MLLM 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

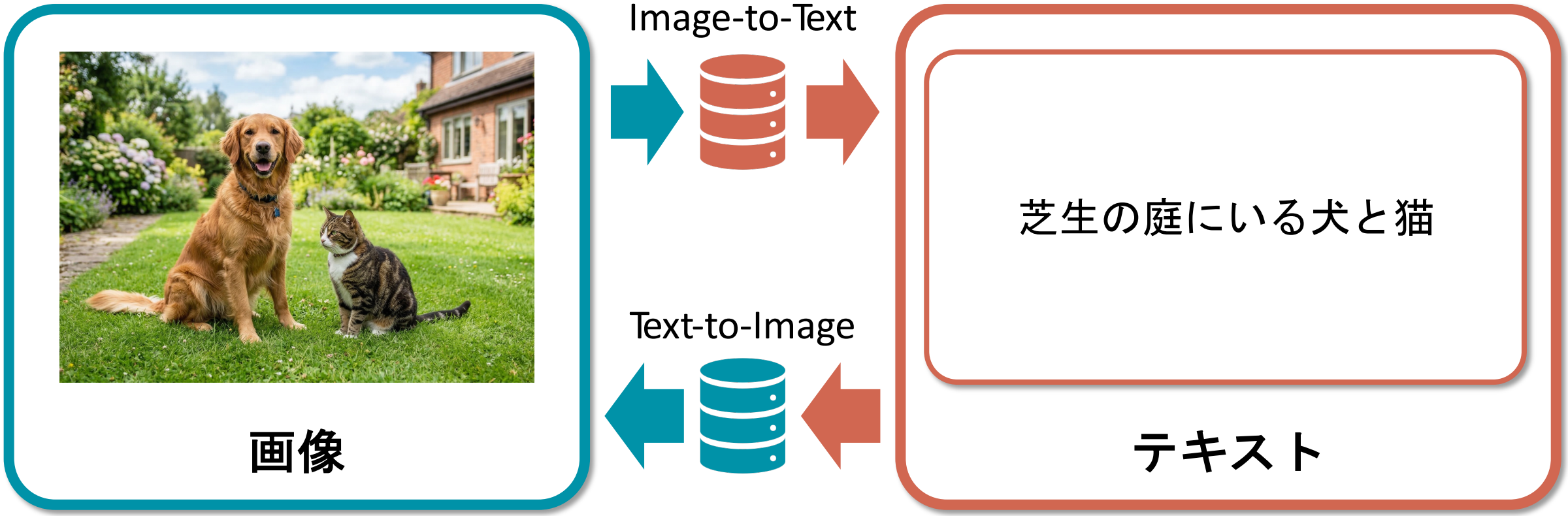
五味 京祐, 柳井 啓司

電気通信大学大学院 情報理工学研究科 情報学専攻

クロスモーダル検索とは？

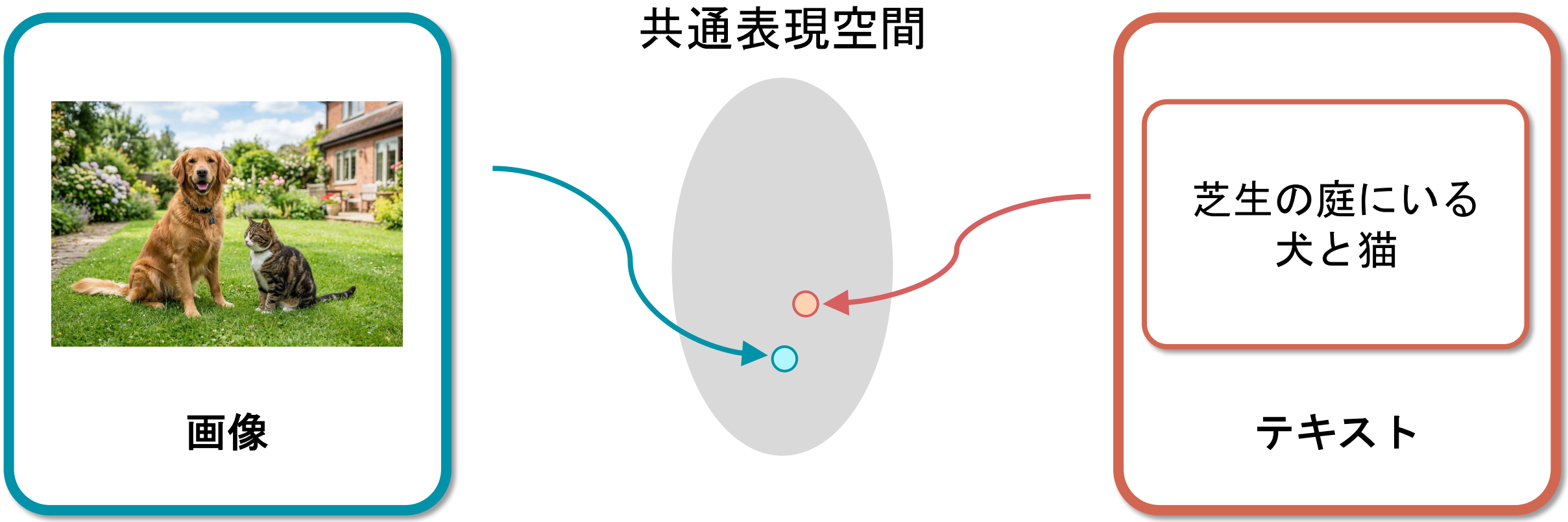
モダリティをまたいだデータの検索

画像とテキストの例



クロスモーダル検索の実現方法

異なるモダリティのデータを共通の表現空間に埋め込む (embedding)



