

LLaVA-Chef: A Multi-modal Generative Model for Food Recipes

Fnu Mohbat
Rensselaer Polytechnic Institute
Troy, New York, USA
mohbaf@rpi.edu

Mohammed J. Zaki
Rensselaer Polytechnic Institute
Troy, New York, USA
zaki@cs.rpi.edu

ABSTRACT

In the rapidly evolving landscape of online recipe sharing within a globalized context, there has been a notable surge in research towards comprehending and generating food recipes. Recent advancements in large language models (LLMs) like GPT-2 [40] and LLaVA [32] have paved the way for Natural Language Processing (NLP) approaches to delve deeper into various facets of food-related tasks, encompassing ingredient recognition and comprehensive recipe generation. Despite impressive performance and multi-modal adaptability of LLMs, domain-specific training remains paramount for their effective application. This work evaluates existing LLMs for recipe generation and proposes LLaVA-Chef, a novel model trained on a curated dataset of diverse recipe prompts in a multi-stage approach. First, we refine the mapping of visual food image embeddings to the language space. Second, we adapt LLaVA to the food domain by fine-tuning it on relevant recipe data. Third, we utilize diverse prompts to enhance the model's recipe comprehension. Finally, we improve the linguistic quality of generated recipes by penalizing the model with a custom loss function. LLaVA-Chef demonstrates impressive improvements over pretrained LLMs and prior works. A detailed qualitative analysis reveals that LLaVA-Chef generates more detailed recipes with precise ingredient mentions, compared to existing approaches.

CCS CONCEPTS

• **Computing methodologies** → **Learning latent representations; Natural language generation.**

KEYWORDS

Food Recipe Generation, Food Computing, Multi-modal Large Language Models, Natural Language Generation

ACM Reference Format:

Fnu Mohbat and Mohammed J. Zaki. 2024. LLaVA-Chef: A Multi-modal Generative Model for Food Recipes. 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, 11 pages. <https://doi.org/10.1145/3627673.3679562>

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.
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/10.1145/3627673.3679562>

1 INTRODUCTION

The significance of food for promoting well-being is growing, as a result understanding food recipes for healthy lifestyles has emerged as a critical research area. The recent growth of recipe data through online platforms and mobile apps has created a rich data resources, driving research efforts towards developing AI-powered solutions for food recognition, ingredient suggestion, and personalizing recipe, all while factoring in dietary restrictions, cultural preferences, and religious considerations [11, 35, 44, 54]. Despite substantial progress, generating recipes or cooking steps solely from food names, images, or ingredients remains a significant challenge. While the computer vision community has leveraged state-of-the-art deep learning techniques to extract ingredients from images, and NLP applications have facilitated recipe generation from food names or ingredients, the recent advances in multi-modal language-vision models offer a promising path towards crafting feasible real-world solutions by fusing visual and textual data.

Large language models (LLMs) [20, 22, 41, 47] have demonstrated a remarkable ability to rapidly learn from vast amounts of text and even multi-modal data [25, 28, 32]. For instance, by incorporating visual features extracted from pretrained vision-language models, several LLMs [5, 28, 32] have shown an enhanced ability to tackle vision-language tasks like image captioning, visual question answering, and visual reasoning. While these models excel in general applications, their expertise plummets when they encounter specialized domains due to insufficient domain-specific training [26, 34]. This deficit often manifests in outputs riddled with hallucinations, inaccuracies, and repetitive text, as Figure 2 demonstrates for food recipes generated by two models.

Initial research focused on computer vision methods for food classification to ingredient detection [8, 19, 23, 43]. Several researchers learned unique food embeddings using text-vision models [35, 43] while others generated food names using image captioning models [11]. Chef Transformer [16] takes a list of ingredients and generates recipes, whereas [11, 17, 46] predict ingredients from food images as an intermediate step towards recipe generation. One recent research [53] fine-tuned the LISA [25] model for a variety of food tasks including food classification, recipe generation and segmentation. Despite various endeavors, none of the models have proven successful in generating effective recipes. Furthermore, most of these models lack robust evaluation or are not publicly available.

In this paper, we address the limitation of the existing methods by proposing LLaVA-Chef, a powerful multi-modal language and vision model for learning food recipes with the help of well curated and diverse set of prompts tailored towards training the model for food domain tasks. Our model extends the LLaVA [32], which consists of Vicuna [12] as a foundation LLM and CLIP [39] as a visual encoder. The architecture of our model is shown in Figure 1. The model concatenates visual and textual embeddings, and inputs them

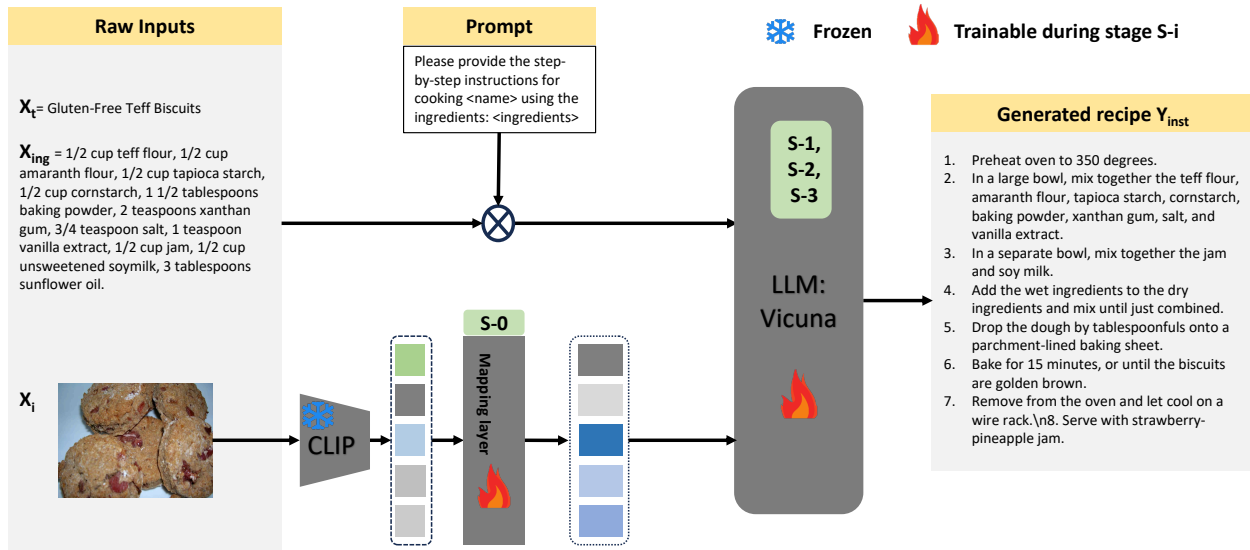


Figure 1: Architecture of LLaVA-Chef and different training stages (as shown in grey). The inputs to the model X_t , X_{ing} , and X_i refer to the recipe title, ingredients and image, respectively. Y_{inst} refers to the generated recipe instructions (which are compared with the ground truth instructions X_{inst} for loss computation). In training Stage-0 (S-0), the image to text mapping layer is fine-tuned. Whereas, in the rest of the training stages S-1, S-2, and S-3 the backbone LLM is fine-tuned. Given a recipe, we sample a prompt, then substitute <name> and <ingredients> with X_t and X_{ing} . Visual features of the image X_i from CLIP are mapped in language space and concatenated with language embeddings before passing through the backbone LLM. The frozen and trainable symbols indicate which layers are fine-tuned (e.g., CLIP is frozen, whereas mapping layer and LLM are trainable).

to the backbone LLM to generate the desired output. Following [26], first we improve the cross-modal representation for food related images by fine-tuning the mapping. Then, the model is fine-tuned on unique prompts that reduce the hallucination and improve the quality of recipe text. In the following training stage, we improve the adaptability of the model for the food domain by introducing more than 100 unique prompts to generate different attributes of a recipe, i.e., title, ingredients and cooking instructions. Finally, we penalize the model with a novel scaling term based on text generation metrics, ultimately leading to improved performance. Thus, gradually involving the augmentation of prompt diversity and task complexity across multiple stages, our model systematically acquires proficiency in handling a wide array of food recipes. We evaluate our model on the Recipe1M dataset [45], specifically on the test samples containing at least one image. Compared to pretrained LLMs, our model consistently achieves higher scores across most metrics. While other models could not get more than 0.1 CIDEr score, our model achieves a remarkable 21-point lead. Qualitative evaluation of the generated recipes confirms the advantages of our model.

