(IS2-202)

CalorieCaMLLM: LiDAR センサと多モーダル大規模言語モデルを用いたオンデバイス食事栄養推定

藤田晴斗, 柳井啓司 (電通大)

概要

背景 食事画像からのカロリー推定には、次の3つの両立が課題となっている：

① プライバシー・リアルタイム性（クラウド非依存）
② 食品カテゴリの汎化性（未知の料理への対応）
③ 量の絶対的把握（相対値でなく実寸法）

貢献 iPhone LiDARによる絶対mmスケール情報(PAVR)を、追加のNN推論なしにMLLMへ直接供給することで①～③を同時に満たすオンデバイス栄養推定を実現

Nutrition5k(n=503)でカロリーMAE 59.6kcal（Qwen3.5-4B+PAVR）を達成

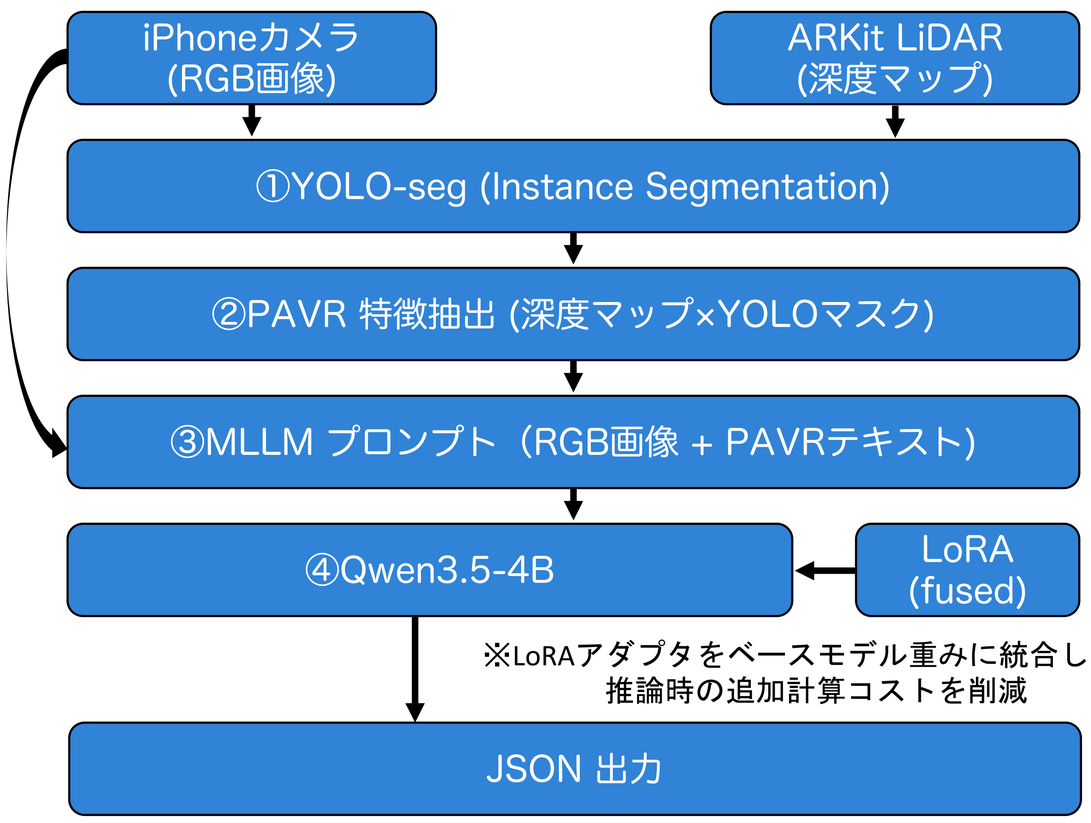
関連研究

表1. 食事栄養推定手法の比較と本研究の位置づけ

	プライバシー/リアルタイム性	カテゴリ非依存	絶対スケール
LiDARCalorieCam[1] (点群+サーバー回帰)	△ (サーバー処理)	✗	✓ (カテゴリ別回帰式)
CalorieVoL[2] (単眼深度+LLM)	✓	✓	✗ (相対スケールのみ)
本研究 CalorieCaMLLM	✓	✓	✓

[1] LiDARCalorieCam:
点群から体積推定→食品カテゴリ別回帰式でカロリー変換（外部サーバー処理が必須）
[2] CalorieVoL: 単眼深度推定値をMLLMプロンプトに注入し精度改善（推定深度は相対スケールに留まる）

