meme?
This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets

- **Pre-trained on dataset including food knowledge**
- **Zero-shot reasoning ability for various visual tasks**

😊 Large potential to overcome data deficiency issue in image-based calorie estimation

[2] OpenAI, G

[3] 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.

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In Advances in Neural Information Processing Systems, 2023.

Our Contribution

1. CalorieLLaVA

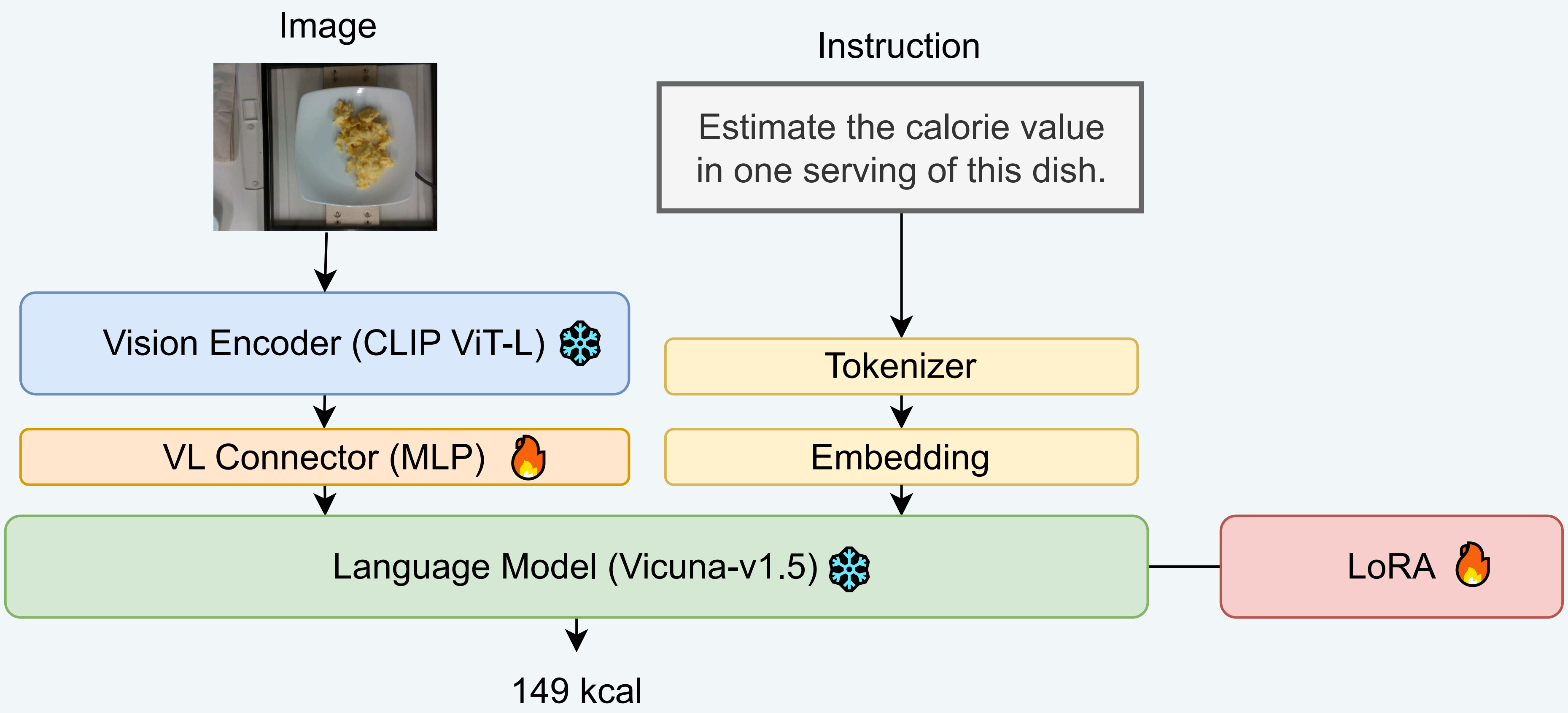
MLLM baseline for image-based calorie estimation

2. Volume Injection

(extra commitment from the paper)

Providing volume context to MLLM for zero-shot calorie estimation

CalorieLLaVA



Image

Instruction

Estimate the calorie value in one serving of this dish.

