CaLoRAify: Calorie Estimation with Visual-Text Pairing and LoRA-Driven Visual Language Models

Saman Khamesian

01/08/2025

INTRODUCTION

- Obesity affects **42%** of adults in the U.S. and is a major cause of **chronic diseases.**
- Over the years, a variety of tools and methods have been developed to aid in **calorie management**, ranging from mobile applications to AI-powered systems.
- Traditional methods for calorie estimation from food images have followed a **multi-step pipeline** involving food classification, portion size estimation, and caloric calculation.

INTRODUCTION

- While effective under **controlled conditions**, these methods face several **limitations**.
- First, the reliance on **specific metadata**, such as reference objects or depth images, makes them impractical for **general users**.
- Second, the **multi-module** nature of traditional pipelines introduces significant **error propagation**, as tasks like segmentation, classification, and volume estimation are handled separately.

INTRODUCTION

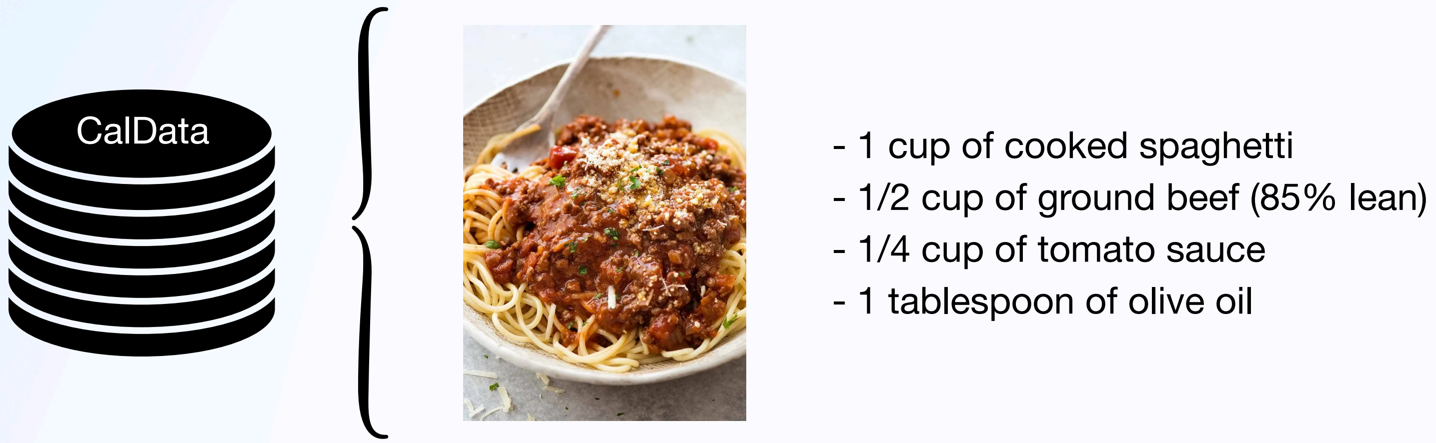
- Recent advancements in vision-language models (VLMs) like LLAVA and MiniGPT-4 have improved how AI processes images and text.
- These models can perform tasks such as answering questions about images or generating detailed descriptions.
- **CaLoRAify uses VLMs and RAG to estimate calories and recognize ingredients from a single food image.**

NOVELTY

1) **Single Image-Based Calorie Estimation:** It requires only a food image for inference, making it highly practical for real-world applications.

2) **Domain-Specific Dataset:** CalData is a comprehensive dataset consisting of 330K image-text pairs, designed specifically for tasks like ingredient recognition and calorie estimation.

