Fine-Grained Alignment for Cross-Modal Recipe Retrieval

Muntasir Wahed
Virginia Tech
mwahed@vt.edu

Xiaona Zhou
Virginia Tech
xzhou1@vt.edu

Tianjiao Yu
Virginia Tech
tianjiao@vt.edu

Ismini Lourentzou
Virginia Tech
ilourentzou@vt.edu

Abstract

Vision-language pre-trained models have exhibited significant advancements in various multimodal and unimodal tasks in recent years, including cross-modal recipe retrieval. However, a persistent challenge in multimodal frameworks is the lack of alignment between the encoders of different modalities. Although previous works addressed image and recipe embedding alignment, the alignment of individual recipe components has been overlooked. To address this gap, we present Fine-grained Alignment for Recipe eMbeddings (FARM), a cross-modal retrieval approach that aligns the encodings of recipe components, including titles, ingredients, and instructions, within a shared representation space alongside corresponding image embeddings. Moreover, we introduce a hyperbolic loss function to effectively capture the similarity information inherent in recipe classes. FARM improves Recall@1 by 1.4% for image-to-recipe and 1.0% for recipe-to-image retrieval. Additionally, FARM achieves up to 6.1% and 15.1% performance improvement in image-to-recipe retrieval tasks, when just one and two components of the recipe are available, respectively. Comprehensive qualitative analysis of retrieved images for various recipes showcases the semantic capabilities of our trained models. Code is available at https://github.com/PLAN-Lab/FARM.

1. Introduction

Recent advances in vision-language models have enabled state-of-the-art performance in several multi-modal and unimodal downstream tasks, *e.g.*, visual question answering, visual entailment, visual reasoning, cross-modal retrieval, *etc*. One such area of research is computational cooking, where the vision-language models can learn from procedural text recipes, associated images, and instructional videos. This enables various downstream subtasks, *e.g.* recipe retrieval [1, 4, 6, 29, 30, 36], recipe generation [5, 21], ingredient substitution [5, 13], and recipe recommendation [16, 31]. A fundamental prerequisite for these downstream subtasks is the development of an efficient and robust cross-modal recipe encoding mechanism.

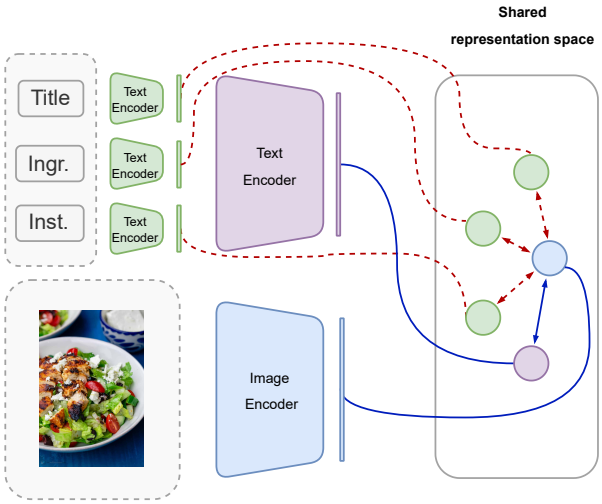

Figure 1. Previous works have focused on aligning the recipe and image embeddings (**blue lines**). FARM further aligns the image embeddings with the associated title, ingredient, and instruction embeddings in a shared representation space (**red lines**). Additionally, we employ a hyperbolic loss function to better capture the similarity among recipe categories.

A common challenge faced by multimodal frameworks is the lack of alignment between the encoders of different modalities. While previous works have made efforts to address the alignment issue between image and recipe embeddings [8, 11, 12, 24, 29, 30], they have predominantly overlooked the alignment of fine-grained information within the various recipe components, such as titles, ingredients, and instructions. Neglecting these individual components disregards their potential to convey useful information. For instance, a recipe's title, such as "caesar salad", can provide substantial insight into how the associated image would look like. Similarly, lists of ingredients and instructions offer valuable cues for understanding the image context. Moreover, some parts of the recipe may be missing in many downstream tasks. For example, in some applications, it may be useful to be able to retrieve recipes given just the title. Consequently, it is important to

learn robust representations that can handle missing components effectively. Previous works have primarily relied on semantic loss functions calculated solely based on class labels [1, 24]. However, class labels also hold valuable similarity attributes. For instance, a cheesecake is expected to be more similar to a brownie than to a salad. Hyperbolic loss functions can capture and incorporate such similarity information, enhancing the overall alignment process.

In this work, we introduce **F**ine-grained **A**lignment for **R**ecipe e**M**beddings (**FARM**), a cross-modal retrieval framework that enhances the alignment between different components of recipes (titles, ingredients, and instructions) and image embeddings in a shared vector space. This alignment is achieved through a projection layer that minimizes the distance between embeddings in a shared embedding space. We introduce two loss functions, namely the triplet loss and the hyperbolic embedding loss, to guide the training process by contrasting embeddings and capturing class similarity information. To evaluate the effectiveness of our approach, we conduct experiments on image-to-recipe and recipe-to-image retrieval tasks on the Recipe1M dataset [23]. FARM yields substantial improvements in Recall@1 performance for image-to-recipe and recipe-to-image retrieval tasks. In addition to performance evaluation, our ablation studies and qualitative analysis highlight that FARM learns meaningful component-level embeddings that capture valuable semantic information. Our contribution can be summarized as follows:

(1) We introduce FARM, a new cross-modal retrieval framework that addresses the lack of fine-grained alignment between recipe components (titles, ingredients, and instructions) and image embeddings. FARM leads to robust representations that improve performance even when parts of the recipe are missing.

(2) We propose the use of a hyperbolic embedding loss to leverage class label information and capture varying levels of similarity between recipes. Through ablation studies, we demonstrate the benefits of incorporating the hyperbolic loss in FARM for both image-to-recipe and recipe-to-image retrieval tasks.

(3) We experimentally show that the alignment of individual recipe components enhances the semantic understanding of recipes, improves robustness in handling missing information, and boosts the performance of cross-modal tasks by capturing the nuanced relationships between different components and images.

