



# ChefFusion: Multimodal Foundation Model Integrating Recipe and Food Image Generation

Peiyu Li\*  
pli9@nd.edu  
University of Notre Dame  
Notre Dame, IN, USA

Yijun Tian  
ytian5@nd.edu  
University of Notre Dame  
Notre Dame, IN, USA

Xiaobao Huang\*  
xhuang2@nd.edu  
University of Notre Dame  
Notre Dame, IN, USA

Nitesh V. Chawla  
nchawla@nd.edu  
University of Notre Dame  
Notre Dame, IN, USA

## Abstract

Significant work has been conducted in the domain of food computing, yet these studies typically focus on single tasks such as t2t (instruction generation from food titles and ingredients), i2t (recipe generation from food images), or t2i (food image generation from recipes). None of these approaches integrate all modalities simultaneously. To address this gap, we introduce a novel food computing foundation model that achieves true multimodality, encompassing tasks such as t2t, t2i, i2t, it2t, and t2ti. By leveraging large language models (LLMs) and pre-trained image encoder and decoder models, our model can perform a diverse array of food computing-related tasks, including food understanding, food recognition, recipe generation, and food image generation. Compared to previous models, our foundation model demonstrates a significantly broader range of capabilities and exhibits superior performance, particularly in food image generation and recipe generation tasks. We open-sourced ChefFusion at GitHub.

## CCS Concepts

• **Applied computing** → **Consumer health; Consumer health;**  
• **Computing methodologies** → **Computer vision; Computer vision; Natural language processing.**

## Keywords

LLMs, Multimodal, Recipe Generation, Food Image Generation

### ACM Reference Format:

Peiyu Li, Xiaobao Huang, Yijun Tian, and Nitesh V. Chawla. 2024. ChefFusion: Multimodal Foundation Model Integrating Recipe and Food Image Generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*, October 21–25, 2024, Boise, ID, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3627673.3679885>

\*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CIKM '24, October 21–25, 2024, Boise, ID, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0436-9/24/10

<https://doi.org/10.1145/3627673.3679885>

## 1 Introduction

Given the fundamental role of food in human life, the field of food computing has recently attracted considerable academic interest [21–23]. This growing area of research has led to numerous studies, each typically focusing on a specific task. For instance, some works [2, 5] focus on generating instructions from food titles and ingredients, as well as generating ingredients from recipe titles and cooking instructions, which fall under text-to-text (t2t) tasks. Other studies [3, 19] concentrate on generating recipes based on food images, which belong to image-to-text (i2t) tasks. Additionally, some research [6, 13] contributes to generating food images from recipes, categorized as text-to-image (t2i) tasks.

Despite these advancements, no approach has yet combined all these modalities into an integrated system, highlighting a significant gap. Moreover, recent developments in Transformer-based large language models (LLMs) [25] and diffusion models [18] have shown exceptional performance in various vision and language tasks. However, current methods in food computing have not kept pace with these state-of-the-art (SotA) techniques in natural language processing (NLP) and computer vision (CV).

To address this gap, we present ChefFusion, a novel food computing foundation model that achieves true multimodality, encompassing tasks such as t2t, t2i, i2t, it2t, and t2ti. ChefFusion integrates these SotA models by employing a pretrained Transformer-based LLM [30] for processing and generating recipes, a visual encoder [16] for extracting image features, and an image generation model [18] for generating food images. This integration enables ChefFusion to perform a diverse array of food computing-related tasks, including food understanding, food recognition, recipe generation, and food image generation (see Figure 3).

The contributions of this paper can be summarized as follows:

- (1) To the best of our knowledge, we present the first general food computing foundation model, which demonstrates a wide suite of multimodal capabilities, including food understanding, food recognition, recipe generation, and food image generation.
- (2) Our work pioneers the integration of multimodal dialogue capability into the field of food computing. This innovation enhances user interaction and engagement, leading to more user-friendly and intuitive systems for assisting users with cooking tasks.
- (3) We perform a comparative analysis of our results with other prominent methods in food computing. Despite the broader scope of our approach, encompassing multimodal capabilities

and functionalities, we demonstrate superior performance, particularly in food image generation and recipe generation tasks.