2 RELATED WORK

Large Foundational Models: The emergence of LLMs like BERT [24] and GPT-2 [40] marked a significant leap in text understanding from summarization to reasoning. This success spurred exploration of even better LLMs and their application to visual-language tasks, including image captioning and visual question answering. Building on the success of LLMs like the 175B parameter model GPT-3.5 [7],

recent smaller counterparts like Mistral [22] and Phi-2 [20] demonstrate promising performance on various language tasks, suggesting potential benefits in efficiency and resource usage. Furthermore, recent proprietary models like GPT-4 [1] and BARD [2] have garnered significant attention for their multi-modal capabilities, but their proprietary nature restricts accessibility and computational feasibility.

On the other hand, open-source multi-modal LLMs [5, 14, 26, 28, 39, 55] have demonstrated their effectiveness in various visual-language tasks. At the core of these multi-modal models lies a foundational LLM fine-tuned for understanding visual data. A common approach involves a pretrained vision-language encoder (e.g., CLIP [39]) to extract visual features, which are then integrated with language embeddings through mapping layers [32, 55] or cross-attention modules [5, 14, 28]. This approach has led to successful applications in domains like medicine [26, 34], finance [33, 52], and law [4, 13]. While some research has explored applying these models to the food domain [11, 18, 27, 53], their performance remains limited due to ineffective or inadequate training strategies.

Recipe Understanding: Early research in the food domain primarily focused on food image classification [19, 23]. Following this, interests shifted towards more intricate tasks including ingredient detection [8, 11, 43], recipe retrieval [10, 45, 50], ingredient substitution recommendations [27, 37], and automatic recipe generation [6, 11, 16, 46, 53]. Notable attempts at recipe generation include Chef Watson’s [48] Bayesian network approach over a knowledge representation schema. Wang et al. [51] proposed a structure-aware generation method for recipes from food images.

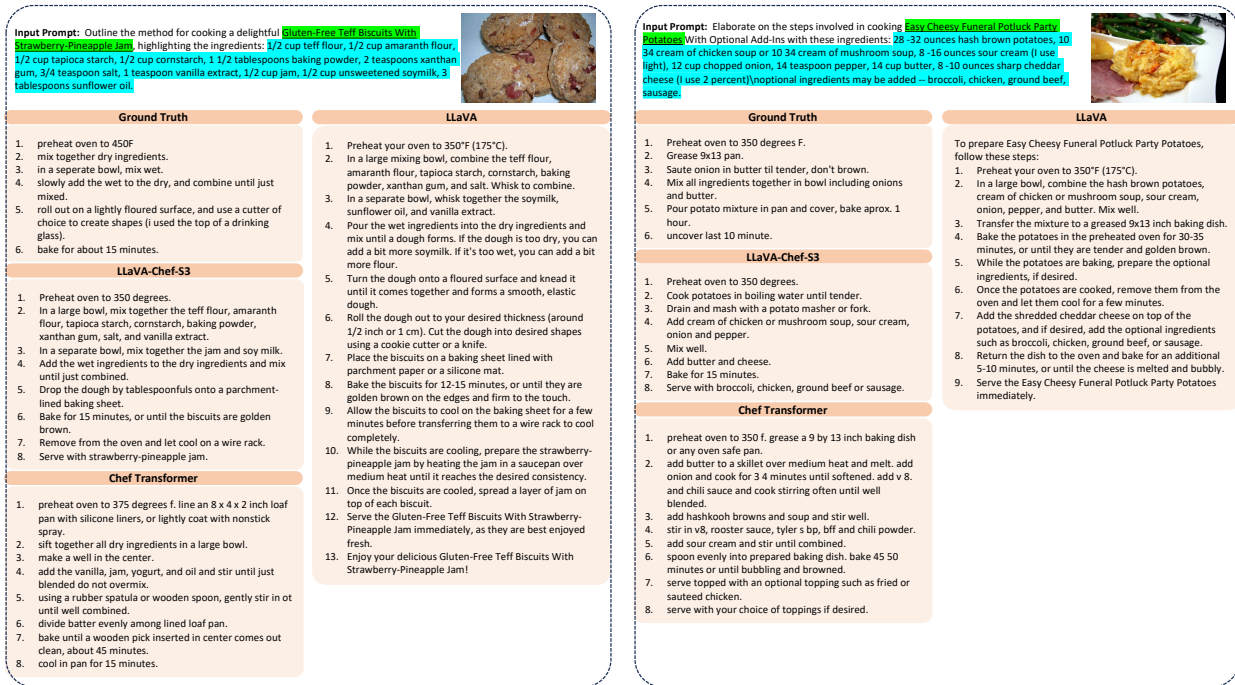


Figure 2: Sample recipes generated by LLaVA-Chef model, Chef-Transformer [16] (open source recipe generation model) and LLaVA [26] (best pretrained model). We can see issues of hallucination, repetitive test, and inaccuracies for previous models.

DoD [43] explored food recognition by learning fine-grained embeddings of food names and ingredients using BLIP-2 [28] and Falcon 7B [3]. RecipeGPT [18] leveraged the GPT-2 [40] architecture, while RecipeMC [46] employed Monte Carlo Tree Search on top of GPT-2 for recipe generation.

More recent works such as RecipeGM [42] and Chef Transformer [16] focused on generating recipes from pre-specified ingredient lists. FIRE [11] utilizes BLIP [29] model for food title generation and a ViT-based multi-class classifier for extracting ingredient lists, followed by the model T5 [41] for recipe generation. FoodLMM [53] fine-tuned LISA [25], a multi-modal model, for diverse food-related tasks including classification, ingredient detection, segmentation and recipe generation. While FoodLMM demonstrates improved performance across multiple tasks compared to baselines, its recipe generation capabilities remain a subject for further improvement.

3 VISUAL INSTRUCTION-FOLLOWING DATA

Building upon the success of LLaVA [32] for visual instruction tuning, we adapt it to food recipe generation. Food recipes encompass both textual elements (title X_t , ingredients X_{ing} , and cooking instructions X_{inst}) and visual information (food image X_i), as illustrated in Figure 3. Despite several efforts to estimate cooking instructions from food images, none could produce good recipes compared to human. Furthermore, a dearth of research exists regarding the generation of complete recipes solely from images, titles, ingredients, or combinations thereof. To bridge this gap, we develop instruction tuning prompts specifically designed to predict

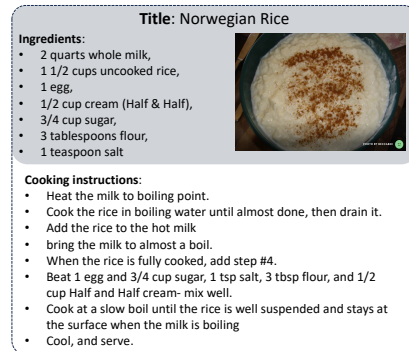


Figure 3: Sample recipe from the Recipe1M dataset. Title is denoted X_t , image X_i , ingredients X_{ing} , and instructions X_{inst} .

Y_t , Y_{ing} , Y_{inst} , or their combination. Our approach entails aligning food image embeddings with corresponding textual attributes by partially fine-tuning the model, followed by fine-tuning the complete model to estimate the desired food attributes through multi-modal fusion.

