一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

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　〒182–8585 東京都調布市調布ケ丘一丁目 5 番地 1
E-mail: †gomi-k@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　インターネット上のレシピデータの増加に伴い、食事画像とレシピテキスト間のクロスモーダル検索の実現が求められている。本研究では、MLLM ベースの Multimodal Embedding モデルをレシピデータでファインチューニングすることで、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構造を必要とせず、高精度で食事画像とレシピテキストのクロスモーダル検索を実現する手法を提案する。そして、Recipe1M データセットを利用した評価実験では先行研究と比較して最高精度を達成し、提案手法の有効性を確認した。

キーワード　Cross-Modal Retrieval, Multimodal Embedding, MLLM, Contrastive Learning, Recipe

1. はじめに

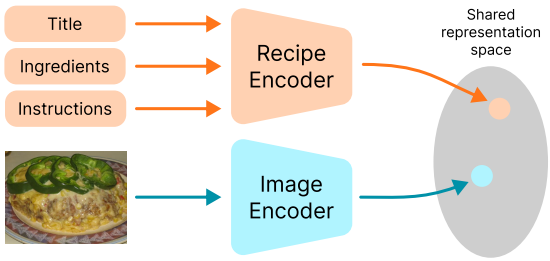
食は人間の日常生活において不可欠な要素であり、健康的な食生活への関心は年々高まっている。近年、インターネット上ではレシピ共有サイトや SNS を通じて、膨大な量の食事画像やレシピテキストが公開されており、これらのマルチモーダルなデータを効果的に活用する技術への需要が増大している。

そこで、食事画像とレシピテキスト間のクロスモーダル検索が重要な研究課題として注目を集めている。このタスクは、食事画像が与えられた際にその食事を調理するためのレシピを検索する、あるいはレシピテキストから対応する食事画像を検索するという双方向の検索を目的としている。この技術は、栄養管理や食事記録、料理支援など幅広い応用が期待される。

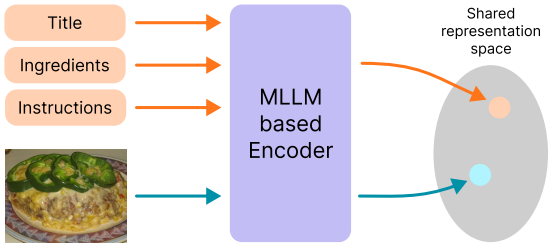
一方、コンピュータビジョンや自然言語処理の分野では、LLaVA [1] や Qwen [2] のようなマルチモーダル大規模言語モデル（Multimodal Large Language Model; MLLM）が急速に発展しており、画像とテキストの統合的な理解能力において顕著な成果を上げている。さらに近年では、MLLM を埋め込みモデルとして活用する研究が進み、VLM2Vec [3] や MM-Embed [4] といった手法により、MLLM の持つ豊富な知識をマルチモーダル埋め込みに活用できる可能性が示されている。

しかし、食事画像とレシピテキストのクロスモーダル検索における既存手法は、図 1a のような画像エンコーダとテキストエンコーダを独立に用いるデュアルエンコーダ型のアプローチが主流であり [5]〜[7]、MLLM は適用されていなかった。また、従来のデュアルエンコーダ型アプローチでは、画像とテキストという異なるモダリティ間の意味的なギャップを埋めるために、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とされてきた。例えば、ACME [8] や R^2GAN [9] では敵対的学習によりモダリティ間の分布を整合させる手法が提案され、H-T [5] や T-Food [6] では構造化されたレシピテキストに対応するために階層的な Transformer が使用されている。

本研究では、図 1b のように事前学習済み MLLM を埋め込みモデルに転換した MLLM ベース埋め込みモデルである VLM2Vec-V1 [3] および VLM2Vec-V2 [10] を食事画像とレシピテキストのクロスモーダル検索に応用する。MLLM は大規模なデータで事前学習されており、画像とテキストの両方を統一的に処理する能力を有している。この事前学習済みの知識を活用することで、従来手法で課題となっていた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を必要とせず、効果的なクロスモーダル埋め込みの獲得が期待できる。提案手法の有効性は大規模なレシピデータセットである Recipe1M [11] による評価実験で確認する。



