CONFIDENTIAL EXTENDED ABSTRACT.
DO NOT DISTRIBUTE ANYWHERE.

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[1,a)]　柳井 啓司[1,b)]

概要

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。

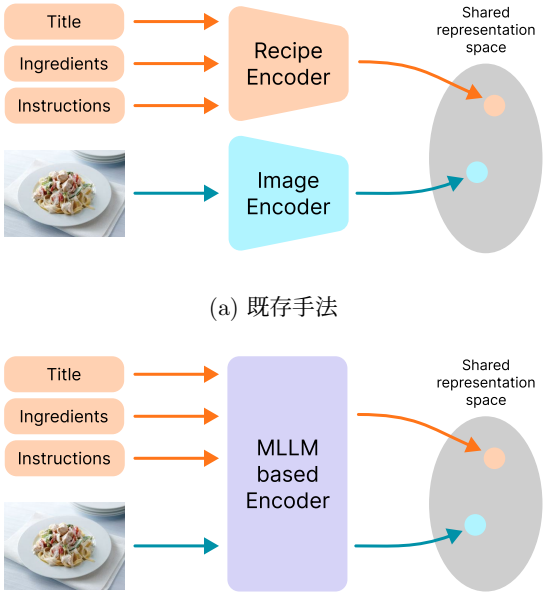
1. はじめに

近年、インターネット上では、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、食事画像から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。

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 [8] や Qwen [24] 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 [7] や MM-Embed [10] 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 [14], [16], [26]、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。



(a) 既存手法

(b) 提案手法

図 1: 既存手法と提案手法の概要

本研究では、図 1b のように MLLM ベース埋め込みモデルである VLM2Vec シリーズ [7], [11] 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 [15] による評価実験で確認する。

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

[1] 電気通信大学
[a)] gomi-k@mm.inf.uec.ac.jp
[b)] yanai@cs.uec.ac.jp

たり重要な研究課題とされている。従来、この分野では CLIP [13], ALIGN [6], BLIP [9], SigLIP [28] 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モデルが標準的なアプローチとして成功を収めてきた。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 [7] では、Qwen2-VL [24] 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 [11] 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 [15] 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。

エンコーダ構造に着目すると、様々な改善が提案されてきた。初期の研究では画像には ResNet [4] が、レシピには Word2Vec [12] と LSTM [5] の組み合わせが用いられていた [15], [22], [29]。その後、レシピに BERT [1] や Tree-LSTM [18] を用いる手法がいくつか発表された [3], [23]。そして Transformer [20] の台頭により、画像には ViT [2]、レシピには階層的 Transformer (H-T [14]) というレシピ用の Transformer を用いる研究が増えた [16], [27]。近年は、CLIP [13] 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 [17], [21], [25]。

このようにエンコーダ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とレシピ用のエンコーダ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。

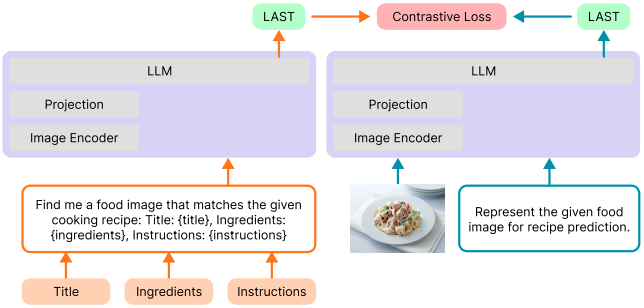


図 2: 提案手法のファインチューニング方法

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.
```

3.2 MLLM による埋め込み

ベースモデルとして利用する VLM2Vec [7] および VLM2Vec-V2 [11] 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。

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

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.
```

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（図 2）。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 [19] 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。

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

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 [15] 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。

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検索候補数が 1,000 個の 1k 設定と 10,000 個の 10k 設定の 2 通りで行ったが、紙面の都合上、本論文では 10k 設定の結果のみを報告し、1k 設定の結果は補足資料に掲載する。

4.1.3 ベースモデル

ベースモデルには VLM2Vec シリーズの 3 つを採用した。VLM2Vec-V1 のサイズの異なる 2 つのモデル (V1-2B[*1], V1-7B[*2]) と、VLM2Vec-V2[*3]である。V1-2B と V2 はパラメータ数が 2B で埋め込み次元数が 1536、V1-7B はパラメータ数が 7B で埋め込み次元数が 3584 である。

4.2 既存手法との比較

表 1 に提案手法と既存手法の比較を示す。提案手法の 3 つのモデルすべてが、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た。

