

A. Implementation Details

For a fair comparison, we follow a similar hyperparameter setup as our baseline, T-Food [24]. While all Transformer models are trained from scratch, we initialize the CLIP-ViT encoder using CLIP weights. The image encoder remains frozen for the first 20 epochs and then all modules are trained using Adam optimizer [9] with a constant learning rate of $1e-5$ (except CLIP-ViT that has $1e-6$ learning rate). The triplet loss dynamic margin is initially set to $\alpha = 0.05$ and is incremented by 0.005 per epoch until it reaches 0.3. Furthermore, for the triplet loss, the associated sample from other modalities is considered positive, and all other samples are considered negatives. For \mathcal{L}_{sem} , the samples from the same class are considered positive, and all the other classes are considered negatives. Similarly, class information is used to define the similarity between samples for the hyperbolic loss function. For the fine-grained alignment loss functions \mathcal{L}_{itc} and \mathcal{L}_{he} , we set the weight equal to $\alpha = 1/m$, where $m = 4$ is the number of different pairs considered in the loss function. We use $\lambda_{itc} = \frac{1}{n_{itc}}$, where n_{itc} is the number of components in \mathcal{L}_{itc} . Similarly, we use $\lambda_{he} = \frac{1}{n_{he}}$, where n_{he} is the number of components in \mathcal{L}_{he} . For all other hyperparameters, we follow a similar setup as T-Food [24]. Unless otherwise specified, we train all models for 120 epochs, with a batch size of 100 using two NVIDIA A100 GPUs.

B. Ablation Studies

Intra-Modality Alignment Following H-T [22], we experiment with the intra-modality loss alignment in Table 4. In this experiment, we modify Eq. (6) to also include intra-modality alignment as shown in Eq. (16).

$$\begin{aligned} \mathcal{L}_{itc}^\dagger = & \mathcal{L}_c(\mathcal{G}_{ttl}, \mathcal{G}_{img}) + \mathcal{L}_c(\mathcal{G}_{ing}, \mathcal{G}_{img}) \\ & + \mathcal{L}_c(\mathcal{G}_{ins}, \mathcal{G}_{img}) + \mathcal{L}_c(\mathcal{G}_{rec}, \mathcal{G}_{img}) \\ & + \underbrace{\mathcal{L}_c(\mathcal{G}_{ttl}, \mathcal{G}_{ing}) + \mathcal{L}_c(\mathcal{G}_{ttl}, \mathcal{G}_{ins}) + \mathcal{L}_c(\mathcal{G}_{ing}, \mathcal{G}_{ins})}_{\text{intra-modality}}. \end{aligned} \quad (16)$$

We observe that intra-modality alignment loss hurts performance, with a 0.8 and 1.5 percentage point drop in R@1 for the image-to-recipe and recipe-to-image tasks, respectively.

Fine-Grained Alignment with Recipe Embedding We further conduct experiments where we align the recipe component embeddings with the recipe embedding, instead of the image embedding. Specifically, we modify Eq. (6) as follows:

$$\begin{aligned} \mathcal{L}_{itc}^\ddagger = & \mathcal{L}_c(\mathcal{G}_{ttl}, \mathcal{G}_{rec}) + \mathcal{L}_c(\mathcal{G}_{ing}, \mathcal{G}_{rec}) \\ & + \mathcal{L}_c(\mathcal{G}_{ins}, \mathcal{G}_{rec}) + \mathcal{L}_c(\mathcal{G}_{rec}, \mathcal{G}_{img}). \end{aligned} \quad (17)$$

We observe that alignment with image embedding performs significantly better than alignment with recipe em-

bedding, achieving 0.6 and 2.8 percentage points performance improvements on the image-to-recipe and recipe-to-image 1k tasks, respectively. Results further indicate that the fine-grained alignment with the image embedding is particularly important for the recipe-to-image retrieval task, possibly due to the component embeddings having additional alignment signals about the corresponding images.