Example of Image-Text pair data



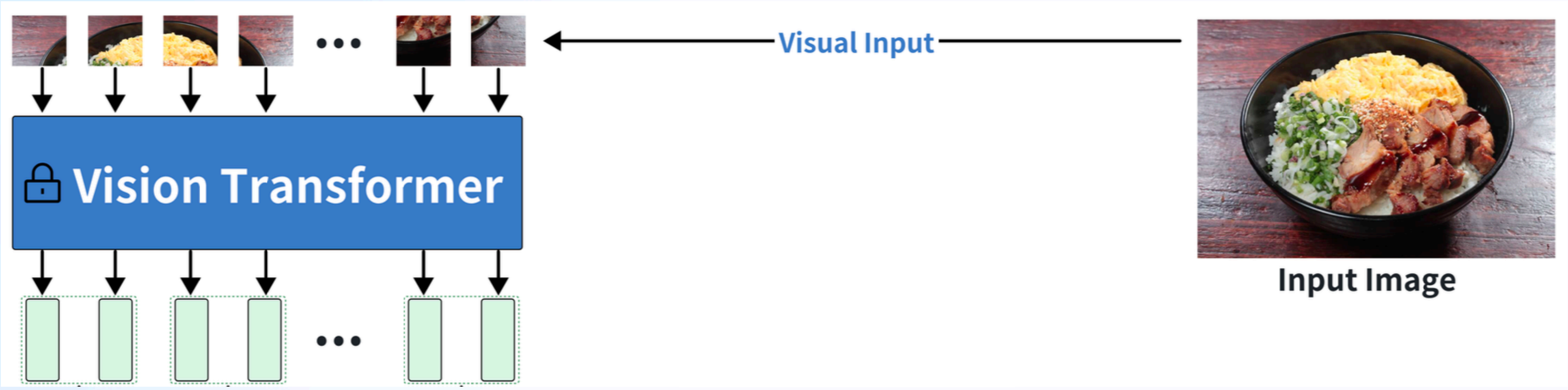
NOVELTY

3) **LoRA (Low-Rank Adaptation):** Adapts the VLM efficiently with the help of the CalData for domain-specific tasks, optimizing computational efficiency without the need for extensive retraining.

4) **RAG (Retrieval-Augmented Generation):** Enhances the model by retrieving external data (e.g., USDA nutritional database) to improve the accuracy of calorie estimations and reduce hallucinations during inference.