従来の最良手法である Yang et al. [27] と Ours (V1-7B) を比較すると、Image-to-Recipe における R@1 を 56.5%から 65.5%へ 9.0 ポイントという大幅な改善をした。

この実験では、改善版である VLM2Vec-V2 よりも VLM2Vec-V1 の方が高い性能を示した。このような結果になった理由として 2 つの要因が考えられる。1 つ目は、VLM2Vec-V2 の改善内容が主に動画と視覚文書での性能向上であるため、今回のタスクでは必ずしも性能向上に寄与していないと考えられること。2 つ目は、4.1.3 節で述べ

[*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

たとおり、VLM2Vec-V2 の方がパラメータ数と埋め込み次元数が小さいことである。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 2 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。

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。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 3 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、Image-to-Recipe の R@1 で 48.9%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（16.1%）や材料のみ（17.2%）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、Image-to-Recipe の R@1 で 63.2%と、全要素使用時の 65.5%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 Image-to-Recipe の R@1 が 13.6%から 16.1%へ、材料のみの場合は 15.1%から 17.2%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 19.4%から 21.4%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。

表 1: Image-to-Recipe 検索と Recipe-to-Image 検索の結果。最良を**太字**、次点を下線で示す。

Method	Venue	Image-to-Recipe				Recipe-to-Image			
		medR↓	R@1↑	R@5↑	R@10↑	medR↓	R@1↑	R@5↑	R@10↑
Salvador et al. [15]	CVPR'17	41.9	-	-	-	39.2	-	-	-
H-T [14]	CVPR'21	4.0	27.9	56.4	68.1	4.0	28.3	56.5	68.1
T-Food [16]	CVPRW'22	2.0	43.4	70.7	79.7	2.0	44.6	71.2	79.7
Yang et al. [27]	ICDAR'24	1.0	56.5	81.0	87.6	1.0	55.7	80.2	87.1
FARM [21]	WACV'24	2.0	44.9	71.8	80.0	2.0	44.3	71.5	80.0
DAR [17]	ECCV'24	2.0	47.8	75.9	84.3	2.0	47.4	75.5	84.1
Wang et al. [25]	MM'25	1.0	51.7	78.2	85.9	1.0	52.2	78.4	86.0
Ours (V1-2B)	-	1.0	59.7	83.5	89.9	1.0	55.8	81.1	88.0
Ours (V1-7B)	-	1.0	**65.5**	**87.4**	**92.5**	1.0	**61.5**	**85.0**	**91.0**
Ours (V2)	-	1.0	59.1	83.3	89.8	1.0	55.7	81.0	88.0

表 2: ファインチューニング前後の性能比較

Method	Image-to-Recipe				Recipe-to-Image			
	medR↓	R@1↑	R@5↑	R@10↑	medR↓	R@1↑	R@5↑	R@10↑
V1-2B (Zero-shot)	40.4	11.4	25.7	33.3	69.9	6.2	14.1	20.1
Ours (V1-2B)	1.0	59.7	83.5	89.9	1.0	55.8	81.1	88.0
V1-7B (Zero-shot)	14.9	17.6	36.0	45.0	14.5	15.8	34.5	44.7
Ours (V1-7B)	1.0	65.5	87.4	92.5	1.0	61.5	85.0	91.0
V2 (Zero-shot)	10.1	20.0	40.6	50.7	9.1	22.3	42.7	52.5
Ours (V2)	1.0	59.1	83.3	89.8	1.0	55.7	81.0	88.0

表 3: 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			Method	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↑
✓			V1-7B w/o Aug.	13.1	13.6	34.0	45.9	13.5	14.9	34.7	45.7
			Ours (V1-7B)	10.0	16.1	38.9	51.0	12.3	15.2	36.0	47.4
	✓		V1-7B w/o Aug.	21.1	15.1	30.1	39.2	8.9	22.4	42.6	52.5
			Ours (V1-7B)	15.5	17.2	33.8	43.7	6.2	25.8	47.8	57.9
		✓	V1-7B w/o Aug.	2.0	48.9	72.3	79.9	2.0	44.0	68.8	77.1
			Ours (V1-7B)	2.0	48.9	72.5	80.1	2.0	43.7	68.4	77.0
✓	✓		V1-7B w/o Aug.	10.0	19.4	39.5	50.6	4.1	30.1	54.1	64.0
			Ours (V1-7B)	8.8	21.4	41.9	53.1	3.1	34.0	58.9	68.9
✓		✓	V1-7B w/o Aug.	1.0	55.1	79.0	86.2	1.3	50.2	75.8	83.8
			Ours (V1-7B)	1.0	54.8	79.3	86.5	1.3	50.4	76.0	84.2
	✓	✓	V1-7B w/o Aug.	1.0	63.2	85.6	91.2	1.0	59.4	83.0	89.4
			Ours (V1-7B)	1.0	63.2	85.7	91.2	1.0	59.1	82.9	89.3
✓	✓	✓	V1-7B w/o Aug.	1.0	65.6	87.3	92.6	1.0	61.8	84.9	90.9
			Ours (V1-7B)	1.0	65.5	87.4	92.5	1.0	61.5	85.0	91.0