3.1 Food Concept Alignment Data

To align food image embeddings with text embeddings, we randomly sample a question prompt X_p for the generation of cooking instructions Y_{inst} from the title X_t , ingredients X_{ing} , and the associated food image X_i . Sample prompts with placeholders are illustrated in Table 1. The prompt X_p contains placeholders tokens

| Input | Output | sample prompt |
|--------------------------------|----------------------------|--|
| Stage 0 and 1 Training Prompts | | |
| $X_i + X_t + X_{ing}$ | Y_{inst} | Given <ingredients>, what are the key steps you need to follow to prepare a perfect <name>? |
| $X_i + X_t + X_{ing}$ | Y_{inst} | Please provide the step-by-step instructions for cooking a delicious <name> from scratch using the following ingredients: <ingredients>. |
| $X_i + X_t + X_{ing}$ | Y_{inst} | Outline the steps to cook a <name> using ingredients: <ingredients> |
| Stage 2 and 3 Training Prompts | | |
| X_i | Y_t | What is the name of the dish in this image? |
| $X_i + X_{ing}$ | Y_t | What is the name of the dish in this image? The ingredients used are: <ingredients> |
| X_i | Y_{ing} | Based on the features of the food in the image, provide a list of possible ingredients. |
| X_i | Y_{inst} | Describe how to prepare the meal shown in the image. |
| X_t | Y_{inst} | Generate cooking instructions for <name>: |
| $X_i + X_t$ | Y_{inst} | Generate cooking steps for <name> shown in this image. |
| $X_i + X_t + X_{ing}$ | Y_{inst} | Elaborate on the steps involved in cooking <name> with these ingredients: <ingredients> |
| X_i | $Y_t + Y_{ing} + Y_{inst}$ | Generate a name, ingredients, and cooking instructions for this dish: |

Table 1: Example prompts utilized at each training stage. We can see that S-0 and S-1 focus on generating cooking instructions, whereas S-2 and S-3 also on additional tasks. During training, we randomly select output task then we select input(s).

<name> and <ingredients> corresponding to the title X_t and ingredients X_{ing} . During training, we substitute these placeholders with their actual values, resulting in the finalized prompt X_q . This refined prompt serves as the query for the model as demonstrated in Figure 1. Throughout the training, we structure inputs into a single-round instructions-following format, as exemplified below:

Human : X_q X_i <STOP> \n
Assistant : Y_{inst} <STOP> \n

During training, optimization focuses solely on the layer that maps visual features to language embeddings. This targeted optimization aims to refine the visual embeddings and enhance their alignment with the food domain, ultimately improving the LLM’s performance for recipe generation.

3.2 Visual Instruction Tuning Data

To adapt our model for food domain, we curated diverse prompts aimed at generating multiple textual attributes of a recipe from a food image and other textual attributes. These prompts effectively leverage the LLM’s ability to perform multi-modal text generation. Specifically, each prompt was designed to elicit a targeted output from the LLM. For instance, one prompt instructed the model to generate the food name based solely on its image. Another prompt tasked the model with predicting the cooking instructions, utilizing both the food image and the provided name. We employed GPT-3.5 to generate prompts for the following target outputs: food name (Y_t), cooking instructions (Y_{inst}), and cooking ingredients (Y_{ing}). Examples of these prompts are presented in Table 1. During training, we randomly select a task and then a prompt specific to the selected task. The chosen prompt either may demand the prediction of a single output (title, ingredients, or instructions) or multiple outputs from the provided inputs. In cases where the recipe lacks an associated image, an empty image is utilized.

Our multi-stage fine-tuning process progressively enhances the model’s understanding of food recipes. Initially (Stage-0), visual embeddings are projected into the language domain, establishing a foundation for subsequent learning. Stage-1 focuses on recipe comprehension by training the model to generate cooking instructions based on the provided food image, title, and ingredients. Subsequent stages (Stages-2 and Stage-3) increase task complexity and reduce input information to promote deeper recipe knowledge acquisition. In the cooking instruction task, diverse prompts expose the model

to varying input modalities (image-only, title-only, image-title, and image-title-ingredients), fostering robustness in recipe generation. Finally, the model is also challenged to predict recipe title, ingredients, and cooking instructions solely from the image, solidifying its ability to infer comprehensive recipe information from limited visual input.

4 LLAVA-CHEF: ADAPTING LLAVA TO FOOD DOMAIN

The performance of LLaVA-Chef is gradually improved by a meticulously designed multi-stage training strategy to unlock its full potential as described below in detail.

4.1 Stage 0: Food domain adaptation

To bridge the gap between visual and language modalities, LLaVA leverages a linear layer to project visual features into the language space. In Stage-0, we concentrate on fine-tuning the mapping layer using image-recipe pairs from the Recipe1M dataset [45]. As illustrated in Figure 1, the food image X_i , name X_t and ingredients X_{ing} are input to the model and the model is asked to generate cooking instructions. Optimization of the mapping layer is achieved through the standard cross-entropy loss function defined as follows:

$$L_{CE} = CE(p(Y_{inst}), p(\hat{Y}_{inst})) \quad (1)$$

Where, $p(Y_{inst})$ is probability of ground truth cooking instruction as one hot-vector, $p(\hat{Y}_{inst})$ indicates probability of the cooking instructions predicted by the model. This fine-tuning aims to optimize the alignment of the visual embeddings with their corresponding language representations, enhancing the model’s ability to capture the nuances of visual information relevant to recipes. Note that this step fine-tunes the mapping layer to better understand the food images.

4.2 Stage 1: Learning the language of recipes

To train our model on predicting cooking instructions from image, title, and ingredients, we curated a dataset of 35 unique prompts. Each prompt incorporates special tokens: <name> representing the food title and <ingredients> signifying the listed ingredients. During training, we randomly sample a prompt, then replace these special tokens with the title and ingredients of the recipe and fine-tune the entire backbone LLM model. This approach allows the

model to learn food-domain embeddings from both visual and textual data seamlessly. Recognizing that not all recipes may have accompanying images, we employed a strategy for handling missing visuals. When an image is unavailable, we substitute it with a black (empty) image as a placeholder. This enables the model to learn from the remaining textual attributes (title and ingredients) and still estimate cooking instructions even without image input. The model is optimized using the default cross entropy loss function as defined above in equation 1.

4.3 Stage 2: Boosting model adaptability via prompt diversity

The Recipe1M dataset [45] offers four attributes for each recipe: image, title, ingredients, and cooking instructions. While image contributes visual information, the latter three act as textual attributes. To diversify our training prompts, we expanded our initial set of 35 prompts by utilizing GPT-3.5 to generate prompts for various recipe-related tasks, bringing the total to 102 prompts, some examples are shown in Table 1. These prompts are task-specific, explicitly defining the input and target output for each prediction scenario. During training, we randomly select a task (what to predict) and a corresponding prompt. We opted to retain cross-entropy as our chosen loss function. This approach fosters model generalizability, enabling it to predict the desired output (e.g., title, ingredients, or instructions) from image, title, or ingredients via fine-tuning as shown in Figure 1. To further improve generalization, we adopted a strategy where at most 50% of the ingredients are omitted from the input during training. This forces the model to infer missing ingredients based on the remaining information, ultimately leading to improved performance across all tasks, including cooking instruction generation from solely image or title.

4.4 Stage 3: Optimizing the recipe language

To enhance the language quality and achieve predictions closer to the ground truth, we extended the training of our model from Stage-2 by introducing an additional penalty loss, based on the commonly used BLEU [36] and Rouge [30] scores that were initially formulated to evaluate machine translation and text summarization tasks. However, one cannot directly optimize these metrics as additional loss terms, since they are non-differentiable (e.g., they are based on n -gram counts). Instead of optimizing them directly, we propose a novel formulation where we use the scores as a multiplicative or scaling factor for the cross-entropy loss. Let Y_{label} denote the ground truth recipe, Y_{pred} the generated recipe (note: *label* can refer to any of the inputs such as title, image, ingredients and/or cooking instructions). Next, define $L_{bleu} = 1 - BLEU(Y_{label}, Y_{pred})$ as the penalty from the SacreBLEU score [38], and $L_{rougeL} = 1 - rougeL(Y_{label}, Y_{pred})$ as the penalty from Rouge-L [31]. Since higher scores are better (with 1 being the maximum score), we penalize by subtracting them from 1. We then combine both into a joint scaling penalty:

$$L_{BR} = \lambda_{bleu}(1 - L_{bleu}) + \lambda_{rougeL}(1 - L_{rougeL}) \quad (2)$$

where λ_{bleu} and λ_{rougeL} are weighting factors. Next, we multiply the (per-sample) scaling penalty L_{BR} with the cross-entropy loss

(L_{CE}), as follows:

$$L_{final} = L_{BR} \times L_{CE} \quad (3)$$

As such L_{BR} , while non-differentiable, works as a (per sample) scaling constant, thus scaling and penalizing the overall loss when the value of either of the metrics goes down; the final loss remains differentiable. This multi-objective approach holds the promise of generating more fluent, accurate, and semantically aligned recipe instructions, as we investigate in the following section.

