National College of Ireland

Beyond Text: How Images Compensate for Missing Quantitative Data in Multimodal Nutrition Estimation

MSc Research Project
MSc in Artificial Intelligence

Bintong Chen
Student ID: 23135808

School of Computing
National College of Ireland

Supervisor: Abdul Shahid

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Bintong Chen
Student ID:	23135808
Programme:	MSc in Artificial Intelligence
Year:	2025
Module:	MSc Research Project
Supervisor:	Abdul Shahid
Submission Due Date:	11/08/2025
Project Title:	Beyond Text: How Images Compensate for Missing Quantitative Data in Multimodal Nutritional Estimation
Word Count:	10116
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Bintong Chen
Date:	12th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Beyond Text: How Images Compensate for Missing Quantitative Data in Multimodal Nutritional Estimation

Bintong Chen
23135808

Abstract

Motivated by practical challenges in obtaining precise ingredient mass data for nutritional estimation, this study investigates how images can compensate for missing quantitative data within a multimodal framework. Three parallel pipelines, text-only, image-only, and multimodal, are implemented using domain-specific models: DistilBERT for text and ResNet18 for images. This study also explores various multimodal fusion architectures to identify the most effective strategy to combine data modalities. Key findings reveal that when precise mass data is available, text-only models achieve the highest accuracy. However, in its absence, multimodal models with gated fusion (per-feature weights) compensate for missing quantitative data effectively, reducing calorie estimation errors by 2.0% MAE. Furthermore, the results highlight fat estimation as the most challenging task and emphasise the importance of tailored fusion architectures in nutritional estimation. Overall, these insights demonstrate the potential of multimodal systems to enhance both the flexibility and accuracy of dietary monitoring.

1 Introduction

Having a balanced diet is a critical part of living a healthy lifestyle. Evidence has shown that many chronic diseases, such as type 2 diabetes and cardiovascular diseases, are closely linked to dietary patterns. In addition, choosing the right food and consuming the appropriate amount can contribute not only to optimal health and certain disease prevention but also to improved mental health (Basiri et al. 2023).

In reality, while the majority of the population only needs to follow general dietary guidelines to achieve beneficial health outcomes, certain demographics require stricter nutritional monitoring. For example, the elderly, people with specific health conditions, and athletes can benefit significantly from carefully monitoring their dietary intake. Tracking the intake of various nutrients can become part of their daily lives.

One common solution for strict nutritional monitoring is to follow a dietitian-approved meal plan specifically tailored to individual needs (Swan et al. 2017). However, this cannot always be followed easily long term, and it can create many inconveniences and challenges. When certain ingredients in the diet plan are not available, or when individuals find themselves in a situation that does not allow them to prepare their food themselves, consuming food outside of the diet plan is inevitable.

Another common approach is to document one's dietary intake in detail for nutritional monitoring (Burke et al. 2011). With this approach, people need to document the ingredients as well as the quantities. Although this approach allows people to have more flexibility in choosing food, and documenting the ingredient list is relatively straightforward, estimating each ingredient's mass without weighing it on a scale can be very challenging. In fact, even nutritionists can make large errors in their estimations when not using quantitative data (Hua et al. 2025). Therefore, finding a solution to compensate for missing quantitative information to improve the efficiency of dietary monitoring without sacrificing accuracy is highly meaningful. With such a solution, instead of weighing every ingredient and documenting precise weights, people only need to document the consumed food and the total mass, with no extra effort required to obtain the individual quantities of ingredients.

Previous studies have shown that using only textual or visual data for nutrition estimations faces different challenges. Image-only models perform well for predicting portion-independent nutritional values but struggle with portion estimation (Thames et al. 2021, Tanabe & Yanai 2025). Meanwhile, text-only approaches require precise quantitative data in order to achieve strong results (Hua et al. 2025). Recent works also suggest that the key focus for further progress in this domain is architectural and fusion strategies (Qi et al. 2023, Yin et al. 2024, Tanabe & Yanai 2025). However, many existing studies in this field have not yet systematically analysed the impact of different fusion techniques on integrating data modalities for nutritional estimation tasks (Thames et al. 2021, Tabassum & Nunavath 2024). To address this gap, this research focuses on leveraging food images in combination with ingredient lists and total dish mass, and systematically comparing and analysing various fusion architectures for integrating textual and visual data for more accurate nutritional estimation.

The aim of this research is to investigate to what extent multimodal fusion techniques can compensate for missing quantitative data in textual inputs using visual data, with a focus on identifying when and how visual information improves estimation accuracy when text lacks mass specifications. To achieve this, three parallel pipelines are implemented: text-only, image-only, and multimodal, using domain-specific models (DistilBERT for text and ResNet18 for images). Four multimodal fusion architectures—early fusion, late fusion, attention fusion, and gated fusion—are systematically evaluated across six model variants to assess performance with and without precise ingredient mass data.

The major contribution of this research is to systematically evaluate how different multimodal fusion techniques and architectures impact the accuracy of calories and macronutrients (carbohydrate, protein, and fat) estimation when combining visual and textual data, in scenarios where textual inputs lack precise mass information for individual ingredients. Furthermore, this research provides insights into nutrition-specific fusion rules, with results demonstrating that: (1) complex fusion architectures do not always outperform simpler ones, and (2) the utility of visual data depends critically on how text and image features are combined.

This paper discusses existing solutions leveraging textual and/or visual data for nutritional estimation in the related work in Section 2. Section 3 outlines the research and evaluation methodology used in this study. The techniques, architecture, and framework used in the implementation are discussed in Section 4. In Section 5, the implementation of the proposed solution is discussed in detail. Section 6 provides a comprehensive analysis of the results and key findings, as well as the implications of these findings from both an academic and a real-world perspective. Last but not least, Section 7 presents

the conclusion of this research as well as proposals for future work.

