# Pic2Prep: A Multimodal Conversational Agent for Cooking Assistance

**Renjith Prasad Kaippilly Mana, Chathurangi Shyalika, Revathy Venkataramanan, Darssan L. Eswaramoorthi, Amit P. Sheth**

Artificial Intelligence Institute, University of South Carolina, USA
kaipplir@mailbox.sc.edu, jayakodc@email.sc.edu, revathy@email.sc.edu, darssan@email.sc.edu, amit@sc.edu

## Abstract

As the demand for healthier, personalized culinary experiences grows, so does the need for advanced food computation models that offer more than basic nutritional insights. However, current food computation models lack the depth to provide actionable insights like ingredient substitution or alternative cooking actions to suit users' dietary goals. To address this, we introduce and demonstrate Pic2Prep, a multimodal conversational system that generates detailed cooking instructions, actions and ingredient lists from both images and text provided by users. The system is developed using a novel dataset generated through Stable Diffusion, where the input consists of recipe titles and ingredient lists from the Recipe1M dataset to create synthesized food images with variations. This dataset is used to fine-tune the Bootstrapping Language-Image Pre-training (BLIP) model to extract cooking instructions and ingredients from food images. Pic2Prep also employs the CookGen model, a small-scale custom generative model to derive specific cooking actions from cooking instructions. A custom mapper, trained on the Mistral model, links these actions to the corresponding ingredients, creating a comprehensive understanding of the cooking process. The system features an interactive user interface that allows users to input images and ask targeted questions, receiving real-time responses.

## Introduction

The saying *"we eat with our eyes first"* highlights a deep connection between visual perception and culinary experience. Food computation has become an increasingly important field, driven by the growing demand for healthier, personalized and sustainable culinary experiences (Min and Jiang 2019). With the rise of platforms like Instagram and YouTube, visual data has become a key source of information, offering unique insights into dishes and ingredients that cannot always be captured through text alone (Tahir and Loo 2021). Recent advancements in computer vision and vision-language models have made it possible to interpret food images for tasks like ingredient recognition and nutritional estimation (Fu and Dai 2024; Dhariwal and Nichol 2021).

However, current vision-based food computation models lack the depth to offer insights on cooking actions and ingredients to do complex tasks like ingredient substitution or recipe modifications (Toledo, Alzahrani, and Martinez 2019). They also fall short of understanding how variations in cooking methods can affect the overall dish. For example, using different cooking methods for the same ingredients can lead to varying nutritional profiles, textures and flavors, such as the difference between roasting vegetables for a caramelized taste versus steaming them for nutrient preservation. Moreover, current models often restrict their scope and do not adopt a comprehensive perspective on recipes, overlooking the integration of both ingredients and cooking processes. For instance, they struggle to link ingredients with cooking actions or generate step-by-step instructions, which are crucial for practical applications in both home and professional kitchens.

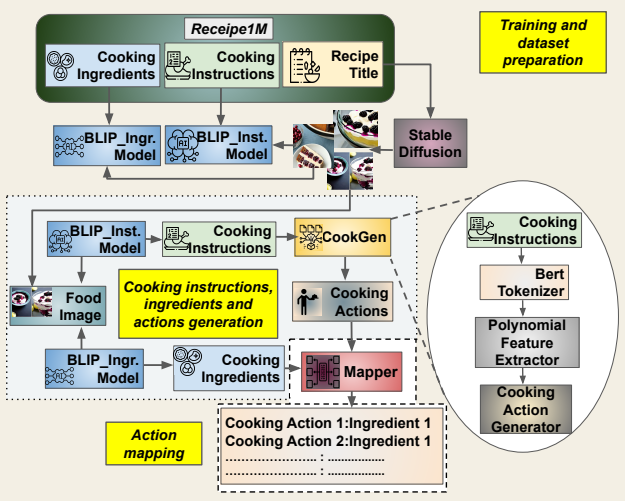


Figure 1: The architecture of Pic2Prep consists of three stages: (1) image generation to train two BLIP models for cooking instructions and ingredient extraction, (2) utilizing these BLIP models with CookGen to generate cooking instructions, ingredient lists and actions and (3) a custom mapper to link the cooking instructions with corresponding ingredients.

To address these limitations, we introduce *Pic2Prep*, a multimodal conversational agent that integrates vision and language models to provide comprehensive cooking instructions and ingredient lists directly from images. The main contributions of this work are as follows: (i) We generate a novel dataset of food images using Stable Diffusion (Rom-

bach et al. 2022), using recipe titles and ingredient lists from the Recipe1M dataset (Salvador et al. 2017) as input. These generated images are used to train and fine-tune BLIP model (Li et al. 2022) for extracting detailed ingredients and cooking instructions; (ii) We employ the CookGen model (Venkataramanan et al. 2023) to generate step-by-step cooking actions based on these instructions; (iii) A custom mapper, trained with the Mistral language model (Jiang et al. 2024), ensures precise linking of actions to ingredients, providing a seamless understanding of the cooking process. (iv) Building on this foundation, *Pic2Prep* offers an interactive chatbot where users can upload images and ask specific questions such as "What are the ingredients?" or "What are the cooking steps?" to receive real-time, personalized responses that enhance the cooking experience.

## System Overview

Pic2Prep utilizes a multimodal framework to generate cooking instructions, actions and ingredient lists from visual and textual inputs. The pipeline consists of three key stages: synthetic image generation, instruction and ingredient modeling and action mapping as shown in Figure 1.