Vision Encoder (CLIP ViT-L) ❄

Tokenizer

VL Connector (MLP) 🔥

Embedding

Language Model (Vicuna-v1.5) ❄

LoRA 🔥

149 kcal

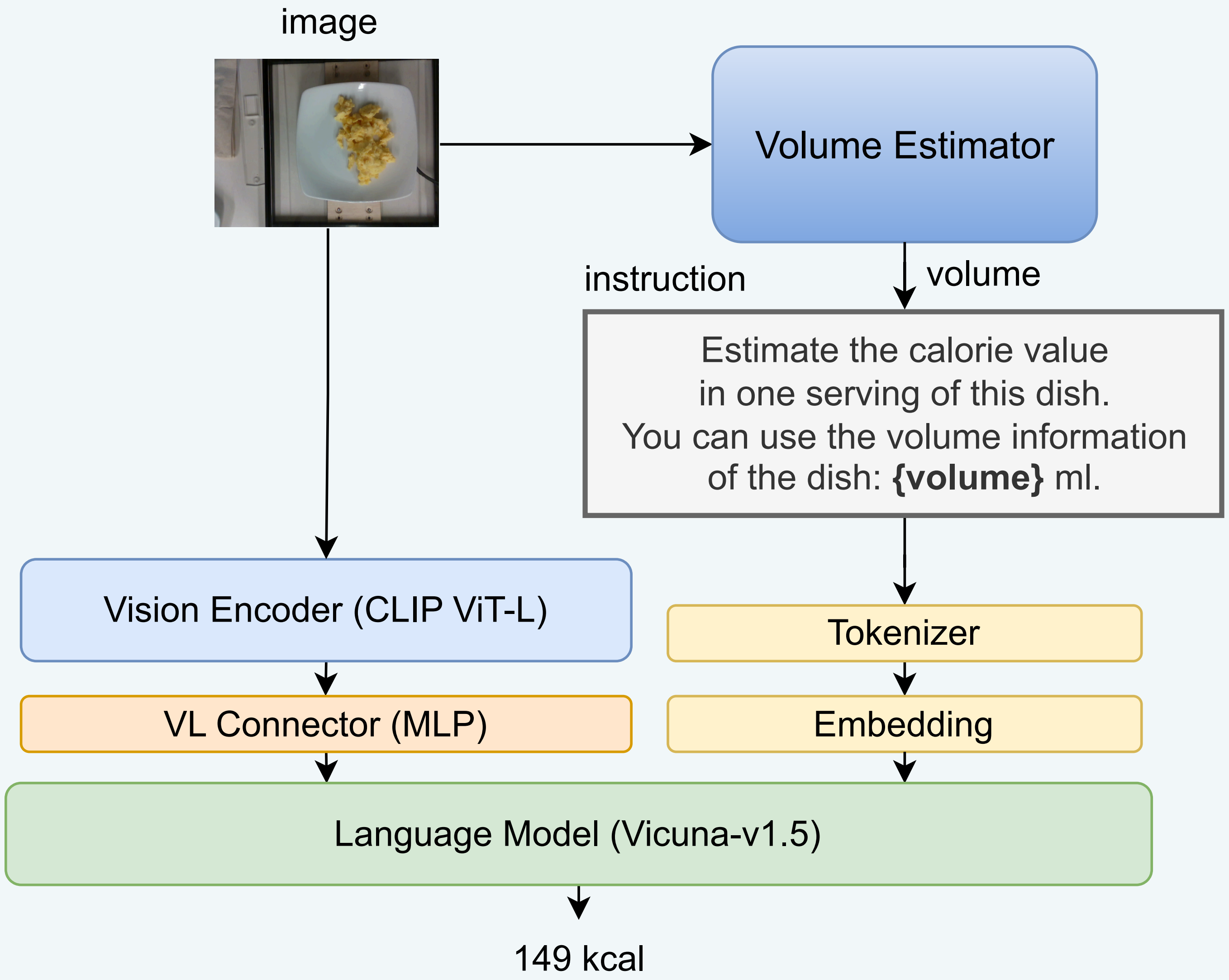
CalorieLLaVA: Experiments

Method	MAE / kcal ↓	MAPE / % ↓	r ↑
GNM [5]	70.6	26.1	-
LLaVA-1.5-7B	123.4	129.5	0.637
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
FoodLMM FT [6]	67.3	**26.6**	-
CalorieLLaVA-7B (Ours)	72.5	41.5	0.927
CalorieLLaVA-13B (Ours)	**64.3**	39.8	**0.934**