2. Related Work

2.1. Recipe Embeddings

Representation learning for recipes involves capturing the textual and structural features of a recipe, which usually consists of a title, ingredients, and step-by-step instructions. Existing works employ a dual-encoder approach for the recipe-image retrieval task, where the text and image are encoded by two different encoders. Most of the works in this area primarily utilized only the ingredients and instructions as inputs [1, 6, 15, 29, 30]. However, subsequent works showed that incorporating the title as an additional input improves retrieval performance [22]. Earlier works utilized LSTM-based architectures to encode the recipes, then passed them onto fully connected layers to map them to a latent space [6, 23, 29, 30]. Other works have also proposed employing Bi-LSTMs [1], hierarchical LSTMs [1], and tree-LSTMs [17] to better capture the inter-dependency between ingredients and instructions. However, such encoding frameworks fail to capture the relative importance of ingredients and instructions. Recent works have consequently shifted focus to Transformer-based encoders [7, 22, 24]. Salvador et al. (2021) [22] proposed the first hierarchical Transformer for recipe embeddings, and observed that the hierarchical structure improved performance when compared to simple average pooling for both Transformer and LSTM encoders. TFood [24] similarly proposed a hierarchical two-layer Transformer setup to capture the interactions between the title, ingredients, and instructions. In our approach, we also adopt a hierarchical Transformer setup to compute recipe embeddings from component embeddings. However, we enforce a finer-grained alignment between the recipe and its component embeddings with the image embedding. Our proposed method enables a more precise and effective alignment between different modalities, leading to improved cross-modal retrieval performance.

2.2. Cross-Modal Alignment

Cross-modal alignment plays a crucial role in integrating and comparing different modalities of data, such as images and text, facilitating various downstream tasks including image captioning, visual question answering, and recipe-image retrieval [8, 22, 24, 25]. Conventional approaches involve aligning the representations of these modalities within a joint embedding space. For instance, some works utilize adversarial loss functions to improve modality alignment by adversarially making it difficult for the discriminator to differentiate between recipe embeddings and image embeddings [12, 29]. Recent work introduces a novel semantic consistency loss that uses KL divergence to bring the output semantic probabilities of image and recipe pairs closer, thereby reducing the intra-class feature distance [30]. Subsequent work proposes a model that learns a more accurate image-recipe similarity by fusing both intra-modality and cross-modality features from local and global aspects [11]. Moreover, recent work introduces a cross-modal implicit relation reasoning module and a similarity distribution matching method to enhance global image-text matching without

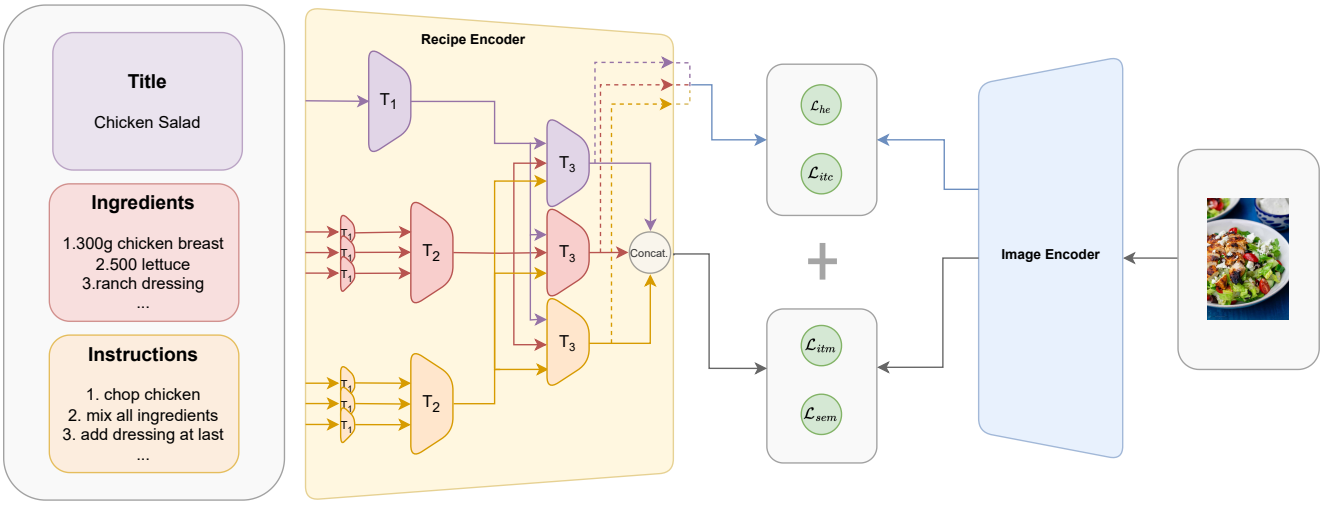

Figure 2. Illustration of the proposed FARM framework. The three components of the recipe, *i.e.*, title, ingredients, and instructions, are passed through hierarchical transformer encoders and are later fused into a single cross-attended recipe embedding. The corresponding image is passed through a pre-trained image encoder. The embeddings from these encoders are passed through a projection layer and then aligned using two fine-grained alignment losses, a proposed hyperbolic embedding loss $\mathcal{L}_{he}$ and an image-text contrastive loss $\mathcal{L}_{itc}$, that align the component embeddings directly with the corresponding image embedding.

additional prior supervision [8]. VLPCook [25] employs a similar cross-modal alignment during finetuning but requires additional data for pretraining. Our work is most similar to H-T [22] and TFood [24]. T-Food [24] does not consider the alignment of component embeddings, while H-T [22] focuses on the intra-modal alignment among the component embeddings themselves. In contrast, our work aligns component embeddings directly with the image embeddings, effectively capturing the semantic relationships between individual recipe components and the corresponding visual information, thereby allowing for a more comprehensive understanding of the recipe content.

3. Proposed Method

3.1. Problem Formulation

Given a set of recipes $\mathcal{R} = \{(r_1, y_1), \ldots, (r_n, y_n)\}$ and an associated set of images $\mathcal{I} = \{i_1, \ldots, i_n\}$, where each recipe r_i is accompanied by a class label $y_i \in \mathcal{Y}$, and $\mathcal{Y}$ denotes the set of class labels, the goal is to learn a text encoder $f_\theta(\cdot)\colon \mathcal{R} \rightarrow \mathbb{R}^d$ and an image encoder $g_{\theta'}(\cdot)\colon \mathcal{I} \rightarrow \mathbb{R}^d$ such that $\mathbf{t}_{rec} = f_\theta(r)$ and $\mathbf{t}_{img} = g_{\theta'}(i)$ for generic $r \in \mathcal{R}$ and $i \in \mathcal{I}$ are the associated recipe and image embeddings. Essentially, f_θ and $g_{\theta'}$ map data points to a shared d-dimensional embedding space and can later on be used in various computational cooking downstream tasks. Each recipe in $\mathcal{R}$ consists of 3 textual components, the title r_{ttl}, the ingredients r_{ing}, and the instructions r_{ins}. Existing works minimize the distance of the learned embeddings for each $r \in \mathcal{R}$ and associated $i \in \mathcal{I}$ in a shared vector space using various alignment strategies. In comparison, we propose aligning the component embeddings, *i.e.*, title $\mathbf{t}_{ttl}$, ingredients $\mathbf{t}_{ing}$, and instructions $\mathbf{t}_{ins}$ embeddings with the image $\mathbf{t}_{img}$ embeddings. The title, ingredients, and instructions contain important semantic information about the recipe, and we hypothesize that aligning them directly with the image embedding can improve performance.

