Retrieval Augmented Recipe Generation

Guoshan Liu[1,2]*, Hailong Yin[1,2]*, Bin Zhu[3], Jingjing Chen[1,2]†, Chong-Wah Ngo[3], Yu-Gang Jiang[1,2]
[1]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]Singapore Management University
{gsliu24, hlyin23}@m.fudan.edu.cn, {chenjingjing, ygj}@fudan.edu.cn
{binzhu, cwngo}@smu.edu.sg

Abstract

The growing interest in generating recipes from food images has drawn substantial research attention in recent years. Existing works for recipe generation primarily utilize a two-stage training method—first predicting ingredients from a food image and then generating instructions from both the image and ingredients. Large Multi-modal Models (LMMs), which have achieved notable success across a variety of vision and language tasks, shed light on generating both ingredients and instructions directly from images. Nevertheless, LMMs still face the common issue of hallucinations during recipe generation, leading to suboptimal performance. To tackle this issue, we propose a retrieval augmented large multimodal model for recipe generation. We first introduce Stochastic Diversified Retrieval Augmentation (SDRA) to retrieve recipes semantically related to the image from an existing datastore as a supplement, integrating them into the prompt to add diverse and rich context to the input image. Additionally, Self-Consistency Ensemble Voting mechanism is proposed to determine the most confident prediction recipes as the final output. It calculates the consistency among generated recipe candidates, which use different retrieval recipes as context for generation. Extensive experiments validate the effectiveness of our proposed method, which demonstrates state-of-the-art (SOTA) performance in recipe generation on the Recipe1M dataset.

1. Introduction

With the rising focus on food and health, food computing [49] has increasingly captured attention and spurred various food related tasks, such as food recognition [3, 7, 12, 22, 28, 30, 43, 48, 59, 68, 75], cross-modal recipe retrieval [6, 51, 57, 66, 76, 77], recipe generation [11, 46, 56, 64, 65], food recommendation [17, 19, 60] and food logging [13, 55]. Previous research on food understanding has primarily focused on classifying food and ingredient recognition [42, 45, 54, 73]. However, due to the limited availability of detailed information on prepared foods, a comprehensive visual food recognition system should not only be able to identify the type of diet or its ingredients but also generate cooking instructions. Therefore, the task of recipe generation has become a significant task in the field of food computing.

Previous methods for recipe generation [9, 47, 56] typically use a two-stage approach: first extracting ingredients from images, then generating instructions based on the embeddings of those ingredients and the images, which is shown in Figure 1 (a). Due to limited training data and poor multi-modal alignment, traditional methods often yield unsatisfactory results. In contrast, Multi-modal Models (LMMs) [1, 2, 8, 14, 44] can directly generate recipes from images in one stage. FoodLMM [70] improved recipe generation performance but still suffers from hallucinations, affecting recipe quality. Figure 1 (b) compares the recipe generation results of different methods. Compared to the ground truth instructions, the results predicted by one of the state-of-the-art LMMs, LLaVA [44], exhibit clear hallucinations, as it incorrectly identifies crumbs as 'beef' and erroneously detects 'tomatoes' and 'taco seasoning' that are not present in the image. Although the two-stage method [56] accurately identifies the correct temperature, it fails to precisely recognize the ingredients. FoodLMM [70] manages to identify most of the correct ingredients but still hallucinated, mistakenly recognizing 'rice'. The result arises due to inadequate multi-modal understanding and a lack of effective use of context, which prevents the models from learning sufficient information.

This paper addresses the limitation by introducing the first retrieval-augmented large multimodal model to generate recipes from food images. The proposed architecture consists of a retriever and a generator. The retriever leverages an off-the-shelf cross-modal recipe retrieval model [57] to identify semantically similar recipes from the image. The generator is built upon LLaVA [44] with LoRA [26] to

*Equal contribution.
†Jingjing Chen is the corresponding author.

Figure 1. (a) The structural differences between our retrieval-augmented framework and the "two-stage" [11, 56] and "LMMs-based" [44, 70] approaches. "G" refers to the generator. (b) Recipe generation results comparison. "GT" refers to ground truth, "LLaVA-FT" denotes the model using pre-trained LLaVA weights fine-tuned on Recipe1M, "Inverse cooking" [56] represents a model trained with two-stage, "FoodLMM" [70] is the LMMs-based model for recipe generation, and "Ours" refers to our model, where yellow highlights indicate ingredients that match those in the "GT", blue signifies cooking instructions predictions matching the "GT", and red font denotes incorrectly predicted ingredients.

generate recipes based on the image and retrieved reference recipes. On the one hand, we propose Stochastic Diversified Retrieval Augmentation (SDRA) to provide a rich and diverse set of retrieved recipes as references, which could provide relevant knowledge for the generation to reduce hallucination [24, 29, 52]. Section 2 of the supplementary materials compares retrieved recipes with the ground truth, highlighting related content that helps alleviate hallucinations and improve generation. On the other hand, we propose Self-consistency Ensemble Voting strategy to improve generation quality during the inference phase. Specifically, the generator produces multiple recipe candidates using different retrieved recipes and the image. Cosine similarity scores are computed for these candidates to assess their mutual agreement. The recipe with the highest score relative to the others is selected as the final output. The consensus among different candidates is capable of further reducing the hallucination in the generated recipes, as detailed in Section 5 of the supplementary materials, which explains why the final output is superior. On Recipe1M dataset [66], our proposed model significantly outperforms existing methods and fine-tuned LLaVA. Moreover, our model demonstrates strong generalizability, surpassing the state-of-the-art (SOTA) benchmarks on the Recipe1M dataset for ingredient recognition metrics.

