マルチモーダルLLM embeddingを用いた 食事画像とレシピテキストのクロスモーダル検索

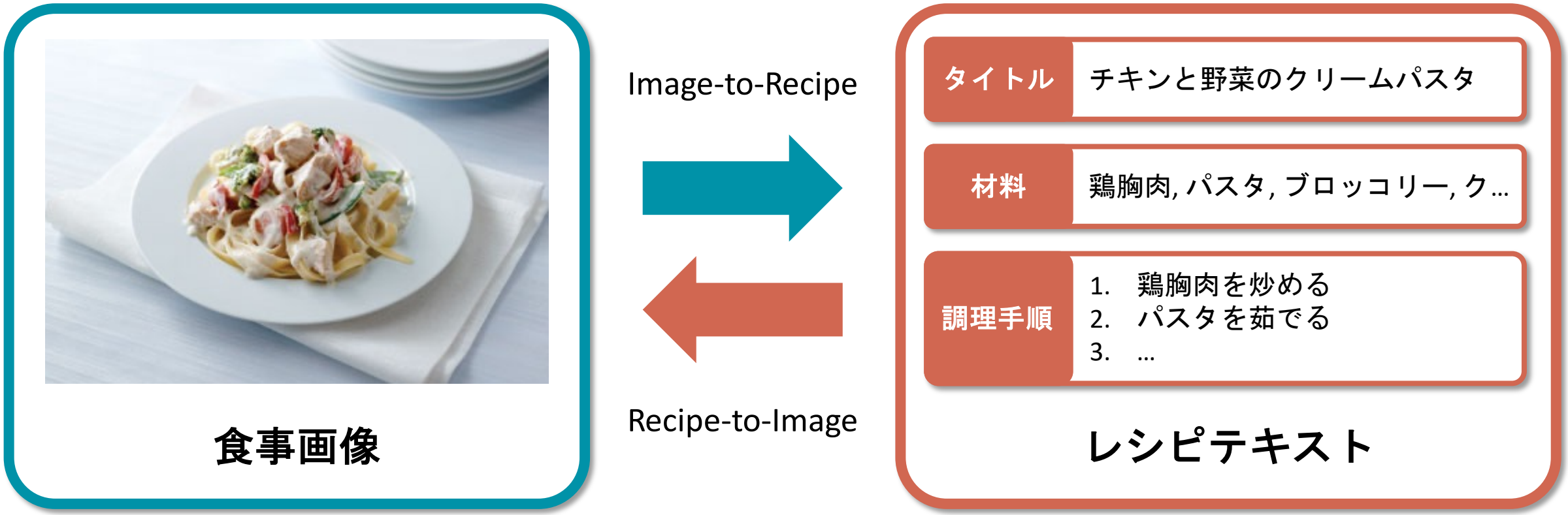
五味 京祐, 柳井 啓司

電気通信大学大学院 情報理工学研究科 情報学専攻



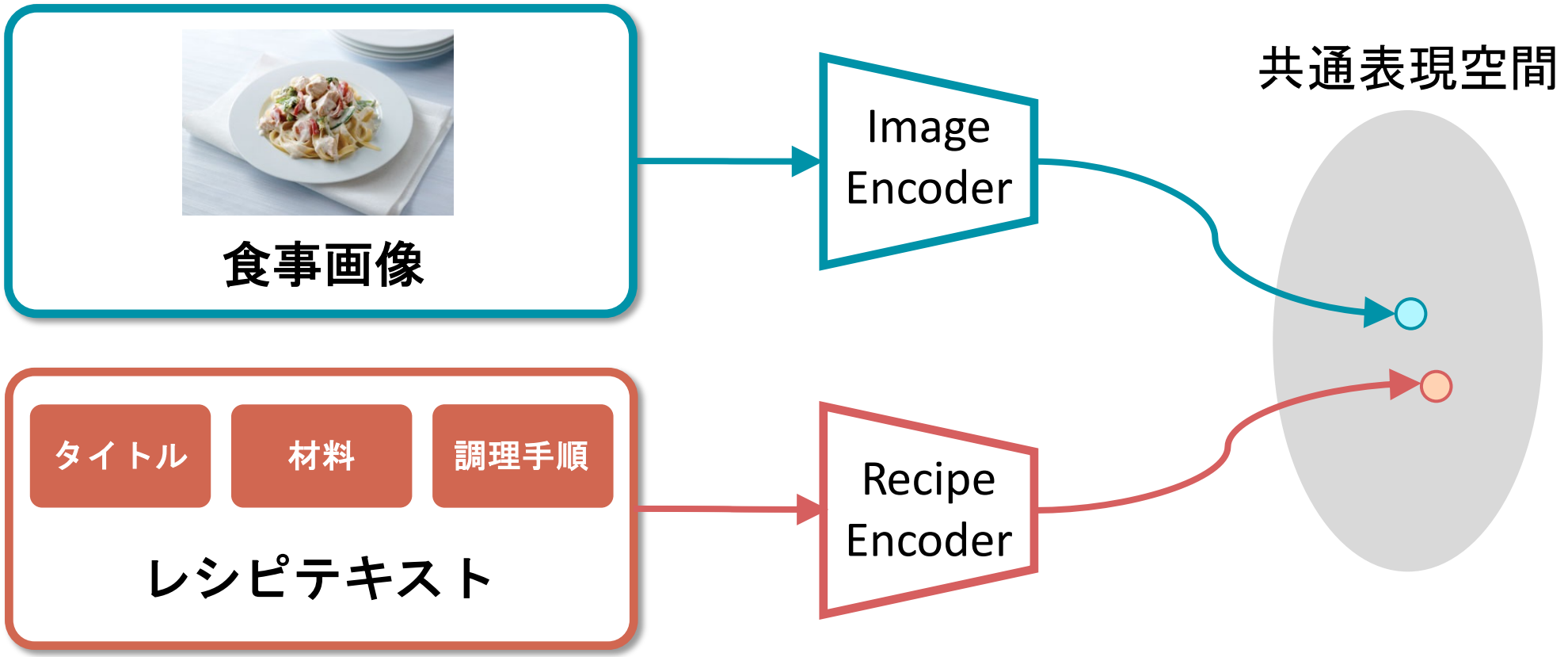
食事画像とレシピテキストのクロスモーダル検索とは？