3.2. Fine-grained Alignment (FARM)

Our proposed method employs a dual-encoder framework. The overall structure is illustrated in Figure 2.

Image Encoder: Inspired by the success of previous vision-language models [3, 18], we use a Vision Transformer (CLIP ViT B/16 [18]) to encode the image input to $\mathbf{t}^1_{img}$. We then pass $\mathbf{t}^1_{img}$ to a linear projection layer to obtain the image embedding $\mathbf{t}_{img}$, as shown in Eq (1).

$$\mathbf{t}_{img} = \mathbf{W}_{img} \cdot \mathbf{t}^1_{img} \tag{1}$$

Recipe Encoder: Given textual data containing titles, ingredients, and instructions, we first learn separate embeddings for each of them, and then fuse them into a shared embedding space. We first employ hierarchical Transformers to encode the recipes [22]. As illustrated in Figure 2, the title, the list of ingredients, and the list of instructions are first processed by a Transformer T_1. Since ingredients and instructions are composed of multiple sentences, we employ a hierarchical Transformer T_2 to capture the intrinsic interactions and dependencies between sentences in each corresponding recipe component, and encode them into multi-

ple sentence embeddings. Recipe components are also dependent on each other, *e.g.*, an instruction can only be performed when the corresponding ingredients are available. To this end, we incorporate an additional hierarchy level with Transformer T_3, which combines the component embeddings into a single recipe embedding. Specifically, T_3 takes the output tokens of each component as a query Q and the concatenation of the other two remaining components as keys K and values V. The final tokens $\mathbf{t}_{rec}$ are obtained by applying linear projection on the concatenated output of T_3, as shown in Eq (2).

$$\begin{aligned}
\mathbf{t}^3_{ttl} &= T_3(\mathbf{t}^1_{ttl}, [\mathbf{t}^2_{ing}; \mathbf{t}^2_{ins}], [\mathbf{t}^2_{ing}; \mathbf{t}^2_{ins}]) \\
\mathbf{t}^3_{ing} &= T_3(\mathbf{t}^2_{ing}, [\mathbf{t}^1_{ttl}; \mathbf{t}^2_{ins}], [\mathbf{t}^1_{ttl}; \mathbf{t}^2_{ins}]) \\
\mathbf{t}^3_{ins} &= T_3(\mathbf{t}^2_{ing}, [\mathbf{t}^1_{ttl}; \mathbf{t}^2_{ing}], [\mathbf{t}^1_{ttl}; \mathbf{t}^2_{ing}]) \\
\mathbf{t}_{rec} &= \mathbf{W}_{rec} \cdot [\mathbf{t}^3_{ttl}; \mathbf{t}^3_{ing}; \mathbf{t}^3_{ins}]
\end{aligned} \tag{2}$$

3.3. Aligning Recipe Components

Early experiments show that using the cross-attended encoders (*i.e.*, T_3) is better than the individual encoders (*i.e.*, T_1 and T_2). We also observe that aligning with $\mathbf{t}_{img}$ yields better performance compared to $\mathbf{t}_{rec}$. In order to align the component embeddings with $\mathbf{t}_{img}$, we apply linear projection to $\mathbf{t}^3_{ttl}$, $\mathbf{t}^3_{ing}$, $\mathbf{t}^3_{ins}$ as shown in Eq (3).

$$\begin{aligned}
\mathbf{t}_{ttl} &= \mathbf{W}_{ttl} \cdot \mathbf{t}^3_{ttl} \\
\mathbf{t}_{ing} &= \mathbf{W}_{ing} \cdot \mathbf{t}^3_{ing} \\
\mathbf{t}_{ins} &= \mathbf{W}_{ins} \cdot \mathbf{t}^3_{ins}
\end{aligned} \tag{3}$$

To further align the component embeddings with the corresponding image embedding, we train with two loss functions, *i.e.*, triplet loss and hyperbolic embedding loss.

Image-Text Contrastive Loss: Triplet loss aligns samples in a shared representation space by minimizing the distance between an anchor sample and a positive sample while maximizing the distance between the anchor and a negative sample. This has been shown to be an effective method for cross-modal recipe retrieval tasks [22, 23, 29]. We utilize this triplet loss to align the image embeddings $\mathbf{t}_{img}$ with component embeddings $\mathbf{t}_{ttl}$, $\mathbf{t}_{ing}$, and $\mathbf{t}_{ins}$ in a shared representation space. By using different component embeddings (*e.g.*, $\mathbf{t}_{ttl}$) as anchors, this loss function encourages dissimilar pairs to be distant from any similar pairs by a certain margin value. For instance, the triplet loss between a title and an image embedding is calculated as follows:

$$\mathcal{L}_t(\mathbf{t}^a_{ttl}, \mathbf{t}^p_{img}, \mathbf{t}^n_{img}) = \left[d(\mathbf{t}^a_{ttl}, \mathbf{t}^p_{img}) + \alpha - d(\mathbf{t}^a_{ttl}, \mathbf{t}^n_{img})\right]_+, \tag{4}$$

where α is the margin, $d(\cdot, \cdot)$ is a distance function, and superscripts a, p, n refer to the anchor, positive, and negative samples, respectively. Here, $\mathbf{t}_{ttl}$ is the output the title encoder, $\mathbf{t}^3_{ttl}$, and $\mathbf{t}_{img}$ is the output of the projection layer $\mathbf{t}^1_{img}$ corresponding to the image input. The margin α, proposed in [24], dynamically adjusts the difficulty of the task. Initially, assuming a difficult task, α is set to a small number that is incremented at each iteration, until it reaches a maximum value. In addition, we also use an adaptive weighting strategy for the triple loss, where we add a dynamic term δ to overcome the vanishing update when most of the triplets are inactive [1]. Then, the image-text contrastive loss can be written using Eq (5).