Overall, our main contributions can be summarized as follows:

- We propose the first retrieval-augmented large multimodal model tailored for recipe generation. Our model introduces a stochastic diversified retrieval-augmentation technique to enhance the diversity and richness of retrieval. Additionally, we employ a Self-consistency Ensemble Voting strategy during the inference stage, using different retrieved recipes to ensure agreement and consistency in the generated recipe.
- Our proposed model achieves SOTA recipe generation performance on Recipe1M dataset. We conduct comprehensive ablation studies to validate the effectiveness of our design choices and demonstrate the contributions of each component to the overall performance. Our proposed model exhibits exceptional adaptability, outperforming current SOTA results in ingredient recognition on the Recipe1M dataset.

2. Related work

2.1. Recipe Generation

The significance of food and the accessibility of comprehensive food datasets, including Recipe1M [66], Vireo Food-172 [6] and Food2K [50], have facilitated computational studies in the domain of food-related computing tasks [49]. Recipe generation poses a significant challenge due to the presence of multiple sentences in cooking instructions. It entails the complex task of generating food recipes based on provided food images [25, 64, 65]. Accurate recipe generation requires understanding food components, images, and processes. Early methods generated ingredients from images first, then instructions from these ingredients and images. [56] used transformers for recipe generation but missed some steps due to a lack of comprehensive structure. Previous efforts did not use Large Multimodal Models (LMMs). Recently, [70] fine-tuned the LMM LISA [33] using multiple datasets, creating the first unified food computing model, including recipe generation. However, FoodLMM still suffers from hallucination issues. This paper aims to mitigate these issues with retrieval augmentation.

Q_{title}: Can you predict the food category of this image?
A_{title}: The food is…
$Q_{ingredients}$: Can you list the ingredients present in this dish?
$A_{ingredients}$: The ingredients are:…
$Q_{instructions}$: Can you provide the preparation instructions for this image?
$A_{instructions}$: Here are the instructions:…

Figure 2. Templates for Recipe Generation.

2.2. Vision-language Multimodal Models

Due to the increasing demand for versatile deep learning models, various large pre-trained models like BERT [15], ViT [16], and GPT [67] have emerged. However, their single-modality limits generalization, leading to the development of multimodal models. Autoregressive language models are now popular for vision-language tasks [1, 2, 4, 31, 38, 39, 41, 71, 72]. For example, LLaVA [44] integrates visual encoder output with LLaMA [61] using synthetic data, while Vicuna [74] uses LLaMA for conversational interactions. The rise of LMMs has expanded their application in various domains. However, hallucination remains an issue. [40] treats it as a binary classification problem, and [20] uses models to generate data for annotators to identify hallucinations. Our model is built upon LLaVA for recipe generation, equipped with a diversified stochastic retrieval augmentation, that boosts the recipe generation capabilities by introducing additional relevant contextual information, enabling the model to learn more specialized and comprehensive information.

2.3. Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) enhances language models (LMs) by incorporating knowledge from an external datastore [35]. This involves fetching relevant documents from external storage to improve the LM's predictive accuracy [5, 21, 24, 29, 36, 52]. While RAG is popular in NLP tasks [34], it is less explored in multimodal models [23, 32]. Some relevant works in retrieval-augmented multimodal language modeling focus on caption generation based on the encoded input image, as well as a collection of retrieved texts which are used as a task demonstration, input to the decoder as a prompt [53]. Recent works [27, 69] train generators with external multimodal information. While RAG enhances language models by incorporating external data, its use in vertical domains is limited [37, 62]. We propose the first Retrieval-Augmented LMMs for recipe generation using a Stochastic Diversified Retrieval Augmentation (SDRA) method. This method employs a pretrained retriever to fetch diverse recipes as supplemental inputs, improving LMM capabilities in recipe generation.

3. Method

3.1. Preliminary

3.1.1 Retrieval Augmented Generation (RAG) Models

RAG models [24, 52, 69] are usually composed of a retrieval model R and a generator (language model) G. The retriever R, based on the query (input sequence) $x_1, ..., x_n$, vectorizes the query to tokens and searches the Top K documents with the highest similarity to the query sequence within the datastore M, denoted as $M' = (m_1, ..., m_K)$. M' is then concatenated with the query to form a context-rich prompt, which is fed into the language model to generate enhanced outputs. Next-token prediction is widely adopted in language models. During the training process, the autoregressive model aims to maximize the conditional probability of the next word given the preceding sequence of words, namely, by optimizing the parameters θ to maximize:

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p_\theta(x_i | x_{<i}), \tag{1}$$

where $x_{<i}$ is the sequence of tokens preceding x_i. Retrieval-augmented language models make predictions conditionally based on the retrieved documents M'. Specifically, by simply merging the retrieved documents into the input query to predict the continuation of the input:

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p_\theta(x_i | [\mathcal{M}'(x_{<i}); x_{<i}]), \tag{2}$$

where $[a; b]$ denotes the concatenation of strings a and b.