[5] Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., Sim, J.: Nutrition5k: Towards automatic nutritional understanding of generic food. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 8903–8911 (2021)

[6] 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 preprint arXiv:2312.14991 (2023)

Volume Injection Approach



image

Volume Estimator

instruction

volume

Estimate the calorie value
in one serving of this dish.
You can use the volume information
of the dish: **{volume}** ml.

Vision Encoder (CLIP ViT-L)

Tokenizer

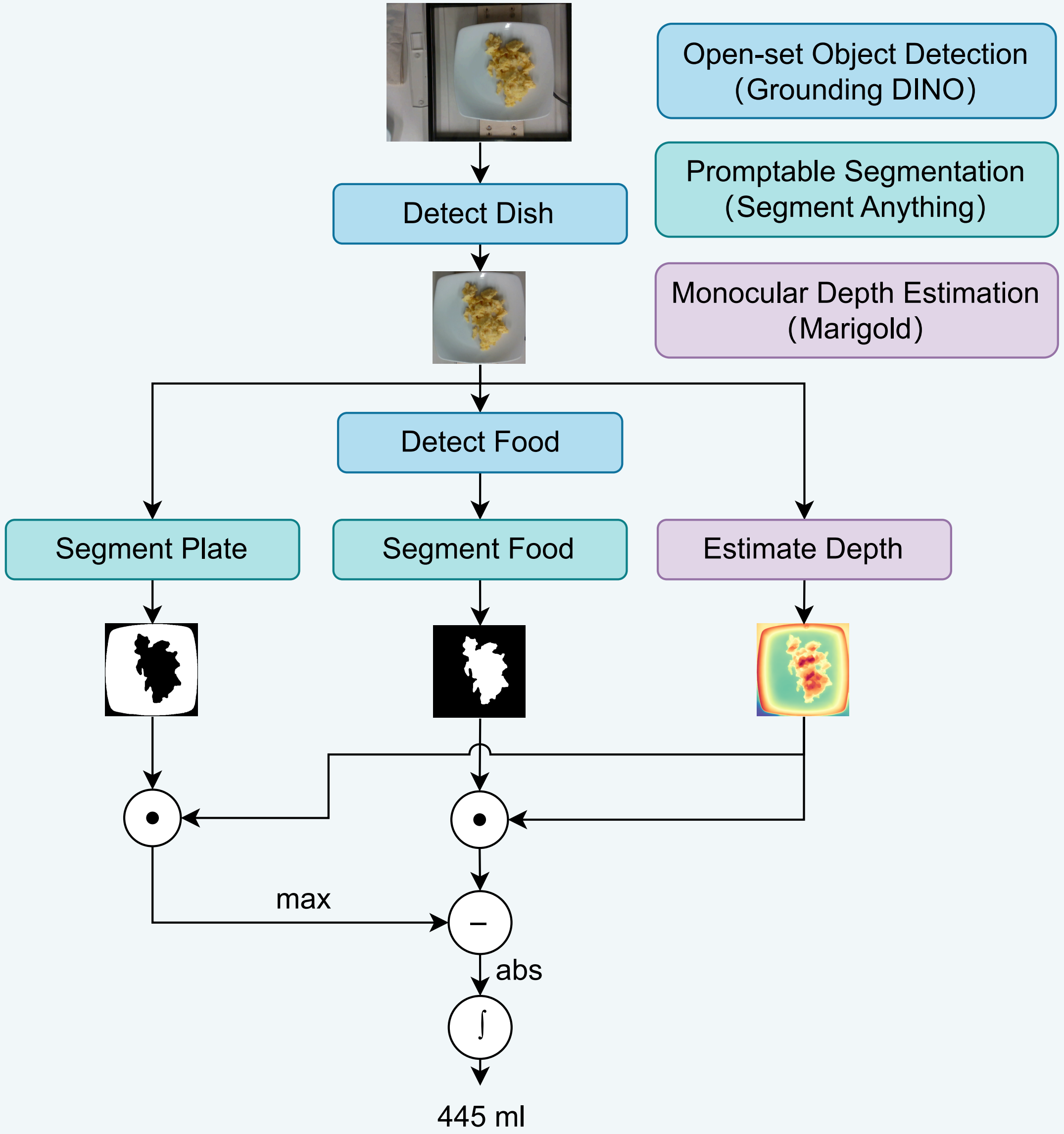
VL Connector (MLP)

