CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

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[1,a)]　柳井 啓司[1,b)]

概要

本稿では，iPhone の LiDAR センサデータと多モーダル大規模言語モデル（MLLM）を統合したオンデバイス食事栄養推定システム CalorieCaMLLM を提案する．本システムは，LiDAR から得られるプレート直径・食材面積・高さなどの絶対スケール情報を構造化テキストとして MLLM プロンプトへ注入する **Plate-Anchored Volumetric Reasoning (PAVR)** フレームワークである．Nutrition5k データセットを用いた 4 条件のアブレーション実験により，LiDAR 絶対スケール情報が追加のニューラルネットワーク推論を必要とする単眼深度推定と同等の精度（カロリー $R^2 = 0.804$，MAE 約 60 kcal）を達成することを実証した．さらに複数の MLLM モデルとの比較実験を行い，Qwen3.5-4B が MAE 59.6 kcal と最良の精度を達成することを確認した．また，MLLM モデルを 4 ビット量子化により iPhone 上でリアルタイム実行を可能にし，クラウド非依存なプライバシー保護型栄養推定を実現した．

1. はじめに

食事の栄養情報を自動推定することは，個人の栄養管理やヘルスケアアプリケーションにおいて重要な課題であるが，既存システムの多くはクラウドへの画像送信を要し，プライバシーおよびレイテンシの問題を抱えている．

LiDARCaloriCam [1] は，DeepLabV3 によるセグメンテーション，LiDAR 点群による体積推定，線形回帰によるカロリー算出などを可能としているが，体積計算の際にサーバーが必要である点や，食品カテゴリ別の回帰式に依存するため未知カテゴリへの汎化などに限界があった．本研究はこの限界を克服するため，MLLM の視覚的意味理解能力に LiDAR の絶対スケール情報を融合する新たなアプローチを採用する．CalorieVoL [2] は DepthAnything V2 による単眼深度推定値を MLLM プロンプトに注入することで精度を改善したが，推定深度は相対スケールに留まり，カメラキャリブレーションの誤差が精度に影響するという制約が残る．対して本研究の PAVR は，iPhone ネイティブ LiDAR センサから直接得られる絶対 mm 単位の空間情報を，追加の NN 推論なしに MLLM へ供給する．

本稿の主な貢献を以下に示す：(1) LiDAR 絶対スケール情報を MLLM プロンプトへ注入する PAVR フレームワークの提案，(2) Nutrition5k 上での 4 条件アブレーションによる各要素の効果の定量的検証，(3) 複数の MLLM モデルでの精度比較，(4) Qwen3-VL-4B の 4 ビット量子化と LoRA 融合による iPhone 実機へのオンデバイス実装．

2. 関連研究

食事カロリー推定の深層学習的アプローチとして，Myers ら [3] は Im2Calories を提案し，セマンティックセグメンテーションと単眼深度推定を組み合わせてカロリーを推定した．Thames ら [4] は Nutrition5k データセットを構築し，単一 RGB 画像から栄養素を直接回帰する CNN を提案した．丹野ら [5] は AR アンカーによる実寸計測を利用した AR DeepCalorieCam V2 を開発したが，複数の操作ステップを要するユーザビリティ上の課題があった．LiDARCaloriCam [1] は iPhone の LiDAR で取得した 3 次元点群から Convex Hull・メッシュ法等を統合して体積を算出し，唐揚げ・コロッケ等 10 カテゴリで高精度を達成したが，カテゴリ別回帰式への依存が汎化の制約となっていた．

MLLM の応用として，FoodLMM [13] は食事特化モデルを提案し，量的推定にはスケール情報が不可欠であることを指摘している．CalorieVoL [2] は単眼深度推定値を MLLM プロンプトに注入することで精度を改善したが，推定深度の相対性という根本的制約が残る．本研究は LiDAR 絶対スケールを用いてこの制約を解消する．

3. 提案手法

3.1 システム概要

CalorieCaMLLM は以下の 4 段階パイプラインで構成される（図 1）．

(1) **セグメンテーション**：Ultralytics YOLO26n-seg [10]

[1] 電気通信大学大学院
[a)] fujita-h@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