Hyperbolic Loss without Fine-Grained Alignment We also experiment with aligning the recipe embeddings with the image embedding without the use of any fine-grained alignments in the hyperbolic space. More specifically, we modify Eq. (13) as follows:

$$\begin{aligned} \mathcal{L}_{he}^\pm = & \mathcal{L}_h(\mathcal{G}_{ttl}, \mathcal{G}_{img}) + \mathcal{L}_h(\mathcal{G}_{ing}, \mathcal{G}_{img}) \\ & + \mathcal{L}_h(\mathcal{G}_{ins}, \mathcal{G}_{img}) + \mathcal{L}_h(\mathcal{G}_{rec}, \mathcal{G}_{img}). \end{aligned} \quad (18)$$

In Table 6, we observe that hyperbolic loss is much more effective when using fine-grained alignment, achieving 1.7 and 2.3 percentage points on the image-to-recipe and recipe-to-image 1k tasks, compared to FARM (+ \mathcal{L}_{he}^\pm) where only the recipe embedding and image embedding are aligned in the hyperbolic space. This further shows the effectiveness of the proposed fine-grained alignment.

C. Qualitative Analysis

We provide qualitative analysis examples of retrieved results on recipe-to-image (Figure 6) and image-to-recipe (Figure 7) retrieval for both FARM and T-Food. In Figure 6 we observe that, while both methods retrieve the correct recipe image, all top-5 retrieved images by FARM are semantically related, consistently featuring noodles or baking. In contrast, T-Food often retrieves irrelevant images, suggesting a tendency to emphasize on peripheral recipe components such as incidental ingredients within the dish or subtext ('whole wheat', 'pasta', 'lemon', etc). These results demonstrate FARM's ability in associating textual recipe descriptions with visually aligned images. Similarly, in Figure 7, both FARM and T-Food retrieve the correct recipe based on the image, however T-Food results contain irrelevant recipes such as 'sweet potato fries' ranked top, 'bean sprout omelet', etc. In contrast, all retrieved items by FARM are semantically relevant and pertain to salmon recipes for the first example, or contain chicken and past for the second example, or shrimp for the third example. This observation highlights that FARM effectively leverages class information to improve the relevance and consistency of retrieval results.

Table 4. Ablation study on intra-modality loss alignment, on image-to-recipe and recipe-to-image retrieval tasks with 1k and 10k pairs. FARM with the proposed \mathcal{L}_{itc} (top row) outperforms a variation with intra-modality alignment (bottom row).

Method	$\mathcal{L}_{itc}^\dagger$	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 (w/o \mathcal{L}_{he})	✓	1.0	71.7	89.7	93.0	1.0	71.3	89.7	92.9	2.0	44.5	71.1	79.4	2.0	43.2	70.6	79.2
FARM (w/o \mathcal{L}_{he})	✗	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

Table 5. Ablation study on fine-grained alignment, on image-to-recipe and recipe-to-image retrieval tasks with 1k and 10k pairs. FARM with the proposed \mathcal{L}_{itc} (top row) outperforms a variant with fine-grained alignment with recipe instead of image embeddings (bottom row).

Method	$\mathcal{L}_{itc}^\dagger$	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 (w/o \mathcal{L}_{he})	✓	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
FARM (w/o \mathcal{L}_{he})	✗	1.0	71.9	89.9	92.9	1.0	70.0	89.8	92.5	2.0	44.0	70.8	79.1	2.0	41.9	69.6	78.4

Table 6. Ablation study on hyperbolic loss aligning only the image embedding with the recipe embedding, on image-to-recipe and recipe-to-image retrieval tasks with 1k and 10k pairs. FARM with the proposed \mathcal{L}_{he} (top row) outperforms its simplified counterpart without fine-grained alignment in the hyperbolic space (bottom row).

Method	\mathcal{L}_{he}	\mathcal{L}_{he}^\pm	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 (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
FARM	✗	✓	1.0	72.0	89.7	92.9	1.0	71.3	89.4	92.9	2.0	44.5	71.6	79.9	2.0	44.0	71.4	79.9