- 2017年のRecipe1Mデータセットの公開から活発に取り組まれるタスク
- すでに40本以上の論文が発表されている



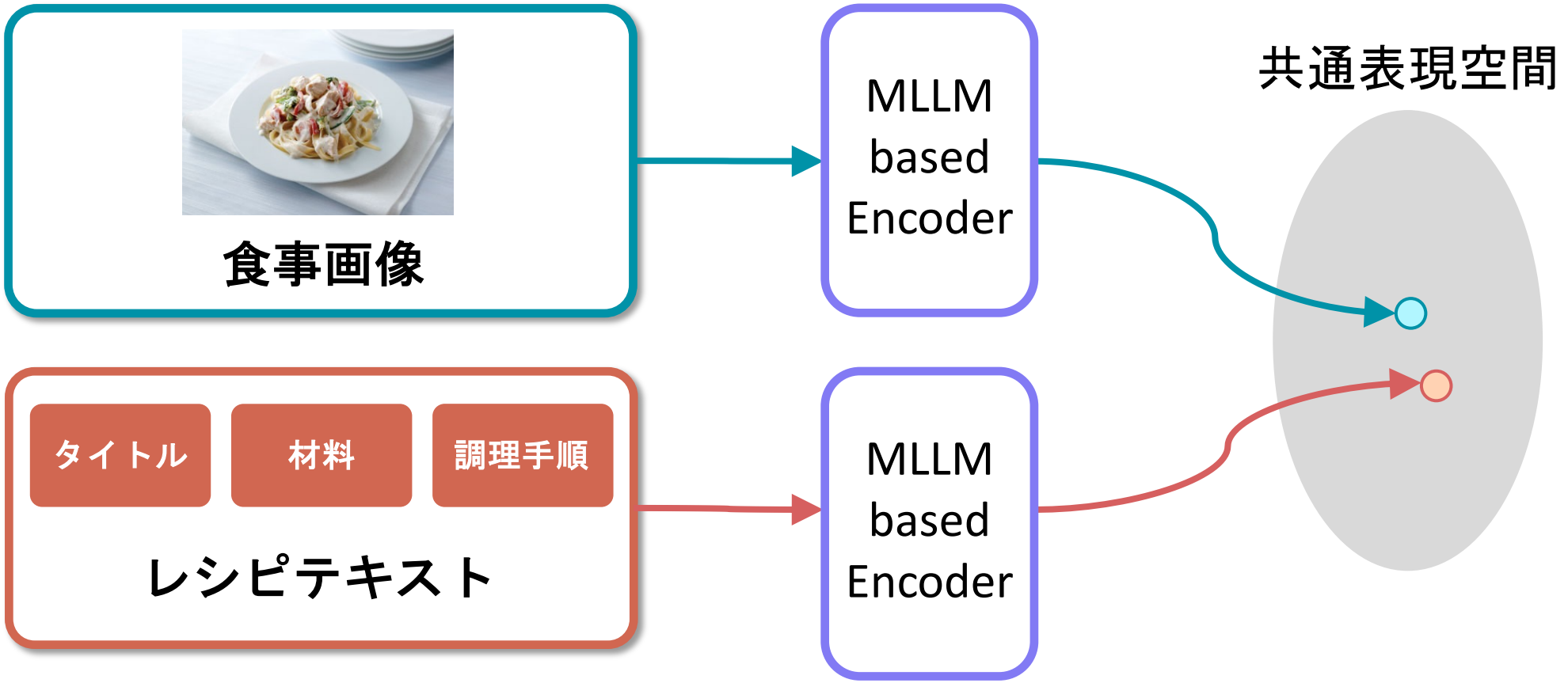
既存のレシピ検索手法

- デュアルエンコーダ型
- 複雑なアライメント学習やタスク固有のネットワーク構造が必要



提案手法

- MLLMベース埋め込みモデルをレシピ・画像共通のエンコーダとして利用