2 Related Work

This section examines four key aspects of nutritional estimation: (1) the strengths and limitations of image-only approaches, (2) challenges in text-only methods and LLMs, (3) current gaps in multimodal LLM architectures and fusion strategies, and (4) the effectiveness of DistilBERT and ResNet for domain-specific text and image processing. Through this structured analysis, this review identifies opportunities to enhance accuracy when quantitative text data is missing — a gap this study addresses through targeted multimodal fusion approaches.

2.1 Image-Based Nutrition Estimation:

The Nutrition5K dataset used in this research was created in 2021. Thames et al. (2021) gathered RGB images, depth data, ingredient lists, weights, and nutritional labels for over 5,000 dishes to study nutrition prediction from images of generic food. To handle the difficulty of portion estimation, the authors integrated RGB and depth data in a hybrid pipeline, employing two separate models with InceptionV2 and InceptionV3 (both pretrained on JFT-300M) as the backbone. The portion-independent model focuses on predicting calories and macronutrients per gram using only RGB data, while the mass regression model predicts the total mass of the dish using both RGB and volume estimates derived from depth data. The final nutritional prediction is obtained by multiplying the per-gram nutrition estimates by the mass prediction. The portion-independent model achieved a 9.5% MAE, and the mass regression model achieved 13.7% MAE. Combining both models gave the best result in this study, with an overall 16.5% MAE for calories (Thames et al. 2021).

Thames et al.'s study reveals that RGB images alone can enable recognition of different food types and their nutrition. However, the disparity between using images to predict calories per gram (9.5% MAE) and total calories (26.1% MAE) suggests that while image data alone is sufficient to recognize food types and their nutritional properties, the image-only approach struggles with predicting total portion sizes without additional cues. When volume information is available, as in this case where depth data bridges the gap, models can successfully estimate nutritional values based on the mass-independent nutritional estimation and the provided mass. The importance and difficulty of addressing accurate portion estimation is further discussed by Tanabe and Yanai, who demonstrate in their study that portion estimation remains a core bottleneck in nutrition estimation tasks, even when leveraging advanced Multimodal LLMs (MLLMs) (Tanabe & Yanai 2025).

In Thames et al.'s study, while the hybrid pipeline discussed above achieves the best results for calorie prediction, the authors also experimented with a native multimodal model. The 4-channel (RGB + depth as the 4th channel) multimodal model achieved an 18.8% MAE, which is significantly better than the RGB-only approach for predicting overall nutritional values (26.1% MAE) and very close to the best-performing hybrid pipeline (16.5% MAE). However, the authors only tested one multimodal approach using early fusion to concatenate raw depth data. This 4-channel multimodal model underperformed, likely due to the naive fusion method (Thames et al. 2021) and the unprocessed noise from the depth data. Other advanced fusion techniques were not explored to enable

cross-modal learning. This represents a key limitation of Thames et al.'s study. Furthermore, Tanabe and Yanai's study highlights that besides prompting strategies, advanced fusion techniques are critical for enabling MLLMs to effectively integrate visual and textual cues, especially when key quantitative details are missing from the text (Tanabe & Yanai 2025).

Finally, it is worth noting that although depth data significantly helps improve the results by assisting images to estimate mass values in Thames et al.'s study, from a real-world perspective, depth data collection is hardware-dependent and not easy to access. Obtaining such data in daily life settings is challenging outside of the lab. Therefore, adopting an alternative method to provide the overall volume would improve real-world applicability.

2.2 Text-Only and LLMs for Nutrition Estimation

The NutriBench study reveals that pure text models, even powerful Large Language Models (LLMs) such as GPT-4o, require precise quantitative data in the text to achieve their best performance on nutritional predictions, while the best result is still underwhelming with an accuracy of 66.8% (Hua et al. 2025).

Hua et al. systematically evaluated 12 Large Language Models (LLMs) on carbohydrate estimation from natural language meal descriptions, with GPT-4o using Chain-of-Thought (CoT) achieving the highest accuracy. One key finding is that text format has a significant impact on model performance, as all models struggled with natural serving descriptions (e.g., 'a cup of rice') but achieved lower error rates with precise metric descriptions (e.g., '80g rice') (Hua et al. 2025). This finding suggests that models require precise quantitative data to provide accurate nutritional estimation when using textual data alone, especially when complementary modalities are not present to assist portion estimation.

The underperformance of LLMs in Hua et al. (2025)'s study, even those with the best results, suggests LLMs' limitations in handling domain-specific text. Anisuzzaman et al. suggest that LLM performance for domain-specific tasks heavily relies on the quantity and quality of domain-specific data for fine-tuning purposes. However, the lack of such data is a common challenge in many fields (Anisuzzaman et al. 2025).

O'Hara et al. support this idea by showing that ChatGPT makes large errors when estimating the nutritional composition of real meals due to the lack of quantitative modelling in the nutrition domain (O'Hara et al. 2025). Therefore, the underperformance of LLMs in Hua et al. (2025)'s study is likely due to: (1) their training data being too broad to capture the specific language and concepts used in food and nutrition, and (2) their lack of training on precise portion-size relationships.

Although fine-tuning LLMs is a possible solution for domain-specific studies, as Anisuzzaman et al. suggest, even fine-tuning smaller LLMs requires large amounts of domain-specific data to achieve strong performance in a specific domain (Anisuzzaman et al. 2025). In addition, most LLMs are mainly designed for text-based tasks and therefore are not ideal for integration with other modalities (Yin et al. 2024, Qi et al. 2023). As a result, the nature of LLMs limits their ability to leverage other data types, such as visual data, in a multimodal setting.

2.3 Multimodal LLM Limitations and Architectural Challenges

Although advanced multimodal LLMs (MLLMs) are designed to handle multiple data types, integrating and reasoning these modalities to allow the model to truly understand the connection between them remains a technical challenge (Yin et al. 2024). Currently, MLLM performance is inconsistent across tasks and domains (Małkiński et al. 2025). Even the progress made in the multimodal LLMs domain has mainly come from careful architectural choices with specialised fusion strategies, such as early, late, attention-based, and gated fusions, rather than LLMs' native capabilities (Yin et al. 2024, Qi et al. 2023). This distinction is particularly important in nutrition estimation with both textual and visual data, as successful predictions heavily depend on the alignment of fine-grained text descriptions and visual information.