METHODOLOGY

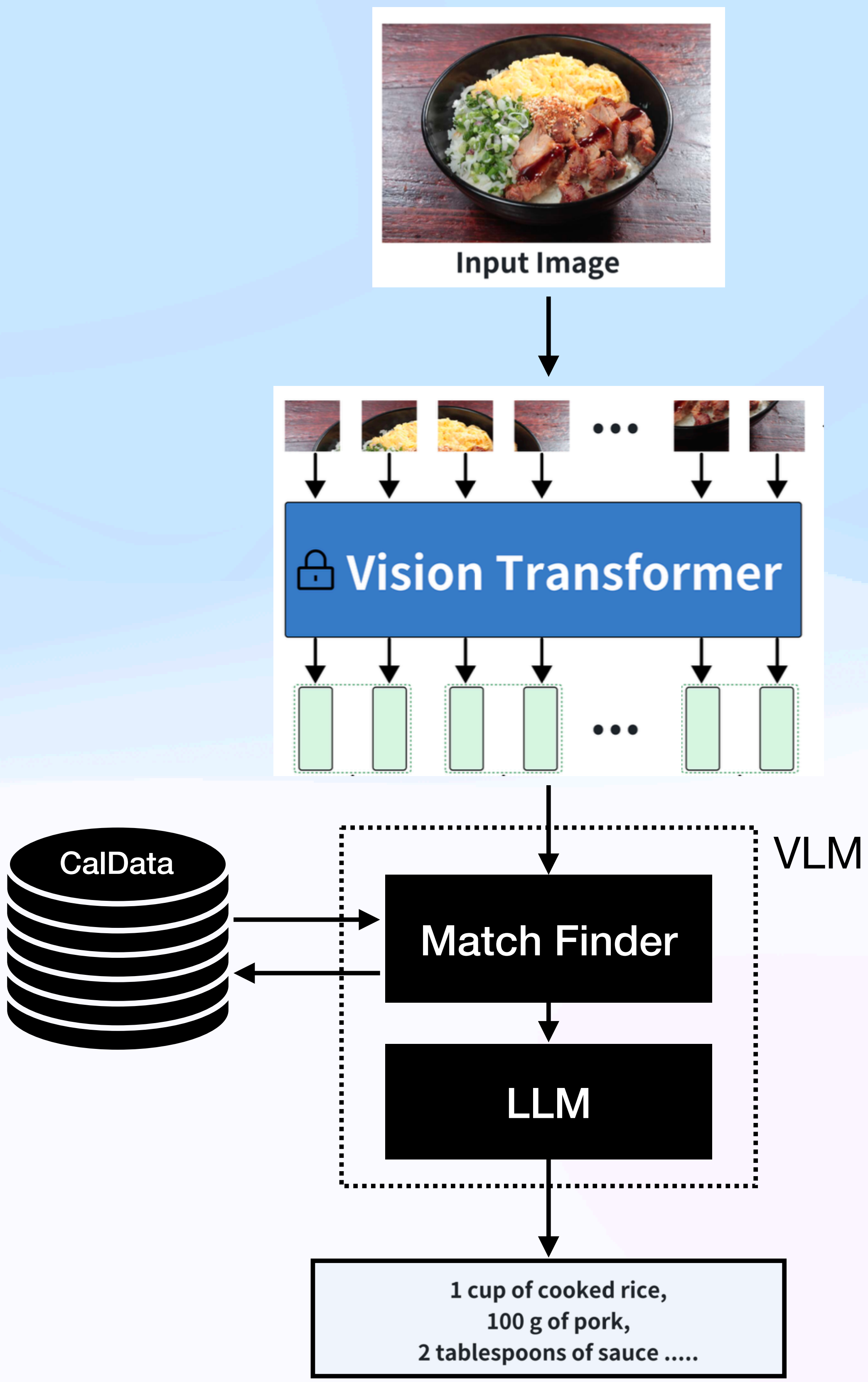
- The system uses a **single input image** to estimate calories.
- The input image is processed by a **Vision Transformer (ViT)**, which extracts **visual features** from the image and encodes them into a **vision vector** (a fixed-high-dimensional feature vector of that image).