- 複雑なアライメント学習やタスク固有のネットワーク構造が**不要**



既存手法と提案手法の比較

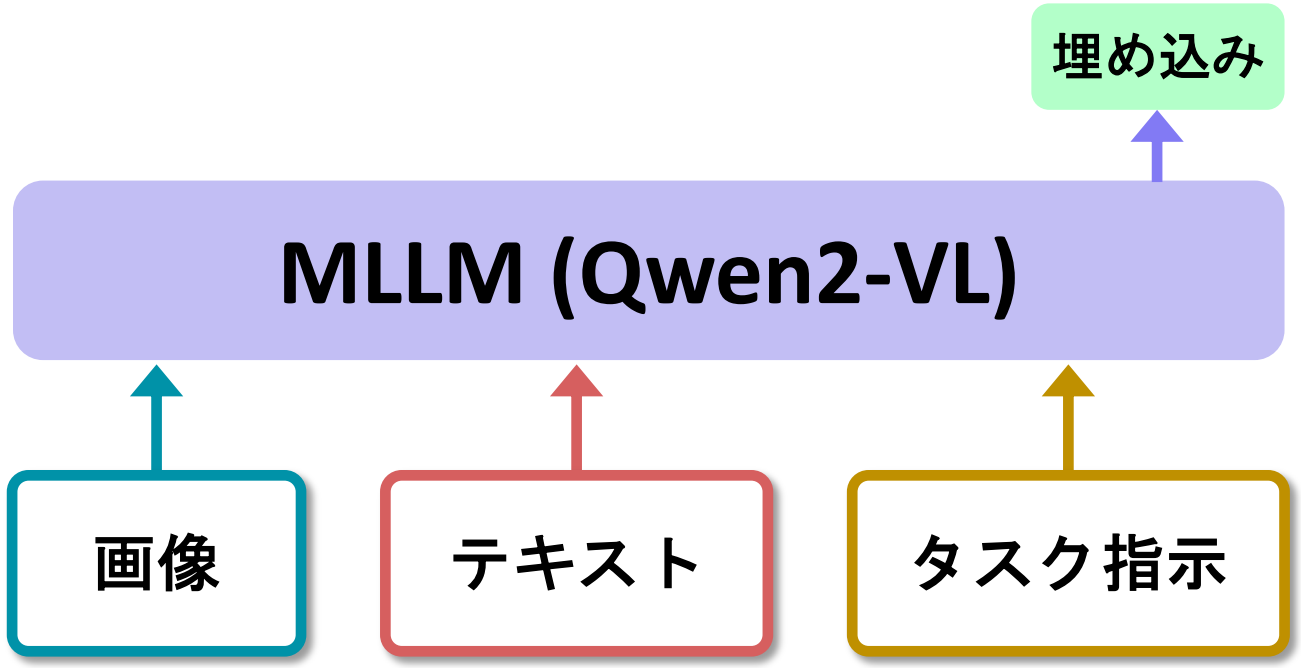
	既存手法	提案手法
エンコーダ構成	2つ （画像用＋レシピ用）	1つ （共通のMLLM）
学習の複雑さ	複雑 （敵対的学習など）	シンプル
ネットワーク構造	タスク固有な設計	汎用的