Yin et al.'s study emphasises that the key in the MLLMs domain is the architectural strategies and the chosen fusion techniques. Tanabe & Yanai's results further support this point. They introduced a food volume estimation module into the MLLM pipeline in their study, and the results suggest that although an explicit module for volume estimation helps improve accuracy to a certain degree, there are still significant challenges for achieving strong performance (Tanabe & Yanai 2025). Fine-tuning helps improve MLLM performance, but advanced fusion strategies are more critical. Furthermore, Yin et al. directly demonstrate in their study that general MLLM are not sufficient for food reasoning tasks, and careful architecture, prompt engineering, and multi-stage training are required for performance improvement (Yin et al. 2024). The findings from these three studies reveal that improvements in nutrition estimation should focus on architectural and cross-modal fusion strategies, rather than the capability of MLLMs themselves.

2.4 Practical Multimodal Solutions: DistilBERT and ResNet for Domain-Specific Text and Image Processing

Multiple studies confirm that lightweight transformer models, such as DistilBERT, have unique advantages in tasks utilising textual data (Sanh et al. 2019, Rohanian et al. 2024, Amandeep & Suresh 2025).

Rohanian et al. state that lightweight transformers are efficiently fine-tuned for clinical language in their study, and they require less training data and computational resources compared to larger models (Rohanian et al. 2024). Their findings highlight the strong adaptation of domain-specific lightweight models to specialised data. In addition, the authors specifically highlight the efficiency and deployability of DistilBERT, which is a lighter alternative to BERT but with 97% of BERT's effectiveness. Sanh et al., the designers of DistilBERT, suggest that DistilBERT is able to achieve high performance even with very limited fine-tuning data (Sanh et al. 2019). Furthermore, Amandeep and Suresh highlight DistilBERT's strong performance and lower computational cost in their study on fake news detection (Amandeep & Suresh 2025). Khan et al. also report in their study on AI-generated content that DistilBERT is a practical choice for real-world applications due to its strong domain-specific task performance and low requirements of computational resources (Khan et al. 2025).

Tabassum & Nunavath (2024) systematically investigate how late fusion of text and image models can be used to detect cyberbullying in social media content. Their experiments compare unimodal approaches (text-only and image-only) with multimodal approaches (text + image), with six text models and three image models. The results

confirm that DistilBERT performed strongly for text-only classification (accuracy: 0.991), while ResNet-50 achieved a high accuracy (0.98) for image-only input. In addition, the hybrid model combining DistilBERT and ResNet-50 achieved a strong multimodal performance with 0.9655 accuracy. Although the best overall performance was obtained by a hybrid of RoBERTa (text) and Vision Transformer (image), the DistilBERT+ResNet approach was still highly accurate. It is also worth noting that the computational efficiency of DistilBERT and ResNet is highlighted in the study. Given the lighter architectures of DistilBERT and ResNet relative to RoBERTa and Vision Transformer, their combination seems to provide a better balance between accuracy and resource requirements.

These findings together demonstrate that DistilBERT and ResNet are strong choices for handling textual data and visual data respectively, and the DistilBERT + ResNet multimodal approach is valid as a practical and effective multimodal solution to leverage both textual and visual data in domain-specific tasks (Tabassum & Nunavath 2024). However, one key limitation of Tabassum and Nunavath's study is that only a late fusion technique was investigated for integrating textual and visual features, and no alternative fusion strategies were explored.

In summary, although previous studies have shown that both textual and visual data alone can provide reasonably good nutrition estimation, each approach struggles with different difficulties. Image-only models perform well for portion-independent nutritional value estimation but struggle with accurate portion estimation unless additional cues, such as depth data, are available (Thames et al. 2021). However, it is also worth noting the accessibility of data to assist images, for the consideration of real-world applicability. LLMs require detailed quantitative information, yet still underperform on domain-specific nutrition tasks. Even with recent MLLMs, achieving strong cross-modal reasoning to capture the connection across modalities remains a challenge. Furthermore, many studies in this field have fallen short in comparing and analyzing different fusion techniques for integrating different data modalities. These gaps highlight the need for lightweight and adaptable multimodal solutions with careful fusion design. To address these issues, the present study investigates whether combining lightweight models, such as DistilBERT and ResNet, with carefully designed fusion strategies can enable more accurate and practical nutrition estimation.

3 Methodology

Figure 1 illustrates the research methodology of this study. This research follows a supervised regression approach to estimate the nutritional values of dishes using text descriptions and RGB images. The process begins with data preparation. Next, three parallel pipelines are presented: a text-based pipeline, an image-based pipeline, and a multimodal pipeline. Each pipeline follows five steps: (1) data transformation, (2) train/test split, (3) model architecture design, (4) model training/fine-tuning, and (5) model evaluation.