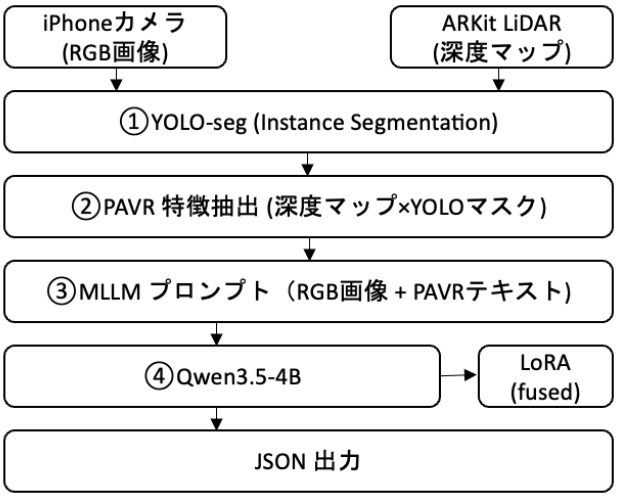


図 1 CalorieCaMLLM のシステムアーキテクチャ

をベースにプレート・食品の 2 クラスでカスタム学習した Instance Segmentation モデルにより，プレート領域と食材領域をそれぞれ検出する．

(2) **LiDAR 特徴抽出（PAVR）**：ARKit LiDAR から取得した深度マップとセグメンテーションマスクを組み合わせ，プレート直径・食材面積・最大高さ・平均高さの 4 パラメータを mm 単位で計算する．

(3) **MLLM 推論**：Qwen3-VL（4B）に対して，RGB 画像とともに PAVR パラメータを構造化テキストとして注入したプロンプトを入力する．

(4) **出力**：カロリー・タンパク質・脂質・炭水化物を JSON 形式で出力する．

3.2 Plate-Anchored Volumetric Reasoning (PAVR)

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，以下の 4 パラメータを MLLM のテキストプロンプトへ埋め込む：

`[PAVR] plate_diameter:`d`mm food_area:`A`cm`^{2}
`food_max_height:`$h_{\max}$`mm food_mean_height:`h_{mean}`mm`
`Estimate the nutritional content of this meal.`

3.2.1 3D 逆投影と実スケール計算

深度マップの各ピクセル (x, y) に対し，ピンホールカメラモデルに基づき 3D 座標を以下のように計算する：

$$X = \frac{(x - c_x) \cdot Z}{f_x}, \quad Y = \frac{(y - c_y) \cdot Z}{f_y}, \quad Z = d(x, y) \tag{1}$$

プレート直径は輪郭点の全点対間の 3D ユークリッド距離の最大値として求める：

$$d_{\text{plate}} = \max_{i \neq j} \|\mathbf{p}_i^{\text{3D}} - \mathbf{p}_j^{\text{3D}}\|_2 \tag{2}$$

学習時は RealSense D435i の公称焦点距離（$f = 615\,\text{px}$, 640×480 解像度時）を，iPhone 推論時は ARKit が毎フレーム提供する実測 intrinsics を使用する．

3.2.2 カメラ傾き補正

カメラが傾くとプレート面に深度勾配が生じ（例：3.4° × 180 mm で約 10.7 mm），スカラー閾値による食品/プレートの分離が破綻する．そのため，プレートマスク内の深度点に対して最小二乗平面フィッティング（$Z = ax + by + c$）を行い，予測深度との差分を食品高さと定義する：

$$h(x, y) = (ax + by + c) - d(x, y) \tag{3}$$

$h(x, y) > 3\,\text{mm}$ を満たすピクセルを食品領域と判定する．また, YOLO マスクの境界誤差を抑制するため, プレートマスクを 6 px モルフォロジー収縮（`MORPH_ELLIPSE`, 13×13 カーネル）した内側マスクを平面フィッティングに用いる．比較条件として，LiDAR の代わりに DepthAnything V2 [12] による単眼深度推定値を同一フォーマットで注入する条件（Cond3）も設けた．Cond3 では推定深度をプレートサイズで正規化することで擬似的な絶対スケールを構成する．この設計により，LiDAR 絶対スケールと単眼深度推定の精度差を直接比較できる．