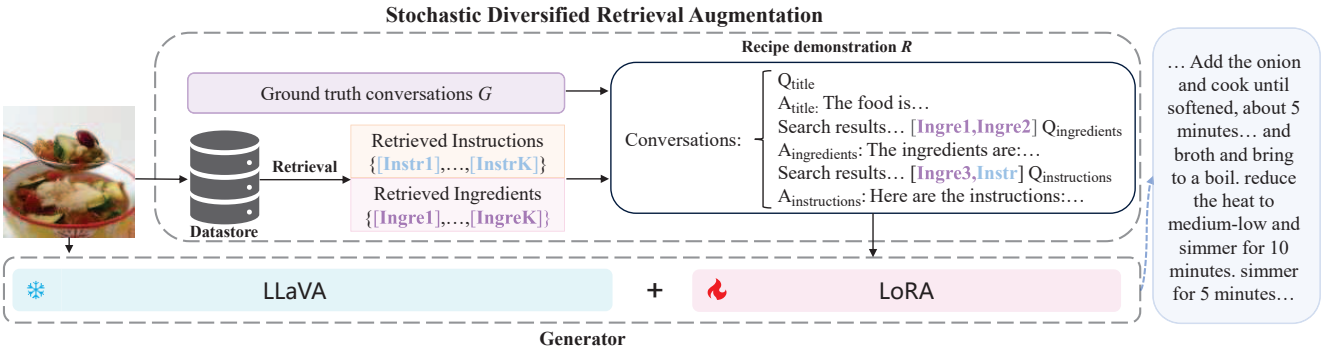
3.1.2 Data Organization Strategy

We organize image-recipe pairs as dialogues, with each image linked to three question-answer pairs (titles, ingredients, instructions). In the retrieval model, the image x is the query; in the generator, questions about titles, ingredients, and instructions serve as queries. During the training process, the model formats each multimodal document as "[<Image> Conversations: Q_{titles}, A_{titles}, $Q_{ingredients}$, $A_{ingredients}$, $Q_{instructions}$, $A_{instructions}$]", which Q_{titles} denotes the title's questions and A_{titles} denotes the title's answers of this image, and similarly for ingredients and instructions. We define these dialogues as ground truth conversations G. See Figure 2 for more details of the templates for our task.

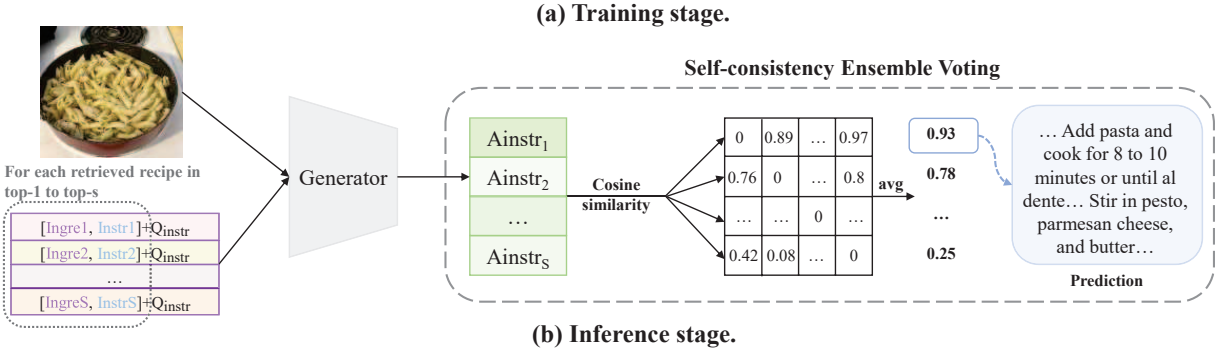
3.2. Retrieval Augmented Recipe Generation

3.2.1 Stochastic Diversified Retrieval Augmentation

As depicted in Figure 3 (a), we utilize image-to-recipe retrieval model [57], to retrieve the top K most similar ingredients and instructions from the data storage to the image x. Unlike existing retrieval augmentation methods [29, 35, 53]



(a) Training stage.



(b) Inference stage.

Figure 3. Overview of our proposed model architecture. Our model consists of a retriever to search semantically similar recipes from the image as reference, and a generator based on a frozen LLaVA [44] with a trainable LoRA [26] to generate recipe with the image and retrieved recipes. **Stochastic Diversified Retrieval Augmentation** is introduced by using retrieved ingredients and instructions, to form Recipe demonstration R, and fed into the generator for training. **Self-consistency Ensemble Voting** is proposed to select the final recipe output based on mutual agreement among the recipe candidates, which are produced by using each recipe from top 1 to top s retrieved recipes as context.

which directly use the retrieved results as context, to ensure the diversity of retrieval information, we randomly sample from the top K retrieval results, selecting three sets of ingredients and one set of instructions as the final retrieval information, as shown in Figure 3 (a). 1 set of ingredients refers to the ingredients part retrieved from an image, specifically, such as 'oil, egg, milk, vanilla'. Two sets of retrieved ingredients are concatenated in sequence before $Q_{ingredients}$, and the instructions and remaining set of retrieved ingredients are concatenated in sequence before $Q_{instructions}$ while informing the model like this that it is a reference result: "Search results for reference is 'retrieved information'. The search results are only for referring, please focus on the image." Finally, they are normalized as recipe demonstration R according to the prompt in Figure 3 (a). Here, the prompts "The food is", "The ingredients are:" and "Here are the instructions:" are similar to the simple, fixed prompts used in other research [39].

3.2.2 Recipe Generation with Retrieval Recipes

Our generation model is built upon a Large Multi-modal Model LLaVA [44] which takes image and text prompts as input. The text prompt (i.e., questions) and the retrieved recipes are concatenated and processed through a tokenizer, then fed into a text encoder to obtain textual features, while the image is processed through an image encoder to obtain image features. The image features are further mapped into the same embedding space as text features via a MLP. Upon receiving both text and image embeddings, the decoder proceeds to produce caption tokens, which are contingent on the image features X and the recipe demonstration R. To reduce the compute requirements for training and to preserve their generalization capabilities, we freeze the generator LLaVA model and only train its patch LoRA [26], which allows the focus to then be on fine-tuning specific, smaller aspects of the model to adapt to recipe generation without the need for extensive computation. The model is trained by minimizing the cross-entropy loss for next token prediction as follows:

$$L_\theta = -\sum_{i=1}^{n} p_\theta(y_i|[X; R; y_{<i}]; \theta), \tag{3}$$

where n is the index of the current tokens, and $y_{<i}$ denotes represents the tokens in the sequence before position i.