## 2 Related work

**Recipe Generation.** Compared to other i2t tasks, generating detailed recipe information or cooking instructions from a food image presents a considerable challenge. To accomplish this, models need to have comprehensive knowledge of food composition, ingredients, and cooking procedures to ensure accuracy. Constrained by limited model capacity and structure, initial attempts in recipe generation relied heavily on information retrieval techniques [28, 29]. More recent approaches employ encoder-decoder architectures in multimodal settings to generate recipes [3, 19, 26]. [19] introduced a framework that uses encoded representations of images and ingredients in the recipe generation process. [26] incorporated tree structures into the encoder-decoder process to include structure-level information. [3] uses images as input to generate titles and ingredients as intermediate representations, which are then used to create complete recipes with an encoder-decoder model. Instead, we leverage a frozen LLM and CLIP image encoder to generate recipes.

**Food Image Generation.** Most prior work in image-to-text (i2t) tasks assumes that visual categories are well-structured singular objects, such as birds or flowers. In contrast, food images exhibit significant variability in appearance depending on the ingredients, making them more challenging to generate accurately. Recent approaches often rely on Generative Adversarial Networks (GANs) to generate food images, as seen in studies like [6, 13, 14, 27, 32]. For instance, [27] and [32] use generative neural networks to produce food images as a constraint to enhance cross-modal recipe retrieval, but these methods typically generate only low-resolution images (e.g.,  $128 \times 128$  pixels). [6] and [13] improves on this by generating higher resolution food images ( $256 \times 256$  pixels) based on the ingredients. In contrast to these methods, our approach utilizes a diffusion model to generate food images, achieving even higher resolution ( $512 \times 512$  pixels).

## 3 Methodology

The training process consists of two primary components: (1) training the model to generate recipe, and (2) training the model to generate food images. Additionally, the model must determine whether to produce text or images at each step. The detailed architecture is illustrated in Figure 1.

### 3.1 Training to Generate Recipe

Given an image  $x$  and its paired recipe  $y$  (tokenized as  $(t_1, \dots, t_N)$ ), our object is to adapt a frozen LLM to handle sequences of interleaved image and text inputs. We follow previous research [4, 8, 9, 11, 24] in learning translation parameters that convert image features into the text embedding space.

We start by extracting visual embeddings  $v_\phi(x) \in R^d$  using a pretrained visual backbone, while keeping its weights  $\phi$  and the LLM weights  $\theta$  fixed. We then develop a linear mapping  $\mathbf{W}_{recipe} \in R^{d \times ke}$  to transform  $v_\phi(x)$  into a sequence of  $k$   $e$ -dimensional vectors, which serve as inputs to the LLM (see Figure 1, left). Here,  $e$  denotes the LLM’s input embedding dimension.

We train  $\mathbf{W}_{recipe}$  on pairs of food image and recipe by minimizing the negative log-likelihood loss of the token sequence  $t_1, \dots, t_N$ :

$$l_r(x, y) = - \sum_{n=1}^N \log p_\theta(t_n | v_\phi(x)^T \mathbf{W}_{recipe}, t_1, \dots, t_{n-1}) \quad (1)$$

### 3.2 Training to Generate Food Image

Following a method similar to [8, 9, 31], we introduce special  $[IMG]$  tokens into the LLM’s vocabulary to enable the model to produce image outputs. Specifically, we add a trainable matrix  $\mathbf{E}_{img} \in R^{m \times e}$  to the embedding matrix of the frozen LLM, which represents the  $m$   $[IMG]$  token embeddings. According to the experiments of [8], as the number of  $[IMG]$  tokens increases, generation generally improves since the inputs to LLM are longer and more expressive. Therefore, we use  $m = 8$   $[IMG]$  tokens to enhance the expressivity of the frozen LLM for novel image generation. Our objective is to train the model to recognize when to generate  $[IMG]$  tokens. This is achieved by minimizing the negative log-likelihood of producing the first  $[IMG]$  token, conditioned on the previously generated tokens:

$$l_p(y) = - \log p_{\{\theta \cup \mathbf{E}_{img}\}}([IMG_1] | t_1, \dots, t_n) \quad (2)$$

During training, the  $[IMG]$  tokens are appended to each recipe. During inference, whenever the first  $[IMG_1]$  token is generated, the subsequent  $m - 1$   $[IMG]$  tokens are always produced.