5 EXPERIMENTS

5.1 Experimental setup

Dataset: We leveraged Recipe1M [45], a large-scale recipe dataset boasting 1 million recipes and 819,000 food images. Each recipe comprises a title, ingredients list, and cooking instructions, with several samples also accompanying one or more images. Recipe1M already provides training, validation, and test splits. For the training phase, we utilized the entire training set consisting of 720,639 recipes (with 619,508 images). However, during testing, we focused on recipes with at least one image. After cleaning the test set by removing samples lacking images or containing corrupted ones, we obtained two curated testing subsets:

- *test50k*: All 50,507 test samples from Recipe1M test that contain at least one image.
- *test1k*: We selected another 1,000 samples (randomly) as *test1k* set for detailed qualitative analysis.

Metrics: To evaluate the generated text quality compared to the ground truth, we employed several image caption and language translation metrics. These metrics include BLEU [36], a precision-based metric specifically designed for machine translation, Rouge [30], a recall-oriented metric for text summarization, METEOR [15] and CIDEr [49], which were specifically developed for assessing image caption quality and exhibit strong correlation with human subjective judgments. Perplexity [21], a measure of language model uncertainty, was also included to provide additional insights into fluency and coherence of the generated text.

Model Training: Our model, LLaVa-Chef, was trained in four consecutive stages on four NVIDIA RTX A6000 48G GPUs with a batch size of 32. We set learning rate to $2e-5$ with a cosine learning scheduler at a warmup ratio of 0.03. Stages 0, 1 and 2 employed the standard cross-entropy loss function. In Stage-3, loss was scaled based on BLEU ($\lambda_{bleu} = 1.01$), and Rouge-L ($\lambda_{rougeL} = 1$). This multi-objective approach prioritized language quality, ultimately leading to improved performance in generated text fidelity when compared to ground-truth recipes. Our model and data is publicly available at <https://github.com/mohbattharani/LLaVA-Chef>.

5.2 LLaVA fine-tuning

Our investigation into recipe generation compared multiple high-performing open-source general-purpose LLMs. We also evaluated Chef Transformer [16] (T5 [41] fine-tuned on the RecipeNLG dataset [6]), the sole publicly available open-source recipe generation model at the time (December 2023). Evaluation on a 1000 sample test set (*test1k*) drawn from the Recipe1M dataset (as detailed in Table 2) revealed LLaVA [32], a multi-modal LLM, to outperform

| Method | Inputs | BLEU-1 | BLEU-2 | SacreBLEU | METEOR | ROUGE-1 | ROUGE-2 | ROUGE-L | CIDEr | Perplexity ↓ |
|--------------------------|-----------------------|-------------|--------------|-------------|--------------|--------------|--------------|--------------|-------------|--------------|
| Chef Transformer [16] | X_{ing} | 0.271 | 0.128 | 0.037 | 0.117 | 0.259 | 0.057 | 0.133 | 0.046 | 54.21 |
| GPT-2 [40] | $X_t + X_{ing}$ | 0.084 | 0.032 | 0.01 | 0.037 | 0.111 | 0.018 | 0.088 | 0.01 | 2.15 |
| Mistral [22] | $X_t + X_{ing}$ | 0.126 | 0.072 | 0.04 | 0.079 | 0.179 | 0.055 | 0.106 | 0.05 | 26.45 |
| Phi-2 [20] | $X_t + X_{ing}$ | 0.143 | 0.07 | 0.027 | 0.147 | 0.202 | 0.047 | 0.108 | 0.002 | 32.41 |
| LLaMA [47] | $X_t + X_{ing}$ | 0.234 | 0.119 | 0.049 | 0.16 | 0.29 | 0.075 | 0.155 | 0.043 | 2.86 |
| InstructBLIP-T5xl [14] | $X_i + X_t + X_{ing}$ | 0.014 | 0.006 | 0.001 | 0.037 | 0.137 | 0.022 | 0.094 | 0.014 | 68.71 |
| InstructBLIP-Vicuna [14] | $X_i + X_t + X_{ing}$ | 0.0013 | 0.0004 | 0.0001 | 0.026 | 0.103 | 0.012 | 0.069 | 0.004 | 135.75 |
| MiniGPTv2 [9] | $X_i + X_t + X_{ing}$ | 0.232 | 0.115 | 0.06 | 0.139 | 0.257 | 0.06 | 0.135 | 0.03 | 157.0 |
| MiniGPT4-LLaMA-2 [55] | $X_i + X_t + X_{ing}$ | 0.2754 | 0.141 | 0.07 | 0.204 | 0.353 | 0.094 | 0.173 | 0.032 | 11.78 |
| LLaVA [32] | $X_i + X_t + X_{ing}$ | 0.29 | 0.155 | 0.06 | 0.2 | 0.366 | 0.105 | 0.184 | 0.041 | 2.6 |

Table 2: Performance of pretrained foundational models on our *test1k*. Notably, pretrained LLaVA, outperforms other evaluated models on most metrics, showcasing its ability to generate food recipes.

| Method | Inputs | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | SacreBLEU | METEOR | ROUGE-1 | ROUGE-2 | ROUGE-L | CIDEr | Perplexity ↓ |
|-----------------------|-----------------------|--------------|--------------|--------------|--------------|--------------|------------|--------------|--------------|--------------|--------------|--------------|
| Chef Transformer [16] | X_{ing} | 0.267 | 0.127 | 0.064 | 0.034 | 0.038 | 0.116 | 0.262 | 0.059 | 0.136 | 0.045 | 30.62 |
| Mistral [22] | $X_t + X_{ing}$ | 0.130 | 0.075 | 0.048 | 0.033 | 0.041 | 0.082 | 0.188 | 0.058 | 0.111 | 0.063 | 75.36 |
| LLaMA [47] | $X_t + X_{ing}$ | 0.252 | 0.129 | 0.072 | 0.043 | 0.053 | 0.156 | 0.293 | 0.077 | 0.156 | 0.031 | 2.86 |
| LLaVA [32] | $X_i + X_t + X_{ing}$ | 0.297 | 0.159 | 0.089 | 0.042 | 0.061 | 0.2 | 0.368 | 0.106 | 0.183 | 0.037 | 2.92 |
| LLaVA-Chef-S1 | $X_i + X_t + X_{ing}$ | 0.322 | 0.19 | 0.117 | 0.075 | 0.096 | 0.159 | 0.404 | 0.141 | 0.217 | 0.187 | 2.62 |
| LLaVA-Chef-S2 | $X_i + X_t + X_{ing}$ | 0.331 | 0.193 | 0.118 | 0.075 | 0.09 | 0.159 | 0.396 | 0.136 | 0.213 | 0.176 | 2.86 |
| LLaVA-Chef-S3 | $X_i + X_t + X_{ing}$ | 0.362 | 0.215 | 0.135 | 0.089 | 0.167 | 0.188 | 0.473 | 0.172 | 0.241 | 0.216 | 2.38 |

Table 3: Results on Recipe1M test set *test50K* (randomly selected 50,507 test samples, fixed for all models). Our model, LLaVA-Chef, gradually improves from Stage-1 to Stage-3 on almost all the metrics.

all contenders, including Chef Transformer. Consequently, LLaVA was chosen for further analysis and fine-tuned on the Recipe1M dataset for enhanced performance. Our training protocol employed a multi-stage fine-tuning approach. Initially, during Stage-0, we conducted fine-tuning for the projection layer over the course of two epochs. Subsequently, throughout the remaining three stages (Stage 1-3), the entire model was fine-tuning for two epochs in each stage.

Our analysis of current open-source LLMs presented in Table 2 reveals intriguing performance in the food domain. In case of text-only models, Chef-Transformer shown higher BLEU-1 and BLEU-2 scores but it has lower scores on SacreBLEU, METEOR, and Rouge-L than LLaMA, indicating potential trade-offs in generation quality. Whereas, comparing all the models, LLaVA seems to outperform. The higher perplexity scores suggests that, with the exception of LLaMA, MiniGPT-4 and LLaVA, all models struggle to generate good quality language, potentially generating text exhibiting hallucinations or incomplete sentences. Though Mistral has impressive performance on standard benchmarks, its higher perplexity score and scores for other metrics lower than Phi-2 raises questions about its effectiveness in this specific context. InstructBLIP [14] generated recipes for more like caption rather than cooking steps. The training data of MiniGPT-4 [55] contains food images paired with cooking instructions, hence it is comparable to Chef-Transformer for recipe generation on several metrics. Overall, LLaVA stands out, achieving remarkable performance on most metrics.