Embedding

Language Model (Vicuna-v1.5)

149 kcal

Food Volume Estimator



Open-set Object Detection (Grounding DINO)

Promptable Segmentation (Segment Anything)

Monocular Depth Estimation (Marigold)

Detect Dish

Detect Food

Segment Plate

Segment Food

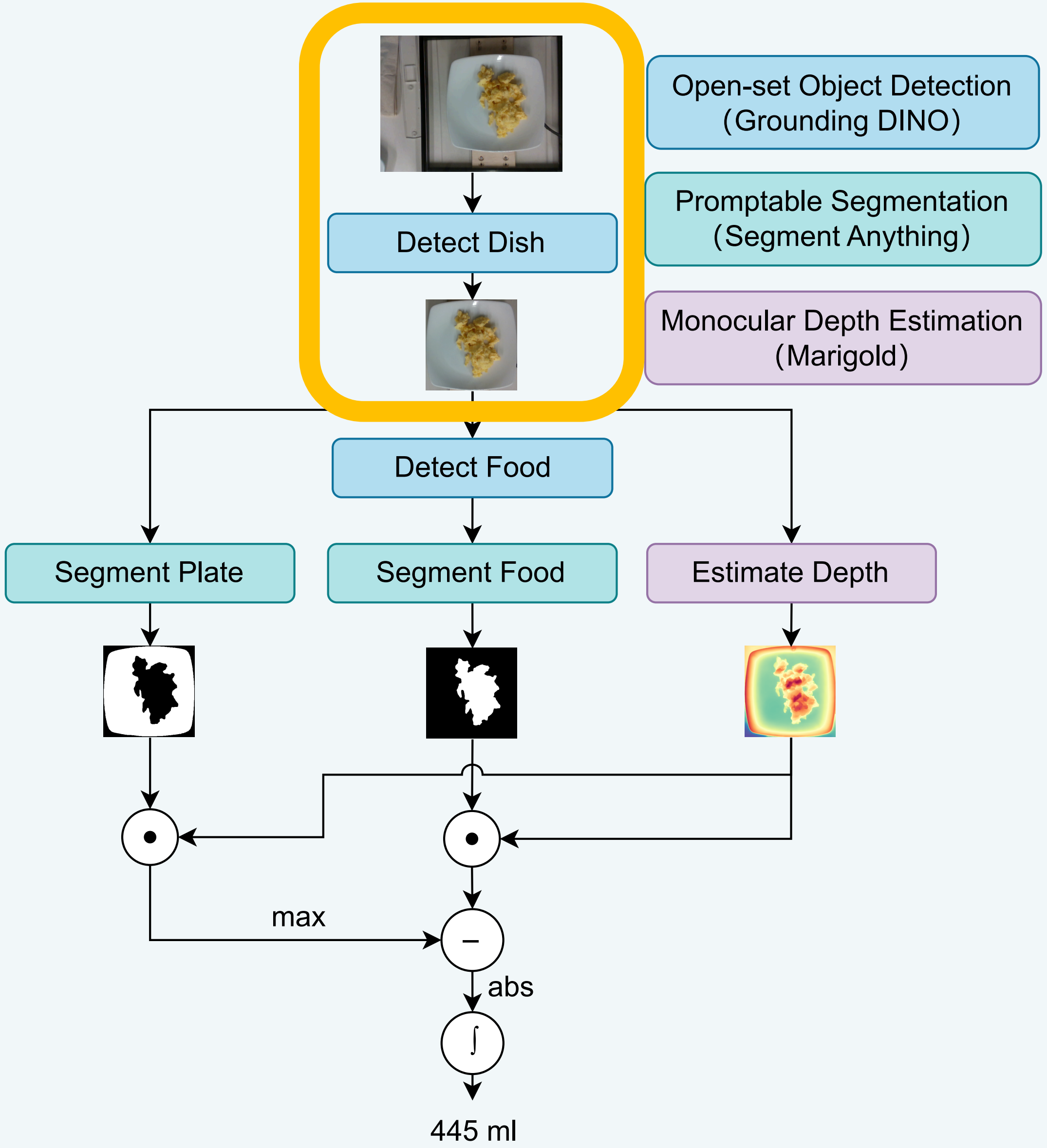
Estimate Depth

max

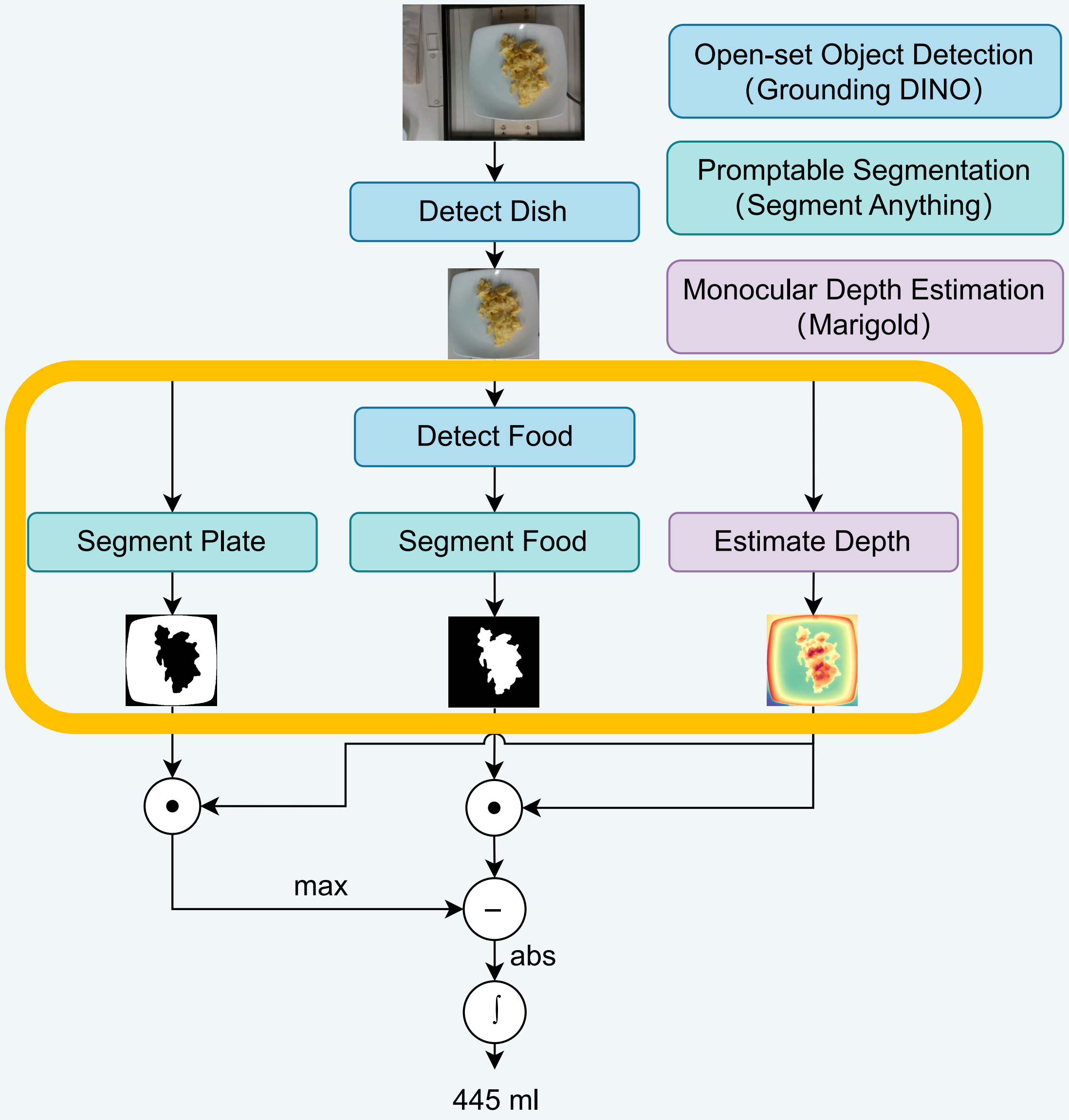
abs

445 ml

Food Volume Estimator


Open-set Object Detection (Grounding DINO)
Promptable Segmentation (Segment Anything)
Monocular Depth Estimation (Marigold)
Detect Dish
Detect Food
Segment Plate
Segment Food
Estimate Depth
max
−
abs
∫
445 ml

Food Volume Estimator



Open-set Object Detection
(Grounding DINO)

Promptable Segmentation
(Segment Anything)

Monocular Depth Estimation
(Marigold)