The text-based pipeline tests four text variants of ingredient descriptions, one of which includes detailed quantitative data that previous work has shown to produce relatively more accurate results (Hua et al. 2025), as well as ingredients listed in descending order. This variant serves as a strong baseline for comparison. Variants that differ in ingredient order (original versus descending by mass) are also examined to explore whether reordering can improve performance when only text data is available, without quantitative

Table 1: Summary of Relevant Studies

Study	Relevant Key Findings	Limitation/Implication
(Thames et al. 2021)	Image-based approaches struggle with portion estimation. RGB + depth improves results.	Only simple early fusion used. Depth data collection is not practical in real-world settings. Supports using RGB-only models and exploring alternative fusion methods.
(Tanabe & Yanai 2025)	Portion estimation is a major challenge for MLLMs. Advanced fusion is essential.	Highlights the need to investigate different fusion and architecture strategies in a multimodal setting.
(Hua et al. 2025)	LLMs underperform on nutrition estimation. Precise quantity data is required.	Supports the use of multiple text variants and highlights the importance of testing performance with and without mass data.
(O'Hara et al. 2025)	LLMs make significant errors due to lack of nutrition-specific modeling.	Demonstrates LLMs alone are not reliable for accurate nutrition prediction on real meals. Justifies model selection over generic LLMs.
(Anisuzzaman et al. 2025)	LLM performance depends on fine tuning with large domain specific datasets.	Demonstrates LLMs are not ideal for small datasets like the one in this study. Supports the choice of lightweight models like DistilBERT.
(Yin et al. 2024) (Qi et al. 2023)	MLLMs require careful architecture and fusion strategies. General models don't perform well across domains.	Demonstrates MLLMs are not suited for this study. Supports the need to design and compare fusion strategies directly.
(Małkiński et al. 2025)	MLLM performance is inconsistent across domains.	Suggests avoiding general-purpose MLLMs. Supports focusing on custom, efficient multimodal fusion models.
(Sanh et al. 2019);(Rohanian et al. 2024); (Amandeep & Suresh 2025); (Khan et al. 2025)	DistilBERT performs well on domain-specific tasks with small datasets and low compute needs.	Justifies using DistilBERT for text processing and model fine-tuning with limited data.
(Tabassum & Nunavath 2024)	DistilBERT and ResNet show strong unimodal performance; their fusion performs well too.	Only late fusion used. Supports using DistilBERT for text and ResNet for images. Justifies investigating multiple fusion strategies beyond late fusion.

details or supporting modalities to assist.

The image-based pipeline establishes a baseline solely using visual data. Given the relatively small dataset size of 3,490 samples, fine-tuning a pretrained CNN model is expected to result in moderate performance. However, this baseline helps identify which nutrients are most challenging to predict from images alone. In addition, this pipeline further investigates how different training subset sizes affect the fine-tuning performance of ResNet18, aiming to determine the minimal data needed for optimal results on the current dataset. These findings provide valuable insights for multimodal fusion design at later stages and cross-pipeline evaluation.

The multimodal pipeline investigates how combining textual and visual modalities impacts nutritional estimation. Six models are implemented to evaluate the effects of four fusion techniques and various architectural strategies in the nutrition-specific domain. The experiments focus on three main objectives. First, the performance of the different fusion models is compared to identify the most effective fusion and architectural design for integrating visual and textual data. Second, the performance of text-only, image-only, and multimodal models is evaluated to understand the contribution of each modality. Third, we investigate how visual information can compensate for missing mass data in text, analyzing how different fusion techniques leverage images to improve ingredient mass estimation when textual ingredient quantitative information is unavailable.

3.1 Data Preparation

Raw dish metadata and RGB images used in this research are retrieved from the Nutrition5k study (Thames et al. 2021). The data preparation step consists of several stages,

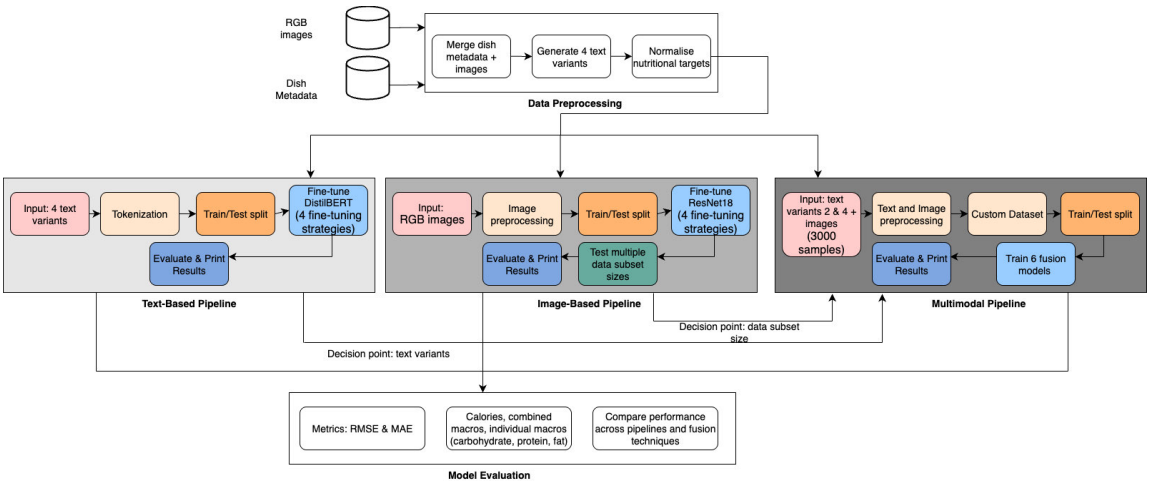

Figure 1: Research methodology framework showing the three parallel pipelines (text-based, image-based, and multimodal) with their respective processing stages.

including data collection, data merging, and data preprocessing.

First, relevant dish metadata and structured ingredient data are retrieved, combined, and nested into a single dataframe based on dish ID. Then, four text description variants are generated, differing in ingredient order (original order vs. descending order by mass) and the inclusion or exclusion of per-ingredient mass.

The decision to create these four text variants is based on key findings in related work. Previous studies have shown that the text format can significantly influence model performance, and models trained on textual data alone require the presence of quantitative ingredient information to achieve strong results (Hua et al. 2025). By varying the inclusion of mass data and ingredient order, this study evaluates how these two factors individually and jointly affect prediction accuracy. In addition, these different formats help determine whether specific ingredient orderings contribute to reducing the impact of missing quantitative data. Finally, the results from these text-based models can serve as a baseline for comparison with multimodal models, and provide insight into whether and to what extent adding visual cues can improve predictions when ingredient quantitative data is not available.