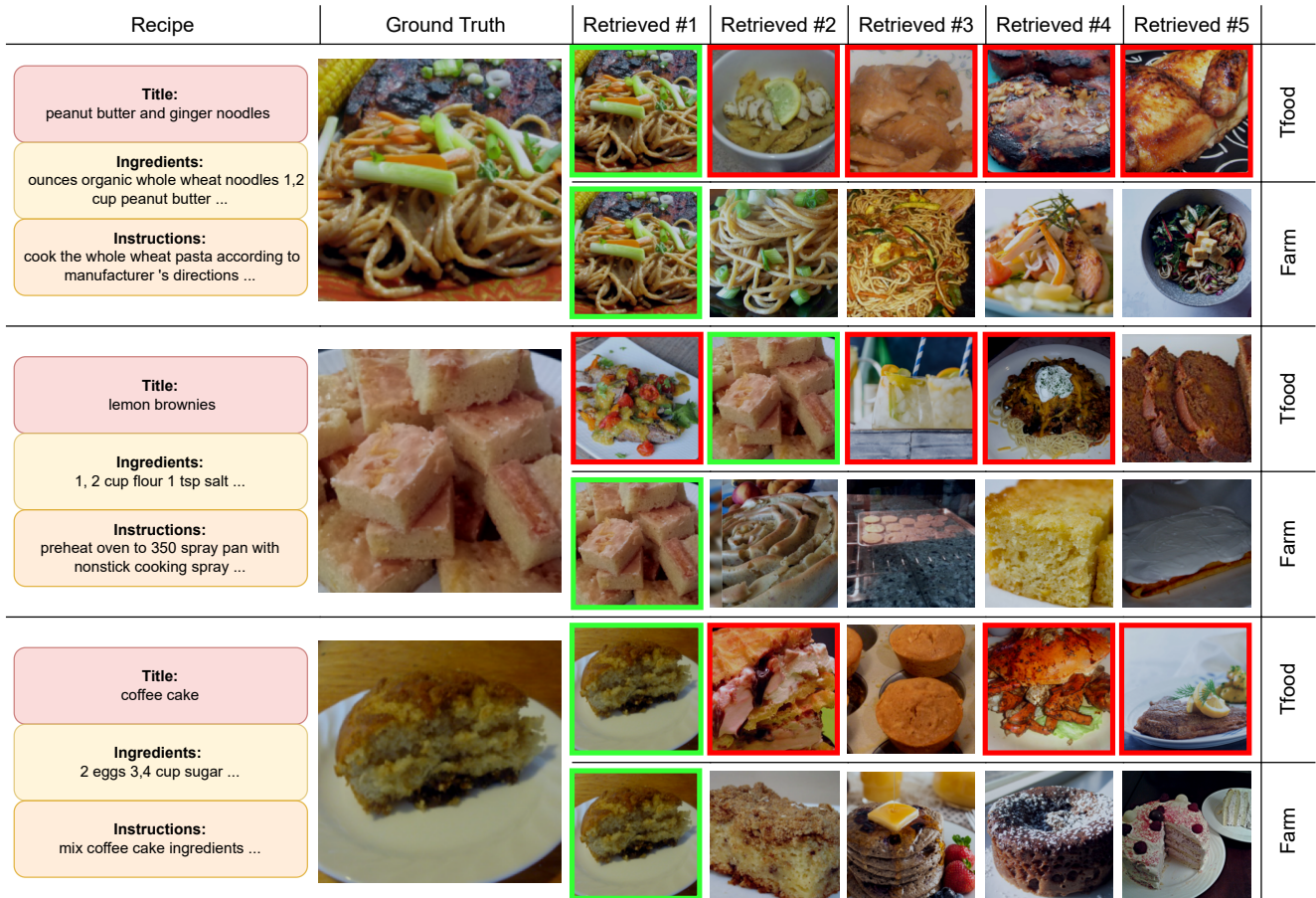


Figure 6. Qualitative examples of FARM on the recipe-to-image task. On each row, the top-5 retrieved images are shown for each recipe.