5.3 Quantitative Results

The results presented in Table 2 for *test1k* set demonstrate that the pre-trained LLaVA [32] outperforms other LLMs including Chef-Transformer [16], despite Chef-Transformer being trained on recipe dataset. A similar trend was also found on test set *test50k* as shown in Table 3, comparing top 4 row, LLaVA has higher scores. Our

LLaVA-Chef model therefore extends the baseline LLaVA model via our novel multi-stage training and fine-tuning framework outlined above. Notably, our model, LLaVA-Chef outperforms other models, with its BLEU and Rouge scores indicating the alignment of the generated cooking instructions with the ground truth.

Open source LLMs: Due to limited benchmarks for recipe generation, we explored the performance of prominent LLMs on the Hugging Face Leader board (December 20, 2023). These include well-established models like GPT-2 [40], and LLaMA [47], as well as recent high-performing options such as Mistral (7B parameters) [22] and Phi-2 [20]. We also considered four multi-modal models in our study including InstructBLIP [14], MiniGPTv2 [9], MiniGPT-4 [55] and LLaVA [32] due to their exceptional performance on visual-language tasks. Additionally, we evaluated Chef Transformer [16], a fine-tuned T5 [41] model specifically designed for recipe generation, offering an open-source option for comparison.

Comparison with existing methods: Direct comparison with the existing literature is challenging due to discrepancies in reported results and limited dataset accessibility. The partial availability of Recipe1M dataset and outdated URLs hinder consistent evaluation. For examples RecipeMC [46] is evaluated on 1000 samples from Recipe1M dataset but they did not share those samples. Similarly, FIRE [11] could get 56K samples from test set of Recipe1M dataset as few URL were no more accessible. In our case, we could get only 50, 507 test samples that contain at least one image per recipe. Although, the test set used by baseline methods and ours might be slightly different, the scores give us a general idea about the performance of the models.

Our model in general outperforms the baseline methods as evident in Table 4 and Table 5. We took the scores for Chef Transformer [16], Inverse Transformer [44], FIRE [11], and FoodLMM [53] from their respective publications. Additionally, we conducted

| Method | SacreBLEU | ROUGE-L |
|-----------------------|--------------|--------------|
| Inverse Cooking[44] | 0.055 | 0.195 |
| FIRE [11] | 0.06 | 0.212 |
| FoodLMM [53] | 0.062 | 0.369 |
| Chef Transformer [16] | 0.046 | 0.175 |
| Chef Transformer* | 0.038 | 0.136 |
| LLaVA* | 0.061 | 0.183 |
| LLaVA-Chef-S3* (Ours) | 0.158 | 0.228 |

Table 4: Results on Recipe1M test set: Due to inconsistency in datasets and lack of publicly available models, results based on our *test50k* benchmark dataset are marked with *.

| Method | Perplexity (gt/pred) ↓ | ROUGE-1 | ROUGE-2 | BLEU |
|---------------|------------------------|--------------|-------------|-------------|
| RecipeMC [46] | 2.934 / 7.337 | 0.362 | 0.115 | 0.08 |
| LLaVA [32] | 6.8 / 2.6 | 0.367 | 0.105 | 0.06 |
| LLaVA-Chef-S3 | 4.14 / 2.4 | 0.473 | 0.17 | 0.17 |

Table 5: Results on 1000 test recipes from Recipe1M dataset (gt: ground truth, pred: predicted or generated text). RecipeMC test recipes are taken from [46].

an evaluation of the publicly available Chef Transformer [16] on our *test50k* set. Intriguingly, our evaluation yielded lower scores for Chef Transformer compared to those reported in its original publication. Notably, the pretrained general-purpose LLaVA [32] marginally surpassed FIRE and is close to FoodLMM in terms of SacreBLEU score. Despite being built upon LLaVA, FoodLMM [53] only achieved a 1-point improvement in SacreBLEU score, although its Rouge-L score is significantly higher.

On the other hand, our model, LLaVA-Chef, as seen in Table 4 demonstrates superior performance, achieving a remarkable nearly 10-point margin over other models in SacreBLEU score, even with second best Rouge-L score. As shown in Table 5, LLaVA-Chef surpasses RecipeMC on both Rouge and BLEU scores. This significant performance gain validates the effectiveness of our approach.

Performance on different cuisines: To evaluate the generalization of LLaVA-Chef, we report the performance of our model on test samples from different cuisines in Table 6, and compare with scores on *test1k*. For most of the cuisines, BLEU and Rouge scores are almost same. Our model shows lowest Rouge scores for French and higher perplexity for German. In general, most of the scores are close to the overall scores on *test1k* set indicating the model generalizes across cuisines, even for those with few training examples (e.g., Japanese or Russian).

5.4 Qualitative Results

Beyond quantitative metrics, evaluating the qualitative aspect of generated recipes is crucial. Figure 2 presents two recipes generated by Chef-Transformer, LLaVA and LLaVA-Chef (Ours). In the left-hand example, all models recommend a lower *temperature* than the ground truth, but the *baking time* remains consistent. In the right-hand example, all models suggest the same *oven temperature* but vary in recommended *cooking time*. LLaVA-Chef generates concise recipes with high accuracy, often surpassing other models and even the ground truth in clarity. When manually looking at the generated recipe, we observe that GPT-2, Mistral and Phi-2

struggle to produce a cohesive recipe, Chef Transformer generated recipe do not have sufficient information, LLaMA sometime fails to generate correct recipes, and InstructBLIP generates text which looks like a caption rather than cooking steps. LLaVA generates detailed recipes but hallucination is common in generated text. However, our LLaVA-Chef generated recipe is concise and closely resembles human generated ground truth recipe.

We also look at how our LLaVA-Chef’s multi-stage approach successively improves the generated recipes. We found that Stage-1 exhibits minor discrepancies, while Stage-3 generates accurate recipes with correct ingredients (see Figure 5). Further analysis reveals that sometimes the recipes are semantically equivalent but linguistically different causing lower scores compared to the ground-truth. Finally, we looked at the impact of combinations of food image X_i , title X_t , ingredients X_{ing} as inputs to our model. We find that solely relying on the image sometimes makes dish prediction difficult, leading to a flawed recipe, though high quality images can provide good results. Providing the title significantly improves the generation. While LLaVA-Chef achieves promising results on the Recipe1M dataset, certain limitations emerged upon closer examination. To summarize, some recipes closely resemble the corresponding ground-truth recipes, while others exhibit significant linguistic divergence resulting in lower Rouge-L scores even though generated recipes are semantically equivalence with the ground truth. For instance, a single step of the ground truth recipe is sometimes split into several steps in the generated recipes, conveying the same information but with different phrasing.

5.5 Ablation Study

Improvement through multi-stage training: LLaVA-Chef’s training in a multi-stage setup demonstrates a gradual improvement in its recipe generation capabilities, as evident from scores in Table 7. Pre-trained LLaVA generates recipes with hallucinations and sometimes discrepancies from the ground truth. However, LLaVA-Chef improves in every stage by a noticeable margin. The example in Figure 5 shows that LLaVA-Chef-S1 correctly estimates the required temperature, but it misjudges the mixing pattern of the ingredients and baking time. In Stage-2, it instructs to combine all ingredients in one step, though it misses an ingredient (garlic). While minor discrepancies in instructions remain, the ability to accurately list all ingredients in Stage-3 highlights the model’s learning trajectory and potential.

Impact of scaling loss: As discussed earlier, after stage-2, we introduce a penalty by scaling the loss based on BLEU and Rouge scores and continued training for 2 epochs. The resulting model is LLaVA-Chef-S3. To evaluate the improvement through this additional penalty, we continued training of S2 for two more epochs with only cross-entropy loss, the resulting model is LLaVA-Chef-S22. As evident in Table 8, although both models have been trained for 2 additional epochs after S2, the difference in performance directly reflects the impact of our novel penalty formulation.