関連研究



関連研究：VLM2Vec[1], VLM2Vec-V2[2]

MLLM (Qwen2-VL)を埋め込みモデルに変換する手法
エンコードしたいデータとタスク指示を同時に入力可能
最後のトークンの最終層の隠れ状態を埋め込みベクトルとして利用



VLM2Vecシリーズをベースモデルとして採用

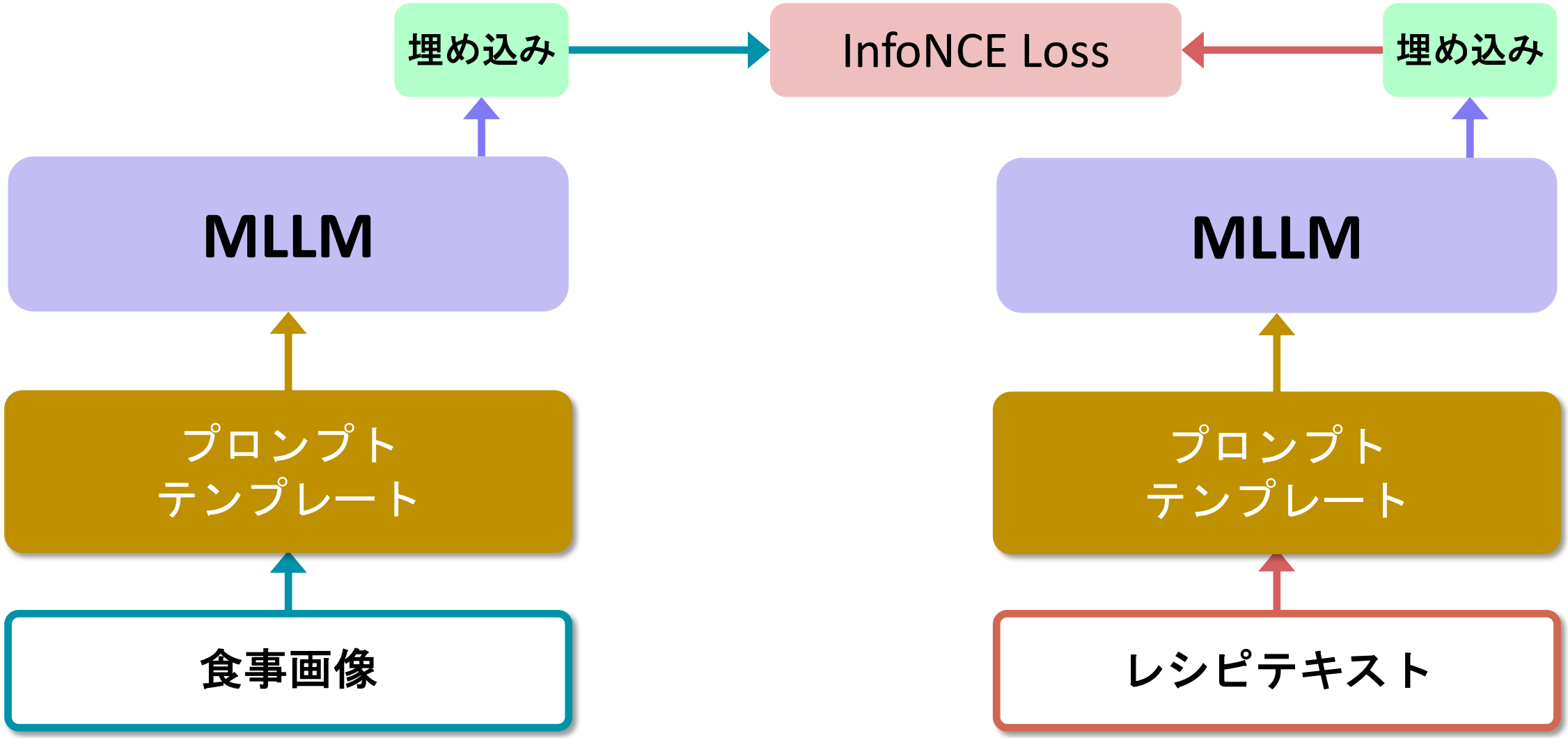
[1] Z. Jiang et al., “VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks,” ICLR, 2025.
[2] R. Meng et al., “VLM2Vec-V2: Advancing Multimodal Embedding for Videos, Images, and Visual Documents,” TMLR, 2026.