3.3. Self-consistency Ensemble Voting

To further improve the quality of recipe generation, we propose self-consistency ensemble voting. As shown in Figure 3 (b), we first retrieve the Top S sets of ingredients and instructions from the food image during the testing phase. Each set is then concatenated before $Q_{instructions}$ in the model input, following the Recipe denomination R

in Figure 3 (a), which concatenates the retrieved data in front of $Q_{instructions}$. The S retrieved sets are sequentially concatenated before $Q_{instructions}$ and input into the model, producing S different outputs. This process generates S different recipes, we note that the generated recipes could be inconsistent with different retrieved recipes as context. As a result, we introduce a score-based ensemble voting method, which selects the best recipe from multiple predictions to maintain self-consistency and improve the quality of the generated recipes. Specifically, denote the generated recipes as $P = \{P_1, ..., P_S\}$ by employing top 1 to top S as retrieved recipes, where S is the number of recipes used for inference. We compute the cosine similarity, BLEU, SacreBLEU scores, or ROUGE L scores among these S recipes, producing a $S \times S$ matrix where each row represents the agreement of all other predictions with the current prediction (excluding diagonal elements). By averaging the sum of agreements in each row, we obtain a confidence score for each recipe. Then, taking the calculation of confidence scores using cosine similarity as an example, we select the recipe with the highest confidence score as our final output as follows:

$$P_{\text{best}} = \arg\max_i \left(\frac{1}{S-1} \sum_{j=1, j\neq i}^{S} \frac{P_i \cdot P_j}{\|P_i\| \|P_j\|} \right), \quad (4)$$

where $\frac{P_i \cdot P_j}{\|P_i\| \|P_j\|}$ refers to cosine similarity between two predictions.

4. Experiments

4.1. Experimental Settings

4.1.1 Dataset and Evaluation Metrics

Following prior works [11, 70], we use the recipes with images in Recipe1M dataset [66] for experiments, including 571,587 pairs for training and 51,304 for testing. In this dataset, each image is associated with a title, ingredients, and instructions. During training, each image corresponds to three question-answer pairs as mentioned in Section 3.1. Ingredients are clustered into 1,488 categories, and quantifiers are removed, in line with [56]. For testing, the first image of each recipe is used, formatted to one question-answer pair per image, to separately test titles, ingredients, and instructions. Generating long recipes word by word is time-consuming, so we randomly select 5,000 samples from test set and fix the seed for all experiments. Similar to [69], we use the Recipe1M training set as our external datastore M to ensure consistency and fairness, avoiding external data.

Following existing works [11, 65], we utilize F1 score and Intersection Over Union (IOU) to evaluate the quality of ingredients, as well as document-level evaluation metrics, specifically BLEU, SacreBLEU, and RougeL, to evaluate the quality of instructions in the generated recipes.

Table 1. Recipe generation performance comparison. "LLaVA-FT" refers to fine-tuned LLaVA model.

Methods	BLEU	SacreBLEU	ROUGE L
Chef Transformer [18]	18.08	4.61	17.54
InverseCooking [56]	7.23	5.48	19.47
TIRG [63]	7.95	—	32.4
VAL [10]	8.83	—	34.20
SGN [65]	12.75	—	36.90
FIRE [11]	—	6.02	21.29
FoodLMM [70]	27.86	6.24	36.96
LLaVA-FT [44]	28.32	5.88	38.18
Ours	**30.11**	**6.42**	**38.93**

4.1.2 Implementation

In our retrieval module R, we use the revamping cross-modal recipe retrieval model [57] based on Transformers. Our generator G utilizes LLaVA [44] augmented with a LoRA [26] patch. During training, we utilize LoRA to fine-tune LLaVa based on the PyTorch framework, employing the pre-trained weights from LISA-7B-v1-explanatory [33]. This process is carried out on four NVIDIA 80G A100 GPUs. For the baseline, which we named "LLaVA-FT," we fine-tuned LLaVA using LoRA without retrieval augmentation, utilizing the same model architecture, training data, and computational resources to ensure a fair comparison.

4.1.3 Training and Inference

During training, we retrieve the top 50 recipes for each image, including titles, ingredients, and instructions. From these, we randomly select three sets of ingredients and one set of instructions to concatenate before $Q_{ingredients}$ and $Q_{instructions}$ as described in Section 3.2. The generator's max sequence length of 4096 encompasses all the information. We optimize the token prediction loss over the entire sequence (Equation 3). Given the strong performance of our cross-modal recipe retriever [57], which uses a transformer to encode recipe components, we keep the retriever constant and focus on training the generator. Future research could explore co-training or fine-tuning the retriever.

During inference, we first use the images from the test set as queries to retrieve from the 571,587 instances in the train set, with the retrieval model and method as mentioned in Section 3.2. Then, using the retrieved instructions and ingredients, we concatenate them in front of $Q_{instructions}$ in the same pattern as during training. For Self-consistency Ensemble Voting, we sequentially use the $tops = \{top1, ..., tops\}$ retrieved information, adding it before $Q_{instructions}$ and then feeding the recipe demonstration R in the form of "[<Image> Conversations: Search results...[Ingre, Instr] $Q_{instructions}$, $A_{instructions}$]" into the generator to obtain s prediction results. The final prediction is made using the Self-consistency Ensemble Voting method.

LLaVA-FT: Preheat oven to 350 degrees…with pork and kimchi in it. mix well. place in a 2 quart casserole dish sprinkle with cheese. Bake for 1 hour. Sprinkle with cheese. Bake for 1 hour.

Inverse cooking: Preheat oven to 350 degrees f…Combine hash browns, sour cream, soup, milk, onion, salt and pepper…Sprinkle with cheddar cheese. Bake in preheated oven for 30 minutes, or until cheese is melted and bubbly.

FoodLMM: Preheat oven to 400 degrees…Combine hash browns, sour cream, soup, milk, onion, salt and pepper…Sprinkle with cheddar cheese. Bake in preheated oven for 30 minutes, or until cheese is melted and bubbly.

Ours: preheat oven to 350 degrees…Bring milk, water, and butter to a boil; mix in the potato …Add corn and…half of the cheese and onions over it…the rest of the potatoes…Bake for 10 to 15 minutes until the cheese has melted.