3.3 出力フォーマット設計

MLLM への出力形式として JSON を採用した．なお，本節の food フィールドアブレーション実験は，後述のパラメータ簡略化（7 パラメータ → 4 パラメータ）以前の暫定設計を用いて実施した（表 1）．食品名を先頭に配置（Exp C）すると食品誤認識がノイズとして後続の栄養推定に伝播し精度が低下した（84.5 kcal）．これは「誤った食品名 × PAVR → 誤った栄養密度の適用」という干渉によるものと解釈できる．末尾配置（Exp B）では Cond2 の精度が向上する一方で Cond4 では改善がみられなかった．これは，PAVR の幾何情報が食品カテゴリの明示なしに機能するため，食品名との干渉が生じないことを示している．最終的に食品名なし（Exp A）・4 パラメータ PAVR を採用し，プロンプト言語を英語に統一することで Cond3/4 で MAE 60 kcal を達成した．

なお，当初は上記 4 パラメータに加え `plate_area`, `camera_distance`, `camera_tilt` を含む 7 パラメータ設計を検討した．しかし，`camera_distance` と `camera_tilt` はノイズとして機能することが実験的に判明したため（4 パラメータ版で MAE が大幅改善），最終的に 4 パラメータへ簡略化した．

3.4 LoRA ファインチューニングと量子化

LoRA ファインチューニングは Linux GPU 環境（NVIDIA RTX 6000 Ada Generation，VRAM 47.7 GB）および M1 Mac 上で実施し，Nutrition5k の 1,423 サンプルを使用した．オプティマイザは AdamW（$\text{lr} = 1 \times 10^{-4}$,

表 1 food フィールド配置による精度比較（Qwen2-VL-2B，パラメータ簡略化前の暫定 7 パラメータ PAVR，テストセット $n = 346$）

実験	food 配置	cond4 MAE		cond2 MAE	
		kcal	質量 g	kcal	質量 g
Exp A	なし	**79.6**	**43.4**	99.7	54.3
Exp B	末尾	79.4	48.1	**75.9**	**45.1**
Exp C	先頭	84.5	50.7	89.8	52.3

weight_decay= 0.01）を用い，ウォームアップ付きコサインスケジューラを適用した（エポック数 2，バッチサイズ 2–4）．最良検証 loss は Cond4（5.421）が Cond2（6.659）を大幅に下回り，PAVR が学習段階から本質的な信号を提供することを確認した．Gemma4-E2B/E4B では vision タワーが専用レイヤでラップされているため，言語モデル部のみを LoRA 対象とした．最終デプロイには Apple Silicon 向け機械学習フレームワーク MLX を用いた推論ライブラリ mlx-vlm [11] により Qwen3-VL-4B を 4 ビット量子化し，iPhone 16 Pro Max 上でのオンデバイス推論を実現した．

4. 実験

4.1 データセットと実験設定

Nutrition5k [4] は Intel RealSense D435i で撮影された 3,490 食品サンプルからなる RGB-D 栄養データセットである．そのうち有効と判断されたサンプル 2,033 件について，訓練 1,423 / 検証 264 / テスト 346 に分割した．学習時センサ（RealSense D435i，構造光）と推論時センサ（Apple LiDAR，ToF）は測定方式が異なるが，いずれも mm 単位の距離を出力するため PAVR 計算の入力形式は共通であり，実機デモで正常動作を確認している．評価指標は MAE・R^2・MAPE を用い，カロリー・タンパク質・脂質・炭水化物の 4 種を推定対象とした．

比較モデルは Qwen2-VL-2B [6]，Qwen3-VL-4B [7]，Qwen3.5-4B [8]，Gemma4-E2B [9]，Gemma4-E4B [9] の 5 種とし，以下の 4 条件を比較した：Cond1（RGB のみ，LoRA なし），Cond2（+LoRA），Cond3（+単眼深度，DepthAnything V2 [12]），Cond4（+LiDAR 絶対スケール，提案法）．Qwen3-VL-4B 以降は Cond1/Cond4 の 2 条件のみ評価した（Cond2/3 は Qwen2-VL-2B で検証済みのため）．

4.2 定量的結果

表 2 に示すように，Cond4（提案法）は Qwen2-VL-2B においてカロリー MAE 60.1 kcal（$R^2 = 0.80$）を達成し，ベースライン Cond1 に対して MAE を 41% 改善した．LoRA のみを付与した Cond2 でも MAE が 102.7 から 62.0 kcal へ大幅に改善されており，LoRA ファインチューニングが