提案手法



提案手法

VLM2Vecの手法に従って、対照学習によるファインチューニング



プロンプト設計概要

VLM2Vecは対照学習を検索タスクとして定式化

	クエリ (Query)	検索候補 (Candidate)
画像→レシピ	クエリ画像	検索候補レシピ
レシピ→画像	クエリレシピ	検索候補画像

{ 画像 / レシピ } × { クエリ / 検索候補 } **の4種類のプロンプトが必要**

プロンプト設計 - 画像

クエリ画像

```
<|image_1|>
Find a cooking recipe describing the given food image.
```

レシピ検索を指示

検索候補画像

```
<|image_1|>
Represent the given food image for recipe prediction.
```

レシピを推測できるような表現をする指示

プロンプト設計 - レシピ

レシピ要素を結合し、単一の文字列として入力

- {title}: タイトル
- {ingredients}: カンマ区切りの材料一覧
- {instructions}: スペース区切りの調理手順

クエリレシピ

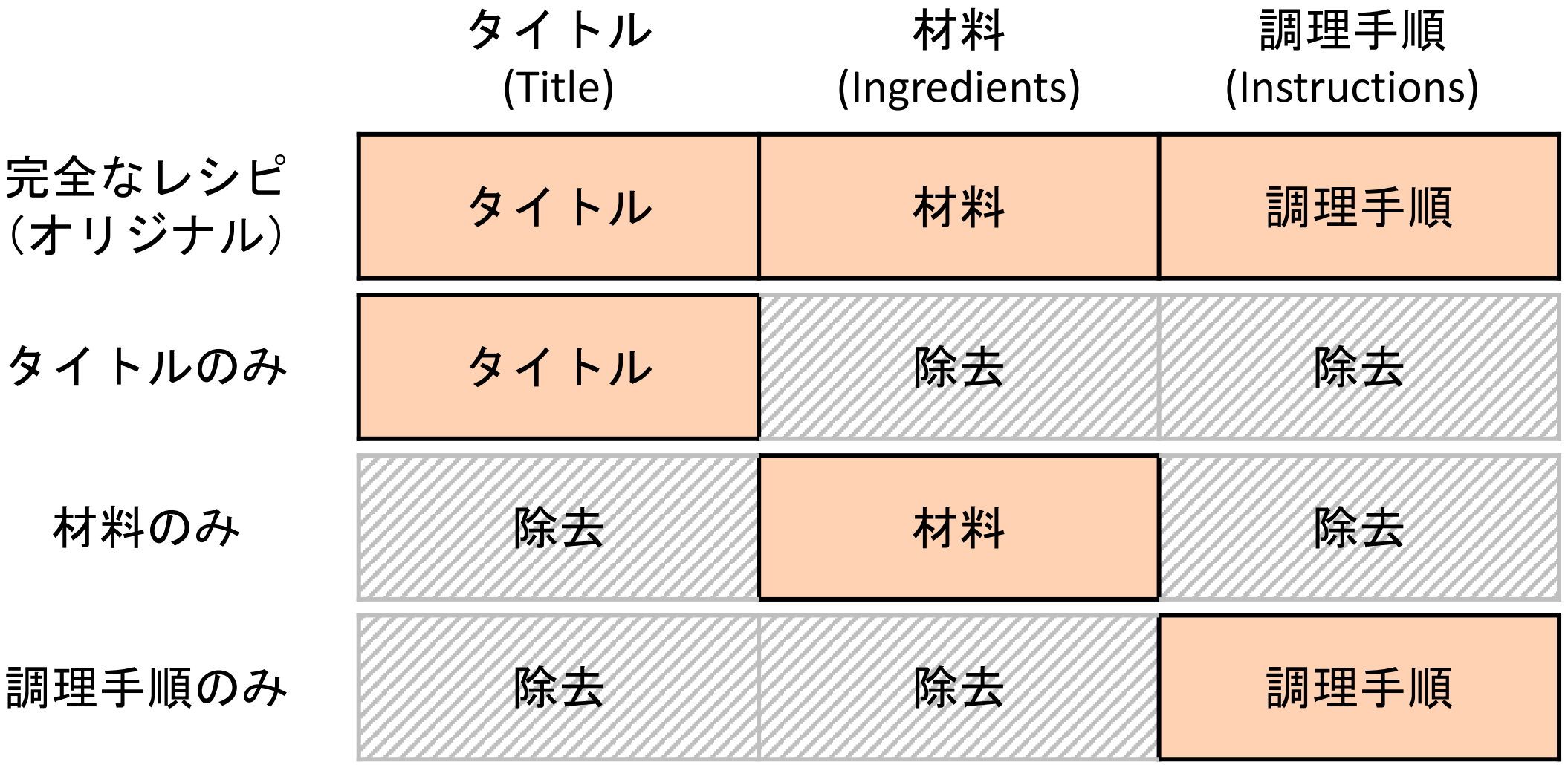
Find me a food image that matches the given cooking recipe:
Title: {title}, Ingredients: {ingredients}, Instructions: {instructions}

検索候補レシピ

A cooking recipe: Title: {title}, Ingredients: {ingredients},
Instructions: {instructions}

レシピ要素の除去によるデータ拡張

現実世界の不完全なレシピデータへの頑健性を向上



実験



実験設定

データセット

Recipe1Mデータセット
- Train: 238,408
- Val: 51,119
- Test: 51,304

評価指標

medR (Median Rank): 正解候補の順位の中央値
Recall@k (k=1, 5, 10): 上位k件の正解率

評価方法

1k設定: 検索候補が1,000件 (低難度)
10k設定: 検索候補が10,000件 (高難度)