(a) 既存手法



(b) 提案手法

図 1: 既存手法と提案手法の概要

2. 関 連 研 究

2.1 画像とテキストのクロスモーダル埋め込み

画像とテキストのクロスモーダル埋め込みは長年にわたり重要な研究課題とされている。従来、この分野では CLIP [12], ALIGN [13], BLIP [14], SigLIP [15] といった、大規模な画像-テキストペアで事前学習されたデュアルエンコーダーモデルが標準的なアプローチとして成功を収めてきた。これらのモデルは画像とテキストを独立したエンコーダーで埋め込み、対照学習によって共通の表現空間で整列させる。しかし、これらの手法はモダリティ間の相互作用が少ないため複雑な視覚と言語の関係性を捉える能力に限界があった。

そこで近年、MLLM の急速な発展に伴い、MLLM を埋め込みモデルとして利用するアプローチが注目されている。VLM2Vec [3] では、Phi-3.5 [16] や LLaVA [1]、Qwen2-VL [2] などの事前学習済み MLLM を、作成した大規模なマルチモーダルデータセットでファインチューニングすることで従来の CLIP ベースのモデルを大幅に上回る性能を達成した。さらに、VLM2Vec-V2 [10] では学習データを追加することで、画像とテキストだけでなく動画と視覚文書にも対応した。

本研究ではこの流れに沿って、今までデュアルエンコーダ型アプローチしか適用されていなかった食事画像とレシピテキストのクロスモーダル検索タスクに、MLLM ベースの埋め込みモデルを適用する。

2.2 食事画像とレシピテキストのクロスモーダル検索

Recipe1M データセット [11] の公開から食事画像とレシピテキストのクロスモーダル検索は活発に取り組まれるタスクとなった。通常の画像-テキスト間の検索との大きな違いは、レシピテキストが構造化されていてタイトル・材料・調理手順の 3 つの要素に分かれていることである。

初期の研究では画像エンコーダーとして ImageNet [17] で事前学習された VGG [18] や ResNet [19] などの CNN、レシピテキストエンコーダーとして Word2Vec [20] と LSTM [21] の組み合わせが用いられていた [8], [9], [11]。その後、レシピエンコーダを改善する手法がいくつか発表された。例えば BERT [22] のような事前学習言語モデルを使う手法や Tree-LSTM [23] によってレシピの階層的構造を捉える手法などである [24], [25]。そして Transformer [26] の台頭により、画像には ViT [27]、レシピには階層的 Transformer (H-T [5]) というレシピ用の Transformer を用いる研究が増えた [6], [28]。近年は、CLIP [12] に代表される VLM を食事ドメインに適応させるアプローチが一般的になっている [29]〜[31]。

このようにエンコーダーのアーキテクチャが変わりながら改善されてきたが、既存手法はすべて画像用のエンコーダーとレシピ用のエンコーダーに分かれたデュアルエンコーダ型のアプローチである。それに対して提案手法は MLLM ベースの埋め込みモデルを食事画像・レシピテキスト共通のエンコーダーとして利用する。

3. 提 案 手 法

3.1 定 式 化

学習用データセットとして、N 組の食事画像とレシピテキスト $(v_i, r_i)_{i=1}^{N}$ が与えられているとする。ここで、v_i は食事画像、r_i は対応するレシピテキストを表す。目的は、両モダリティのデータを共通の埋め込み空間へ写像する統一エンコーダ $\Phi(\cdot)$ を学習することである。各レシピ r_i は、タイトル、材料リスト、調理手順リストから構成される。

Φ は、対応するペアに対しては画像埋め込み $\mathbf{e}_v = \Phi(v)$ とレシピ埋め込み $\mathbf{e}_r = \Phi(r)$ が近くなり、非対応ペアに対しては離れるように学習されることを目指す。