表 2 カロリー推定精度の比較（Nutrition5k テストセット, $n = 346$）

条件	モデル	R^2	MAE (kcal)
Cond1 (RGB のみ)	Qwen2-VL-2B	0.53	102.7
Cond2 (+LoRA)	Qwen2-VL-2B	0.79	62.0
Cond3 (単眼深度)	Qwen2-VL-2B	0.80	60.0
Cond4 (PAVR)	Qwen2-VL-2B	0.80	60.1
Cond1 (RGB のみ)	Qwen3-VL-4B	0.55	103.0
Cond4 (PAVR)	Qwen3-VL-4B	0.80	60.7
Cond1 (RGB のみ)	Qwen3.5-4B	0.57	99.0
Cond4 (PAVR)	Qwen3.5-4B	0.80	**59.6**
Cond1 (RGB のみ)	Gemma4-E2B	0.21	128.9
Cond4 (PAVR)	Gemma4-E2B	0.75	68.0
Cond1 (RGB のみ)	Gemma4-E4B	0.49	112.9
Cond4 (PAVR)	Gemma4-E4B	0.78	62.7

基盤として有効に機能することが確認された．全モデル中最良は Qwen3.5-4B（59.6 kcal）であり，Gemma4-E4B（62.7 kcal）は Qwen3-VL-4B（60.7 kcal）と同等水準で E2B（68.0 kcal）を 8% 上回った．

Cond3（単眼深度）と Cond4（LiDAR）は Qwen2-VL-2B で同等精度（MAE 60 kcal，$R^2 = 0.80$）を達成した．単眼深度推定は DepthAnything V2 の追加推論を必要とするのに対し，LiDAR は iPhone ネイティブセンサのみで絶対スケールを取得できる点が実装上の優位性である．LoRA 適用後は 2B と 4B の精度差が小さく（Qwen2-VL-2B 60.1 vs Qwen3-VL-4B 60.7），LoRA によって小規模モデルが十分な適応能力を獲得することが示された．一方，LoRA なしのベースライン（Cond1）では 4B モデルの方が低い MAE を示しており（Qwen3.5-4B 99.0 vs Qwen2-VL-2B 102.7），より大規模なモデルの事前学習知識が栄養推定の初期性能に寄与していると考えられる．

全栄養素の MAE について，Cond4 は全栄養素で大幅な精度向上をもたらし，炭水化物の改善が 71%（22.0 g → 6.4 g）と最も顕著であった（表 3）．Cond1 では炭水化物 $R^2 = -0.90$ と決定係数が負値となっており，LoRA なしでは炭水化物量の推定がランダム予測以下であることを示している．PAVR による絶対スケール情報の注入が，食品体積と炭水化物量の関連づけを MLLM に学習させる上で特に有効であったと考えられる．

4B モデル間の比較では，テキスト能力強化モデルの Qwen3.5-4B（MAE 59.6 kcal）がビジョン言語特化の Qwen3-VL-4B（MAE 60.7 kcal）をわずかに上回った．これは，PAVR フレームワーク下では数値テキストの解釈・推論能力が視覚エンコーダ性能と同等以上に重要であることを示唆する．Gemma4 ファミリでは，Cond4 適用後の改善率は E4B（62.7 kcal）で 44%，E2B（68.0 kcal）で 47% と，いずれも大きく，絶対スケール情報の有効性はモ

表 3 マクロ栄養素別 MAE 比較（Qwen2-VL-2B，テストセット $n = 346$）

栄養素	単位	Cond1 MAE	Cond4 MAE	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%

デルサイズに依存しないことが確認された．

4.3 実機デモ

図 2 に iOS アプリ CalorieCaMLL のスクリーンショットを示す．Qwen3-VL-4B ベースの Cond4 融合モデルを搭載し，ユーザが iPhone でプレート上の食品を撮影すると，LiDAR 特徴抽出・MLLM 推論が端末内で完結し，約 5 秒でカロリー・三大栄養素が表示される．本デモではこのアプリを通じて参加者が実際に栄養推定を体験できる環境を提供する．