GT: Preheat oven to 350 degrees…Heat milk, water and butter to boiling; stir in contents of both pouches of potatoes…Stir in corn. Spoon half the potato mixture in pan. Sprinkle half each of cheese and onions; top with remaining potatoes…Bake 10 to 15 minutes until cheese is melted. Enjoy !

(a)

LLaVA-FT: …use white frosting and black decorating icing…Add candy corn to the top of each cupcake…beat the eggs and mix in the melted butter, milk, and vanilla extract. Add the vanilla extract and beat until…

Inverse cooking: Preheat oven to 350f…Sift together flour, cocoa, baking powder and salt…Add flour mixture alternately with chocolate, beating until just blended after each addition…Frost cupcakes with white frosting.

FoodLMM: Cream butter and sugar. Add eggs and vanilla. Mix in flour, salt, baking powder, and cocoa. Mix in milk and nuts. Bake at 350 degrees for 15 minutes. Frost with chocolate frosting.

Ours: prepare cupcakes...cut licorice into 3 inch sections...frost the cupcakes with the white frosting…to make the legs of the spider…place two pieces of candy corn on the front of the cupcake for fangs and use two red hots as eyes...

GT: prepare cupcakes…cut licorice into 3 inch pieces…frost the cupcakes with the white frosting. insert licorice pieces into the outer edges of the cupcakes to make the legs of the spider…arrange two pieces of candy corn on the front of the cupcake for fangs and use two red hots as eyes…

(b)

Figure 4. Qualitative results. The ingredients in generated recipes that overlap with ground truth ("GT") are highlighted in yellow, while details in the instructions that match the GT are shown in blue. Otherwise, the incorrect generation results are displayed in red. Best viewed in color.

The input food image 'tuna and tomato cream pasta sauce'.

Ground truth:
-Mince the garlic, and saute in olive oil until fragrant.
-Season with salt and pepper to taste…crush the tomatoes while…
-You can use this sauce not only for pasta, but as sauce for omelets, meat or fish dishes.

Prediction:
-Cook the spaghetti in a large pot of boiling salted water until al dente…add the onion and garlic and cook…
-Add the tomatoes, basil, salt and pepper and cook…
-Add the cooked spaghetti and toss to coat.

Figure 5. Comparison between generated recipes and GT recipes. The highlights in yellow indicate ingredients that match those in the GT, ingredients incorrectly identified by the model are signified in red.

4.2. Performance Comparison

4.2.1 Quantitative Comparison of Recipe Generation

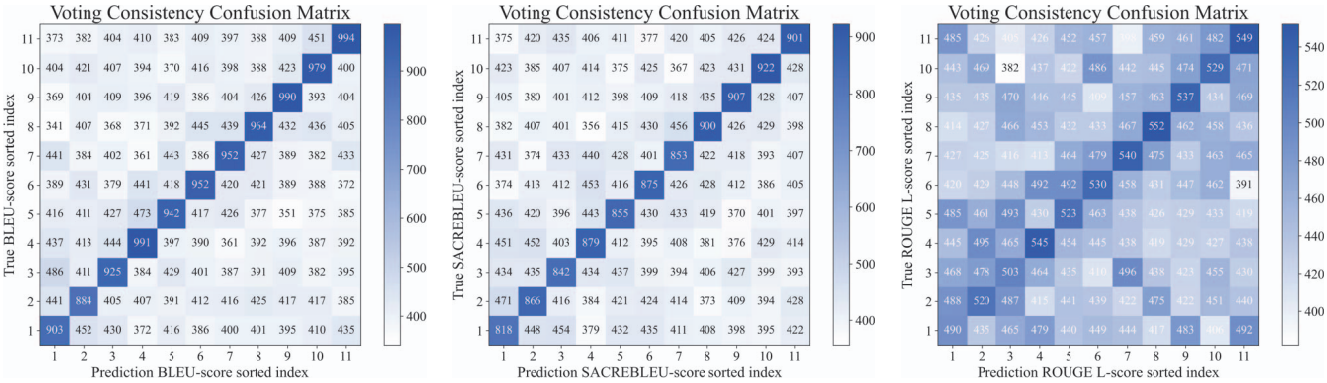
Table 1 demonstrates that our proposed method outperforms all the existing works by a noticeable margin. In particular, our method manages to surpass the recent LMM-based models, including LLaVA-FT (fine-tuned LLaVA) [44] and FoodLMM [70]. We achieve a relative improvement of 2.25%, 0.18%, and 1.97% over FoodLMM in BLEU, SacreBLEU and RougeL scores, respectively. These results demonstrate that our proposed framework can generate more precise and coherent recipes, confirming the effectiveness of our model.

4.2.2 Quantitative Comparison of Ingredients Recognition

For the inference of ingredients, the input for each image is $Q_{ingredients}$, which directly generates the answers for the ingredients. Table 2 lists the performance of ingredient recognition with existing methods in Recipe1M dataset. Our proposed method is superior to all the methods, with 1.05% and 1.03% improvement in terms of both F1 and IOU respectively, compared to FIRE [11]. Note that FIRE and InverseCooking both specifically design an ingredient recognition network. In contrast, our method is capable of generating the ingredients and instructions in a conversational manner.

4.2.3 Qualitative Results

Figure 4 presents the qualitative results comparison between our proposed model and other models including LLaVA-FT, inverse cooking [56], FoodLMM [70], as well as the ground truth "GT". It can be observed that our model, compared to the other models, can predict ingredients, preparation time, and temperature details more accurately, and produce more detailed and precise instructions. For instance, for Figure 4 (a), our model successfully identifies ingredients that the LLaVA-FT model fails to recognize, such as 'milk', 'butter' and 'onions', and it makes more precise predictions regarding the timing, es-



(a) Confusion matrix for BLEU score. (b) Confusion matrix for SacreBLEU score. (c) Confusion matrix for ROUGE L score.