本フレームワークでは、$\Phi(\cdot)$ は視覚入力とテキスト入力の両方を入力可能な MLLM として実装される。

3.2 MLLM による埋め込み

ベースモデルとして利用する VLM2Vec [3] および VLM2Vec-V2 [10] にしたがって、本手法では MLLM の最後のトークンの最終層の隠れ状態を埋め込み表現として採用する。

VLM2Vec では、通常の対照学習のようにペア同士の埋め込みを双方向に近づけるのではなく、クエリから検索候補を探す一方向の検索タスクとして対照学習を定式化する。そのため、同じデータであってもクエリとして入力する場合と検索候補として入力する場合とでは異なるプロンプトが用いられる。クエリとして入力する場合は検索を促すようなプロンプトを、検索候補として入力する場合はよりシンプルに埋め込み表現を求めるようなプロンプトを用いる。

本手法でもこれに従い、Image-to-Recipe 検索と Recipe-to-Image 検索の 2 つの検索方向に応じて、食事画像・レシピテキストそれぞれにクエリ用と検索候補用の 2 種類のプロンプトを設計する。

3.2.1 画像の入力方法

リスト 1、2 に画像を入力する際のプロンプトのテンプレートを示す。プロンプト内の`<|image_1|>`は画像を表すプレースホルダである。

リスト 1: 画像を Image-to-Recipe 検索のクエリとして入力する際のプロンプトのテンプレート

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

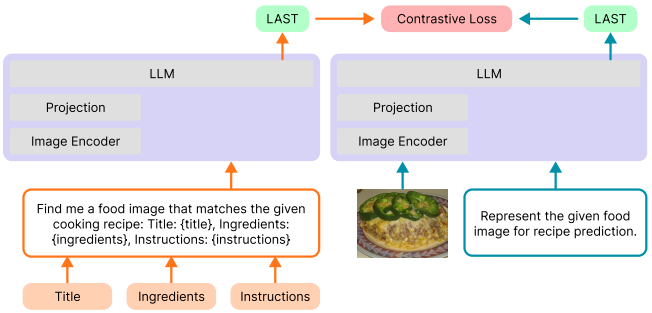


図 2: 提案手法のファインチューニング方法

リスト 2: 画像を Recipe-to-Image 検索の検索候補として入力する際のプロンプトのテンプレート

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

3.2.2 レシピテキストの入力方法

リスト 3、4 にレシピテキストを入力する際のプロンプトのテンプレートを示す。プロンプト内の`{title}`はタイトルを、`{ingredients}`は材料リストをカンマ区切りで結合した文字列を、`{instructions}`は調理手順リストをスペース区切りで結合した文字列をそれぞれ表す。

リスト 3: レシピテキストを Recipe-to-Image 検索のクエリとして入力する際のプロンプトのテンプレート

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

リスト 4: レシピテキストを Image-to-Recipe 検索の検索候補として入力する際のプロンプトのテンプレート

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

3.3 レシピ要素の除去によるデータ拡張

現実世界のレシピデータでは、タイトル、材料、手順のすべてが揃っていない場合がある。そのような不完全なレシピデータに対する頑健性を向上させるために、レシピ要素の除去によるデータ拡張を提案する。このデータ拡張では、タイトル・材料・手順という 3 つのレシピ要素のうち 2 つを欠損させ、1 つの要素しか持たないレシピを学習データに追加する。つまり、元の完全なレシピデータに 3 種類の不完全レシピデータが追加され、以下の合計 4 種類のデータを学習に利用する。

（1） タイトル・材料・手順すべてを含む完全なレシピ
（2） タイトルしか持たない不完全なレシピ
（3） 材料しか持たない不完全なレシピ
（4） 手順しか持たない不完全なレシピ

4. 3. 2 節の実験で本手法の有効性を検証する。

3.4 目 的 関 数

3. 1 節で述べた目標を達成するため、VLM2Vec に従い対照学習によるファインチューニングを行う。VLM2Vec ではクエリから検索候補への一方向の検索タスクとして学習を行う。そこで、同一の画像-レシピペアから Image-to-Recipe 用と Recipe-to-Image 用の 2 つのデータセットを構築し、それぞれに対して一方向の InfoNCE 損失 [32] を計算する。これにより、結果として双方向の検索が可能な埋め込み表現が学習される。