図 2 撮影画面



図 3 カロリー推定結果

5. 考察

5.1 PAVR の有効性と設計の知見

単眼深度推定と同等の精度が，追加の NN 推論を一切要さずに達成されたことについて，オンデバイス展開における実装上の優位性を示している．LiDAR は iPhone ネイティブセンサから付加的な計算コストなしに絶対スケールを取得でき，推論時のメモリ使用量の削減とモデル配布の簡易化に直結する．精度差が MAE で 0.1 kcal（60.0 vs 60.1 kcal）にとどまることは，注入される 4 パラメータ（プレート直径・食品面積・最大高さ・平均高さ）が絶対スケール推定に必要な幾何情報を十分に集約していることを示唆する．

3.3 節で示した food フィールド実験の結果は，PAVR が食品カテゴリの明示的認識に依存せず機能することを示している．食品名の先頭配置（Exp C）が Cond4 の精度低下を招いた（84.5 kcal）ことは，幾何情報と誤った食品認識の干渉を示すものであり，PAVR の価値は食品認識が困難な状況での頑健性にある．

Qwen3.5-4B がビジョン言語特化モデルを上回った結果は，PAVR フレームワーク下では数値テキストを正確に解釈する言語能力が視覚エンコーダ性能と同等以上に重要であることを示唆する．モバイル端末への実装上の制約から，iPhone へのデプロイには現状 Qwen3-VL-4B を選択したが，全モデル中最良精度を示した Qwen3.5-4B の量子化版が実現すれば，より高精度なオンデバイス推論が可能になる可能性がある．

5.2 限界と今後の展望

現状ではプレートを使用した食器のみ PAVR が適用可能であり，Nutrition5k の約 29% を占めるボウル・カップ等の非皿食器には対応できない．また，複数食品が混在する食事を一括で計測することも現状では困難である．今後は（1）iPhone 実撮影データを用いた定量的精度検証，（2）MLLM の常識知識を利用した異常値の事後検証機構（USDA FoodData Central との照合），（3）Gemma4-E4B 等 4B モデルの軽量化と iPhone 上での実運用検証を進める予定である．

6. おわりに

本稿では CalorieCaMLLM を提案した．PAVR フレームワークにより，LiDAR 絶対スケールを MLLM プロンプトへ注入することで，追加の NN 推論なしに単眼深度推定と同等の精度（$R^2 = 0.80$，MAE 約 60 kcal）を達成した．複数モデルの比較では Qwen3.5-4B が最良（MAE 59.6 kcal）であった．4 ビット量子化により iPhone 上でのオンデバイス推論を実現し，プライバシー保護とオンデバイス処理を両立した．

参考文献

[1] H. Fujita and K. Yanai, “Mobile Food Calorie Estimation Using Smartphone LiDAR Sensor,” in *Proc. ACCV Workshop (ACPR)*, 2025.

[2] H. Tanabe and K. Yanai, “CalorieVoL: Volumetric Food Nutritional Calorie Estimation Using Monocular Depth,” in *Proc. MMM*, 2025.

[3] A. Myers et al., “Im2Calories: Towards an Automated Mobile Vision Food Diary,” in *Proc. ICCV*, pp. 1233–1241, 2015.

[4] Q. Thames et al., “Nutrition5k: Towards Automatic

Nutritional Understanding of Generic Food," in *Proc. CVPR*, pp. 8903–8911, 2021.

[5] R. Tanno, T. Ege, and K. Yanai, "AR DeepCalorieCam V2: Food Calorie Estimation with CNN and AR-Based Actual Size Estimation," in *Proc. ACM VRST*, 2018.

[6] Qwen Team, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," *arXiv:2409.12191*, 2024.

[7] Qwen Team, "Qwen3-VL Technical Report," *arXiv:2505.09875*, 2025.

[8] Qwen Team, "Qwen3.5 Technical Report," *arXiv:2509.05695*, 2025.

[9] Gemma Team, Google DeepMind, "Gemma 4 Technical Report," *arXiv:2504.01083*, 2025.

[10] Ultralytics, "Ultralytics YOLO," `https://github.com/ultralytics/ultralytics`, 2026.

[11] Prince Canuma, "mlx-vlm: Vision Language Models for Apple Silicon," `https://github.com/Blaizzy/mlx-vlm`, 2024.

[12] L. Yang et al., "Depth Anything V2," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024.

[13] Y. Yin et al., "FoodLMM: A Versatile Food Assistant Using Large Multi-modal Model," in *Proc. IEEE Int. Workshop Multimedia Signal Processing (MMSP)*, 2025.