ベースモデル

VLM2Vecシリーズの3つのモデルをベースモデルとして採用

モデル	パラメータ数	埋め込み次元数	特徴
V1-2B	2B	1536	小モデル
V1-7B	7B	3584	大モデル
V2	2B	1536	V1の改善版

既存手法との比較

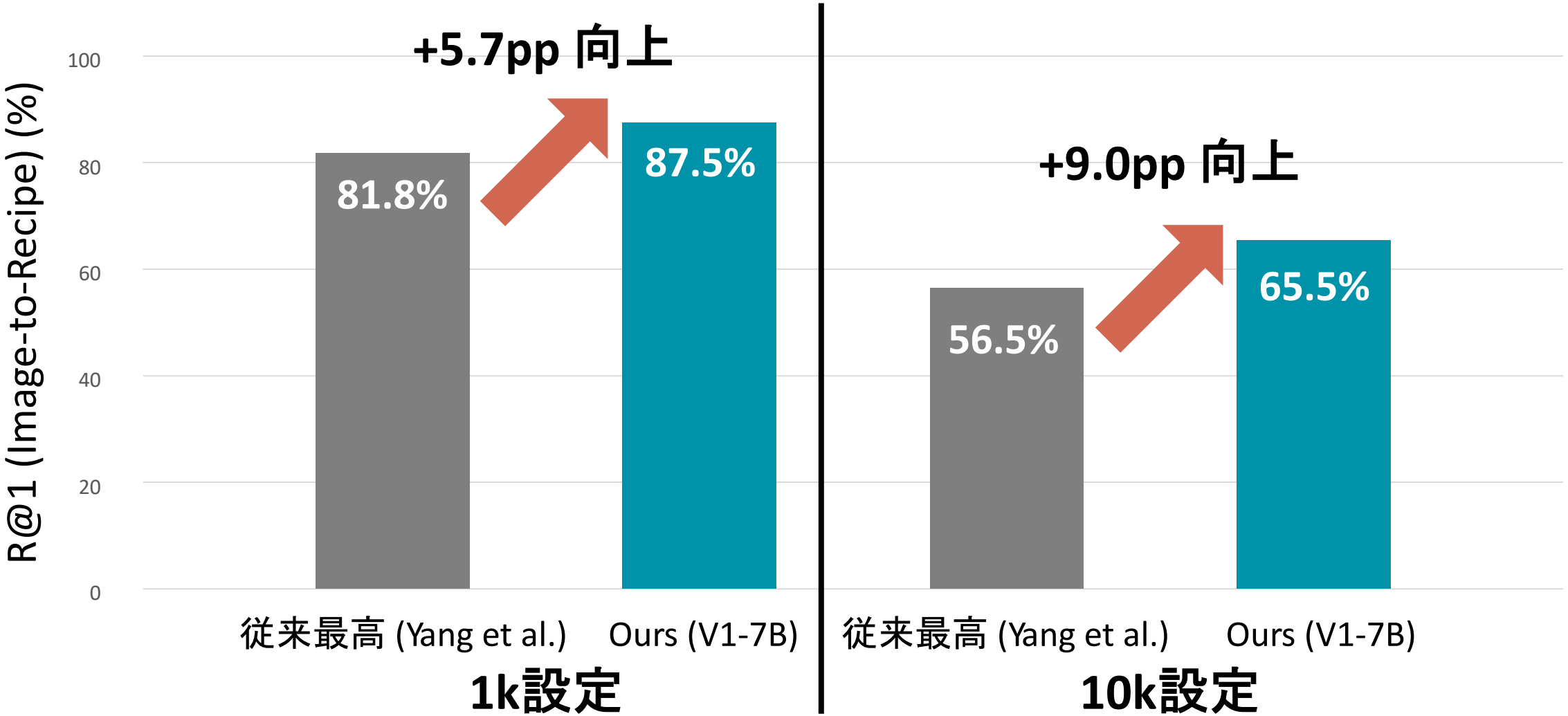
全評価指標でOurs (V1-7B)が最高性能を達成

各実験設定での R@1↑ の値

Method	Image-to-Recipe		Recipe-to-Image	
	1k	10k	1k	10k
Salvador et al. (2017)	24.0	-	25.0	-
H-T (2021)	60.0	27.9	60.3	28.3
T-Food (2022)	72.3	43.4	72.6	44.6
Yang et al. (2024)	81.8	56.5	81.2	55.7
DAR (2024)	77.3	47.8	77.1	47.4
Wang et al. (2025)	79.1	51.7	78.3	52.2
Ours (V1-2B)	84.1	59.7	81.5	55.8
Ours (V1-7B)	**87.5**	**65.5**	**85.1**	**61.5**
Ours (V2)	83.8	59.1	81.7	55.7