$$\begin{aligned}
\mathcal{L}_c(\mathcal{G}_{ttl}, \mathcal{G}_{img}) = &\frac{1}{\delta_{rec}} \sum_{\mathbf{t}_{ttl} \in \mathcal{G}_{ttl}} \mathcal{L}_t(\mathbf{t}^a_{ttl}, \mathbf{t}^p_{img}, \mathbf{t}^n_{img}) \\
&+ \frac{1}{\delta_{img}} \sum_{\mathbf{t}_{img} \in \mathcal{G}_{img}} \mathcal{L}_t(\mathbf{t}^a_{img}, \mathbf{t}^p_{ttl}, \mathbf{t}^n_{ttl}),
\end{aligned} \tag{5}$$

where the superscripts a, p, and n denote anchor, positive and negative examples, respectively. The sets $\mathcal{G}_{ttl}$ and $\mathcal{G}_{img}$ represent the sets of title and image embeddings, and δ_{rec} and δ_{img} correspond to the number of triplets that contribute to the loss, with title and image embeddings as anchors. Similarly, we calculate $\mathcal{L}_c(\mathcal{G}_{ing}, \mathcal{G}_{img})$ and $\mathcal{L}_c(\mathcal{G}_{ins}, \mathcal{G}_{img})$, which are the alignment of $\mathcal{G}_{img}$ with $\mathcal{G}_{ing}$ and $\mathcal{G}_{ins}$, respectively. We include an additional loss term $\mathcal{L}_c(\mathcal{G}_{rec}, \mathcal{G}_{img})$ to align the final recipe embeddings $\mathcal{G}_{rec}$ with the image embeddings $\mathcal{G}_{img}$. The final fine-grained triplet alignment loss is the average of the four alignment losses, *i.e.*,

$$\begin{aligned}
\mathcal{L}_{itc} = &\mathcal{L}_c(\mathcal{G}_{ttl}, \mathcal{G}_{img}) + \mathcal{L}_c(\mathcal{G}_{ing}, \mathcal{G}_{img}) \\
&+ \mathcal{L}_c(\mathcal{G}_{ins}, \mathcal{G}_{img}) + \mathcal{L}_c(\mathcal{G}_{rec}, \mathcal{G}_{img}).
\end{aligned} \tag{6}$$

Hyperbolic Embedding Loss: We additionally apply hierarchical clustering and hyperbolic metric learning [34]. By representing recipes in hyperbolic space, we can better capture the varying degrees of similarity between them. Hyperbolic spaces provide a natural way to represent hierarchical structures due to their negative curvature. Recipes often exhibit a hierarchical organization, where ingredients and cooking steps are grouped into categories and subcategories. In Euclidean space, representing this hierarchical structure can be challenging because distances between ingredients or steps might not accurately reflect their actual relationships. In contrast, in hyperbolic space, the negative curvature allows for a more efficient representation of hierarchies. Ingredients and cooking steps can be placed at appropriate distances from each other, reflecting their similarities and differences more accurately. This means that recipes with similar ingredients and techniques will be positioned closer to each other, while those with distinct elements, *e.g.*, completely dissimilar ingredients, will be farther apart. For instance, within the dessert category, a carrot cake will be closer to a muffin than to an ice cream. Hyperbolic spaces enable capturing of such nuances in recipe similarity. Furthermore, the infinite volume property of hyperbolic spaces allows for a larger number of recipes to be

effectively represented and compared within a finite space [19], which can be advantageous in recipe recommendation, where the goal is to search for similar recipes based on pre-specified criteria or to generate creative variations.

Specifically, for each recipe component, we first map the embeddings to a hyperbolic manifold $\mathbb{D}_\tau^n$ using Eq. (7). For instance, for the title embeddings $\mathbf{t}_{ttl}$, we compute hyperbolic representations as follows:

$$\mathbf{z}_{ttl} = \exp^\tau(\mathbf{t}_{ttl}) := \tanh(\sqrt{\tau}\|\mathbf{t}_{ttl}\|)\frac{\mathbf{t}_{ttl}}{\sqrt{\tau}\|\mathbf{t}_{ttl}\|}. \quad (7)$$

We perform similar calculations for the instructions, ingredients, recipe, and image embeddings. Next, we calculate the distance between two samples, for example, $\mathbf{z}_{ttl}$ and $\mathbf{z}_{img}$, using Eq. (8):

$$d\left(\mathbf{z}_{ttl}, \mathbf{z}_{img}\right) = \cosh^{-1}\left(1 + \frac{2\left\|\mathbf{z}_{ttl} - \mathbf{z}_{img}\right\|^2}{\left(1 - \left\|\mathbf{z}_{ttl}\right\|^2\right)\left(1 - \left\|\mathbf{z}_{img}\right\|^2\right)}\right). \quad (8)$$

To apply hierarchical clustering, we calculate the distance between the samples from two classes $C_a, C_b \in \mathcal{Y}$:

$$d_{C_a,C_b} = \frac{1}{n_{y_a} n_{y_b}} \sum_{\substack{\mathbf{z}_{ttl}\in C_a,\\ \mathbf{z}_{img}\in C_b}} d\left(\mathbf{z}_{ttl}, \mathbf{z}_{img}\right), \quad (9)$$

where $\mathbf{z}_{ttl}, \mathbf{z}_{img}$ are samples from classes C_a, C_b respectively, and n_a, n_b represent the number of samples in C_a, C_b respectively. Then, we use a distance threshold γ to calculate the similarity level between two classes C_a and C_b, *i.e.*,

$$s_{ab} = \frac{d_{C_a,C_b}}{\gamma}. \quad (10)$$

Finally, we employ a log-ratio loss function that encourages dissimilar embeddings to be pushed apart in proportion to their level of similarity [34]. Given triplet sample $\left\{\mathbf{z}_{ttl}^a, \mathbf{z}_{img}^p, \mathbf{z}_{img}^n\right\} \in \mathcal{S}$, the log-ratio loss can be defined as

$$\mathcal{L}_{hyp}(\mathbf{z}_{ttl}^a, \mathbf{z}_{img}^p, \mathbf{z}_{img}^n) = \left(\log\frac{\left\|\mathbf{z}_{ttl}^a - \mathbf{z}_{img}^p\right\|}{\left\|\mathbf{z}_{ttl}^a - \mathbf{z}_{img}^n\right\|} - \log\Omega^{s_{ij}-s_{ik}}\right)^2, \quad (11)$$

where superscripts a, p and n denote anchor, positives and negatives, respectively, and Ω is a hyperparameter that controls the degree of similarity. The variables i, j, and k indicate class indices to which the anchor, positive, and negative samples belong, and the similarity between classes C_i and C_j, denoted as s_{ij}, is calculated as shown in Equation (10). Next, the hyperbolic embedding loss between two embeddings is computed using Eq. (12):

$$\mathcal{L}_h(\mathcal{G}_{ttl}, \mathcal{G}_{img}) = \sum_{\mathbf{z}_{ttl}\in\mathcal{G}_{ttl}} \mathcal{L}_{hyp}(\mathbf{z}_{ttl}^a, \mathbf{z}_{img}^p, \mathbf{z}_{img}^n), \quad (12)$$

where the sets $\mathcal{G}_{ttl}$ and $\mathcal{G}_{img}$ represent the sets of title and image embeddings in the hyperbolic manifold. The final hyperbolic embedding loss is the average of the four loss functions, as shown in Eq. (13).