To enable our LLM to generate image outputs, the  $[IMG]$  tokens must be mapped to a semantically meaningful region within the input space of an image generation model  $D_\psi$ . To achieve this, we use a 4-layer encoder-decoder transformer model [25]  $f_w$  with trainable weights  $w$ . The model  $f_w$  is conditioned on  $h_{\{\theta \cup \mathbf{E}_{img}\}}(y, [IMG])$  and  $L$  learned query embeddings  $(q_1, \dots, q_L) \in R^{L \times r_1}$ , where  $L$  is the maximum input sequence length of the text-to-image generation backbone  $D_\psi$ . We optimize the trainable weights  $((q_1, \dots, q_L)$  and  $w$ ) by minimizing the MSE loss of the model  $f_w$  outputs against the embeddings produced by the text encoder  $T_\psi$  of a frozen text-to-image generation model:

$$l_g(y) = \|f_w(h_{\{\theta \cup \mathbf{E}_{img}\}}(y, [IMG_1]), \dots, h_{\{\theta \cup \mathbf{E}_{img}\}}(y, [IMG_m]), q_1, \dots, q_L) - T_\psi(y)\|^2 \quad (3)$$

During inference, when  $[IMG]$  tokens are generated, we can synthesize an image:

$$GeneratedFoodImage = D_\psi(f_w(h_{\{\theta \cup \mathbf{E}_{img}\}}(y, [IMG_1]), \dots, h_{\{\theta \cup \mathbf{E}_{img}\}}(y, [IMG_m]), q_1, \dots, q_L)) \quad (4)$$

### 3.3 Dataset and Implement Details

We train on Recipe1M [20], which contains more than 1 million recipes and almost 900k images. We use the OPT-6.7B [30] model as the LLM backbone (which produce hidden states  $h_\theta$  with embedding dim  $e = 4096$ ). For the visual model used to extract features  $v_\phi$ , we use the CLIP [16] ViT-L model. For our text-to-image generation backbone  $D_\psi$ , we use the Stable Diffusion [18] v1.5 model (with  $L = 77$  input vectors). We use  $k = 4$  visual tokens, and  $m = 8$  learnt  $[IMG]$  tokens. We set the query embedding dimension  $r = 512$ . All pretrained model weights are kept frozen, and we only train the linear layers  $\mathbf{W}_{recipe}$ , the embedding matrix  $\mathbf{E}_{img}$ , the parameter  $w$  and query vectors  $q_1, \dots, q_L$ . We use bfloat16 precision [1], and optimize using Adam [7] ( $\beta_1 = 0.9, \beta_2 = 0.95$ ) with a learning rate