Impact of input attributes: We also assess LLaVA and LLaVA-Chef models under various input configurations, including scenarios where only the food image is provided (excluding title and ingredients), the food image with the title (excluding ingredients),

| Cuisine | Train Samples | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | SacreBLEU | METEOR | ROUGE-1 | ROUGE-2 | ROUGE-L | CIDEr | Perplexity ↓ |
|----------------|---------------|--------|--------|--------|--------|-----------|--------|---------|---------|---------|-------|--------------|
| North-American | 13119 | 0.377 | 0.216 | 0.13 | 0.082 | 0.127 | 0.185 | 0.445 | 0.133 | 0.198 | 0.202 | 16.36 |
| American | 9026 | 0.376 | 0.215 | 0.129 | 0.081 | 0.125 | 0.184 | 0.411 | 0.124 | 0.183 | 0.2 | 18.79 |
| European | 5683 | 0.384 | 0.22 | 0.133 | 0.084 | 0.143 | 0.187 | 0.411 | 0.125 | 0.204 | 0.198 | 18.60 |
| Asian | 2526 | 0.377 | 0.215 | 0.129 | 0.08 | 0.152 | 0.184 | 0.511 | 0.172 | 0.232 | 0.194 | 10.46 |
| Mexican | 2472 | 0.379 | 0.218 | 0.131 | 0.083 | 0.131 | 0.186 | 0.507 | 0.15 | 0.224 | 0.205 | 11.9 |
| Italian | 2047 | 0.393 | 0.226 | 0.137 | 0.087 | 0.214 | 0.19 | 0.526 | 0.194 | 0.273 | 0.193 | 9.57 |
| Indian | 544 | 0.377 | 0.215 | 0.128 | 0.081 | 0.104 | 0.184 | 0.531 | 0.173 | 0.229 | 0.193 | 11.34 |
| French | 427 | 0.377 | 0.215 | 0.129 | 0.081 | 0.04 | 0.184 | 0.097 | 0.011 | 0.043 | 0.191 | 15.29 |
| English | 248 | 0.377 | 0.216 | 0.13 | 0.083 | 0.106 | 0.185 | 0.241 | 0.038 | 0.148 | 0.226 | 16.42 |
| Middle-Eastern | 267 | 0.376 | 0.215 | 0.128 | 0.08 | 0.134 | 0.184 | 0.487 | 0.103 | 0.218 | 0.197 | 11.43 |
| Thai | 252 | 0.374 | 0.214 | 0.127 | 0.08 | 0.315 | 0.183 | 0.664 | 0.311 | 0.28 | 0.188 | 8.56 |
| German | 247 | 0.378 | 0.216 | 0.129 | 0.081 | 0.068 | 0.185 | 0.366 | 0.091 | 0.173 | 0.192 | 42.98 |
| Russian | 223 | 0.377 | 0.215 | 0.129 | 0.081 | 0.135 | 0.184 | 0.459 | 0.106 | 0.182 | 0.189 | 8.71 |
| Japanese | 132 | 0.38 | 0.218 | 0.13 | 0.082 | 0.055 | 0.186 | 0.362 | 0.101 | 0.2 | 0.198 | 10.48 |
| test1k | - | 0.366 | 0.218 | 0.137 | 0.09 | 0.17 | 0.189 | 0.473 | 0.17 | 0.24 | 0.242 | 17.9 |

Table 6: Performance of LLaVA-Chef on generating recipe that belong to different cuisines

| Ground Truth | LLaVA-Chef |
|--|--|
| <p>1. Prompt: Compose a recipe for Avocado-Hummus Dip incorporating the following ingredients: 2 medium Hass avocados, peeled and chopped, One 15-ounce can chickpeas, drained and rinsed, 1/4 cup fresh lemon juice, 1 1/2 tablespoons tahini (sesame paste), 1/2 cup extra-virgin olive oil, plus more for drizzling, Kosher salt, Freshly ground pepper, Assorted crudites, bread and tortilla chips, for serving.</p> <p>1. In a food processor, puree the avocados with the chickpeas, lemon juice and tahini.\n2. Add the 1/2 cup of oil and puree until smooth; season with salt and pepper.\n3. Transfer the dip to a bowl, drizzle with olive oil and serve with crudites, bread and chips.</p> | <p>1. In a food processor, combine the avocados, chickpeas, lemon juice, tahini, olive oil, and salt and pepper.\n2. Process until smooth and creamy.\n3. Transfer to a serving bowl and drizzle with more olive oil.\n4. Serve with crudites, bread slices, and tortilla chips.</p> |
| <p>2. Prompt: Generate a comprehensive recipe for crafting Anchovy and Capers Dressing with the listed ingredients: 2 tablespoons lemon juice, 5 tablespoons olive oil, 1 teaspoon fresh ground black pepper, 6 anchovy fillets, pounded to a coarse paste, 1 tablespoon capers, chopped.</p> <p>1. Mix together all ingredients.\n2. Salt should not be needed because of the anchovies.</p> | <p>1. In a small bowl, whisk together the lemon juice, olive oil, and pepper.\n2. Add the anchovy paste and capers.\n3. Whisk again to combine.\n4. Taste and adjust seasonings if necessary.\n5. The dressing can be refrigerated for up to 2 days.</p> |
| <p>3. Prompt: Given 1 1/2 cups buttermilk, 1 1/2 tablespoons margarine, 2 tablespoons white sugar, 1 teaspoon salt, 3 cups bread flour, 1 1/3 cups whole wheat flour, 2 1/4 teaspoons active dry yeast, what are the key steps you need to follow to prepare a perfect Buttermilk Bread ?</p> <p>1. Place buttermilk, butter or margarine, sugar, salt, flour, whole wheat flour, and yeast into pan of bread machine.\n2. Bake on White Bread setting.\n3. Cool on wire racks before slicing.</p> | <p>1. Mix all ingredients together.\n2. Let the dough rise for 1 hour.\n3. Punch down the dough.\n4. Let the dough rise for 1 hour.\n5. Shape the dough into loaves.\n6. Let the loaves rise for 1 hour.\n7. Bake at 350 degrees for 30 minutes.</p> |

Figure 4: Sample recipes produced by the LLaVA-Chef-S3 model.