ミニバッチとして B 組のクエリ-候補ペア $(q_i, c_i)_{i=1}^{B}$ が与えられたとき、対応するペアのみを正例、同一バッチ内の他のすべての候補を負例とし、以下の式 1 のように InfoNCE 損失を計算する。

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathrm{sim}(\mathbf{e}_{q_i},\mathbf{e}_{c_i})/\tau)}{\sum_{j=1}^{B}\exp(\mathrm{sim}(\mathbf{e}_{q_i},\mathbf{e}_{c_j})/\tau)} \tag{1}$$

表 1: ベースモデル

モデル	パラメータ数	埋め込み次元数
VLM2Vec-V1-2B[1]	2B	1536
VLM2Vec-V1-7B[2]	7B	3584
VLM2Vec-V2[3]	2B	1536

[1] https://huggingface.co/TIGER-Lab/VLM2Vec-Qwen2VL-2B
[2] https://huggingface.co/TIGER-Lab/VLM2Vec-Qwen2VL-7B
[3] https://huggingface.co/VLM2Vec/VLM2Vec-V2.0

ここで、$\mathbf{e}_{q_i}$ はクエリの埋め込み表現、$\mathbf{e}_{c_i}$ は候補の埋め込み表現、$\mathrm{sim}(\cdot,\cdot)$ はコサイン類似度、τ は温度パラメータを表す。

4. 実　　験

4.1 実 験 設 定

4.1.1 データセット

提案手法の評価には先行研究と同様に Recipe1M データセット [11] を利用した。このデータセットには 100 万件を超えるレシピデータが含まれているが、本研究で利用したのはレシピに対応する食事画像が存在するもののみであり、学習用・検証用・テスト用のデータはそれぞれ 238,408、51,119、51,304 ペアとなっている。各レシピはタイトルの文字列、材料の文字列のリスト、調理手順の文字列のリストで構成される。

4.1.2 評 価 指 標

評価指標は先行研究にならい、median Rank (medR), Recall@k (R@k, $k = 1, 5, 10$) である。medR は検索をしたときの正解候補の順位の中央値で計算され、R@k は正解候補が上位 k 番目以内に含まれる確率で計算される。

また、この評価指標の算出は 1k 設定と 10k 設定の 2 通りで行われる。1k 設定では、テストデータから 1,000 ペアをランダムにサンプリングして評価指標を計算するという手順を 10 回行い、その平均値を算出する。10 回の平均を取ることで、ランダムサンプリングによる評価値のばらつきを抑える。10k 設定では、タスクを更に困難にするためにサンプル数を 10,000 ペアに増やして評価値を算出する。

評価指標の算出に公平を期すため、H-T [5] の公式実装を用いた(注1)。

4.1.3 実 装 詳 細

学習や評価には NVIDIA RTX A6000 を 8 基用いた。ベースモデルには VLM2Vec シリーズの 3 つを採用した。モデルの詳細は表 1 に示す。

学習では LoRA [33] を適用した。LoRA のランク、alpha、ドロップアウト率はそれぞれ 16、64、0.1 である。最適化手法は Adam [34] を用い、学習率は 1×10^{-4}、学習ステップ数は 2000 である。損失関数は InfoNCE [32] を用い、温度パラメータ τ は 0.02、バッチサイズは 128 である。

4.2 既存手法との比較

表 2 に提案手法と既存手法の比較を示す。提案手法の Ours (V1-2B), Ours (V1-7B), Ours (V2) という 3 つモデルすべてが、

(注1)：https://github.com/amzn/image-to-recipe-transformers

全評価指標において既存手法と同等以上の性能を示した。特に V1-7B が最高性能を達成し、1k 設定での Image-to-Recipe と Recipe-to-Image の R@1 はそれぞれ 87.5%, 85.1% となった。