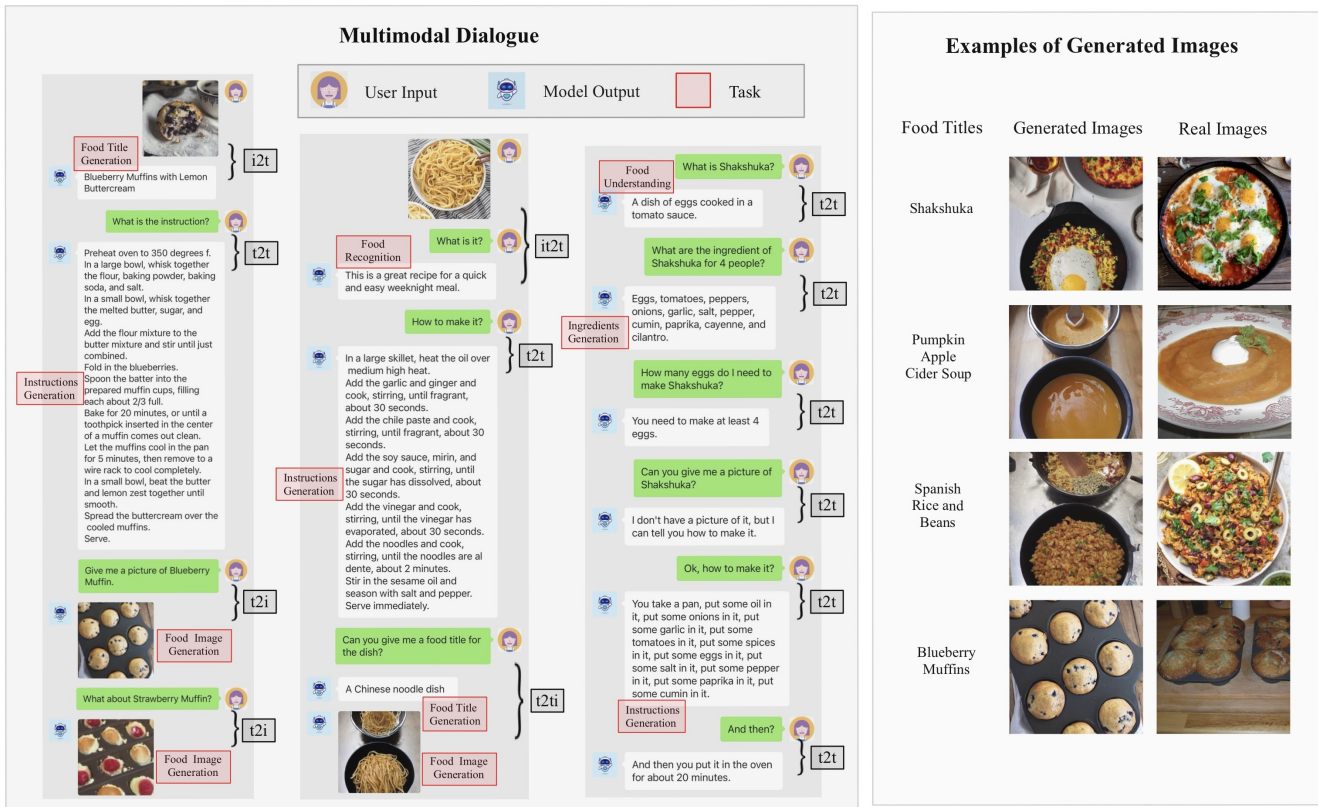


Figure 3: Case Study: ChefFusion demonstrates a wide suite of multimodal capabilities, including food understanding, food recognition, recipe generation, food image generation and multimodal dialogue (left). Example of food images generated by ChefFusion (right).

evaluates the overlap of bigrams between the generated text and the reference text. It calculates the precision, recall, and F1-score of these bigrams. In essence, ROUGE-2 helps assess how well a machine-generated summary or translation captures the important phrases or concepts present in the reference text at the bigram level.

## 4.2 Tasks

**i2t task:** Images in the Recipe1M are utilized as the input for the models and the generated recipes are compared with the ground-truth recipes. In our study, our model shows the best performance both in SacreBLEU and ROUGE-2 of 6.97 and 0.12 respectively compared to the rest of the baseline models, see Table 1. This indicates that the generated recipes closely resemble human-generated references, implying a high degree of translation accuracy. The performance of our model could be attributed to various factors. Firstly, it leverages LLM and CLIP models as its backbone, enabling it to capture intricate relationships between food images and corresponding recipes more effectively. Secondly, the model may have been trained on a larger and more diverse food dataset, facilitating better generalization to unseen food examples. Furthermore, meticulous hyperparameter tuning and optimization strategies could have contributed to its superior performance.

**t2i task:** The recipes in the Recipe1M dataset are used as the input for the models and the generated images are compared with the ground-truth images. In Table 2, our model shows the best performance 0.74 compared to the rest of the models. This suggests that the images generated by this model exhibit a strong alignment with the provided textual descriptions, indicating high fidelity and relevance. To be specific about exceeding the performance of Stable Diffusion, our model enhanced the semantic capturing capability by introducing trainable matrix  $E_{img}$ , which enables the CLIP model in our backbone to capture more accurate and relevant information within the recipe context.