| Method | Inputs | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | SacreBLEU | METEOR | ROUGE-1 | ROUGE-2 | ROUGE-L | CIDEr | Perplexity ↓ |
|---------------|-----------------------|--------|--------|--------|--------|-----------|--------|---------|---------|---------|-------|--------------|
| LLaVA | X_i | 0.152 | 0.057 | 0.024 | 0.011 | 0.015 | 0.096 | 0.178 | 0.026 | 0.1 | 0.004 | 16.76 |
| | X_t | 0.213 | 0.101 | 0.051 | 0.027 | 0.03 | 0.144 | 0.262 | 0.059 | 0.136 | 0.019 | 32.94 |
| | $X_i + X_t$ | 0.158 | 0.061 | 0.025 | 0.011 | 0.016 | 0.104 | 0.195 | 0.029 | 0.109 | 0.005 | 14.62 |
| | $X_i + X_{ing}$ | 0.277 | 0.144 | 0.079 | 0.045 | 0.054 | 0.182 | 0.349 | 0.095 | 0.177 | 0.036 | 2.77 |
| | $X_t + X_{ing}$ | 0.293 | 0.157 | 0.088 | 0.051 | 0.061 | 0.2 | 0.367 | 0.106 | 0.183 | 0.046 | 2.38 |
| | $X_i + X_t + X_{ing}$ | 0.29 | 0.154 | 0.087 | 0.051 | 0.06 | 0.2 | 0.367 | 0.105 | 0.182 | 0.041 | 2.6 |
| LLaVA-Chef-S1 | X_i | 0.144 | 0.059 | 0.027 | 0.014 | 0.021 | 0.067 | 0.207 | 0.035 | 0.13 | 0.014 | 3.54 |
| | X_t | 0.225 | 0.111 | 0.06 | 0.035 | 0.048 | 0.107 | 0.278 | 0.069 | 0.158 | 0.065 | 3.54 |
| | $X_i + X_t$ | 0.227 | 0.115 | 0.063 | 0.037 | 0.05 | 0.108 | 0.283 | 0.073 | 0.162 | 0.065 | 2.89 |
| | $X_i + X_{ing}$ | 0.253 | 0.144 | 0.088 | 0.055 | 0.074 | 0.137 | 0.357 | 0.113 | 0.196 | 0.168 | 3.0 |
| | $X_t + X_{ing}$ | 0.325 | 0.191 | 0.119 | 0.076 | 0.097 | 0.16 | 0.404 | 0.14 | 0.218 | 0.201 | 2.63 |
| | $X_i + X_t + X_{ing}$ | 0.327 | 0.192 | 0.116 | 0.074 | 0.096 | 0.16 | 0.405 | 0.14 | 0.219 | 0.198 | 3.54 |
| LLaVA-Chef-S2 | X_i | 0.188 | 0.082 | 0.04 | 0.021 | 0.047 | 0.078 | 0.223 | 0.042 | 0.136 | 0.016 | 2.3 |
| | X_t | 0.253 | 0.13 | 0.072 | 0.042 | 0.076 | 0.11 | 0.294 | 0.079 | 0.166 | 0.078 | 2.71 |
| | $X_i + X_t$ | 0.256 | 0.131 | 0.072 | 0.042 | 0.076 | 0.111 | 0.298 | 0.078 | 0.167 | 0.081 | 2.74 |
| | $X_i + X_{ing}$ | 0.308 | 0.175 | 0.106 | 0.067 | 0.111 | 0.151 | 0.378 | 0.12 | 0.204 | 0.174 | 2.77 |
| | $X_t + X_{ing}$ | 0.338 | 0.197 | 0.121 | 0.078 | 0.124 | 0.163 | 0.407 | 0.138 | 0.219 | 0.179 | 2.6 |
| | $X_i + X_t + X_{ing}$ | 0.337 | 0.196 | 0.121 | 0.078 | 0.124 | 0.163 | 0.41 | 0.14 | 0.221 | 0.189 | 2.63 |
| LLaVA-Chef-S3 | X_i | 0.209 | 0.092 | 0.042 | 0.021 | 0.082 | 0.091 | 0.242 | 0.048 | 0.135 | 0.011 | 1.72 |
| | X_t | 0.283 | 0.149 | 0.081 | 0.047 | 0.116 | 0.142 | 0.37 | 0.108 | 0.193 | 0.094 | 2.08 |
| | $X_i + X_t$ | 0.293 | 0.155 | 0.086 | 0.049 | 0.123 | 0.146 | 0.373 | 0.11 | 0.195 | 0.102 | 2.05 |
| | $X_i + X_{ing}$ | 0.337 | 0.197 | 0.12 | 0.077 | 0.156 | 0.177 | 0.45 | 0.156 | 0.232 | 0.203 | 2.43 |
| | $X_t + X_{ing}$ | 0.362 | 0.213 | 0.132 | 0.086 | 0.16 | 0.187 | 0.471 | 0.166 | 0.249 | 0.215 | 2.41 |
| | $X_i + X_t + X_{ing}$ | 0.366 | 0.218 | 0.137 | 0.09 | 0.17 | 0.189 | 0.473 | 0.17 | 0.24 | 0.242 | 2.4 |

Table 7: We analyzed the role of different information sources in generating cooking instructions on the test1K subset of the Recipe1M test set. While food images provide valuable context, our ablation study reveals that food names and ingredients are essential for accurate results.


| | |
|---|--|
|  | Prompt: Generate a comprehensive recipe for crafting Stuffed Artichokes using the following ingredients: 2 lbs lean ground beef, 7 ounces fresh mushrooms chopped, 1 large onion, diced finely, 2 large eggs, 12 cup celery, diced finely, 12 cup tomato puree, 12 cup water, 1 teaspoon salt to taste, 12 teaspoon freshly ground black pepper, 1 teaspoon thyme, 3 cloves garlic, minced, 24 large canned artichoke bottoms or 24 large frozen artichoke bottoms, 2 cups beef broth. |
| Ground Truth | 1. Preheat the oven to 350 degrees F.\n2. Combine all the ingredients except for the artichoke bottoms and the beef broth.\n3. Divide and mound the meat mixture on the artichoke bottoms, about 4 tablespoons each.\n4. In a well greased baking pan, place the filled artichokes, filled side up, in one layer.\n5. Pour the broth over the whole thing.\n6. Bake, lightly covered with foil, for 1 1/2 hours.\n7. Uncover and bake another 10-15 minutes, until browned.\n8. Serve hot. |
| LLaVA-Chef-S3 | 1. Preheat oven to 350 degrees F.\n2. In a large skillet, saute the ground beef, mushrooms, onion, and garlic until the meat is browned.\n3. Drain off the fat.\n4. In a large bowl, combine the meat mixture with the eggs, celery, tomato puree, water, salt, pepper, and thyme.\n5. Stuff the artichoke bottoms with the meat mixture.\n6. Place the stuffed artichokes in a large baking dish.\n7. Pour the beef broth over the artichokes.\n8. Cover the dish with aluminum foil and bake for 1 hour.\n9. Remove the foil and bake for an additional 15 minutes. |
| LLaVA-Chef-S2 | 1. Preheat oven to 350 degrees F.\n2. In a large bowl, combine the ground beef, mushrooms, onion, celery, eggs, tomato puree, water, salt, pepper, and thyme.\n3. Mix well.\n4. Stuff the artichoke bottoms with the mixture.\n5. Place the artichoke bottoms in a large baking dish.\n6. Pour the beef broth over the artichokes.\n7. Cover the dish with aluminum foil and bake for 45 minutes.\n8. Remove the foil and bake for an additional 15 minutes. |
| LLaVA-Chef-S1 | 1. Preheat oven to 350 degrees F.\n2. In a large skillet, brown the ground beef with the onion and garlic.\n3. Add the mushrooms and cook until softened.\n4. Drain the artichoke bottoms and add them to the skillet.\n5. Stir in the tomato puree, water, salt, pepper and thyme.\n6. Bring to a boil, reduce heat, and simmer for 15 minutes.\n7. Place the artichoke bottoms in a baking dish and cover with the meat mixture.\n8. Bake for 30 minutes.\n9. Serve with rice or pasta. |
| LLaVA | To prepare a delicious Stuffed Artichokes, follow these steps:\n1. Preheat your oven to 375°F (190°C).\n2. In a large bowl, combine the lean ground beef, mushrooms, onion, and eggs. Season the mixture with salt, pepper, and thyme to taste.\n3. Add the celery, tomato puree, water, and minced garlic to the bowl. Mix everything together until well combined.\n4. Place the mixture in a large baking dish and bake for about 45 minutes, or until the filling is cooked through and the edges are slightly browned.\n5. While the filling is cooking, prepare the artichoke bottoms. If using fresh artichoke bottoms, trim the ends and remove the tough outer leaves. If using frozen artichoke bottoms, follow the package instructions for preparation.\n6. After the filling is cooked, carefully remove it from the oven and let it cool for a few minutes.\n7. Carefully stuff the cooked filling into the artichoke bottoms, dividing the mixture evenly among the artichokes.\n8. Pour the beef broth over the stuffed artichokes, making sure to cover them completely.\n9. Return the baking dish to the oven and bake for another 25-30 minutes, or until the artichokes are tender and the filling is hot and bubbly.\n10. Serve the Stuffed Artichokes hot, garnished with fresh parsley if desired.\n\nEnjoy your delicious Stuffed Artichokes! |

Figure 5: Example recipes generated by pre-trained LLaVA and each stage of our model. We can see how each stage successively improves the generated recipe, showcasing the effectiveness of our multi-stage training.

| Method | Loss | BLEU-1 | BLEU-4 | SacreBLEU | ROUGE-L | CIDEr |
|---------------|------------------------|--------------|-------------|-------------|-------------|--------------|
| LLaVA-Chef-S2 | L_{CE} | 0.372 | 0.08 | 0.158 | 0.227 | 0.191 |
| LLaVA-Chef-S3 | $L_{BR} \times L_{CE}$ | 0.366 | 0.09 | 0.17 | 0.24 | 0.242 |

Table 8: Effect of language quality penalty loss function.

and title with ingredients. The evaluation is conducted on the *test1k* test set, and the outcomes are summarized in Table 7. Our LLaVA-Chef model improves in each steps, outperforms others, showing the effectiveness of our multi-stage approach. Our initial observations revealed that images alone convey less semantic information about the food compared to food names. This is likely due to the limitations of visual information captured in images. Nevertheless, title and ingredients remain a crucial factor in recipe generation as evident by increase in scores when both are input to the model.