従来の最良手法である Yang et al. [28] と Ours (V1-7B) を比較すると、1k 設定での Image-to-Recipe における R@1 を 81.8% から 87.5% へ、また 10k 設定の Image-to-Recipe における R@1 を 56.5% から 65.5% へと改善した。これらはそれぞれ、5.7 ポイントおよび 9.0 ポイントという大幅な向上を示している。

この実験では、より新しいモデルである VLM2Vec-V2 をベースとした Ours (V2) よりも、VLM2Vec-V1 をベースとした Ours (V1-7B) の方が一貫して高い性能を示した。このような結果になった要因として考えられるのはまず、VLM2Vec-V1 から VLM2Vec-V2 の変更点が学習データセットや学習方法であり、バックボーンモデルは同じ Qwen2-VL [2] であるということ。また、VLM2Vec-V2 では多様なモダリティに対応するため学習データセットに動画と視覚文書が追加されたが、その追加データによる学習が本研究が対象としている食事画像とレシピテキストのクロスモーダル検索タスクにおいては必ずしも性能向上に寄与していないと考えられる。さらに重要なのはモデルサイズと埋め込み次元数の違いである。表 1 に示すとおり、VLM2Vec-V2 は 2B パラメータで 1,536 次元なのに対し、VLM2Vec-V1-7B は 7B パラメータで 3,584 次元である。レシピ検索のような材料、調理手順、見た目といった多様かつ細粒度の情報を整合させる必要があるタスクではモデル容量と高次元の埋め込み表現が重要であると考えられる。

4.3 アブレーション分析

4.3.1 ファインチューニングの効果

Recipe1M データセットにおけるファインチューニングの効果を検証するために、ファインチューニング前後で性能を比較する。結果を表 3 に示す。この実験によって、提案手法の 3 つのモデルすべてにおいてファインチューニングが大幅な性能向上をもたらすことが確認された。例えば、V1-7B モデルはファインチューニングなしのゼロショット設定では 1k Image-to-Recipe の R@1 が 40.8% にとどまっていたが、ファインチューニングによって 87.5% まで 46.7 ポイントの改善が得られた。同様の傾向は他のモデルでも見られ、V1-2B は 29.5% から 84.1% へ、V2 は 45.1% から 83.8% へそれぞれ大幅に性能が向上している。

これらの結果は、VLM2Vec のような汎用的な視覚言語理解能力を持ったモデルであっても、食事画像とレシピという特定ドメインにおける高精度なクロスモーダル検索を実現するためにはタスク特化のファインチューニングが不可欠であることを示している。

4.3.2 レシピの構成要素が検索性能に与える影響

実際の応用場面では、レシピデータのすべての要素が揃っているとは限らない。例えば、タイトルのみが与えられている場合や、材料リストが明示されていない場合が考えられる。本節では、推論時にレシピ要素の一部のみを使用した場合の検索性能を評価し、さらに 3.3 節で提案したデータ拡張の効果を検証する。評価には V1-7B モデルを用い、データ拡張なし（V1-7B w/o Aug.）とデータ拡張あり（V1-7B）の 2 条件を比較する。結果を表 4 に示す。

まず、レシピ要素の組み合わせによる性能の違いについて分析する。単一の要素のみを使用する場合、手順のみを用いた場合が最も高い性能を示し、1k Image-to-Recipe の R@1 で 74.2% を達成した。これは、調理手順が料理の見た目との対応関係を捉える上で最も情報量の多い要素であることを示唆している。一方、タイトルのみ（44.4%）や材料のみ（38.8%）では性能が大幅に低下した。2 つの要素を組み合わせる場合は、材料と手順の組み合わせが最も高い性能を示し、1k Image-to-Recipe の R@1 で 85.9% と、全要素使用時の 87.5% に迫る性能を達成した。また、当然ながら 3 つの要素すべてを使用した場合に最高性能が得られた。

次に、データ拡張の効果について分析する。提案したデータ拡張は、特に要素が欠損した条件において性能向上に寄与している。タイトルのみの場合、データ拡張により 1k Image-to-Recipe の R@1 が 40.4% から 44.4% へ、材料のみの場合は 34.9% から 38.8% へ向上した。データ拡張で追加されたのは 2 要素が除去された 1 要素しか持たないレシピデータであるにもかかわらず、タイトルと材料という 2 要素の組み合わせでも 44.8% から 46.9% への改善が見られた。