**Synthetic Image Generation**: To train the model, we generate a large set of synthetic images using the Recipe1M dataset, which contains 1 million recipes with titles, ingredients, instructions and images. We input recipe titles into a Stable Diffusion model to generate multiple images for each recipe. This diversity in visual representations helps improve the model's ability to generalize and handle variations in real-world scenarios.

**Instruction and Ingredient Modeling**: We train two separate BLIP models: one to generate cooking instructions and another for ingredient extraction from the synthetic data. The combination of these models allows us to create a robust dataset linking visual content with textual descriptions, enhancing the system's understanding of cooking processes.

**Action Mapping**: The instructions generated are fed into the CookGen model to infer detailed cooking actions. Finally, a Mistral model is used to map each cooking action to its corresponding ingredients, creating a complete and accurate representation of the cooking process. This approach ensures that actions and ingredients are linked, to conduct downstream tasks of ingredient or cooking action subsitution effectively

## Demonstration

**User Interface and Interaction**: Pic2Prep features a user-friendly interface divided into two main sections: a section to upload food images and a chatbot interaction panel enabling users to ask cooking-related questions, as shown in Figure 2. This design ensures easy navigation and interaction between the user and the backend models.

**User Inputs and Queries**: Users initiate interactions by uploading a food image and typing their questions into the chatbot panel. The system supports a variety of queries related to the recipe, such as:

- *What are the ingredients in this image?*
- *What are the cooking actions for these instructions?*

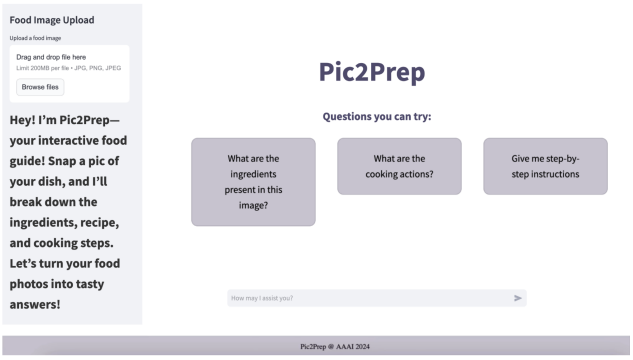


Figure 2: Main interface of Pic2Prep

- *Can you provide step-by-step instructions?*

If the image is unrelated to food or a query falls outside the system's scope, the system provides a fallback response: "I don't know."

**Real-time Interaction and Response**: Once a query is submitted, the system dynamically processes the input in real-time. Depending on the nature of the question, the corresponding model is invoked to generate an appropriate response. For instance, the BLIP model extracts ingredients from food images, while the CookGen and Mistral-based models handle cooking actions and instructions, as illustrated in Figure 3.

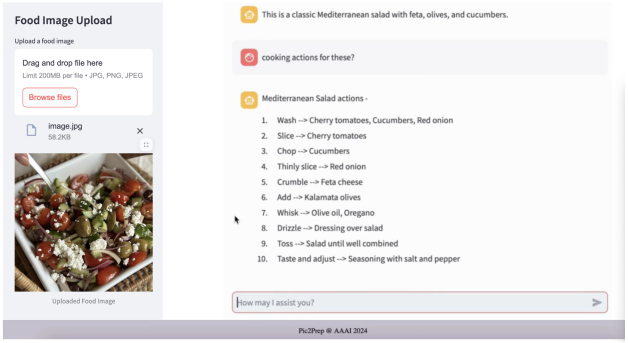


Figure 3: Pic2Prep responding to user queries and providing real-time interaction

## Conclusion

In this paper, we introduced Pic2Prep, a novel multimodal conversational agent that combines vision-language models and interactive AI-driven processing to generate cooking instructions, actions and ingredients from images and text.Pic2Prep addresses several limitations present in existing food computation systems by offering real-time interaction and an intuitive user interface. The system paves the way for more advanced applications, such as ingredient substitution, dietary recommendation systems and even fully automated kitchen assistance. Future work will explore the expansion of Pic2Prep to include more complex cooking processes and additional queries. Code is available at (Prasad 2024a) and demo is available at (Prasad 2024b)

## References


Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Fu, K.; and Dai, Y. 2024. Recognizing Multiple Ingredients in Food Images Using a Single-Ingredient Classification Model. arXiv:2401.14579.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2024. Mixtral of Experts. arXiv:2401.04088.

Min, W.; and Jiang, W. 2019. A comprehensive review on food computational approaches: Models, applications, and challenges. *Trends in Food Science & Technology*, 85: 56–68.

Prasad, R. 2024a. Pic2Prep - Code Repository. https://github.com/renjithk4/Pic2Prep. Accessed: Jan 30, 2025.

Prasad, R. 2024b. Pic2Prep - Demonstration Video. https://youtu.be/M_KRwD9BMQU. Accessed: Jan 30, 2025.

Tahir, G. A.; and Loo, C. K. 2021. A comprehensive survey of image-based food recognition and volume estimation methods for dietary assessment. In *Healthcare*, volume 9, 1676. MDPI.

Toledo, R. Y.; Alzahrani, A. A.; and Martinez, L. 2019. A food recommender system considering nutritional information and user preferences. *IEEE Access*, 7: 96695–96711.

Venkataramanan, R.; Roy, K.; Raj, K.; Prasad, R.; Zi, Y.; Narayanan, V.; and Sheth, A. 2023. Cook-Gen: Robust Generative Modeling of Cooking Actions from Recipes. arXiv:2306.01805.