METHODOLOGY

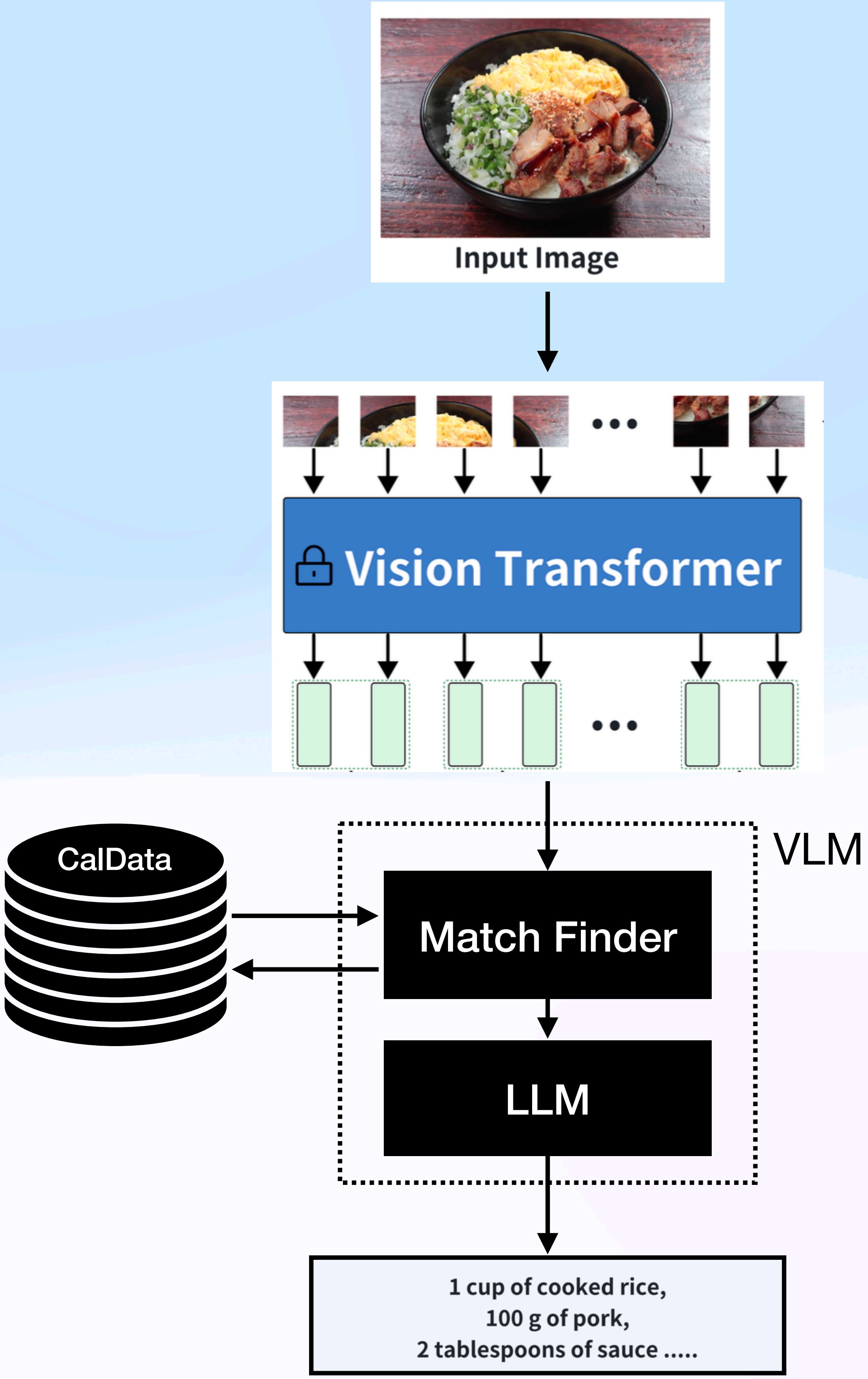
- The system **compares the user's input vision vector** with the pre-computed vectors in the CalData dataset to find the **most similar match.**

- Once a match is found, the system retrieves the **associated text** (e.g., ingredient lists, nutritional facts, or instructions) from the dataset for that matched image.



METHODOLOGY

- MiniGPT-4, also known as MiniGPT-v2, is a vision-language model (VLM) used for generating associated text from visual inputs.
- It combines a large language model (LLaMA-2) as its backbone with visual features extracted by a Vision Transformer (ViT), enabling it to process and align textual and visual data effectively.
- In this work, MiniGPT-4 is fine-tuned using the CalData dataset, with the **Low-Rank Adaptation (LoRA)** technique applied to efficiently adapt the model to the domain-specific task of calorie estimation and ingredient recognition.