Detect Dish

Detect Food

Segment Plate

Segment Food

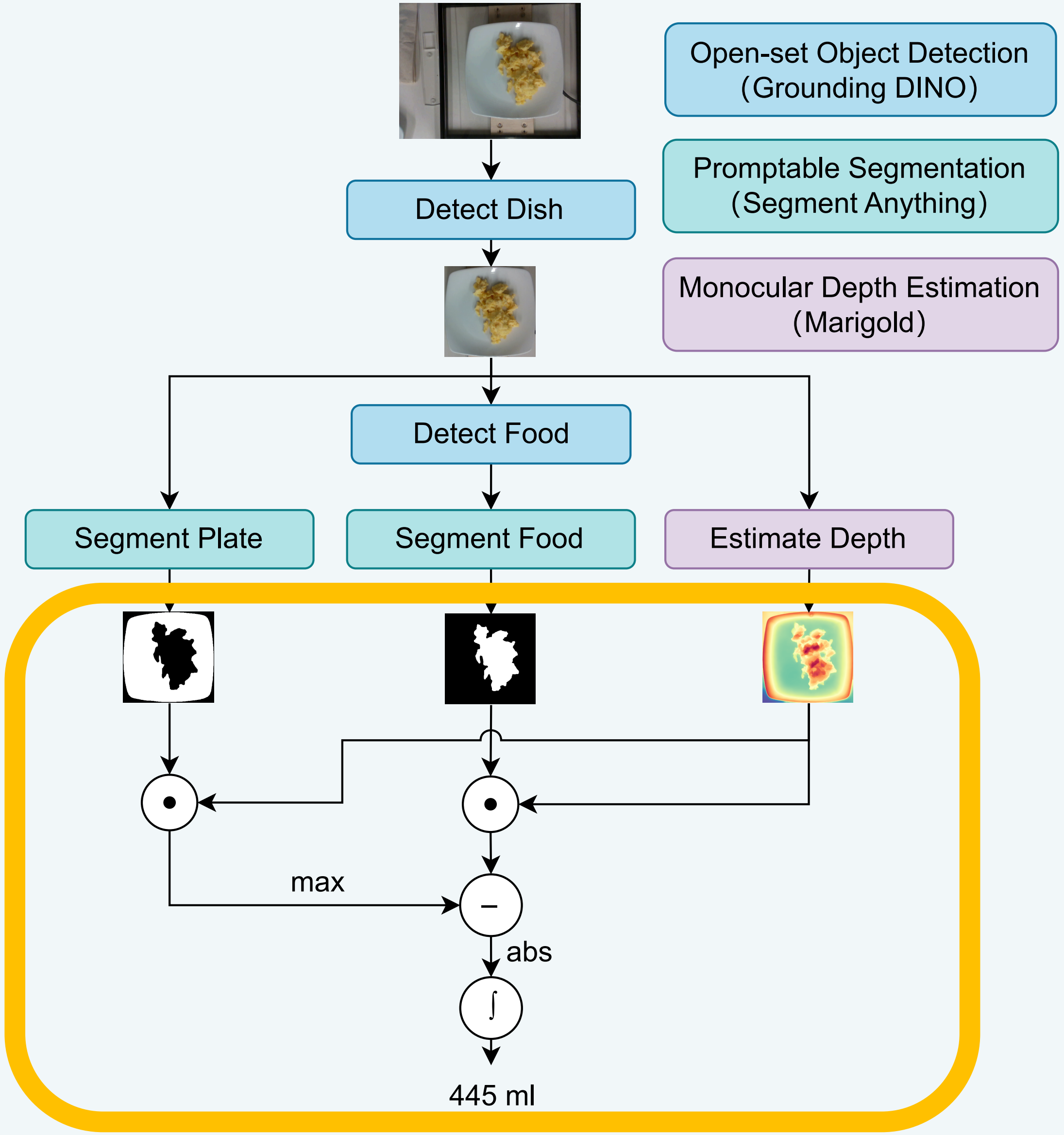
Estimate Depth

max

abs

445 ml

Food Volume Estimator



Open-set Object Detection (Grounding DINO)

Promptable Segmentation (Segment Anything)

Monocular Depth Estimation (Marigold)

Detect Dish

Detect Food

Segment Plate

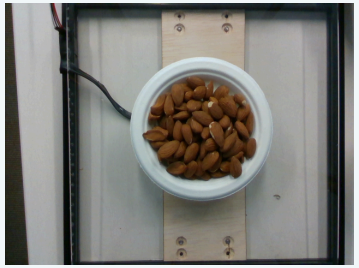
Segment Food

Estimate Depth

Experiments: Zero-shot Volume Injection

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 (Ours)	6122.7	6591.4	-0.041
GPT-4V + vol (Ours)	83.8	54.1	0.816
GPT-4o + vol (Ours)	**78.8**	**43.4**	**0.846**

Experiments: Zero-shot Volume Injection



Answer: **687 kcal**

GPT-4V

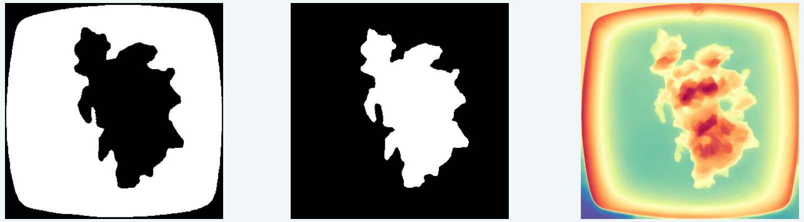
A typical serving size of almonds is about **1 ounce** (28 grams).
However, without a reference for the exact amount in this bowl, it's difficult to provide an accurate calorie count.
The calorie estimate would be around **[[160]] calories**.