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

食事画像

Image-to-Recipe

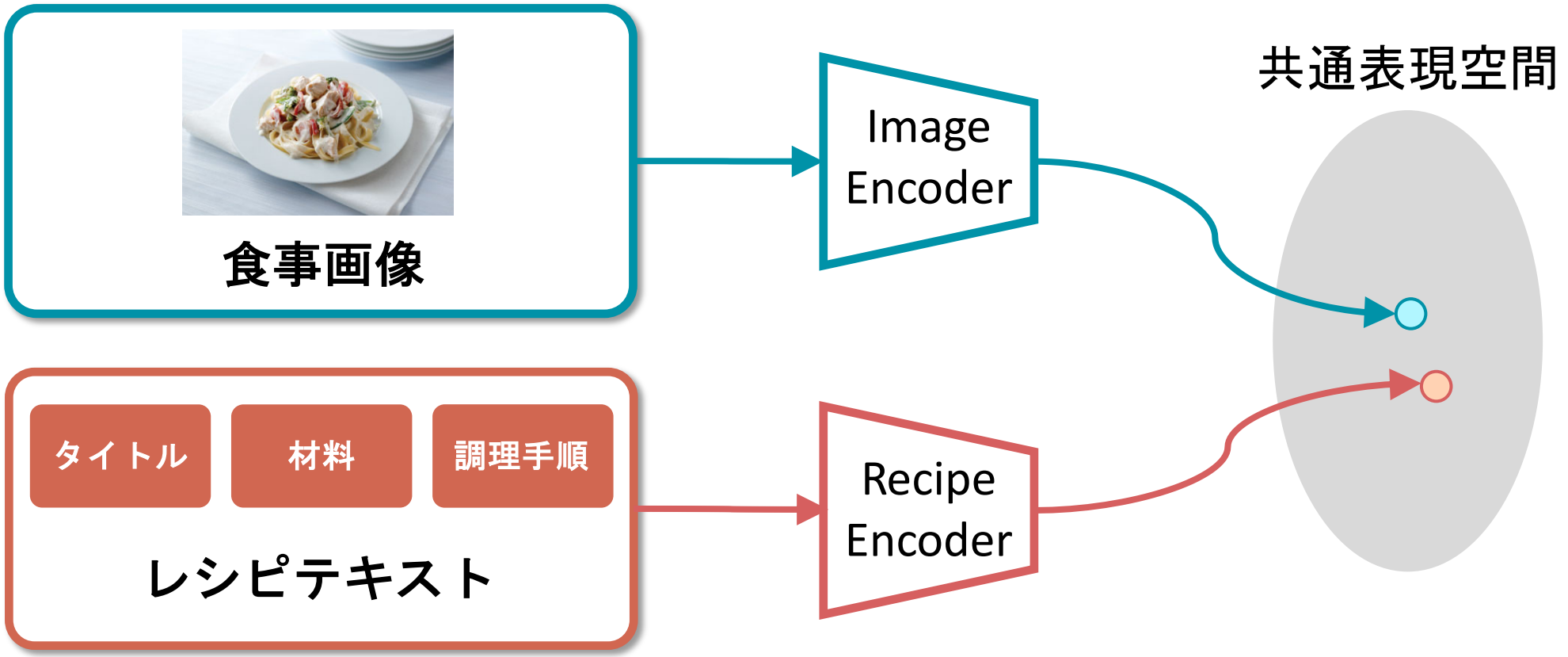
Recipe-to-Image

タイトル	チキンと野菜のクリームパスタ
材料	鶏胸肉, パスタ, ブロッコリー, ク…
調理手順	1. 鶏胸肉を炒める 2. パスタを茹でる 3. …

レシピテキスト

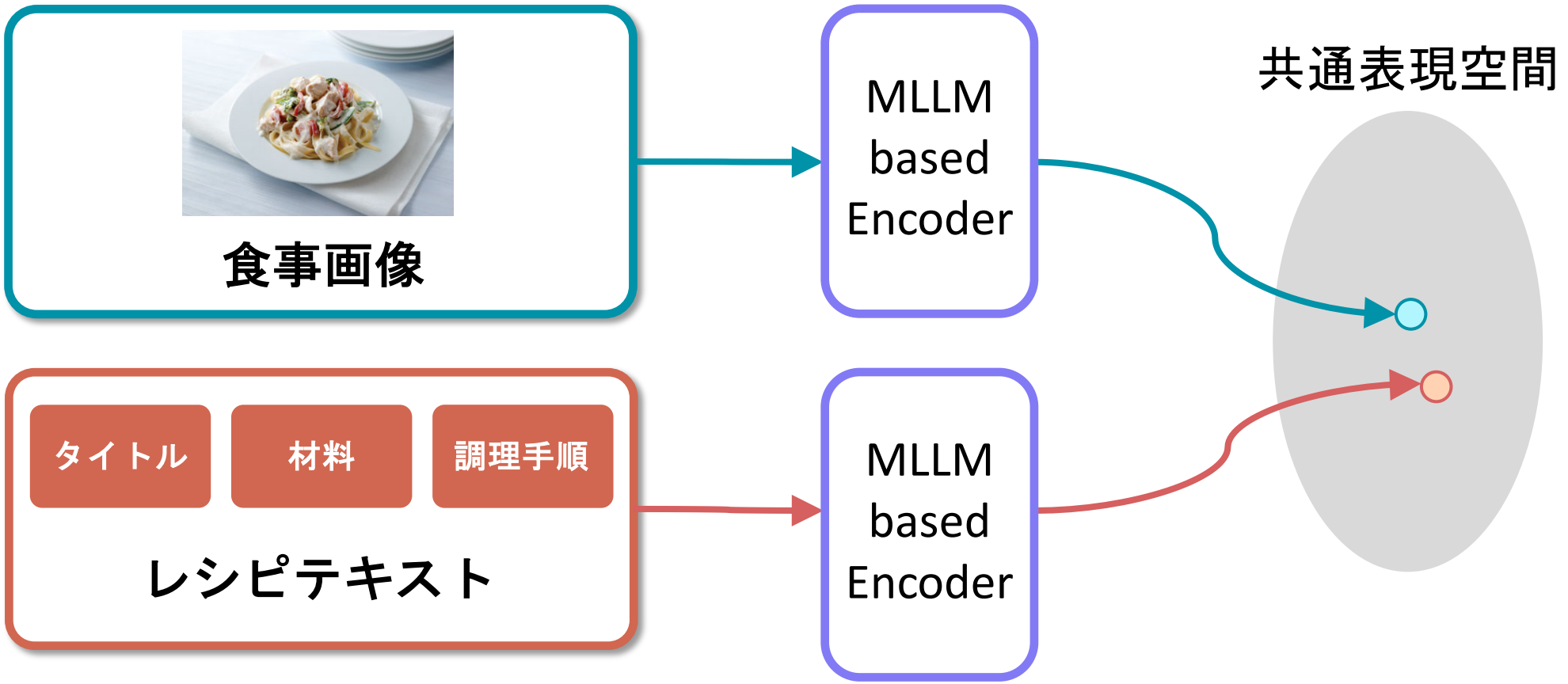
既存のレシピ検索手法

- 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：画像とテキストのクロスモーダル埋め込み

従来

デュアルエンコーダ型
例: CLIP, ALIGN, BLIP, SigLIP