METHODOLOGY

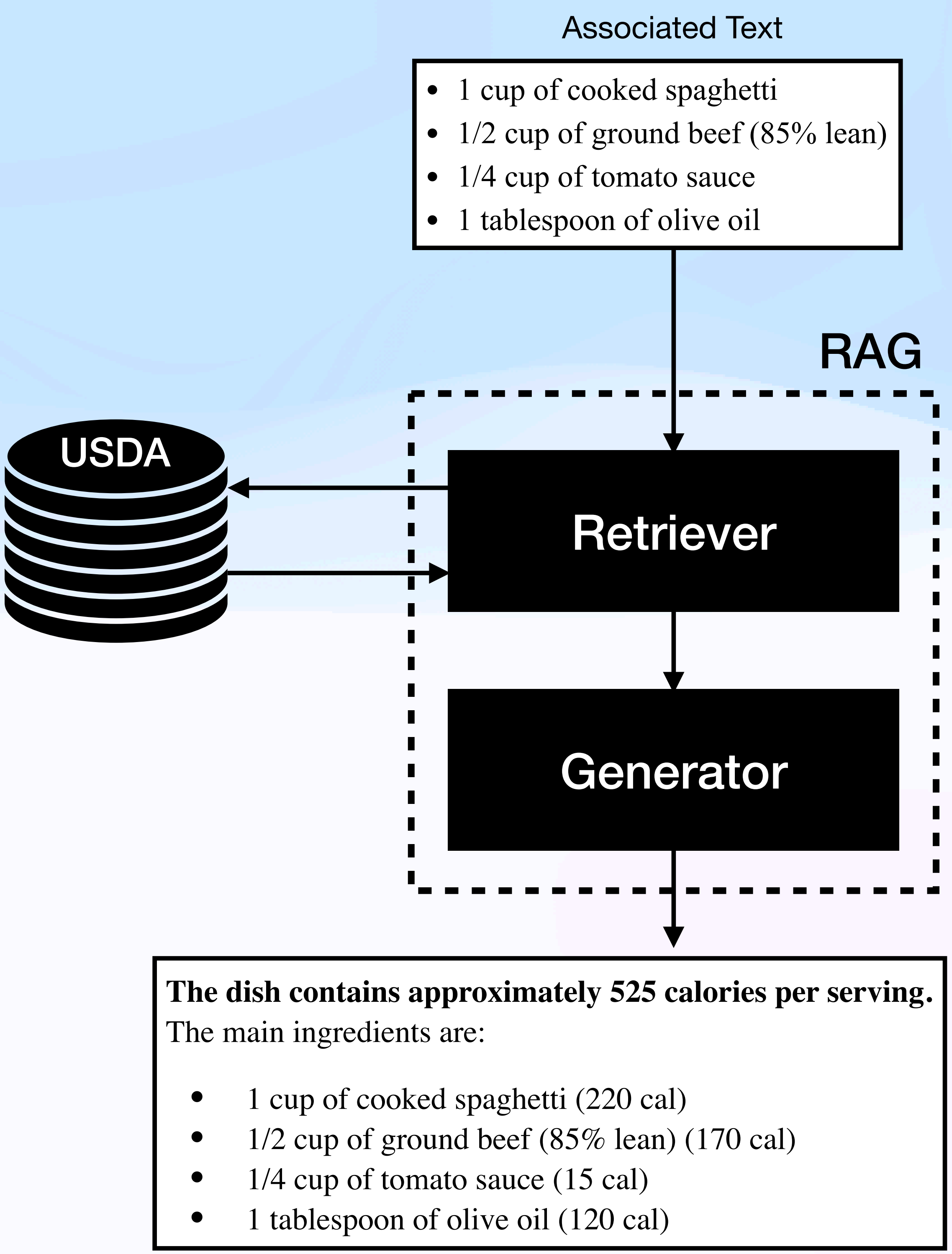
- The text is further **enhanced using RAG**, which:
  - Queries an external knowledge base (e.g., **USDA nutritional database**) for additional details about the ingredients.
  - Incorporates these details into the retrieved text to improve accuracy and completeness.
- *ADDITIONAL INFO: What is RAG and How it works?*