既存手法との比較

Ours (V1-7B)がSOTAを大幅に更新



定性的比較：Image-to-Recipe Retrieval

クエリ画像



正解レシピ

タイトル	easy okara burger
材料	• 1/2 carrot • 1 onion (large) • 200 grams ground pork • 100 grams fresh okara • 1 egg …
調理手順	1. mince the carrot and onion. 2. saute both until the onions are tender, and then season with salt and pepper. 3. when finished, turn off the heat and let cool. 4. combine the egg, panko, and milk in a bowl. …

DAR[1]

	Top-1	Top-2	Top-3
タイトル	mild teriyaki ham burger steak	teriyaki chicken burgers with a teriyaki mushroom sauce	teriyaki hamburger steak
材料	• 4 tbsp water • 1 tbsp soy sauce • 1 tbsp mirin • 1 heaping teaspoon sugar …	• 1 lb ground chicken breast • 1 medium onion, finely chopped • 3 garlic cloves, minced …	• 200 grams ground meat • 1 egg • 1 onion • 1 panko …
調理手順	1. mix together all of the ingredients. 2. when the hamburgers are grilled, blot off excess oil, …	1. heat oil in skillet. 2. mix ground chicken with the next 10 ingredients. 3. if mixture feels too moist …	1. chop the onion finely, and sweat in a frying pan. 2. let the onion cool, and mix with the ingredients. …

提案手法

	Top-1	Top-2	Top-3
タイトル	chicken meatloaf burgers for bentos	easy okara burger	all-purpose balsamic sauce with mushrooms
材料	• 300 grams ground chicken • 1/2 japanese leek • 1 medium egg • 3 tbsp ground yamaimo …	• 1/2 carrot • 1 onion (large) • 200 grams ground pork • 100 grams fresh okara …	• 2 tbsp balsamic vinegar • 2 tbsp red wine • 1 tbsp soy sauce • 1 tsp honey …
調理手順	1. ground the ginger and the yamaimo. 2. mince the leek. 3. shred the shimeji …	1. mince the carrot and onion. 2. saute both until the onions are tender, and then season with salt and pepper. …	1. mix all the ingredients together. 2. cut the mushrooms in half, then into 3mm thick …

[1] F. Song et al., “Enhancing Recipe Retrieval with Foundation Models: A Data Augmentation Perspective,” ECCV, 2024.

定性的比較：Recipe-to-Image Retrieval

クエリレシピ

タイトル	nutmeg-maple cream pie
材料	• 3/4 cup maple syrup • 2 1/4 cups heavy cream • 4 egg yolks • 1 whole egg …
調理手順	1. preheat oven to 300 degrees. 2. in a medium saucepan over medium-high heat, reduce maple syrup by a quarter, 5 to 7 minutes. 3. stir in cream and bring to a simmer. 4. remove from heat. …

正解画像



DAR[1]



Top-1



Top-2



Top-3

提案手法



Top-1



Top-2



Top-3

[1] F. Song et al., “Enhancing Recipe Retrieval with Foundation Models: A Data Augmentation Perspective,” ECCV, 2024.

おわりに

まとめ

- 複雑なアライメントの学習やタスク固有のネットワークが不要で高精度なレシピ検索手法を提案
- 従来のデュアルエンコーダ型のアプローチを廃止し、MLLMベース埋め込みモデルを共通のエンコーダとして利用
- データ拡張によって不完全レシピへの頑健性を向上

今後の展望

- 画像とレシピの同時入力による検索
- ユーザーの指示に基づいた検索