$$\begin{aligned}\mathcal{L}_{he} = \mathcal{L}_h(\mathcal{G}_{ttl}, \mathcal{G}_{img}) + \mathcal{L}_h(\mathcal{G}_{ing}, \mathcal{G}_{img}) \\ + \mathcal{L}_h(\mathcal{G}_{ins}, \mathcal{G}_{img}) + \mathcal{L}_h(\mathcal{G}_{rec}, \mathcal{G}_{img})\end{aligned} \quad (13)$$

3.4. Final Training Objective

Image-Text Matching (ITM) Loss: To facilitate the alignment process between the dual encoders, we incorporate multimodal regularization using Transformer decoders along with an Image-Text Matching loss (ITM) [2, 14]. ITM is a binary cross-entropy loss designed to classify the most suitable image-text pair. By optimizing this loss, we encourage the encoders to effectively match and align corresponding images and texts. The loss can be written as

$$\mathcal{L}_{itm} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)\right), \quad (14)$$

where n is the total number of positive and negative pairs of recipe and image embeddings, y denote the labels $\{0, 1\}$ indicating whether the image and text belong to the same pair, and $\hat{y}$ denotes the output of the Multimodal Regularization (MMR) layer. To construct negative pairs, we sample the closest negative image using cosine similarity.

Semantic Loss: We additionally employ a semantic triplet loss [1] so that the embeddings focus on semantically interesting features. The semantic loss $\mathcal{L}_{sem}$ is similar to $\mathcal{L}_{itc}$, except for the selection of positive and negative samples. Here, the positive samples are those that share the same class with the anchor and the negative samples are the samples with different classes.

Combined Final Loss: The final total loss function combines all the aforementioned losses together, *i.e.*,

$$\mathcal{L} = \lambda_{itc}\mathcal{L}_{itc} + \lambda_{he}\mathcal{L}_{he} + \lambda_{sem}\mathcal{L}_{sem} + \lambda_{itm}\mathcal{L}_{itm} \quad (15)$$

where λ_{itc}, λ_{he}, λ_{sem} and λ_{itm} are the weights for the corresponding losses. Additionally, to further improve model performance, we employ hard-negative sampling based on the multinomial probability distribution [20, 26]. More specifically, the negative samples for $\mathcal{L}_{itc}$ and $\mathcal{L}_{he}$ are selected such that they are closest to the positive samples.

4. Experimental Results

4.1. Dataset

We make use of the Recipe1M [23] dataset in our experiments. Recipe1M is a large-scale, structured corpus consisting of over 1 million cooking recipes and 800k food items. The dataset includes rich information about each recipe and

Table 1. Experimental results on image-to-recipe and recipe-to-image retrieval with 1k and 10k pairs on the Recipe1M [23] dataset. Best-performing method for each metric is highlighted in bold and the second-best is underlined.

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. [23]	5.2	24.0	51.0	65.0	5.1	25.0	52.0	65.0	41.9	-	-	-	39.2	-	-	-
Adamine [1]	2.0	40.2	68.1	78.7	2.0	39.8	69.0	77.4	13.2	14.8	34.6	46.1	14.2	14.9	35.3	45.2
R2GAN [36]	2.0	39.1	71.0	81.7	2.0	40.6	72.6	83.3	13.9	13.5	33.5	44.9	12.6	14.2	35.0	46.8
MCEN [6]	2.0	48.2	75.8	83.6	1.9	48.4	76.1	83.7	7.2	20.3	43.3	54.4	6.6	21.4	44.3	55.2
ACME [29]	1.0	51.8	80.2	87.5	1.0	52.8	80.2	87.6	6.7	22.9	46.8	57.9	6.0	24.4	47.9	59.0
SN [35]	1.0	52.7	81.7	88.9	1.0	54.1	81.8	88.9	7.0	22.1	45.9	56.9	7.0	23.4	47.3	57.9
IMHF [10]	1.0	53.2	80.7	87.6	1.0	54.1	82.4	88.2	6.2	23.4	48.2	58.4	5.8	24.9	48.3	59..4
Wang et. al [28]	1.0	53.5	81.5	88.8	1.0	55.0	82.0	88.8	6.0	23.4	48.8	60.1	5.6	24.6	50.0	61.0
SCAN [30]	1.0	54.0	81.7	88.8	1.0	54.9	81.9	89.0	5.9	23.7	49.3	60.6	5.1	25.3	50.6	61.6
HF-ICMA [11]	1.0	55.1	86.7	92.4	1.0	56.8	87.5	93.0	5.0	24.0	51.6	65.4	4.2	25.6	54.8	67.3
MSJE [32]	1.0	56.5	84.7	90.9	1.0	56.2	84.9	91.1	5.0	25.6	52.1	63.8	5.0	26.2	52.5	64.1
SEJE [33]	1.0	58.1	85.8	92.2	1.0	58.5	86.2	92.3	4.2	26.9	54.0	65.6	4.0	27.2	54.4	66.1
M-SIA [12]	1.0	59.3	86.3	92.6	1.0	59.8	86.7	92.8	4.0	29.2	55.0	66.2	4.0	30.3	55.6	66.5
X-MRS [7]	1.0	64.0	88.3	92.6	1.0	63.9	87.6	92.6	3.0	32.9	60.6	71.2	3.0	33.0	60.4	70.7
H-T [22]	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
H-T (ViT) [22]	1.0	64.2	89.1	93.4	1.0	64.5	89.3	**93.8**	3.0	33.5	62.1	72.8	3.0	33.7	62.2	72.7
T-Food (ViT) [24]	1.0	68.2	87.9	91.3	1.0	68.3	87.8	91.5	2.0	40.0	67.0	75.9	2.0	41.0	67.3	75.9
T-Food [24]	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
FARM (Ours)	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**

its accompanying image. In addition to the title, ingredients, and instructions, the recipe data includes the source URL and partitions (train, test, or validation), with 238,999, 51,119, and 51,303 pairs for training, validation and testing, respectively. For each 224×224 image, the dataset contains the image path, the class name (*e.g.*, blue cheese, buttermilk biscuits, chocolate, cheesecake *etc.*), and a class ID. A unique identifier allows for easy cross-referencing between images and their corresponding recipes.

4.2. Cross-Modal Retrieval

We evaluate FARM using the Recipe1M dataset [23] for image-to-recipe and recipe-to-image retrieval tasks in both 1k and 10k settings. In image-to-recipe retrieval, the task is to retrieve the correct recipe given an image, while in recipe-to-image retrieval, the goal is to retrieve the image given a recipe. We compare FARM against several state-of-the-art baselines. Following previous works, we report medR, which represents the median index of the retrieved samples. We also provide retrieval accuracy metrics R@1, R@5, and R@10, which indicate the number of correct retrieved items when considering only the top K samples.