GPT-4V+Vol (Ours)

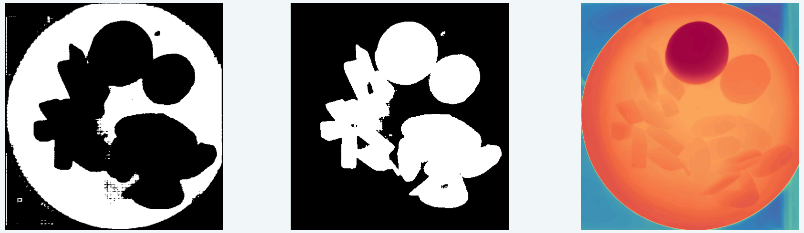
A typical serving size of almonds is about 1 ounce (28 grams). This serving size generally contains about **160 to 170 calories**. Given the volume information of the dish **(527 ml)**, it could potentially hold around **4 to 5 ounces** of almonds.
For the purpose of providing a single calorie value,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

Food Volume Estimator

Estimated Volume: 445.3 mL
GT Caloirie: 149 kcal
Estimated Calorie: 144 kcal



Estimated Volume: 600.3 mL
GT Calorie: 156 kcal
Estimated Calorie: 222 kcal



Conclusion

- **CalorieLLaVA**: MLLM baseline for image-based calorie estimation
- **Volume Injection (extra):** Providing volume context to MLLM for zero-shot calorie estimation

Future Work

- Evaluation on volume estimation and more various foods
- More efficient injection method to recognize food size