モダリティ間の相互作用が少ない
関係性を捉える能力に限界

- Multimodal Large Language Model (MLLM)の発展
- MLLMの埋め込みモデルへの転用

最近

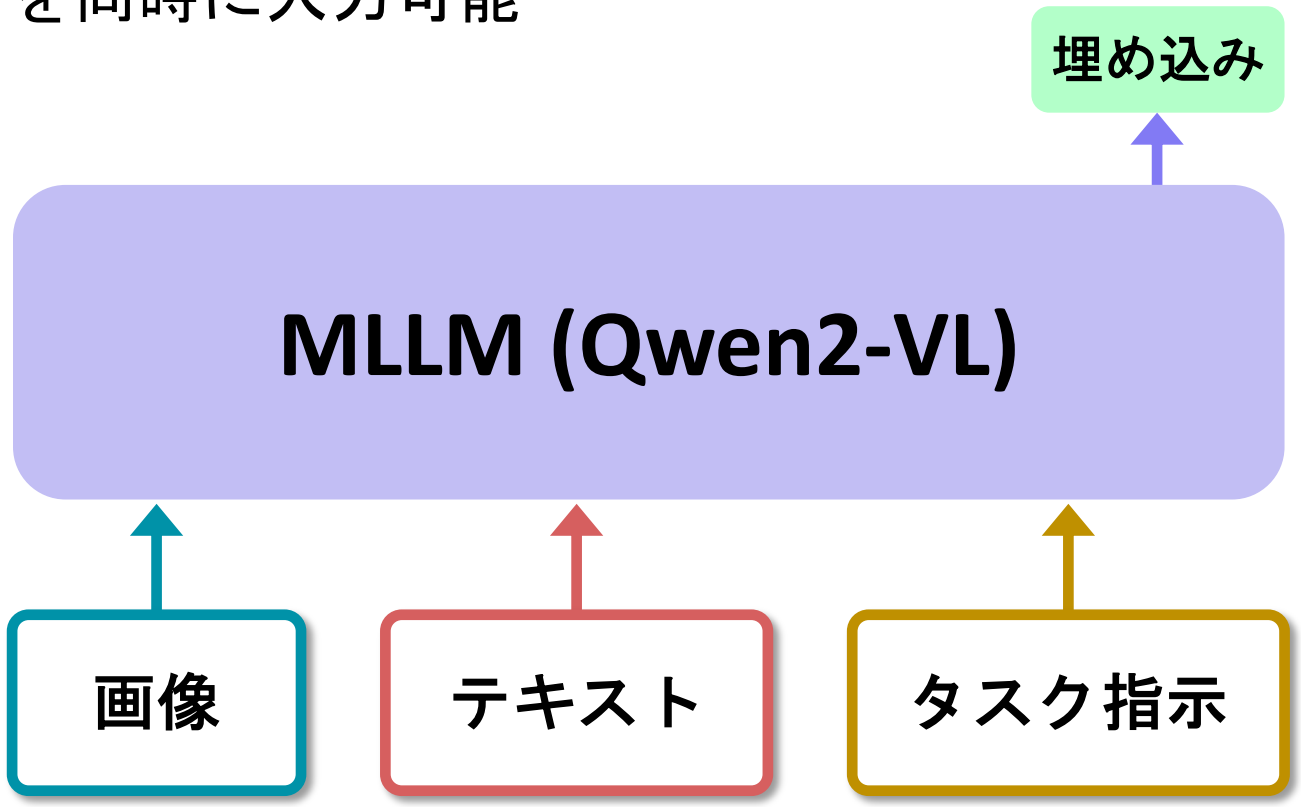
MLLMベース
例: VLM2Vec, MM-Embed

MLLMの事前知識を利用可能
指示テキストを入力可能
より複雑なタスクに強い

本研究もこの流れに沿って、MLLMベース埋め込みモデルをレシピに応用

関連研究：VLM2Vec, VLM2Vec-V2

Qwen2-VLを埋め込みモデルとしてファインチューニング
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採用
タスク指示とデータを同時に入力可能