In Table 1, we observe that FARM outperforms all baselines on the image-to-recipe retrieval task. More specifically, FARM outperforms the best baseline, TFood [24], by 1.4 and 1.5 percentage points on R@1 on the 1k and 10k sets, respectively. In the recipe-to-image retrieval task, FARM outperforms T-Food [24] by 1.0 percentage points on the 1k setup, but T-Food outperforms FARM by 0.3 percentage points on the 10k setup. This may indicate that the visual cues provided by the images alone are sufficient to establish meaningful associations, without the need for explicit alignment of recipe components. On the other hand, in image-to-recipe retrieval, FARM benefits from the rich textual cues and achieves better performance in aligning images with textual recipes. Notably, FARM demonstrates superior performance on the R@5 and R@10 metrics on both 1k and 10k settings in the recipe-to-image retrieval task.

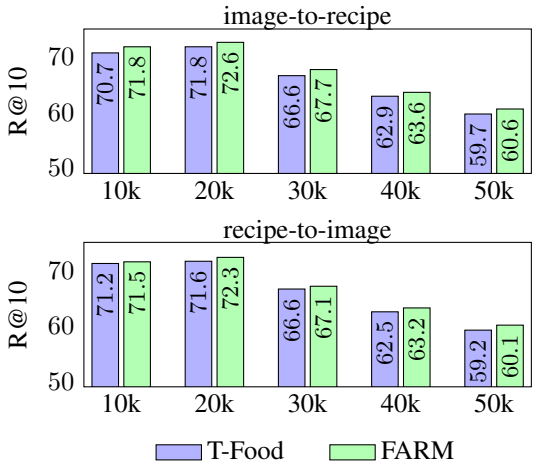

Figure 3. FARM scalability analysis.

To further illustrate scalability, we present a comparative analysis in Figure 3, where FARM's performance is compared against T-Food on settings larger than 10k. We observe that FARM consistently outperforms T-Food in both image-to-recipe and recipe-to-image tasks.

4.3. Ablation Studies

Loss Components: We present an ablation study comparing the contribution of various FARM loss components. Table 2 results reveal that the $\mathcal{L}_{sem}$ component significantly boosts performance in the 1k setting, while improvements in 10k are notable but smaller. This observation emphasizes our model's robustness and ability to extract semantic information from limited data. Additionally, $\mathcal{L}_{sem}$ is

Table 2. Ablation study on $\mathcal{L}_{sem}$ and $\mathcal{L}_{he}$ loss components, on image-to-recipe and recipe-to-image retrieval with 1k and 10k pairs.

Method	$\mathcal{L}_{sem}$	$\mathcal{L}_{he}$	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 ↑
FARM	✗	✗	1.0	71.8	89.6	92.7	1.0	70.9	89.3	92.5	2.0	43.9	71.1	79.7	2.0	43.7	71.0	79.6
	✓	✗	1.0	72.5	90.2	93.0	1.0	72.8	90.7	93.2	2.0	44.2	71.0	79.4	2.0	43.9	71.0	79.5
	✓	✓	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**

Table 3. Experimental results on the image-to-recipe and recipe-to-image retrieval on the 1k and 10k settings, when part of the recipe is missing. The best-performing method is highlighted in bold.

Experiment	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 ↑
Title	T-Food	84.5	3.4	10.3	15.8	12.0	16.2	36.6	48.0	810.9	0.8	2.5	4.1	114.4	4.5	12.7	18.6
	FARM (✗ $\mathcal{L}_{he}$)	75.8	3.9	11.9	18.11	**9.9**	18.0	**40.8**	**51.0**	769.6	0.9	2.9	4.5	**88.5**	**4.9**	**14.3**	**20.9**
	FARM	**39.8**	**6.8**	**18.5**	**26.6**	10.4	**18.2**	40.2	50.4	**390.9**	**1.4**	**4.7**	**7.4**	96.0	4.8	13.7	20.2
Ingredients	T-Food	3.6	32.1	57.9	68.5	2.0	44.7	71.5	80.3	28.0	11.4	27.3	36.1	10.2	19.7	40.6	50.6
	FARM (✗ $\mathcal{L}_{he}$)	**3.0**	**35.4**	62.3	72.2	2.0	**47.2**	73.1	80.7	19.8	13.2	30.9	40.6	**8.0**	**21.2**	**43.5**	**53.7**
	FARM	**3.0**	**35.4**	**64.3**	**75.3**	2.0	46.9	**74.0**	**81.9**	**18.4**	**13.6**	**31.4**	**41.2**	8.2	21.1	43.3	53.6
Instructions	T-Food	40.5	6.5	19.3	27.3	8.0	19.5	42.8	54.6	373.1	1.3	4.5	7.5	74.0	5.5	15.2	21.7
	FARM (✗ $\mathcal{L}_{he}$)	37.9	8.7	22.7	31.2	6.3	**22.9**	**48.0**	**59.3**	356.5	2.1	6.7	10.4	**49.6**	**7.4**	**19.6**	**27.5**
	FARM	**16.5**	**12.6**	**31.0**	**41.2**	**6.2**	22.7	47.4	57.5	**168.3**	**2.8**	**8.9**	**13.7**	59.0	6.8	18.0	25.5
✗ Title	T-Food	1.0	64.4	86.9	91.4	1.0	65.3	87.92	**91.8**	3.0	35.2	62.3	71.9	3.0	36.0	63.4	73.2
	FARM (✗ $\mathcal{L}_{he}$)	1.0	67.4	86.8	90.4	1.0	67.7	87.3	91.3	**2.0**	38.3	**65.3**	**74.6**	**2.0**	38.4	**65.9**	**75.2**
	FARM	1.0	**67.9**	**87.6**	**91.5**	1.0	**68.3**	**88.5**	**91.8**	2.2	**38.4**	65.1	74.4	**2.0**	**38.7**	**65.9**	75.1
✗ Ingredients	T-Food	6.0	21.2	48.2	61.2	2.4	37.7	67.4	76.6	48.2	5.5	16.1	24.4	15.8	14.0	32.7	43.4
	FARM (✗ $\mathcal{L}_{he}$)	4.8	26.5	52.5	63.6	**2.0**	**43.1**	**71.0**	**79.6**	38.6	9.0	22.3	30.7	**10.8**	**17.7**	**38.9**	**49.6**
	FARM	**2.8**	**36.2**	**64.7**	**74.7**	**2.0**	42.5	70.3	78.8	**18.4**	**12.1**	**29.9**	**49.4**	12.2	16.4	36.6	47.3
✗ Instructions	T-Food	1.0	56.1	80.3	86.7	1.0	61.7	84.1	89.0	5.0	27.4	52.3	62.6	3.2	31.8	58.7	68.6
	FARM (✗ $\mathcal{L}_{he}$)	1.0	**60.5**	83.3	88.2	1.0	**64.1**	85.2	**89.7**	**3.8**	**31.7**	**57.7**	**67.6**	**3.0**	**34.0**	**61.1**	**71.1**
	FARM	1.0	59.7	**83.4**	**88.6**	1.0	63.0	**85.3**	**89.7**	4.0	31.1	56.6	66.7	**3.0**	33.6	60.4	70.3

more effective on recipe-to-image retrieval, achieving 1.9% improvement (1k), compared to a modest 0.7% improvement on the image-to-recipe task. The relatively smaller impact of $\mathcal{L}_{sem}$ on image-to-recipe retrieval indicates that the richness and informativeness of visual features can often serve as dominant cues for retrieving textual recipe descriptions. Enhancing semantic alignment improves the ability to extract meaningful information from textual descriptions, leading to more accurate retrieval.