METHODOLOGY

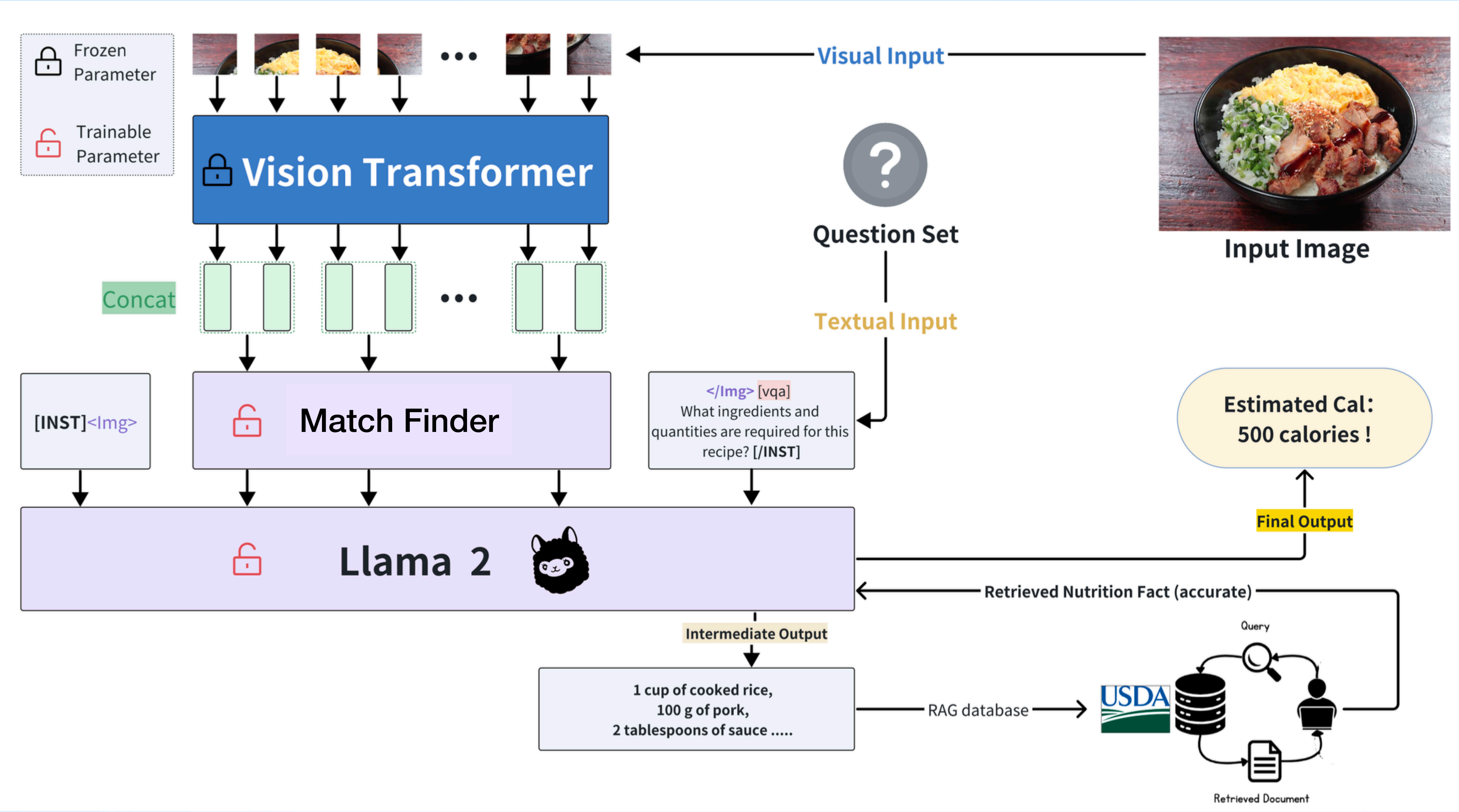
❖ RAG has two main components:

1) **Retriever:** Searches a large, external knowledge base (e.g., USDA nutritional database) to find relevant documents or information based on a given query.

2) **Generator:** A pre-trained language model (e.g., LLaMA-2 or BART) takes the retrieved documents and generates a response.



METHODOLOGY



DATASET

- For this study, they created an open-source comprehensive dataset (CalData) fitted to the task of food calorie estimation.
- The dataset was derived by combining multiple sources, including a large-scale receipt dataset (1M+ entries) and a nutrition instruction dataset containing detailed food amounts.
- Following a class-balanced sampling strategy [29], they identified 5801 unique samples.
- Each sample was associated with multiple images, resulting in an initial pool of 76,767 images. After augmentation, they

RESULTS

- Table 1 shows the metrics results of our experiment, where the baseline model is before fine-tuning, with the backbone of MiniGPT-4, and the fine-tuned is the model we trained based on the baseline.

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** is a set of metrics commonly used to measure how much overlap there is between the generated text and the reference text, focusing on shared words or sequences of words.