VLM2Vecシリーズをベースモデルとして採用

関連研究：レシピ検索

複雑な学習手法

- 敵対的学習
  - 例: ACME, R²GAN
- KLDを用いた意味的一貫性ロス
  - 例: SCAN

タスク固有のネットワーク

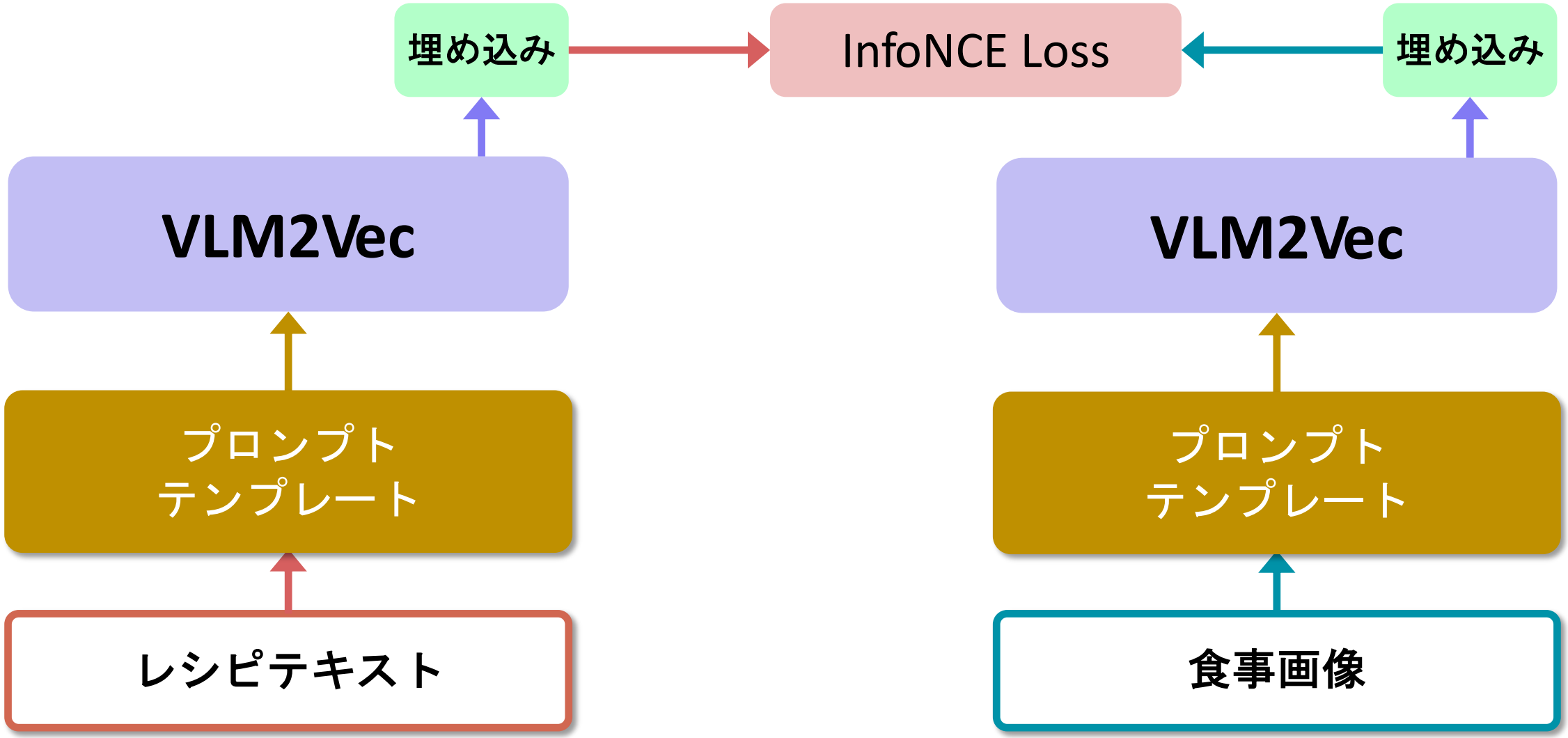
- 木構造LSTM
  - 例: CHEF
- 階層型Transformer
  - 例: H-T, T-Food