Recipe	Ground Truth	Retrieved #1	Retrieved #2	Retrieved #3	Retrieved #4	Retrieved #5	
	<p>Title: tandoori salmon</p> <p>Ingredients: 2 pieces salmon filets 1 tsp cooking sake ...</p> <p>Instructions: in a skillet saute the veggies in 2 3 tbs of cooking oil ...</p>	Title: sweet potato fries	Title: tandoori salmon	Title: seared salmon	Title: bean sprout omelet	Title: steamed swai fish filets	Tfood
		Ingredients: 3 bananas mashed ...	Ingredients: 2 pieces salmon filets ...	Ingredients: 4 salmon filets ...	Ingredients: 2 slice thinly sliced pork belly ...	Ingredients: 2 swai filets ...	
		Instructions: cook the beans with the water till tender ...	Instructions: heat oil in skillet ...	Instructions: peel off the skin from the cucumber ...	Instructions: in a large skillet saute the veggies ...	Instructions: combine all ingredients except the cheese ...	
		Tfood	Title: tandoori salmon	Title: blackened salmon	Title: baked salmon	Title: salmon with bourbon	Title: one pan salmon
			Ingredients: 2 pieces salmon filets ...	Ingredients: 4 whole filets ...	Ingredients: 1 piece salmon ...	Ingredients: 4 tablespoons butter ...	Ingredients: 4 salmon filets ...
			Instructions: in a skillet saute the veggies ...	Instructions: preheat oven to 325 ...	Instructions: preheat oven to 325 ...	Instructions: using a mortar and pestle mash the garlic ...	Instructions: rub sugar on both sides of the chicken breast ...
	<p>Title: smoky spicy tomatillo salsa</p> <p>Ingredients: 2 pieces salmon filets ...</p> <p>Instructions: combine onion vinegar ...</p>	Title: smoky spicy tomatillo salsa	Title: pressure cooker double dhal	Title: curried beef	Title: grilled zucchini hummus	Title: simple genovese sauce	Tfood
		Ingredients: 2 pieces salmon filets ...	Ingredients: 1,2 cups channa dal ...	Ingredients: 2 tablespoons vegetable oil ...	Ingredients: 1 lb zucchini ...	Ingredients: 30 grams fresh basil ...	
		Instructions: combine onion vinegar ...	Instructions: measure 1 cup beet juice ...	Instructions: cut avocados in half ...	Instructions: fry one finely chopped onion ...	Instructions: measure 1 cup beet juice ...	
		Tfood	Title: simple spinach dip	Title: smoky spicy tomatillo salsa	Title: anchovy salad dressing	Title: cherry avocado smoothie	Title: black bean & poblano dip
			Ingredients: 4 cups baby spinach ...	Ingredients: 1 lb tomatillo cut into quarters ...	Ingredients: 6 tablespoons olive oil ...	Ingredients: 1 avocado without skin ...	Ingredients: 2 cups poblano chiles ...
			Instructions: cut pork into 8 pieces ...	Instructions: heat the oil over moderate heat ...	Instructions: pour milk into blender ...	Instructions: trim and discard the ends of the eggplant ...	Instructions: heat the oil over moderate heat ...
	<p>Title: shrimp in garlic sauce</p> <p>Ingredients: 1, 3 cup olive oil ...</p> <p>Instructions: season with salt and pepper ...</p>	Title: shrimp in garlic sauce	Title: simple rice pilaf	Title: mild teriyaki hamburger steak	Title: chewy oatmeal cookies	Title: corn salad	Tfood
		Ingredients: 1, 3 cup olive oil ...	Ingredients: 1 2 sticks unsalted butter ...	Ingredients: 1, 2 cups butter softened...	Ingredients: 4 tpsn water ...	Ingredients: 3 cups cooked corn kernels ...	
		Instructions: season with salt and pepper ...	Instructions: preheat oven to 325 degrees ...	Instructions: mix together honey mustard and cayenne spread ...	Instructions: mix together all of the ingredients ...	Instructions: combine all ingredients in large non stick skillet ...	
		Tfood	Title: shrimp in garlic sauce	Title: poached shrimp with thai basil	Title: spicy baked shrimp	Title: shrimp & bacon pasta	Title: shrimp scampi
			Ingredients: 1, 3 cup olive oil ...	Ingredients: kosher salt as needed...	Ingredients: vegetable cooking spray ...	Ingredients: salt and pepper ...	Ingredients: 1 env . good seasons garlic ...
			Instructions: season with salt and pepper ...	Instructions: combine pie filling pecans and cinnamon ...	Instructions: combine pie filling pecans and cinnamon ...	Instructions: rub sugar on both sides of the chicken breast ...	Instructions: stir white chocolate and pineapple juice ...

Figure 7. Qualitative analysis of FARM on the image-to-recipe task. On each row, the top-5 retrieved recipes are shown for each image.