5.3. Metrics Results

Metric	Baseline	Fine-tune	Increase %
ROUGE-1	0.209	0.2173	3.97%
ROUGE-2	0.0611	0.0947	55.01%
ROUGE-L	0.1643	0.1734	5.53%
ROUGE-Lsum	0.1643	0.1733	5.48%
BLEU	0.0135	0.0218	61.48%
SacreBLEU	1.3518	2.1845	61.60%
BERTScore (P)	0.8441	0.846	0.23%
BERTScore (R)	0.8117	0.8135	0.22%
BERTScore (F1)	0.8273	0.8289	0.19%
Aggregate Metrics	0.431	0.4662	8.16%

Table 1. Performance comparison between Baseline and Fine-tune models.

RESULTS

5.4. Qualitative Results

In Figure 3 we show some examples of our model performing different food-related VQA tasks.

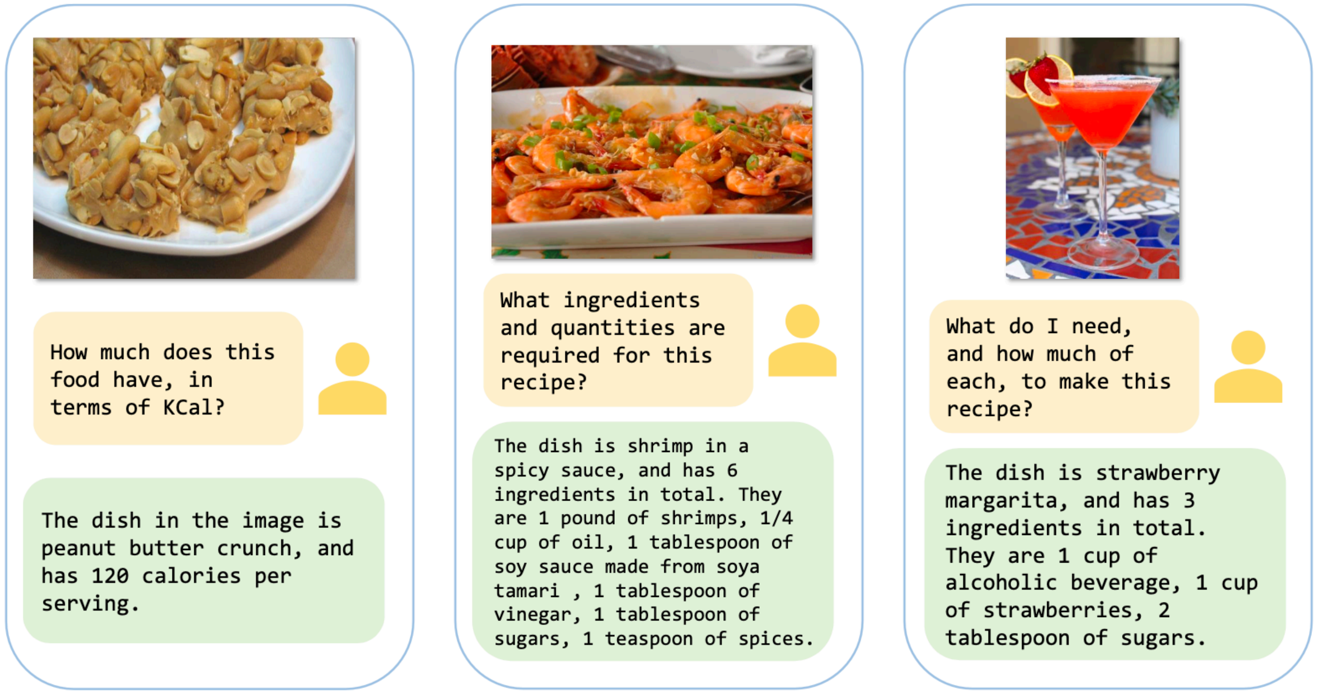


Figure 3. Qualitative results of the model output

Thank you for your attention