We also observe that the hyperbolic embedding loss significantly improves model performance on both image-to-recipe and recipe-to-image retrieval, outperforming FARM (w/o $\mathcal{L}_{he}$) across all metrics. In contrast to the observations for $\mathcal{L}_{sem}$, the impact of the hyperbolic loss is more pronounced in the image-to-recipe retrieval task compared to the recipe-to-image retrieval task. This suggests that the hyperbolic embedding loss effectively captures and leverages the inherent hierarchical relationships and similarities between recipes, making it easier to retrieve the most relevant textual descriptions for a given food image.

Missing Information: To demonstrate the effectiveness of our component-level alignment, we present experiments when one or more recipe components are missing (Table 3). For example, in the Title setup, only the title information is available, and ingredients and instructions are set to empty strings. Similarly, in the Title+Ingredient setup, the tile and ingredient are available, while the instructions are not. We compare T-Food with two versions of our proposed approach, with and without the hyperbolic loss, *i.e.*, FARM and FARM (w/o $\mathcal{L}_{he}$), respectively.

We observe that both versions of FARM outperform T-Food across all settings, proving the effectiveness of the proposed component-level alignment. In terms of R@1, in image-to-recipe 1k retrieval, performance improvements vary between $[3.3\%, 6.1\%]$ when only one component of the recipe is available, and between $[3.5\%, 15.1\%]$ when two components are available. Similarly, in the recipe-to-image 1k retrieval, performance improvements vary between $[2.0\%, 3.2\%]$ when one component is available, and between $[1.3\%, 4.8\%]$ when two components are available.

We also observe better performance on image-to-recipe retrieval. Between the two variants of our approach, FARM performs better than FARM (w/o $\mathcal{L}_{he}$) in the image-to-recipe retrieval task in most cases, however, FARM (w/o $\mathcal{L}_{he}$) sometimes performs better than FARM in the recipe-to-image retrieval task. We further observe that, among the different components of the recipe, the inclusion of the title seems to have the least effect on performance gains. This may point toward the need to design a better approach to encoding titles. On the other hand, ingredients seem to be the most useful component, with FARM achieving 35.4 R@1 in the only-ingredient setup, which is close to the combination of title and instructions (36.2 R@1).

4.4. Qualitative Analysis

We conduct qualitative analysis to understand FARM's semantic capabilities. In Figure 4, we present a t-SNE [27] comparison of the FARM and T-Food recipe embeddings. We observe that FARM achieves better separation, which

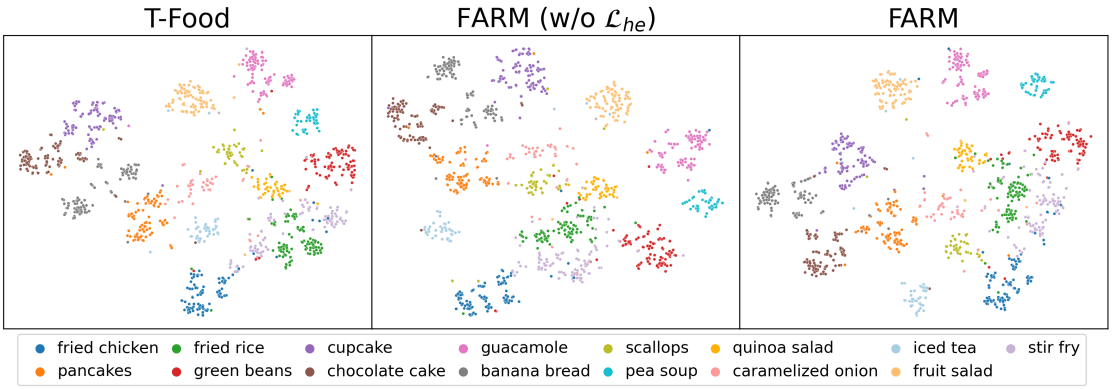

Figure 4. Recipe embeddings t-SNE visualization from 15 randomly sampled classes from the Recipe1M dataset.

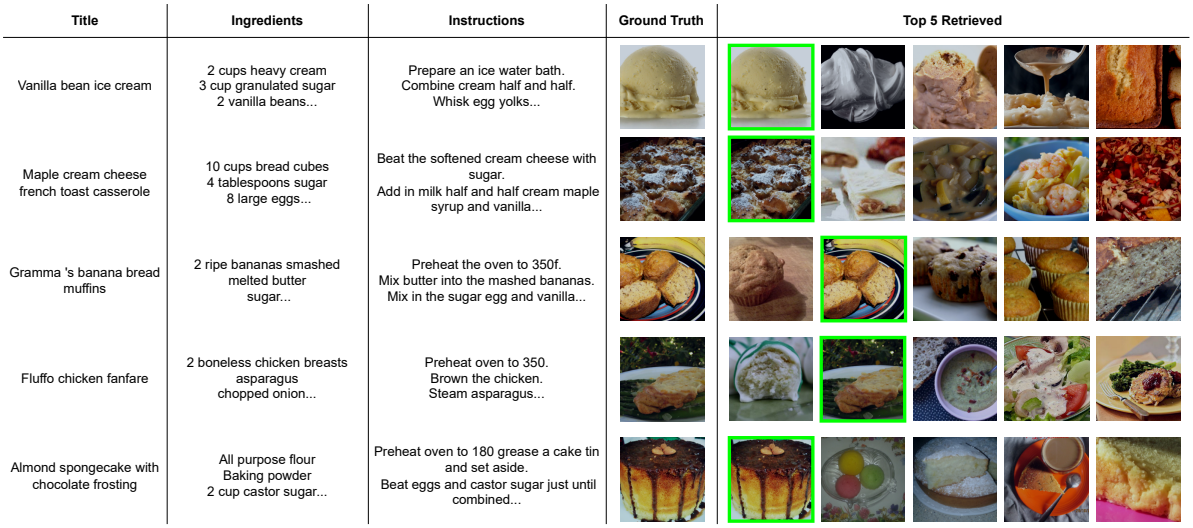

Figure 5. FARM qualitative analysis on recipe-to-image retrieval. Each row presents the top-5 retrieved images, alongside their associated textual components for each recipe.

leads to less overlap among the various classes. For example, 'stir fry' has a lot of overlap with fried rice in the recipe embeddings learned using T-Food. FARM reduces this overlap significantly. In Figure 5, we demonstrate a few retrieved samples in the image-to-recipe retrieval task. We observe that FARM retrieves semantically similar images for each of the recipes. We further observe that sometimes the results are similar to the ground truth, although they do not exactly look like the ground truth. For example, for 'banana bread muffin', FARM retrieves a different image of the bread muffin instead of the ground truth, while the actual ground truth is also in the top-3 results. This explains the lower R@1 compared to R@5.