Secondly, the paths to corresponding RGB images for each dish are identified and catalogued. A new dataframe containing dish IDs and image paths is created. Then, the metadata dataframe with text input variants and the image path dataframe are merged and filtered. Only dishes with both valid metadata and associated images are retained for further processing. This approach ensures that both the text-based and image-based pipelines experiment on an identical set of dishes, allowing for a fair and direct comparison of their performance. After merging and filtering, the dataset includes 3,490 dishes with both image and text data.

Finally, four target variables (total calories, fat, carbohydrate, and protein values) are normalised using standard scaling. Both the processed dataframe and the scaler are saved for reproducibility and usage in later modelling steps.

3.2 Text-Based Pipeline

The text-based pipeline uses only structured text descriptions to predict nutritional values of dishes. The process begins by loading the filtered dataset generated during the data preparation step.

First, the data transformation step tokenises the selected text variant using the DistilBERT tokenizer. After tokenisation, the data is split into training and test sets. Both the input tensor and the standardised labels are partitioned accordingly.

Then, in the model architecture design step, a regression model is customised based on a pretrained DistilBERT architecture, with an added linear layer for nutrient prediction. DistilBERT is selected because of its strength in leveraging limited data to achieve strong performance on domain-specific tasks, as highlighted in multiple studies (Sanh et al. 2019, Rohanian et al. 2024, Amandeep & Suresh 2025, Khan et al. 2025).

Four fine-tuning strategies are explored to identify the most effective fine-tuning settings for adapting DistilBERT to this task. Finally, the last step in the text-based pipeline is evaluation, which will be further discussed in Section 3.4 Model Evaluation.

3.3 Image-Based Pipeline

The image-based pipeline uses only RGB images of dishes for nutritional prediction. The process begins by loading the filtered list of image paths, each paired with its corresponding dish ID.

First, in the data transformation step, each image is preprocessed according to a predefined pipeline. A custom PyTorch dataset is then created to link each image with its nutritional labels. After data preprocessing, the dataset is split into training and test sets. PyTorch DataLoaders are constructed for batching and shuffling during training and evaluation.

Then, for the model architecture design step, a pretrained ResNet18 convolutional neural network (CNN) is employed. The final fully connected layer is replaced to enable regression over four nutritional outputs. ResNet is selected based on the strong performance of ResNet-50 alone and the DistilBERT+ResNet approach reported in Tabassum and Nunavath's study (Tabassum & Nunavath 2024). ResNet18 is chosen over other ResNet variants due to the computational resources available for this research.

During the model fine-tuning step, as in the text-based pipeline, four fine-tuning strategies are explored. In addition, various dataset subset sizes (300, 600, 1000, 2000, 3000, and 3,490) are tested for each training strategy to assess the impact of dataset size on model performance with different fine-tuning settings. This approach aims to identify the most effective fine-tuning settings and data size for adapting ResNet18 to this task.

Finally, in the model evaluation step, results for all training strategies and subset sizes are systematically recorded.

3.4 Multimodal Pipeline

The multimodal pipeline leverages both image and text for nutritional estimation. The best-performing fine-tuning strategies for DistilBERT and ResNet are selected based on the results from text-based and image-based pipelines, along with the optimal data subset size. Two text variants are selected for comparison in the multimodal setting.

After data transformation, the custom dataset is split into training and test sets, and can be further sampled to leverage the optimal data size determined in the image-based

pipeline.

For the model architecture design step, several fusion strategies are explored to integrate visual and textual data: (1) Early Fusion, (2) Late Fusion, (3) Attention Fusion, and (4) Gated Fusion. Attention Fusion and Gated Fusion each have two different architecture designs, resulting in six fusion models in total. This approach addresses the gap in previous studies, which often lack a comprehensive comparison of different fusion techniques when integrating data modalities (Thames et al. 2021, Tabassum & Nunavath 2024). These strategies are designed to systematically explore different ways of combining image and text features to improve prediction accuracy.

After the model training phase, results for all fusion models are systematically recorded.

3.5 Model Evaluation

Thames et al. (2021)'s study serves as a strong baseline for the evaluation approach in this study for two reasons. To start, both Thames et al. (2021)'s study and this research leverage the Nutrition5K dataset. Secondly, both studies focus on nutritional estimation from visual data with one additional data modality. In Thames et al. (2021)'s study, Mean Absolute Error (MAE) and Percentage Mean Absolute Error (PMAE) are employed to measure the accuracy of calorie and macronutrient predictions (Thames et al. 2021, Tabassum & Nunavath 2024).

In my study, all three pipelines use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) as metrics to evaluate the results, reported in both standardised and original units. As in Thames et al. (2021)'s study, MAE is chosen as it represents the average size of errors. In addition, RMSE is selected because comparing MAE and RMSE helps reveal whether errors are generally consistent or are dominated by a few large outliers (when RMSE is significantly higher than MAE). Percentage Mean Absolute Error (PMAE) is not used in this study, but is replaced with the method of reporting MAE as a percentage relative to the average true value across samples. This approach allows one to summarize the overall error magnitude relative to the dataset's average value, avoiding problems that PMAE can face when actual values are very small or zero, which could distort percentage calculations. This approach provides a better balance between simplicity and interpretability.

Metrics are calculated for both calories and each individual macronutrient (fat, carbohydrate, protein), the same as in Thames et al. (2021)'s study. In addition, a combined macronutrient metric is reported for a more straightforward comparison of macronutrient estimates across models. As previously discussed, MAE values in original units are also presented as a percentage relative to the average value across samples, replacing PMAE, for better interpretability.

The final output includes:

(1) Calories and Combined Macros: RMSE and MAE (Standardised)

(2) Calories and Combined Macros: RMSE and MAE (Original Units) with percentage values

(3) Individual Macros (Fat, Carbs, Protein): MAE (Original Units) with percentage values

The reason for reporting RMSE and MAE across multiple prediction targets—total calories, combined macronutrients, and individual macronutrients—is that, beyond evaluating model accuracy from an academic perspective, this approach provides practical

insights into the specific challenges of predicting certain macronutrients using machine learning. These findings can offer valuable insights into applying these models in real-world scenarios, helping specific demographic groups with specific nutritional concerns.