Incorporating images alongside textual prompts failed to improve the performance of a pre-trained LLaVA model for recipe generation tasks. This might be attributed to limitations in the model’s ability to map visual features of food images effectively into the language space. Conversely, our fine-tuned LLaVA-Chef-S1 exhibits minimal performance enhancement from image integration, regardless of its placement alongside the title or in conjunction with both title and ingredients. LLaVA-Chef-S2 exposed to a wider variety of prompts during training, demonstrates significant improvement over LLaVA when presented with solely an image. Although titles and ingredients remain essential for generating accurate cooking instructions. Our final model, LLaVA-Chef-S3, generally achieves superior scores. Interestingly, LLaVA-Chef-S3, when prompted solely with an image (X_i), achieved the lowest perplexity score, but it has underwhelming performance on other metrics. Notably, while all models, including Chef-Transformer,

exhibited CIDEr scores lower than 1, our final model achieved an impressive improvement of nearly 24 points in this metric.

6 CONCLUSION

This work presents LLaVA-Chef, a multi-modal model trained for recipe generation. Through systematic evaluation of prominent open-source LLMs, we identified LLaVA as the optimal starting point. Subsequent fine-tuning utilized specially curated prompts to progressively guide the model’s adaptation to the food domain. Our multi-stage method incorporated diverse prompts and a novel language quality penalty loss function, leading to significant performance gains that surpass existing methods by noticeable margins yielding state-of-the-art performance for this task. Notably, the final model, LLaVA-Chef-S3, generates recipes that are demonstrably more accurate and detailed than its predecessors, often featuring precise ingredient mentions that enhance understandability and sometimes even surpasses the quality of human-authored ground truth recipes. These findings highlight the effectiveness of our stage-wise fine-tuning approach and paves the way for further advancements for food-related tasks. While LLaVA-Chef outperforms other models in recipe generation tasks, it lacks the capability to suggest ingredient substitutions with accompanying justifications regarding health impacts. Future research will focus on expanding LLaVA-Chef’s functionalities beyond recipe generation to incorporate ingredient substitution while considering dietary constraints. Another interesting direction is to consider numeric information in evaluating generated recipes, such as cooking time or temperature, ingredient quantities, and so on.

REFERENCES

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report.
- [2] Google AI. 2023. Bard: An experimental large language model. <https://ai.google/research/projects/bard/>. Accessed: February 5, 2024.
- [3] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, et al. 2022. Falcon-40B: An open large language model with state-of-the-art performance.
- [4] Dang Hoang Anh, Dinh-Truong Do, Vu Tran, and Nguyen Le Minh. 2023. The Impact of Large Language Modeling on Natural Language Processing in Legal Texts: A Comprehensive Survey. In *15th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 1–7.
- [5] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models.
- [6] 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*.
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [8] Jingjing Chen, Bin Zhu, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. 2021. A Study of Multi-Task and Region-Wise Deep Learning for Food Ingredient Recognition. *IEEE Transactions on Image Processing* 30 (2021), 1514–1526. <https://doi.org/10.1109/tip.2020.3045639>
- [9] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning.
- [10] Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. 2018. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM. <https://doi.org/10.1145/3240508.3240627>
- [11] Prateek Chhikara, Dhiraj Chaurasia, Yifan Jiang, Omkar Masur, and Filip Ilievski. 2023. FIRE: Food Image to REcipe generation. *arXiv preprint arXiv:2308.14391* (2023).
- [12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023) (2023).
- [13] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *arXiv preprint arXiv:2401.01301* (2024).
- [14] W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500 2* (2023).
- [15] Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1292–1302.
- [16] Mehrdad Farahani, Kartik Godawat, Haswanth Aekula, Deepak Pandian, and Nicholas Broad. Dec 16, 2023. Chef Transformer.
- [17] Bahare Fatemi, Quentin Duval, Rohit Girdhar, Michal Drozdal, and Adriana Romero-Soriano. 2023. Learning to Substitute Ingredients in Recipes. *arXiv preprint arXiv:2302.07960* (2023).
- [18] 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*.
- [19] Jiangpeng He and Fengqing Zhu. 2021. Online continual learning for visual food classification. In *Proceedings of the IEEE/CVF international conference on computer vision*.
- [20] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Jan 10, 2024. Phi-2: The surprising power of small language models. <https://huggingface.co/microsoft/phi-2> (Jan 10, 2024).
- [21] Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* (1977).
- [22] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
- [23] Rajdeep Kaur, Rakesh Kumar, and Meenu Gupta. 2023. Deep neural network for food image classification and nutrient identification: A systematic review. *Reviews in Endocrine and Metabolic Disorders* (2023).
- [24] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, Vol. 1. 2.
- [25] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692* (2023).
- [26] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890* (2023).
- [27] Diya Li and Mohammed J Zaki. 2022. Food Knowledge Representation Learning with Adversarial Substitution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*.
- [28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.
- [29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.
- [30] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.
- [31] Chin-Yew Lin and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*. 605–612.
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).
- [33] Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. 2023. FinGPT: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485* (2023).
- [34] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*. PMLR.
- [35] Dim P. Papadopoulos, Enrique Mora, Nadiia Chepurko, Kuan Wei Huang, Ferda Ofli, and Antonio Torralba. 2022. Learning Program Representations for Food Images and Cooking Recipes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*.
- [37] Chantal Pellegrini, Ege Ozsoy, Monika Wintergerst, and Georg Groh. 2021. Exploiting Food Embeddings for Ingredient Substitution. In *HEALTHINF*.
- [38] 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, Brussels, Belgium, 186–191.
- [39] 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.
- [40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* (2019).
- [41] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* (2020).
- [42] Anja Reusch, Alexander Weber, Maik Thiele, and Wolfgang Lehner. 2021. RecipeGM: A Hierarchical Recipe Generation Model. In *2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW)*. IEEE, 24–29.
- [43] Jesús M Rodríguez-de Vera, Pablo Villacorta, Imanol G Estepa, Marc Bolaños, Ignacio Sarasúa, Bhalaji Nagarajan, and Petia Radeva. 2023. Dining on Details: LLM-Guided Expert Networks for Fine-Grained Food Recognition. In *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*.
- [44] 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*.
- [45] 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 Conference on Computer Vision and Pattern Recognition*. IEEE.
- [46] Karan Taneja, Richard Segal, and Richard Goodwin. 2024. Monte Carlo Tree Search for Recipe Generation using GPT-2. *arXiv preprint arXiv:2401.05199* (2024).

- [47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models.
- [48] Lav R Varshney, Florian Pinel, Kush R Varshney, Debarun Bhattacharjya, Angela Schörgendorfer, and Yi-Min Chee. 2019. A big data approach to computational creativity. *IBM Journal of Research and Development* (2019).
- [49] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 4566–4575.
- [50] Muntasir Wahed, Xiaona Zhou, Tianjiao Yu, and Ismini Lourentzou. 2024. Fine-Grained Alignment for Cross-Modal Recipe Retrieval. In *Proceedings of the Winter Conference on Applications of Computer Vision*. IEEE, 5584–5593.
- [51] Hao Wang, Guosheng Lin, Steven C. H. Hoi, and Chunyan Miao. 2020. *Structure-Aware Generation Network for Recipe Generation from Images*. Springer International Publishing, 359–374. https://doi.org/10.1007/978-3-030-58583-9_22
- [52] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).
- [53] Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. 2023. FoodLM: A Versatile Food Assistant using Large Multi-modal Model. *arXiv preprint arXiv:2312.14991* (2023).
- [54] Qing Zhang, David Elswiler, and Christoph Trattner. 2023. Understanding and predicting cross-cultural food preferences with online recipe images. *Information Processing & Management* 60, 5 (2023), 103443.
- [55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing vision-language understanding with advanced large language models.