複雑な学習手法
タスク固有のネットワーク
} **が不要な手法を提案**

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抽出

プロンプト設計 - レシピ

クエリレシピ

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

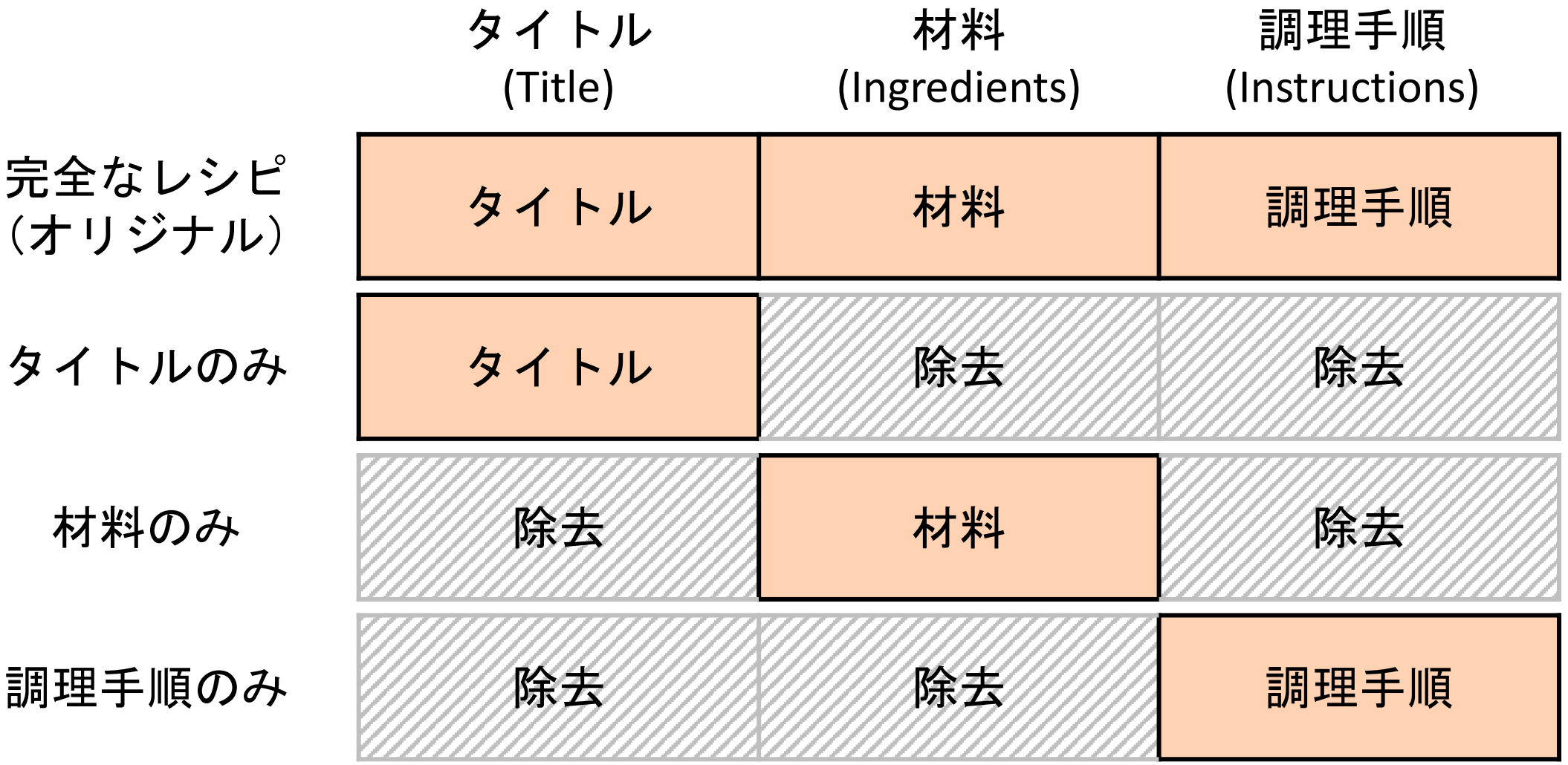
検索候補レシピ

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

{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

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

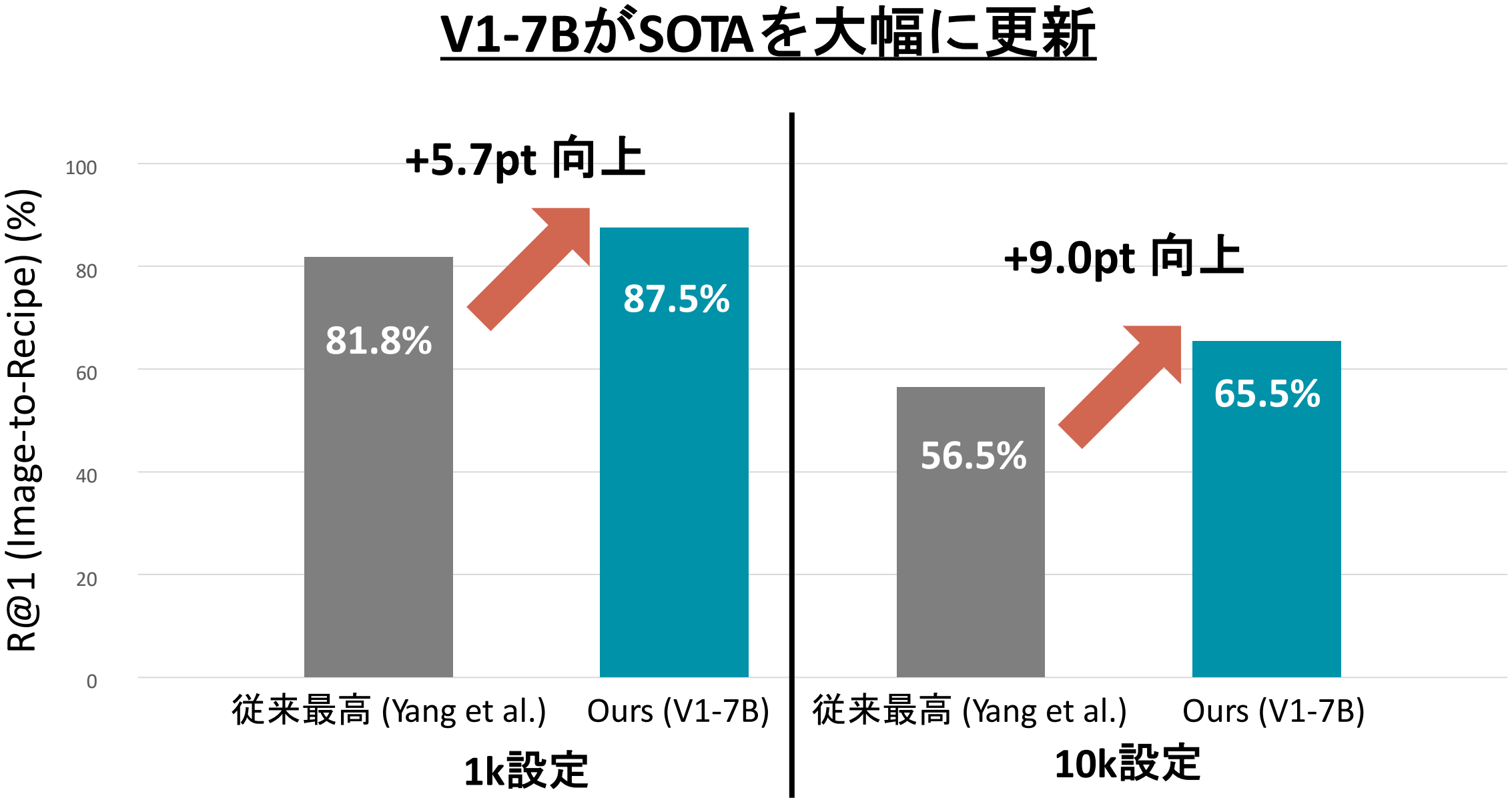
モデル	パラメータ数	埋め込み次元数	特徴
VLM2Vec-V1-2B	2B	1536	V1の小モデル
VLM2Vec-V1-7B	7B	3584	V1の大モデル
VLM2Vec-V2	2B	1536	V1の改善版（動画・書類画像へ対応）

既存手法との比較

- 提案手法の3つのモデルすべてが
  全評価指標で既存手法と同等以上の性能
- Ours (V1-7B)が最高性能を達成

Method	1k								10k							
	Image-to-Recipe				Recipe-to-Image				Image-to-Recipe				Recipe-to-Image			
	medR↓	R@1↑	R@5↑	R@10↑	medR↓	R@1↑	R@5↑	R@10↑	medR↓	R@1↑	R@5↑	R@10↑	medR↓	R@1↑	R@5↑	R@10↑
Salvador et al. (2017)	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0	41.9	-	-	-	39.2	-	-	-
H-T (2021)	1.0	60.0	87.6	92.9	1.0	60.3	87.6	93.2	4.0	27.9	56.4	68.1	4.0	28.3	56.5	68.1
T-Food (2022)	1.0	72.3	90.7	93.4	1.0	72.6	90.6	93.4	2.0	43.4	70.7	79.7	2.0	44.6	71.2	79.7
Yang et al. (2024)	1.0	81.8	95.9	97.8	1.0	81.2	96.0	97.9	1.0	56.5	81.0	87.6	1.0	55.7	80.2	87.1
FARM (2024)	1.0	73.7	90.7	93.4	1.0	73.6	90.8	93.5	2.0	44.9	71.8	80.0	2.0	44.3	71.5	80.0
DAR (2024)	1.0	77.3	95.3	97.7	1.0	77.1	95.4	97.9	2.0	47.8	75.9	84.3	2.0	47.4	75.5	84.1
Wang et al. (2025)	1.0	79.1	94.6	97.0	1.0	78.3	95.0	97.2	1.0	51.7	78.2	85.9	1.0	52.2	78.4	86.0
Ours (V1-2B)	1.0	84.1	97.3	98.8	1.0	81.5	96.6	98.4	1.0	59.7	83.5	89.9	1.0	55.8	81.1	88.0
Ours (V1-7B)	1.0	**87.5**	**98.0**	**99.2**	1.0	**85.1**	**97.6**	**99.1**	1.0	**65.5**	**87.4**	**92.5**	1.0	**61.5**	**85.0**	**91.0**
Ours (V2)	1.0	83.8	96.9	98.7	1.0	81.7	96.5	98.3	1.0	59.1	83.3	89.8	1.0	55.7	81.0	88.0