一方、手順を含む組み合わせや全要素使用時ではデータ拡張の効果は限定的であった。これは、手順が十分な情報を持つ場合にはデータ拡張による追加の頑健性が不要であることを示している。以上の結果から、提案したデータ拡張は情報量の少ない要素のみが利用可能な場面で特に有効であることが確認された。

5. おわりに

本研究では、近年登場した MLLM ベースの Multimodal Embedding 手法である VLM2Vec をレシピデータセットでファインチューニングすることで、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構造を必要とせず、高精度で食事画像とレシピテキストのクロスモーダル検索を実現する手法を提案した。そして、レシピという構造化されたデータを MLLM で扱うためのプロンプトの設計や、不完全なレシピデータへの頑健性を向上させるためのデータ拡張などを実施した。Recipe1M データセットを利用した評価実験では先行研究と比較して最高精度を達成し、提案手法の有効性を確認した。

今後の展望として、MLLM をエンコーダーとして利用することで画像とレシピの同時入力やタスク指示を考慮した埋め込みも可能になったことを活かして、画像とレシピの組からレシピを検索する新たな検索タスクに取り組むことや、単に似た料理を検索するのではなくユーザーの指示に基づいた検索を実現することなどを考えている。

文　献

[1] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models," arXiv preprint arXiv:2407.07895, 2024.

[2] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-VL: Enhancing Vision-Language

表 2: Recipe1M における Image-to-Recipe 検索と Recipe-to-Image 検索の結果。最良の結果を**太字**で、次点の結果を下線で示す。

Method	Venue	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. [11]	CVPR'17	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0	41.9	-	-	-	39.2	-	-	-
H-T [5]	CVPR'21	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 [6]	CVPRW'22	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. [28]	ICDAR'24	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 [30]	WACV'24	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 [29]	ECCV'24	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. [31]	MM'25	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

表 3: ファインチューニング前後の性能比較

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↑
V1-2B (Zero-shot)	4.65	29.5	51.9	61.2	7.7	18.4	42.8	57.9	40.4	11.4	25.7	33.3	69.9	6.2	14.1	20.1
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
V1-7B (Zero-shot)	2.1	40.8	66.5	76	2	39.8	69	79	14.9	17.6	36	45	14.5	15.8	34.5	44.7
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
V2 (Zero-shot)	2	45.1	73.4	82.4	2	47.3	74.7	83.3	10.1	20	40.6	50.7	9.1	22.3	42.7	52.5
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

表 4: 不完全なレシピデータの推論におけるデータ拡張の効果

推論で使用した要素			手法	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

Model's Perception of the World at Any Resolution," arXiv preprint arXiv:2409.12191, 2024.

[3] Z. Jiang, R. Meng, X. Yang, S. Yavuz, Y. Zhou, and W. Chen, "VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks," Proc. of International Conference on Learning Representations, 2025.

[4] S.-C. Lin, C. Lee, M. Shoeybi, J. Lin, B. Catanzaro, and W. Ping, "MM-Embed: Universal Multimodal Retrieval with Multimodal LLMs," Proc. of International Conference on Learning Representations, 2025.

[5] A. Salvador, E. Gundogdu, L. Bazzani, and M. Donoser, "Revamping Cross-Modal Recipe Retrieval with Hierarchical Transformers and Self-supervised Learning," Proc. of IEEE Computer Vision and Pattern Recognition, pp.15470–15479, 2021.

[6] M. Shukor, G. Couairon, A. Grechka, and M. Cord, "Transformer Decoders with MultiModal Regularization for Cross-Modal Food Retrieval," Proc. of IEEE Computer Vision and Pattern Recognition Workshops, pp.4566–4577, 2022.

[7] J. Yang, J. Chen, and K. Yanai, "Transformer-Based Cross-Modal Recipe Embeddings with Large Batch Training," Proc. of International Conference of Multimedia Modeling, pp.471–482, 2023.