All pipelines use the same evaluation metrics and report results in a consistent format throughout this research. This consistency in evaluation across pipelines ensures fairness and direct comparability for later comparison and analysis.

4 Design Specification

This research adopts two deep learning models, DistilBERT for textual data and ResNet18 for image data, and explores four fusion techniques to integrate these two data modalities.

4.1 Data Preparation

4.1.1 Text-Based Pipeline

Four text input variants are generated for each dish during data preprocessing. All quantity data is rounded to two decimal places to reduce noise before the text generation. This process results in four text variants, which are presented in Table 2.

After text generation, texts are tokenised using the DistilBERT tokenizer, which converts each text description into input IDs and attention masks for the transformer model. Sequences are padded to a maximum length of 128 tokens.

Table 2: Text Variants and Descriptions

Text Variant	**Description**	**Example**
Variant 1	A dish description with ingredients in their original order, without per-ingredient mass	"Dish mass: 193.0g. Ingredients: soy sauce, garlic, white rice, parsley, onions, brown rice, vinegar, apple, mixed greens, sugar, salt, lemon juice, olive oil, pork, bok choy, pepper, millet."
Variant 2	A description with ingredients sorted by mass (descending), without per-ingredient mass	"Dish mass: 193.0g. Ingredients: brown rice, pork, mixed greens, white rice, bok choy, sugar, apple, millet, soy sauce, olive oil, garlic, onions, vinegar, lemon juice, salt, parsley, pepper."
Variant 3	A description with ingredients in their original order, including per-ingredient mass	"Dish mass: 193.0g. Ingredients: soy sauce 3.40g, garlic 2.12g, white rice 8.50g, parsley 0.21g, onions 1.71g, brown rice 68.00g, vinegar 0.85g, apple 4.27g, mixed greens 21.34g, sugar 6.37g, salt 0.53g, lemon juice 0.85g, olive oil 3.24g, pork 59.47g, bok choy 8.50g, pepper 0.21g, millet 3.41g."
Variant 4	A description with ingredients sorted by mass (descending), including per-ingredient mass.	"Dish mass: 193.0g. Ingredients: brown rice 68.00g, pork 59.47g, mixed greens 21.34g, white rice 8.50g, bok choy 8.50g, sugar 6.37g, apple 4.27g, millet 3.41g, soy sauce 3.40g, olive oil 3.24g, garlic 2.12g, onions 1.71g, vinegar 0.85g, lemon juice 0.85g, salt 0.53g, parsley 0.21g, pepper 0.21g."

4.1.2 Image-Based Pipeline

For the image-based pipeline, each input image is first resized to 224×224 pixels, converted into a tensor, and normalised using the ImageNet dataset means and standard deviation values to match the expected input distribution of the ResNet18 model. This processing is encapsulated in a torchvision transform pipeline, which is applied to all images.

After image data processing, a custom PyTorch dataset is used to manage image file loading using their paths along with nutritional labels matched by dish ID from metadata stored in a CSV. This design enables the retrieval of image-label pairs at runtime.

4.1.3 Multimodal Pipeline

For the multimodal pipeline, the process begins by merging the filtered text and image datasets based on dish IDs, which ensures each entry contains both an RGB image and selected corresponding text variants. A custom dataset class is employed, within which each sample consists of an image tensor, the tokenised text inputs, and the label tensor.

4.2 Model Architecture

4.2.1 DistilBERT Model

For text-based predictions, a pretrained DistilBERT model is adopted as the core text encoder. The model is extended with a linear regression layer to map the [CLS] token to four quantitative nutritional outputs. Table 3 shows the detailed architecture of the model. Fine-tuning strategies for this model are detailed in Section 4.3 Fine-Tuning Strategies.

Table 3: DstilBERT Model Architecture

Layer / Component	**Quantity / Details**
Text Encoder	Pretrained DistilBERT
Transformer encoder layers	6
Hidden size	768
Attention heads	12
Regression head	1 linear layer (output dimension 4)
Dropout	Present (default DistilBERT dropout)

4.2.2 ResNet18 Model

For image-based predictions, a pretrained ResNet18 CNN is adopted as the model backbone. The original final classification layer is replaced with a fully connected regression head for four nutritional value outputs. This design allows the model to leverage pretrained feature representations and adapt the output layer for the regression task. Table 4 shows the detailed architecture of the model. Fine-tuning strategies for this model are detailed in Section 4.3 Fine-Tuning Strategies.

Table 4: ResNet18 Model Architecture

Layer / Component	**Quantity / Details**
Image Encoder	Pretrained ResNet18
Residual Layers	4 (named layer1 to layer4)
Number of Basic Blocks	2 blocks per residual layer
Hidden Channels	64, 128, 256, 512 (varies across layers)
Fully Connected Regression Head	1 linear layer (output dimension 4)
Pretrained Weights	ImageNet weights

This study implements four multimodal fusion strategies to combine information from both text and images: early fusion, late fusion, gated fusion, and attention fusion. In addition, six fusion models are implemented to explore different architectural strategies. All experiments are conducted using a consistent setup: a dataset of 3,000 samples, with

each sample containing a text description and a corresponding food image, and a fixed training duration of five epochs. This consistent setup ensures fair comparisons across models and pipelines.

4.2.3 Early Fusion Model

The early fusion model combines data modalities at the feature level. It begins with extracting embeddings separately from the ResNet18 image encoder and the DistilBERT text encoder. Then, these two feature vectors are concatenated into a single vector, which is passed through a multi-layer perceptron (MLP) to predict four nutritional values. This approach allows the model to learn joint relationships between textual and visual features from the start. Table 5 shows the detailed architecture of the model.