既存手法との比較

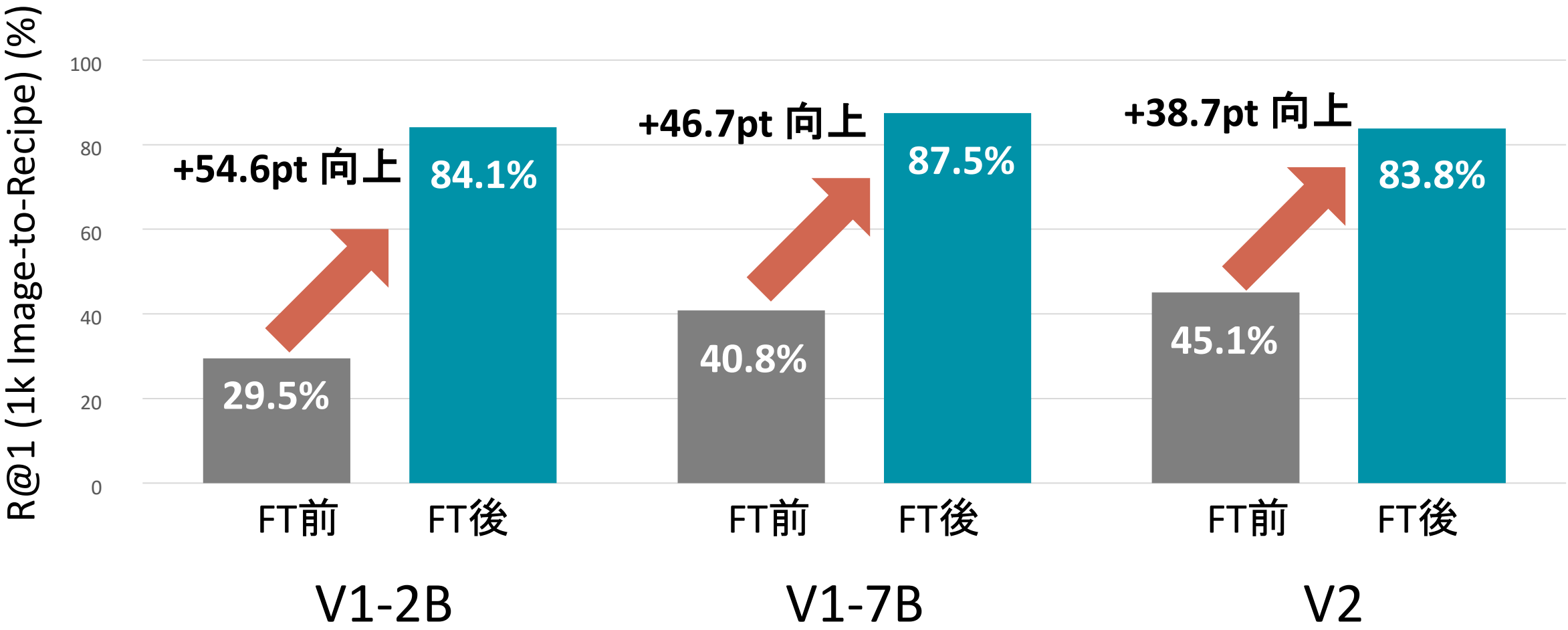

V1-7BがSOTAを大幅に更新
R@1 (Image-to-Recipe) (%)
100
80
60
40
20
0
+5.7pt 向上
81.8%
87.5%
従来最高 (Yang et al.)
Ours (V1-7B)
1k設定
+9.0pt 向上
56.5%
65.5%
従来最高 (Yang et al.)
Ours (V1-7B)
10k設定

ファインチューニング前後の比較

ファインチューニングによる大幅な性能向上を確認

Method	1k							
	Image-to-Recipe				Recipe-to-Image			
	medR↓	R@1↑	R@5↑	R@10↑	medR↓	R@1↑	R@5↑	R@10↑
V1-2B (Zero-shot)	4.65	29.5	51.9	61.2	7.7	18.4	42.8	57.9
Ours (V1-2B)	1.0	84.1	97.3	98.8	1.0	81.5	96.6	98.4
V1-7B (Zero-shot)	2.1	40.8	66.5	76.0	2.0	39.8	69.0	79.0
Ours (V1-7B)	1.0	87.5	98.0	99.2	1.0	85.1	97.6	99.1
V2 (Zero-shot)	2.0	45.1	73.4	82.4	2.0	47.3	74.7	83.3
Ours (V2)	1.0	83.8	96.9	98.7	1.0	81.7	96.5	98.3