## 5 Conclusion

In this study, we introduce a novel multimodal food computing foundation model that integrates a Transformer-based LLM for recipes, a visual encoder for image features, and an image generation model. This model excels in diverse tasks such as food understanding, recognition, recipe generation, and image generation. Despite the broader scope of our approach, encompassing multimodal capabilities and functionalities, we demonstrate superior performance, particularly in food image generation and recipe generation tasks.

## References

- [1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).
- [2] Michał Bień, Michał Gilski, Martyna Maciejewska, Wojciech Taisner, Dawid Wisniewski, and Agnieszka Lawrynowicz. 2020. RecipeNLG: A cooking recipes dataset for semi-structured text generation. In *Proceedings of the 13th International Conference on Natural Language Generation*. 22–28.
- [3] Prateek Chhikara, Dhiraj Chaurasia, Yifan Jiang, Omkar Masur, and Filip Ilievski. 2024. Fire: Food image to recipe generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 8184–8194.
- [4] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2021. MAGMA—Multimodal Augmentation of Generative Models through Adapter-based Finetuning. *arXiv preprint arXiv:2112.05253* (2021).
- [5] Helena H. Lee, Ke Shu, Palakorn Achananuparp, Philips Kokoh Prasetyo, Yue Liu, Ee-Peng Lim, and Lav R Varshney. 2020. RecipeGPT: Generative pre-training based cooking recipe generation and evaluation system. In *Companion Proceedings of the Web Conference 2020*. 181–184.
- [6] Fangda Han, Ricardo Guerrero, and Vladimir Pavlovic. 2020. CookGAN: Meal image synthesis from ingredients. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1450–1458.
- [7] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [8] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems* 36 (2024).
- [9] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*. PMLR, 17283–17300.
- [10] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. <https://aclanthology.org/W04-1013>
- [11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
- [12] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
- [13] Siyuan Pan, Ling Dai, Xuhong Hou, Huating Li, and Bin Sheng. 2020. ChefGAN: Food image generation from recipes. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4244–4252.
- [14] Dim P Papadopoulos, Youssef Tamaazousti, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2019. How to make a pizza: Learning a compositional layer-based gan model. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8002–8011.
- [15] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, 186–191. <https://www.aclweb.org/anthology/W18-6319>
- [16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- [18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [19] Amaia Salvador, Michal Drozdal, Xavier Giró-i Nieto, and Adriana Romero. 2019. Inverse cooking: Recipe generation from food images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10453–10462.
- [20] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning cross-modal embeddings for cooking recipes and food images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3020–3028.
- [21] Yijun Tian, Chuxu Zhang, Zhichun Guo, Chao Huang, Ronald Metoyer, and Nitesh V. Chawla. 2022. RecipeRec: A Heterogeneous Graph Learning Model for Recipe Recommendation. In *IJCAI*.
- [22] Yijun Tian, Chuxu Zhang, Zhichun Guo, Yihong Ma, Ronald Metoyer, and Nitesh V Chawla. 2022. Recipe2Vec: Multi-modal Recipe Representation Learning with Graph Neural Networks. In *IJCAI*.
- [23] Yijun Tian, Chuxu Zhang, Ronald Metoyer, and Nitesh V. Chawla. 2022. Recipe Recommendation With Hierarchical Graph Attention Network. *Frontiers in Big Data* (2022).
- [24] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.
- [25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [26] Hao Wang, Guosheng Lin, Steven CH Hoi, and Chunyan Miao. 2022. Learning structural representations for recipe generation and food retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3363–3377.
- [27] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven CH Hoi. 2019. 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*. 11572–11581.
- [28] Liping Wang, Qing Li, Na Li, Guozhu Dong, and Yu Yang. 2008. Substructure similarity measurement in chinese recipes. In *Proceedings of the 17th international conference on World Wide Web*. 979–988.
- [29] Haoran Xie, Lijuan Yu, and Qing Li. 2010. A hybrid semantic item model for recipe search by example. In *2010 IEEE International Symposium on Multimedia*. IEEE, 254–259.
- [30] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).
- [31] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.
- [32] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11477–11486.