Figure 6. Confusion matrix of Self-consistency Ensemble Voting for 5,000 test samples. The horizontal axis represents the index sorted from smallest to largest based on the scores calculated between each sample's top 11 retrieval-augmented prediction results and the ground truth, while the vertical axis represents the index sorted from smallest to largest based on the confidence levels obtained from voting among these 11 predictions for each sample, using cosine similarity for the voting process.

Table 2. Comparison of ingredient recognition results in terms of IOU and F1.

Methods	IOU(%)	F1(%)
R_{I2L} [58]	18.92	31.83
R_{I2LR} [58]	19.85	33.13
FF_{TD} [56]	29.82	45.94
InverseCooking [56]	32.11	48.61
FIRE [11]	32.59	49.27
Ours	**33.62**	**50.32**

pecially noting "10 to 15 minutes." For Figure 4 (b), the generated recipe by our model is quite semantically similar to the GT, whereas the other models erroneously predicted unnecessary ingredients, such as 'flour', 'vanilla' and 'nuts' in FoodLMM. These results demonstrate the effectiveness of our model to alleviate the issue of hallucination in recipe generation. Figure 5 shows that our model, while able to identify some correct ingredients like 'tomatoes' and 'garlic' for the food 'tuna and tomato cream pasta sauce', still performs poorly in generating comprehensive instructions. A significant reason is our textual evaluation metrics cannot recognize 'pasta' and 'spaghetti' as the same food; hence, even if the model correctly identifies 'pasta' as 'spaghetti', the difference in expression leads to low textual metric scores. Similarly, leading models also struggle to predict accurate recipes. We aim to improve our model's ingredient recognition and text generation to match the training dataset's vocabulary and style for more robust application across various recipes.

4.3. Ablation Study

4.3.1 Stochastic Diversified Retrieval Augmentation (SDRA)

We first investigate the effect number of retrieved recipes K for randomization for SDRA. Note that we do not use Self-consistency Ensemble Voting during inference to ensure the fairness of the experiment. Instead, SDRA is directly added to LLaVA-FT by increasing K from 1, 10, 50 to 100 to investigate the effect of the number of retrieved recipes for randomization. The results displayed in Table 3 indicate that SDRA (top 50) performs the best, demonstrating that our SDRA enhances its effectiveness by increasing the diversity and richness of the retrieved recipes within a certain retrieval scope. However, expanding the scope of the search too broadly can introduce noise into the model, thereby diminishing its performance. Specifically, for the $k = 1$ case, the top 1 and top 2 retrieved sets of ingredients are added before $Q_{ingredients}$, while the top 3 ingredients and retrieved instructions are concatenated before $Q_{instructions}$. The results verify the effectiveness of our SDRA by increasing the diversity and richness of the retrieved recipes.

Table 3. Ablation study of Stochastic Diversified Retrieval Augmentation (SDRA). "LLaVA-FT" denotes fine-tuned LLaVA, which is used as our baseline. "SDRA(fixed top 1)" refers to the approach where specifically, the top 2 retrieved ingredients sets and the top 3 ingredients along with top 1 instruction are pre-appended to their respective query placeholders. "SDRA(top k)" refers to the augmentation of the model by randomly selecting top k retrieval information when using the SDRA method.

Methods	BLEU	SacreBLEU	ROUGE L
LLaVA-FT	28.32	5.88	38.18
+SDRA(fixed top 1)	28.79	6.08	**38.46**
+SDRA(top 10)	28.52	6.07	**38.46**
+SDRA(top 50)	**29.23**	**6.21**	38.43
+SDRA(top 100)	28.67	6.04	38.36

Furthermore, we adopted two methods for concatenating retrieved information to verify the impact of the amount of retrieved information on the model's performance. The first method is as shown in Figure 3 (a), where 3 sets of ingredients and 1 set of instructions are added. The second

Table 4. Ablation study of the way concatenating retrieved information. "(1 set)" and "(2 sets)" indicate Recipe demonstration R as shown in Figure 7 and Figure 3 (a) respectively.

Methods	BLEU	SacreBLEU	ROUGE L
LLaVA-FT	28.32	5.88	38.18
+SDRA(1 set)	28.79	6.08	**38.46**
+SDRA(2 sets)	**29.23**	**6.21**	38.43

Table 5. Ablation study of Self-consistency Ensemble Voting. 'S' refers to the number of generated recipes for ensemble voting. 'Sum' is the sum of cosine similarity scores.

Scoring metric	Number	BLEU	SacreBLEU	ROUGE L	Sum
Cosine Similarity	S=1	29.23	6.21	38.43	73.87
	S=3	29.68	6.31	38.68	74.67
	S=5	30.12	6.39	38.66	75.17
	S=7	30.07	6.41	38.84	75.32
	S=9	**30.11**	**6.42**	38.91	75.44
	S=11	**30.11**	**6.42**	**38.93**	**75.47**

method involves adding only 1 set of retrieved ingredients before $Q_{ingredients}$, and similarly, adding only 1 set of instructions before $Q_{instructions}$ as shown in Figure 7. Figure 7 shows the format of the training data. Although the title is not required during the inference process, we include it during training to increase the amount of information used for training. During the inference phase, only 1 set of instructions is added before $Q_{instructions}$. In this way, we compare the use of the traditional RAG method, which involves adding fixed retrieval information before $Q_{ingredients}$ and $Q_{instructions}$. Specifically, the top 1 and top 2 retrieved sets of ingredients are added before $Q_{ingredients}$, while the top 3 ingredients and retrieved instructions are concatenated before $Q_{instructions}$. Table 4 indicates that incorporating 2 sets of retrieved recipes is more beneficial for the model's predictions than that of 1 set. However, due to the length limitations of model input, we are not able to examine more than two sets before each of the question.