5. おわりに

本研究では、MLLM ベースの埋め込みモデルである VLM2Vec を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。

MLLM をエンコーダ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構造上可能になった。これを活かして、今後は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。

参考文献

[1] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Burstein, J., Doran, C. and Solorio, T., eds.), pp. 4171–4186 (2019).

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

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

[4] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).

[5] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780 (1997).

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

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

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

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

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

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

[12] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *arXiv preprint arXiv:1301.3781* (2013).

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

[14] Salvador, A., Gundogdu, E., Bazzani, L. and Donoser, M.: Revamping Cross-Modal Recipe Retrieval with Hierarchical Transformers and Self-supervised Learning, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15470–15479 (2021).

[15] Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I. and Torralba, A.: Learning Cross-Modal Embeddings for Cooking Recipes and Food Images, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3068–3076 (2017).

[16] Shukor, M., Couairon, G., Grechka, A. and Cord, M.: Transformer Decoders with MultiModal Regularization for Cross-Modal Food Retrieval, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4566–4577 (2022).

[17] Song, F., Zhu, B., Hao, Y. and Wang, S.: Enhancing Recipe Retrieval with Foundation Models: A Data Augmentation Perspective, *Proceedings of European Conference on Computer Vision* (Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T. and Varol, G., eds.), pp. 111–127 (2024).

[18] Tai, K. S., Socher, R. and Manning, C. D.: Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks, *Proceedings of Annual Meeting of the Association for Computational Linguistics* (Zong, C. and Strube, M., eds.), pp. 1556–1566 (2015).

[19] van den Oord, A., Li, Y. and Vinyals, O.: Representation Learning with Contrastive Predictive Coding, *arXiv preprint arXiv:1807.03748* (2018).

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ukasz Kaiser, Ł. and Polosukhin, I.: Attention Is All You Need, *Advances in Neural Information Processing Systems*, Vol. 30 (2017).

[21] Wahed, M., Zhou, X., Yu, T. and Lourentzou, I.: Fine-Grained Alignment for Cross-Modal Recipe Retrieval, *Proceedings of IEEE Winter Conference on Applications of Computer Vision*, pp. 5572–5581 (2024).

[22] Wang, H., Sahoo, D., Liu, C., Lim, E.-p. and Hoi, S. C. H.: Learning Cross-Modal Embeddings With Adversarial Networks for Cooking Recipes and Food Images, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11564–11573 (2019).

[23] Wang, H., Sahoo, D., Liu, C., Shu, K., Achananuparp, P., Lim, E.-p. and Hoi, S. C. H.: 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).

[24] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J. and Lin, J.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, *arXiv preprint arXiv:2409.12191* (2024).

[25] Wang, Q., Ngo, C.-W., Cao, Y. and Lim, E.-P.: Mitigating Cross-modal Representation Bias for Multicultural Image-to-Recipe Retrieval, *Proceedings of ACM International Conference on Multimedia*, pp. 6223–6231 (2025).

[26] Yang, J., Chen, J. and Yanai, K.: Transformer-Based Cross-Modal Recipe Embeddings with Large Batch Training, *Proceedings of International Conference on Multimedia Modeling*, pp. 471–482 (2023).

[27] Yang, J., Chen, J. and Yanai, K.: Improving Cross-

Modal Recipe Embeddings with Cross Decoder, *Proceedings of ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, pp. 1–4 (2024).

[28] Zhai, X., Mustafa, B., Kolesnikov, A. and Beyer, L.: Sigmoid Loss for Language Image Pre-Training, *Proceedings of IEEE International Conference on Computer Vision*, pp. 11975–11986 (2023).

[29] Zhu, B., Ngo, C.-W., Chen, J. and Hao, Y.: R^2GAN: Cross-Modal Recipe Retrieval With Generative Adversarial Network, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11469–11478 (2019).