ファインチューニング前後の比較



VLM2Vecのような高性能で汎用的な手法でも
レシピ検索ではタスク特化のファインチューニングが不可欠

不完全なレシピデータでの結果

2つ観点で分析

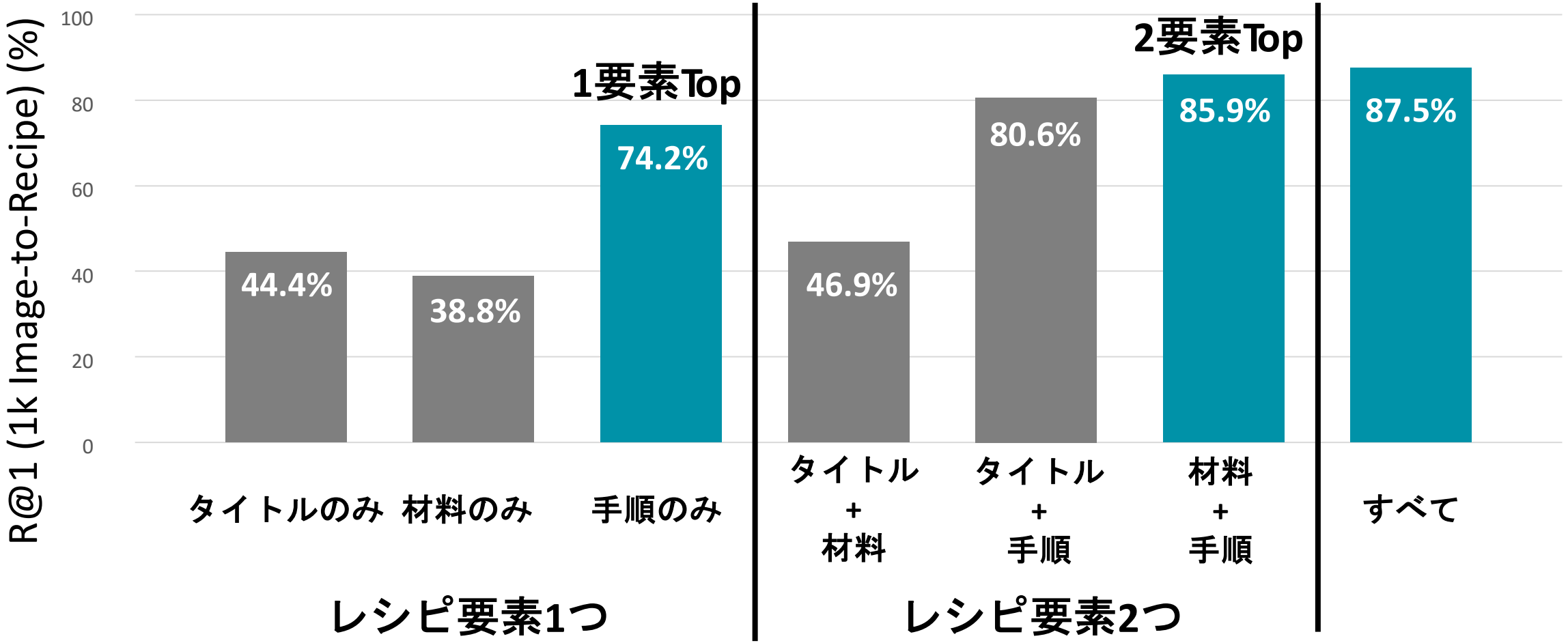
1. レシピ要素の組み合わせによる性能の違い
2. データ拡張の効果

推論で使用した要素			Method	1k								10k							
				Image-to-Recipe				Recipe-to-Image				Image-to-Recipe				Recipe-to-Image			
タイトル	材料	手順		medR↓	R@1↑	R@5↑	R@10↑	medR↓	R@1↑	R@5↑	R@10↑	medR↓	R@1↑	R@5↑	R@10↑	medR↓	R@1↑	R@5↑	R@10↑
✓			V1-7B w/o Aug.	2.0	40.4	71.7	81.7	2.0	40.0	70.9	80.9	13.1	13.6	34.0	45.9	13.5	14.9	34.7	45.7
			V1-7B	2.0	44.4	75.6	84.5	2.0	41.5	72.8	82.6	10.0	16.1	38.9	51.0	12.3	15.2	36.0	47.4
	✓		V1-7B w/o Aug.	3.0	34.9	62.8	74.8	1.9	47.4	75.0	83.5	21.1	15.1	30.1	39.2	8.9	22.4	42.6	52.5
			V1-7B	2.3	38.8	68.0	79.7	1.2	52.4	78.9	86.6	15.5	17.2	33.8	43.7	6.2	25.8	47.8	57.9
		✓	V1-7B w/o Aug.	1.0	74.0	91.6	95.0	1.0	70.8	90.3	94.2	2.0	48.9	72.3	79.9	2.0	44.0	68.8	77.1
			V1-7B	1.0	74.2	91.6	95.3	1.0	70.9	90.0	94.3	2.0	48.9	72.5	80.1	2.0	43.7	68.4	77.0
✓	✓		V1-7B w/o Aug.	2.0	44.8	74.5	84.4	1.0	58.3	83.9	90.3	10.0	19.4	39.5	50.6	4.1	30.1	54.1	64.0
			V1-7B	2.0	46.9	78.2	87.1	1.0	62.3	87.1	92.4	8.8	21.4	41.9	53.1	3.1	34.0	58.9	68.9
✓		✓	V1-7B w/o Aug.	1.0	80.2	95.7	97.9	1.0	77.5	94.8	97.6	1.0	55.1	79.0	86.2	1.3	50.2	75.8	83.8
			V1-7B	1.0	80.6	95.6	98.0	1.0	77.8	95.1	97.6	1.0	54.8	79.3	86.5	1.3	50.4	76.0	84.2
	✓	✓	V1-7B w/o Aug.	1.0	85.9	97.5	98.8	1.0	83.6	97.0	98.6	1.0	63.2	85.6	91.2	1.0	59.4	83.0	89.4
			V1-7B	1.0	85.9	97.4	98.8	1.0	83.2	96.9	98.5	1.0	63.2	85.7	91.2	1.0	59.1	82.9	89.3
✓	✓	✓	V1-7B w/o Aug.	1.0	87.6	98.2	99.1	1.0	85.3	97.7	99.1	1.0	65.6	87.3	92.6	1.0	61.8	84.9	90.9
			V1-7B	1.0	87.5	98.0	99.2	1.0	85.1	97.6	99.1	1.0	65.5	87.4	92.5	1.0	61.5	85.0	91.0