Table 5: Early Fusion Model Architecture

Layer / Component	**Quantity / Details**
Image Encoder	Pretrained ResNet18 (output feature size 512), final layer removed (identity)
Text Encoder	Pretrained DistilBERT (hidden size 768)
Feature Fusion	Concatenation of image (512) + text (768) features
Fusion MLP	Linear(1280 $\rightarrow$ 512) $\rightarrow$ ReLU $\rightarrow$ Dropout(0.3) $\rightarrow$ Linear(512 $\rightarrow$ 128) $\rightarrow$ ReLU $\rightarrow$ Linear(128 $\rightarrow$ 4)

4.2.4 Late Fusion Model

In the late fusion model, image and text inputs are processed independently through separate feature extractors and MLPs. After both modalities produce their own embeddings, the two embeddings are concatenated. This concatenated embedding is then passed to and further processed by a final fusion network to predict four nutritional values. This method processes image and text features separately in the early stages and combines them at a later step. Table 6 shows the detailed architecture of the model.

Table 6: Late Fusion Model Architecture

Layer / Component	**Quantity / Details**
Image Encoder	Pretrained ResNet18 (output feature size 512)
Removed classification head	Replaced with identity layer
Image MLP	Linear (512 $\rightarrow$ 256) $\rightarrow$ ReLU $\rightarrow$ Dropout (0.3)
Text Encoder	Pretrained DistilBERT (hidden size 768)
Text MLP	Linear (768 $\rightarrow$ 256) $\rightarrow$ ReLU $\rightarrow$ Dropout (0.3)
Fusion MLP	Linear (512 $\rightarrow$ 128) $\rightarrow$ ReLU $\rightarrow$ Dropout (0.3) $\rightarrow$ Linear (128 $\rightarrow$ 4)

4.2.5 Attention Fusion Model with Global Weights

This model uses an attention mechanism to dynamically weight the importance of each modality. This is achieved using a small two-layer MLP that learns attention scores to determine the contribution of each modality before combining the image and text embeddings. One overall weight is decided, then these two weighted embeddings are combined to form a fused vector. This approach allows the model to focus more on the more informative modality over the other one when making a prediction. Table 7 shows the detailed architecture of the model.

Table 7: Attention Fusion Model with Global Weights Architecture

Layer / Component	**Quantity / Details**
Image Encoder	Pretrained ResNet18 (output feature size 512), final layer removed (identity)
Image Projection	Linear (512 → 256)
Text Encoder	Pretrained DistilBERT (hidden size 768)
Text Projection	Linear (768 → 256)
Attention Layer	2-layer MLP: Linear (512 → 128) → ReLU → Linear (128 → 2) → Softmax (weights for image & text)
Fusion	Weighted sum of projected image and text embeddings
Output Layer	Linear (256 → 128) → ReLU → Dropout (0.3) → Linear (128 → 4)

4.2.6 Attention Fusion Model with Cross-Attention

This model enables the interaction between image and text features by allowing them to attend to each other. After extracting features separately from the ResNet18 image encoder and DistilBERT text encoder, these features are projected into a shared space and then normalised. This model adopts a cross-attention mechanism, so text features can focus on relevant parts of the image features and vice versa. The model then uses learned weights to combine these attended features into a single fused representation, which is used to predict four nutritional values. Table 8 shows the detailed architecture of the model.

Table 8: Attention Fusion Model with Cross-Attention Architecture

Layer / Component	**Quantity / Details**
Image Encoder	Pretrained ResNet18 (output feature size 512), final layer removed (identity)
Image Projection	Sequential: Linear (512 → 256) → LayerNorm (256)
Text Encoder	Pretrained DistilBERT (hidden size 768)
Text Projection	Sequential: Linear (768 → 256) → LayerNorm (256)
Cross-Attention	MultiheadAttention with embed_dim=256, num_heads=4, batch_first=True
Modality Weighting	2-layer MLP: Linear (512 → 128) → ReLU → Linear (128 → 2) → Softmax (weights for image & attended text)
Fusion	Weighted sum of image embedding and attention output
Output Layer	Linear (256 → 128) → ReLU → Dropout (0.3) → Linear (128 → 4)

4.2.7 Gated Fusion Model with Per-Feature Weights

The gated fusion model with per-feature weights combines image and text features by learning separate importance weights for each element in the feature vectors. After extracting features from ResNet18 and DistilBERT and projecting them into a shared space, a gating network predicts a weight for every feature dimension independently, with values between zero and one. These weights decide how much each feature from the image and text contributes to the final combined representation. Then, the model multiplies each feature by its gating weight and sums them to produce a fused embedding. This fine-grained gating allows the model to emphasise certain important features from both modalities. Table 9 shows the detailed architecture of the model.

Table 9: Gated Fusion Model with Per-Feature Weights Template

Layer / Component	**Quantity / Details**
Image Encoder	Pretrained ResNet18 (output feature size 512), final layer removed (identity)
Image Projection	Linear (512 → 256)
Text Encoder	Pretrained DistilBERT (hidden size 768)
Text Projection	Linear (768 → 256)
Gating Mechanism	1-layer gate: Linear (512 → 256) → Sigmoid (per-feature gating weights)
Fusion	Per-feature weighted sum: gate weights * image + (1 - gate weights) * text
Output Layer	Linear (256 → 128) → ReLU → Dropout (0.3) → Linear (128 → 4)

4.2.8 Gated Fusion Model with Scalar Weight

The gated fusion model with scalar weight combines image and text features by learning a single weight value that balances their overall contributions. After extracting features separately from the ResNet18 image encoder and the DistilBERT text encoder, the model projects them into a common space with normalisation. Then the model uses a gating network to produce a scalar weight between zero and one. This weight determines how much emphasis to give the image features and the text features in the final fused representation. Then, the fused embedding is computed by multiplying the image features by this weight and the text features by its complement, and then summing the two. This approach requires fewer parameters than per-feature gating and can be more stable when the dataset is smaller. Table 10 shows the detailed architecture of the model.