4.3.2 Self-consistency Ensemble Voting

Table 5 shows the variation in recipe generation quality when calculating the confidence of candidate recipes using cosine similarity. The table shows that as more candidate recipes are generated, the text metrics for recipe quality steadily improve. In addition to cosine similarity, we also used BLEU, SacreBLEU, and ROUGE L as scoring metrics, with results presented in Section 4 of the supplementary materials. Our best result, as displayed in Table 1, is based on Cosine Similarity due to its consistent performance. We choose S=11 because the BLEU and SacreBLEU stopped improving at S=11. These results verify the effectiveness of our Self-consistency Ensemble Voting. As cosine similarity shows more steady and robust results in improving BLEU, SacreBLEU, and ROUGE L scores with the increase of S, we report the results of cosine similarity with $S = 11$ in Table 1. As the parameter S increases, the computational cost scales linearly. We consistently observe performance improvements as S grows. However, this comes with a trade-off between performance gains and computational expense. For instance, while higher values like S = 11 may yield better results, selecting S = 5 could be a more practical choice when computational resources are constrained, as it strikes a balance between efficiency and performance.

Additionally, in Figure 6, we plot a confusion matrix comparing the confidence ranking of seven predictions—obtained through voting with cosine similarity on the top 11 retrieval results from 5,000 test samples—with their actual BLEU, SacreBLEU, and ROUGE L score rankings. The shade of color represents the number of samples, for example, in Figure 6 (a), (1,3) indicates the number of samples with the highest confidence yet ranked third in BLEU scores. The results show that the final prediction selected by confidence and the actual prediction scores are consistent, leading to consistently better performance, further emphasizing the importance of introducing the voting mechanism during inference.

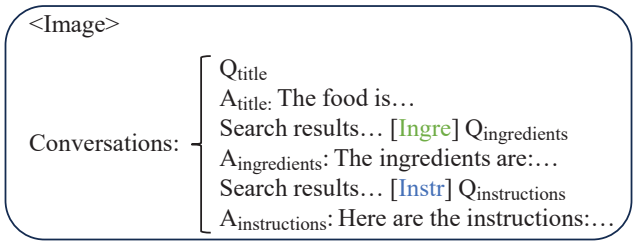


Figure 7. One set for Recipe demonstration R.

5. Conclusion

We have presented the first retrieval augmented large multimodal mode to mitigate the hallucination issue for recipe generation. We introduce the Stochastic Diversified Retrieval Augmentation to enable the model to better acquire useful knowledge from retrieved retrieval and propose Self-consistency Ensemble Voting, which optimizes the final instructions by scoring predictions obtained from different retrieval information against each other. Experimental results validate our method's effectiveness and potential for widespread application in food computing. Future work will introduce a self-reflection strategy to refine incorrect generation results, improving recipe accuracy and reliability.

Acknowledgements

This research/project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Proposal ID: T2EP20222-0046). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 3

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 1

[4] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 3

[5] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024. 3

[6] Jingjing Chen and Chong-Wah Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 32–41, 2016. 1, 2

[7] Jingjing Chen, Bin Zhu, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. A study of multi-task and region-wise deep learning for food ingredient recognition. *IEEE Transactions on Image Processing*, 30:1514–1526, 2020. 1

[8] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1

[9] Jing-jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1771–1779, 2017. 1

[10] Yanbei Chen, Shaogang Gong, and Loris Bazzani. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3001–3011, 2020. 5

[11] Prateek Chhikara, Dhiraj Chaurasia, Yifan Jiang, Omkar Masur, and Filip Ilievski. Fire: Food image to recipe generation. pages 8184–8194, 2024. 1, 2, 5, 6, 7

[12] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. Food recognition: a new dataset, experiments, and results. *IEEE journal of biomedical and health informatics*, 21(3):588–598, 2016. 1

[13] Andrew Martin Cox, Pamela McKinney, and Paula Goodale. Food logging: an information literacy perspective. *Aslib Journal of Information Management*, 69(2):184–200, 2017. 1

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. 3

[16] 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. 2020. 3

[17] David Elsweiler, Christoph Trattner, and Morgan Harvey. Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, pages 575–584, 2017. 1

[18] Mehrdad Farahani, Kartik Godawat, Haswanth Aekula, Deepak Pandian, and Nicholas Broad. Chef transformer, 2023. 5

[19] Jill Freyne and Shlomo Berkovsky. Intelligent food planning: personalized recipe recommendation. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 321–324, 2010. 1

[20] Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. Go figure: A meta evaluation of factuality in summarization. 2020. 3

[21] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. 3

[22] Yinxuan Gui, Bin Zhu, Jingjing Chen, Chong Wah Ngo, and Yu-Gang Jiang. Navigating weight prediction with diet diary. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 127–136, 2024. 1

[23] Shir Gur, Natalia Neverova, Chris Stauffer, Ser-Nam Lim, Douwe Kiela, and Austin Reiter. Cross-modal retrieval augmentation for multi-modal classification. 2021. 3

[24] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. pages 3929–3938, 2020. 2, 3

[25] Helena H. Lee, Ke Shu, Palakorn Achananuparp, Philips Kokoh Prasetyo, Yue Liu, Ee-Peng Lim, and Lav R Varshney. Recipegpt: Generative pre-training based cooking recipe generation and evaluation system. In *Companion Proceedings of the Web Conference 2020*, pages 181–184, 2020. 2

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021. 1, 4, 5

[27] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379, 2023. 3

[28] Shuqiang Jiang, Weiqing Min, Linhu Liu, and Zhengdong Luo. Multi-scale multi-view deep feature aggregation for food recognition. *IEEE Transactions on Image Processing*, 29:265–276, 2019. 1

[29] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. 2023. 2, 3

[30] Pengkun Jiao, Xinlan Wu, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yugang Jiang. Rode: Linear rectified mixture of diverse experts for food large multi-modal models. *arXiv preprint arXiv:2407.12730*, 2024. 1