[8] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, and S.C.H. Hoi, "Learning Cross-Modal Embeddings With Adversarial Networks for Cooking Recipes and Food Images," Proc. of IEEE Computer Vision and Pattern Recognition, pp.11564–11573, 2019.

[9] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R^2GAN: Cross-Modal Recipe Retrieval With Generative Adversarial Network," Proc. of IEEE Computer Vision and Pattern Recognition, pp.11469–11478, 2019.

[10] R. Meng, Z. Jiang, Y. Liu, M. Su, X. Yang, Y. Fu, C. Qin, Z. Chen, R. Xu, C. Xiong, Y. Zhou, W. Chen, and S. Yavuz, "VLM2Vec-V2: Advancing Multimodal Embedding for Videos, Images, and Visual Documents," arXiv preprint arXiv:2507.04590, 2025.

[11] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," Proc. of IEEE Computer Vision and Pattern Recognition, pp.3068–3076, 2017.

[12] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Proc. of International Conference on Machine Learning, pp.8748–8763, 2021.

[13] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," Proc. of International Conference on Machine Learning, pp.4904–4916, 2021.

[14] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Proc. of International Conference on Machine Learning, pp.12888–12900, 2022.

[15] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid Loss for Language Image Pre-Training," Proc. of IEEE International Conference on Computer Vision, pp.11975–11986, 2023.

[16] M. Abdin, J. Aneja, et al., "Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone," arXiv preprint arXiv:2404.14219, 2024.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," Proc. of IEEE Computer Vision and Pattern Recognition, pp.248–255, 2009.

[18] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv preprint arXiv:1409.1556, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proc. of IEEE Computer Vision and Pattern Recognition, pp.770–778, 2016.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Proc. of International Conference on Learning Representations, 2013.

[21] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computing, vol.9, no.8, pp.1735–1780, 1997.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proc. of North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.4171–4186, 2019.

[23] K.S. Tai, R. Socher, and C.D. Manning, "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks," Proc. of Annual Meeting of the Association for Computational Linguistics, pp.1556–1566, 2015.

[24] R. Guerrero, H.X. Pham, and V. Pavlovic, "Cross-modal Retrieval and Synthesis (X-MRS): Closing the Modality Gap in Shared Subspace Learning," Proc. of ACM International Conference Multimedia, pp.3192–3201, 2021.

[25] H. Wang, D. Sahoo, C. Liu, K. Shu, P. Achananuparp, E.-p. Lim, and S.C.H. Hoi, "Cross-Modal Food Retrieval: Learning a Joint Embedding of Food Images and Recipes With Semantic Consistency and Attention Mechanism," IEEE Transactions on Multimedia, vol.24, pp.2515–2525, 2022.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. ukaszKaiser, and I. Polosukhin, "Attention is All you Need," Advances in Neural Information Processing Systems, vol.30, 2017.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Proc. of International Conference on Learning Representations, 2021.

[28] J. Yang, J. Chen, and K. Yanai, "Improving Cross-Modal Recipe Embeddings with Cross Decoder," Proc. of ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, pp.1–4, 2024.

[29] F. Song, B. Zhu, Y. Hao, and S. Wang, "Enhancing Recipe Retrieval with Foundation Models: A Data Augmentation Perspective," Proc. of European Conference on Computer Vision, pp.111–127, 2025.

[30] M. Wahed, X. Zhou, T. Yu, and I. Lourentzou, "Fine-Grained Alignment for Cross-Modal Recipe Retrieval," Proc. of IEEE Winter Conference on Applications of Computer Vision, pp.5572–5581, 2024.

[31] Q. Wang, C.-W. Ngo, Y. Cao, and E.-P. Lim, "Mitigating Cross-modal Representation Bias for Multicultural Image-to-Recipe Retrieval," Proc. of ACM International Conference Multimedia, pp.6223–6231, 2025.

[32] A. van denOord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," arXiv preprint arXiv:1807.03748, 2018.

[33] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," Proc. of International Conference on Learning Representations, 2022.

[34] D.P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, 2014.