手法



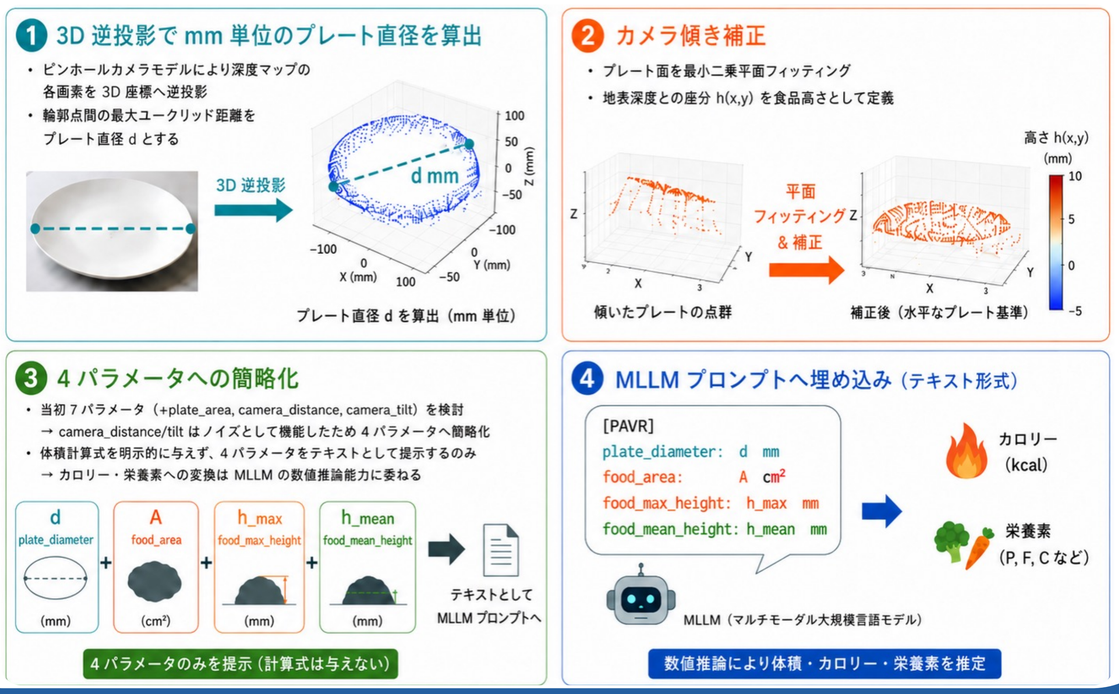
出力項目：
カロリー(kcal)・タンパク質(g)・脂質(g)・炭水化物(g)・総質量(g)

★ 本手法のポイント 体積計算式を明示的に与えず、PAVRの4パラメータをテキストとして 提示するのみ → カロリー・栄養素への変換はMLLMの数値推論能力に委ねる
（LiDARCalorieCam[1]の明示的回帰式 W=aV+b との本質的な違い）

PAVR: Plate-Anchored Volumetric Reasoning

絶対スケールを持つプレートを空間的基準として活用し、4パラメータをテキストとしてMLLMプロンプトへ埋め込む

```
[PAVR] plate_diameter: d mm
food_area: A cm²
food_max_height: h_max mm
food_mean_height: h_mean mm
```



結果

表2. 比較実験の条件と入力情報　Nutrition5k (n=503) での実験

実験環境	追加情報	追加学習
① Zero-shot (RGB)	なし	✗
② RGB + LoRA	なし	✓
③ + 単眼深度(CalorieVoL)	相対体積	✓
④ + LiDAR (Ours / PAVR)	**絶対 4 パラメータ**	✓

表3. 各モデル・条件における栄養推定精度（MAE / R^2）

MLLM モデル	① Zero-shot	② RGB + Lora	③ + 単眼深度	④ + LiDAR (Ours)
Qwen2-VL-2B	102.7 / 0.53	62.0 / 0.79	60.0 / 0.80	60.1 / 0.80
Qwen3-VL-4B	103.0 / 0.55	-	-	60.7 / 0.80
Qwen3.5-4B	99.0 / 0.57	-	-	**59.6 / 0.80** *(Best)*
Gemma4-E2B	128.9 / 0.21	-	-	68.0 / 0.75
Gemma4-E4B	112.9 / 0.49	-	-	62.7 / 0.78
LLaVA-1.5-13B [3]	177.1 / 0.43*	64.3 / 0.87*		
GPT-4o [3]	82.7 / 0.67*		78.8 / 0.72*	

※ R^2 (Coefficient of Determination / 決定係数) :モデルの適合度を示す指標
※ MAE (Mean Absolute Error / 平均絶対誤差):誤差の絶対値の平均
※LLaVA-1.5-13BおよびGPT-4oのR^2値は、相関係数 r から R^2 ≒ r^2 として換算した推定値

表4. Qwen2-VL-2Bにおける栄養素別の推定誤差および改善率（MAE）

栄養素	① RGB	④ + LiDAR (Ours)	改善率
カロリー (kcal)	102.7	60.1	-41%
タンパク質 (g)	8.8	4.9	-44%
脂質 (g)	10.7	6.1	-43%
炭水化物 (g)	**22.0**	**6.4**	**-71%**
総質量 (g)	70.2	37.2	-47%

考察・今後の課題

- PAVRの有効性：追加のNN推論を要さず単眼深度推定と同等精度を達成（Qwen2-VL-2B: paired t-test p=0.791, Wilcoxon p=0.833、有意差なし）
→ 実装上の優位性
- Qwen3.5-4Bがビジョン言語特化モデルを上回る→ 数値テキスト解釈能力の重要性を示唆
- 限界：
プレート食器のみ対応（Nutrition5kの約29%を占める非皿食器は非対応）／複数食品の一括計測は困難
- 今後：
iPhone実撮影データでの定量的精度検証／Chain-of-Thought などを活用した異常値の事後検証／Gemma4-E4B等の軽量化とiPhone実運用検証／PFC値も活用した栄養素アドバイス

[1] H. Fujita and K. Yanai, "Mobile Food Calorie Estimation Using Smartphone LiDAR Sensor," in Proc. of ACPR, 2026.
[2] H. Tanabe and K. Yanai, "CalorieVoL: Integrating Volumetric Context into Multimodal Large Language Models for Image-Based Calorie Estimation," in Proc. of MMM, 2025.
[3] H. Tanabe and K. Yanai, "Reasoning-Driven Food Energy Estimation via Multimodal Large Language Models," Nutrients, vol. 17, no. 7, art. no. 1128, 2025.