[31] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Lumen: Unleashing versatile vision-centric capabilities of large multimodal models. *arXiv preprint arXiv:2403.07304*, 2024. 3

[32] Yang Jiao, Zequn Jie, Weixin Luo, Jingjing Chen, Yu-Gang Jiang, Xiaolin Wei, and Lin Ma. Two-stage visual cues enhancement network for referring image segmentation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1331–1340, 2021. 3

[33] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2, 5

[34] Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. Pre-training via paraphrasing. volume 33, pages 18470–18481, 2020. 3

[35] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. volume 33, pages 9459–9474, 2020. 3

[36] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. volume 33, pages 9459–9474, 2020. 3

[37] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[38] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[39] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3, 4

[40] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. 2023. 3

[41] Yian Li, Wentao Tian, Yang Jiao, and Jingjing Chen. Eyes can deceive: Benchmarking counterfactual reasoning abilities of multi-modal large language models. *arXiv preprint arXiv:2404.12966*, 2024. 3

[42] Chengxu Liu, Yuanzhi Liang, Yao Xue, Xueming Qian, and Jianlong Fu. Food and ingredient joint learning for fine-grained recognition. *IEEE transactions on circuits and Systems for Video Technology*, 31(6):2480–2493, 2020. 1

[43] Guoshan Liu, Yang Jiao, Jingjing Chen, Bin Zhu, and Yu-Gang Jiang. From canteen food to daily meals: Generalizing food recognition to more practical scenarios. *IEEE Transactions on Multimedia*, pages 1–10, 2024. 1

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. volume 36, 2024. 1, 2, 3, 4, 5, 6

[45] Xinda Liu, Lili Wang, and Xiaoguang Han. Transformer with peak suppression and knowledge guidance for fine-grained image recognition. *Neurocomputing*, 492:137–149, 2022. 1

[46] Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. Generating personalized recipes from historical user preferences. *arXiv preprint arXiv:1909.00105*, 2019. 1

[47] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Wide-slice residual networks for food recognition. In *2018 IEEE Winter conference on applications of computer vision (WACV)*, pages 567–576. IEEE, 2018. 1

[48] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. 2022. 1

[49] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys (CSUR)*, 52(5):1–36, 2019. 1, 2

[50] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[51] Hai X Pham, Ricardo Guerrero, Vladimir Pavlovic, and Jiatong Li. Chef: cross-modal hierarchical embeddings for food domain retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2423–2430, 2021. 1

[52] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. 2, 3

[53] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023. 3

[54] Javier Ródenas, Bhalaji Nagarajan, Marc Bolaños, and Petia Radeva. Learning multi-subset of classes for fine-grained food recognition. In *Proceedings of the 7th International Workshop on Multimedia Assisted Dietary Management*, pages 17–26, 2022. 1

[55] Doyen Sahoo, Wang Hao, Shu Ke, Wu Xiongwei, Hung Le, Palakorn Achananuparp, Ee-Peng Lim, and Steven CH Hoi. Foodai: Food image recognition via deep learning for smart food logging. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2260–2268, 2019. 1

[56] Amaia Salvador, Michal Drozdzal, Xavier Giró-i Nieto, and Adriana Romero. Inverse cooking: Recipe generation from food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10453–10462, 2019. 1, 2, 5, 6, 7

[57] Amaia Salvador, Erhan Gundogdu, Loris Bazzani, and Michael Donoser. Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15475–15484, 2021. 1, 3, 5

[58] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3020–3028, 2017. 7

[59] Fangzhou Song, Bin Zhu, Yanbin Hao, and Shuo Wang. Enhancing recipe retrieval with foundation models: A data augmentation perspective. In *European Conference on Computer Vision*, pages 111–127, 2024. 1

[60] Chun-Yuen Teng, Yu-Ru Lin, and Lada A Adamic. Recipe recommendation using ingredient networks. In *Proceedings of the 4th annual ACM web science conference*, pages 298–307, 2012. 1

[61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[62] Tom van Sonsbeek and Marcel Worring. X-tra: Improving chest x-ray tasks with cross-modal retrieval augmentation. In *International Conference on Information Processing in Medical Imaging*, pages 471–482. Springer, 2023. 3

[63] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019. 5

[64] Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan Miao. Structure-aware generation network for recipe generation from images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 359–374. Springer, 2020. 1, 2

[65] 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*, 45(3):3363–3377, 2022. 1, 2, 5

[66] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11572–11581, 2019. 1, 2, 5

[67] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. Xgpt: Cross-modal generative pre-training for image captioning. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*, pages 786–797. Springer, 2021. 3

[68] Shulin Yang, Mei Chen, Dean Pomerleau, and Rahul Sukthankar. Food recognition using statistics of pairwise local features. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2249–2256. IEEE, 2010. 1

[69] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. 2022. 3, 5

[70] Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. Foodlmm: A versatile food assistant using large multi-modal model. *arXiv preprint arXiv:2312.14991*, 2023. 1, 2, 5, 6

[71] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eagle: Towards efficient arbitrary referring visual prompts comprehension for multimodal large language models. *arXiv preprint arXiv:2409.16723*, 2024. 3

[72] Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024. 3

[73] Mengyang Zhang, Guohui Tian, Ying Zhang, and Hong Liu. Sequential learning for ingredient recognition from images. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1

[74] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. volume 36, 2024. 3

[75] Bin Zhu, Chong-Wah Ngo, and Wing-Kwong Chan. Learning from web recipe-image pairs for food recognition: Problem, baselines and performance. *IEEE Transactions on Multimedia*, 24:1175–1185, 2021. 1

[76] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11477–11486, 2019. 1

[77] Bin Zhu, Chong-Wah Ngo, and Jing-jing Chen. Cross-domain cross-modal food transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3762–3770, 2020. 1