5. Conclusion

In this work, we introduce FARM, a cross-modal fine-grained alignment framework for recipe retrieval that aligns individual recipe components (title, ingredients, and instructions) with image embeddings. FARM incorporates a hyperbolic loss that captures the hierarchical structure and relationships within the embedding space, leading to improved retrieval performance. Experimental results demonstrate improvements compared to the current state-of-the-art methods in both image-to-recipe and recipe-to-image retrieval tasks. FARM enhances robustness in handling missing information and consistently surpasses baselines when one or more recipe components are missing.

6. Acknowledgements

This work is supported by the Amazon – Virginia Tech Initiative for Efficient and Robust Machine Learning and the Amazon Alexa Prize TaskBot Challenge 2. Any findings or conclusions in this material are solely those of the authors and do not necessarily reflect views of the sponsors. The authors would also like to thank Dr. Thomas for the valuable discussions held during the Multimodal Vision class.

References

[1] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 35–44, 2018.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120, 2020.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[4] Mikhail Fain, Niall Twomey, Andrey Ponikar, Ryan Fox, and Danushka Bollegala. Dividing and conquering cross-modal recipe retrieval: from nearest neighbours baselines to SoTA. *arXiv:1911.12763*, 2019.

[5] Bahare Fatemi, Quentin Duval, Rohit Girdhar, Michal Drozdzal, and Adriana Romero-Soriano. Learning to substitute ingredients in recipes. *arXiv:2302.07960*, 2023.

[6] Han Fu, Rui Wu, Chenghao Liu, and Jianling Sun. MCEN: Bridging cross-modal gap between cooking recipes and dish images with latent variable model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14570–14580, 2020.

[7] Ricardo Guerrero, Hai X Pham, and Vladimir Pavlovic. Cross-modal retrieval and synthesis (X-MRS): Closing the modality gap in shared subspace learning. In *ACM International Conference on Multimedia*, pages 3192–3201, 2021.

[8] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[10] Jiao Li, Jialiang Sun, Xing Xu, Wei Yu, and Fumin Shen. Cross-modal image-recipe retrieval via intra-and inter-modality hybrid fusion. In *International Conference on Multimedia Retrieval*, pages 173–182, 2021.

[11] Jiao Li, Xing Xu, Wei Yu, Fumin Shen, Zuo Cao, Kai Zuo, and Heng Tao Shen. Hybrid fusion with intra-and cross-modality attention for image-recipe retrieval. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 244–254, 2021.

[12] Lin Li, Ming Li, Zichen Zan, Qing Xie, and Jianquan Liu. Multi-subspace implicit alignment for cross-modal retrieval on cooking recipes and food images. In *ACM International Conference on Information & Knowledge Management*, pages 3211–3215, 2021.

[13] Shuyang Li, Yufei Li, Jianmo Ni, and Julian McAuley. SHARE: A system for hierarchical assistive recipe editing. In *Conference on Empirical Methods in Natural Language Processing*, pages 11077–11090, 2022.

[14] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] Javier Marın, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203, 2021.

[16] Lei Meng, Fuli Feng, Xiangnan He, Xiaoyan Gao, and Tat-Seng Chua. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In *ACM International Conference on Multimedia*, pages 3460–3468, 2020.

[17] Hai X Pham, Ricardo Guerrero, Vladimir Pavlovic, and Jiatong Li. CHEF: Cross-modal hierarchical embeddings for food domain retrieval. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 2423–2430, 2021.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[19] John G Ratcliffe, S Axler, and KA Ribet. *Foundations of hyperbolic manifolds*, volume 149. Springer, 1994.

[20] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

[21] Amaia Salvador, Michal Drozdzal, Xavier Giró-i Nieto, and Adriana Romero. Inverse cooking: Recipe generation from food images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10453–10462, 2019.

[22] Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15475–15484, 2021.

[23] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3020–3028, 2017.

[24] Mustafa Shukor, Guillaume Couairon, Asya Grechka, and Matthieu Cord. Transformer decoders with multimodal regularization for cross-modal food retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4567–4578, 2022.

[25] Mustafa Shukor, Nicolas Thome, and Matthieu Cord. Structured vision-language pretraining for computational cooking. *arXiv:2212.04267*, 2022.

[26] Afrina Tabassum, Muntasir Wahed, Hoda Eldardiry, and Ismini Lourentzou. Hard negative sampling strategies for contrastive representation learning. *arXiv:2206.01197*, 2022.

[27] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

[28] Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan

Miao. Learning structural representations for recipe generation and food retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[29] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11572–11581, 2019.

[30] Hao Wang, Doyen Sahoo, Chenghao Liu, Ke Shu, Palakorn Achananuparp, Ee-peng Lim, and Steven CH Hoi. Cross-modal food retrieval: learning a joint embedding of food images and recipes with semantic consistency and attention mechanism. *IEEE Transactions on Multimedia*, 24:2515–2525, 2021.

[31] Wenjie Wang, Ling-Yu Duan, Hao Jiang, Peiguang Jing, Xuemeng Song, and Liqiang Nie. Market2Dish: health-aware food recommendation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(1):1–19, 2021.

[32] Zhongwei Xie, Ling Liu, Yanzhao Wu, Lin Li, and Luo Zhong. Learning tfidf enhanced joint embedding for recipe-image cross-modal retrieval service. *IEEE Transactions on Services Computing*, 2021.

[33] Zhongwei Xie, Ling Liu, Yanzhao Wu, Luo Zhong, and Lin Li. Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering. *ACM Transactions on Information Systems*, 40(4):1–27, 2021.

[34] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Unsupervised hyperbolic metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12465–12474, 2021.

[35] Zichen Zan, Lin Li, Jianquan Liu, and Dong Zhou. Sentence-based and noise-robust cross-modal retrieval on cooking recipes and food images. In *International Conference on Multimedia Retrieval*, pages 117–125, 2020.

[36] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. R2GAN: Cross-modal recipe retrieval with generative adversarial network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11477–11486, 2019.