不完全なレシピデータでの結果

レシピ要素の組み合わせによる性能の違い

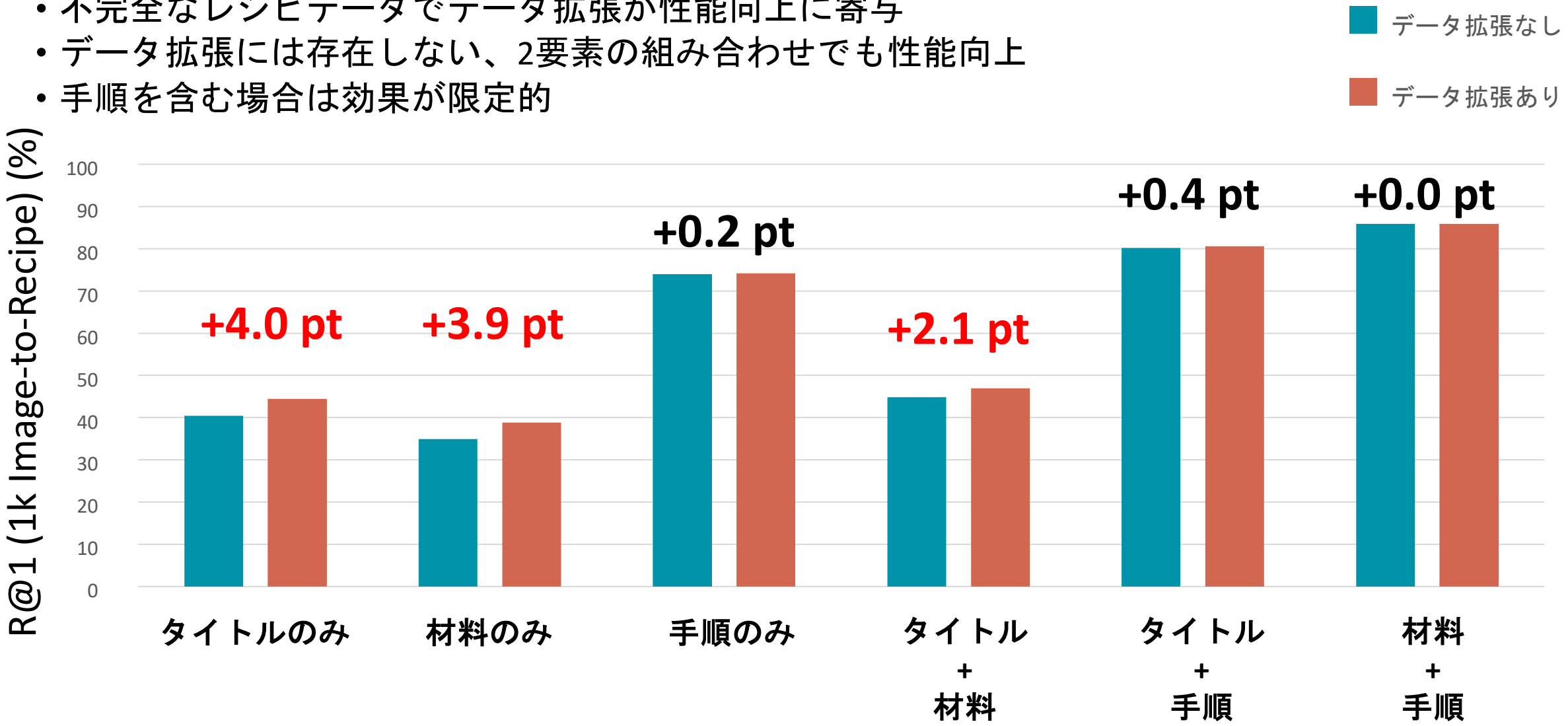
- レシピ要素１つ：手順のみが最も高い性能
- レシピ要素２つ：材料と手順の組み合わせが最も高い性能



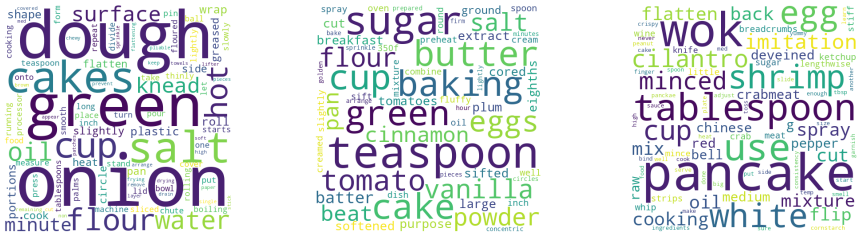
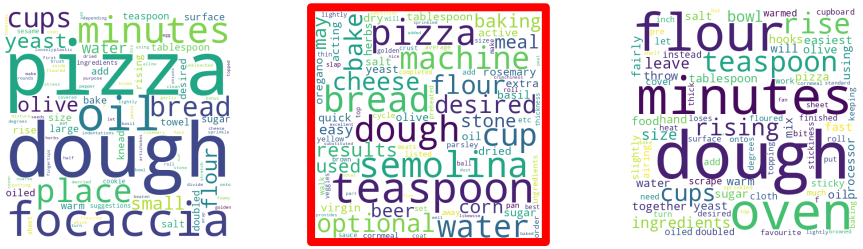
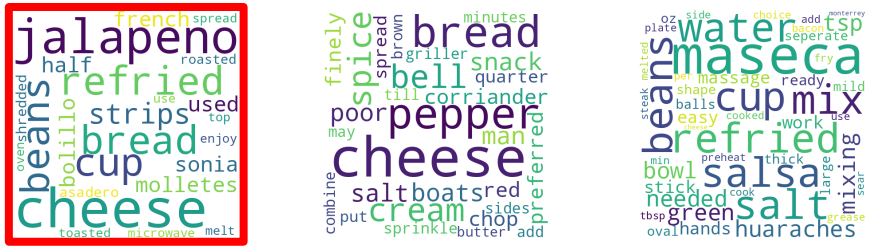
不完全なレシピデータでの結果

データ拡張の効果の分析

- 不完全なレシピデータでデータ拡張が性能向上に寄与
- データ拡張には存在しない、2要素の組み合わせでも性能向上
- 手順を含む場合は効果が限定的



定性的比較：Image-to-Recipe Retrieval

クエリ画像	Ground Truth	Top3	
	**easy pizza dough for bread machine** 1 cup water 1 tablespoon extra virgin olive oil 3/4 teaspoon salt … add ingredients to pan in order li… after the cycle is completed, ligh… top with desired sauce, cheeses, … …	DAR	
		提案手法	
	**sonia's molletes** 2 french bread or bolillo in half 1 cup refried beans 1 cup cheese i used asadero shre… 1 jalapeno roasted in strips i toasted the bread and spread r… you can melt the cheese in oven… also you can use jalapenos from… enjoy ?	DAR	
		提案手法	

定性的比較：Recipe-to-Image Retrieval

クエリレシピ	Ground Truth	Top3		
nutmeg-maple cream pie 3/4 cup maple syrup 2 1/4 cups heavy cream 4 egg yolks 1 whole egg … preheat oven to 300 degrees. in a medium saucepan over med… stir in cream and bring to a simmer. remove from heat. …		DAR			
		提案手法			
tropical baked bananas 2 bananas 1 lime, juiced 1 tablespoon dark rum 1/4 teaspoon ground cinnamon … preheat the oven to 400f peel the bana… lay the pieces into a buttered 9x13x2-inch… pour the lime juice and rum evenly over t… sprinkle on the cinnamon, nutmeg, and b... …		DAR			
		提案手法			

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