Table 10: Gated Fusion Model with Scalar Weight Architecture

Layer / Component	**Quantity / Details**
Image Encoder	Pretrained ResNet18 (output feature size 512), final layer removed (identity)
Image Projection	Sequential: Linear (512 → 256) → LayerNorm (256)
Text Encoder	Pretrained DistilBERT (hidden size 768)
Text Projection	Sequential: Linear (768 → 256) → LayerNorm (256)
Gating Mechanism	1-layer gate: Linear (512 → 1) → Sigmoid (scalar gating weight)
Fusion	Scalar-weighted sum: $gate\_weight \times image + (1 - gate\_weight) \times text$
Output Layer	Linear (256 → 128) → ReLU → Dropout (0.3) → Linear (128 → 4)

4.3 Fine-Tuning Strategies

To evaluate how the degree of adaptation impacts model performance, four fine-tuning strategies are explored in the text-based and image-based pipelines.

(1) full fine-tuning (all model layers trainable),
(2) fixed feature extraction (all layers frozen except the regression head),
(3) partial fine-tuning with only the last one transformer layer unfrozen,
(4) partial fine-tuning with only the last two transformer layers unfrozen.

Full fine-tuning achieves the best performance in the text-based and image-based pipelines, therefore full fine-tuning is applied to both DistilBERT and ResNet in all multimodal models. Further evaluation and analysis of these fine-tuning strategies are provided in Section 6 Evaluation.

5 Implementation

5.1 Hardware and Computational Resources

All models were trained, fine-tuned, and tested on a 2022 Apple MacBook Air, which is equipped with an Apple M2 chip. The M2 integrates an 8-core CPU (4 performance and 4 efficiency cores) and an 8-core GPU, supported by 16 GB of unified memory. The system ran macOS Ventura (Version 13.4). In addition, the GPU's Metal 3 framework optimises parallel computation for machine learning tasks. This configuration provided sufficient processing power and memory to handle the tasks required for this research.

5.2 Training Hyperparameters

For the text-based pipeline, models are trained using the AdamW optimizer with a learning rate of 3e-5. For the image-based pipeline, the Adam optimizer with a learning rate of 1e-4 is adopted. The multimodal pipeline uses the AdamW optimizer with two separate learning rates: 3e-5 for the text model parameters and 1e-4 for all other parts of the model. The loss function is mean squared error for all pipelines, which reflects the regression task. All models are trained for five epochs, as loss logs suggest that this provides the right balance and allows the models to learn sufficient information from the training data without overfitting or wasting computational resources. Models are trained on GPU when available, otherwise on CPU.

5.3 Additional Consistency Settings

To ensure reproducibility and fair comparison, a fixed random seed (42) is used for dataset splitting and sampling across all pipelines. The training set size is set to 80% and the test set size is set to 20% of the data consistently.

6 Evaluation

This research systematically compares and evaluates the performance of the text-based, image-based, and multimodal pipelines, as well as cross-pipeline analyses, from both academic and practitioner perspectives. The evaluation metrics employed are discussed in detail in Section 3. Research Methodology (see Section 3.5. Model Evaluation). Throughout this study, percentage errors are reported relative to the average value across samples. The key experimental results are presented in Tables 11-13 and Figure 2.

Table 11: Text-Based Pipeline: Full Fine-Tuning Performance (Text Variant 2 vs. 4)

Text Variant	Calorie RMSE (Std)	Calories MAE (Std))	Calorie RMSE (kcal))	Calories MAE (kcal, %)	Macros RMSE (Std)	Macros MAE (Std)	Macros RMSE (g)	Macros MAE (g, %)	Fat (%)	Carbs (%)	Protein (%)
v2	0.7022	0.2140	155.50	47.40, 19.9%	0.8992	0.2095	18.78	3.79, 24.3%	25.9%	26.4%	20.9%
v4	0.6419	0.1444	142.17	31.98, 13.4%	0.8487	0.1381	18.02	2.53, 16.2%	15.7%	18.4%	14.1%

Interpreting Table 11: The first row (v2) shows results for text variant 2 (ingredients in descending order without mass) using the fully fine-tuned DistilBERT model. The standardized calorie errors are RMSE 0.7022 and MAE 0.2140, while in original units this translates to 47.40 kcal MAE (19.9% error). The table includes both combined and individual macronutrient errors. Comparing both variants, variant 4 (with mass data) consistently gives better results across all measurements.

Table 12: Image-Based Pipeline: Full Fine-Tuning Performance (3,000-Sample Subset)

Subset size	Calorie RMSE (Std)	Calories MAE (Std))	Calorie RMSE (kcal))	Calories MAE (kcal, %)	Macros RMSE (Std)	Macros MAE (Std)	Macros RMSE (g)	Macros MAE (g, %)	Fat (%)	Carbs (%)	Protein (%)
3000	0.4062	0.2602	89.95	57.63, 24.8%	0.4457	0.2943	7.82	5.24, 34.6%	39.7%	30.5%	35.4%

Table 13: Multimodal Pipeline: Fusion Model Performance (Text Variants 2 & 4, 3,000 Samples)

Fusion Model	Text Variant	Calorie RMSE (Std)	Calories MAE (Std))	Calorie RMSE (kcal))	Calories MAE (kcal, %)	Macros RMSE (Std)	Macros MAE (Std)	Macros RMSE (g)	Macros MAE (g, %)	Fat (%)	Carbs (%)	Protein (%)
Early	2	0.3800	0.2078	84.16	46.03, 19.8%	0.3748	0.2015	6.09	3.51, 23.1%	31.0%	20.4%	20.4%
Early	4	0.2746	0.1737	60.82	38.48, 16.6%	0.2899	0.1751	5.02	3.18, 21.0%	21.6%	24.0%	17.1%
Late	2	0.3173	0.1940	70.26	42.96, 18.5%	0.3443	0.2089	5.92	3.73, 24.6%	27.9%	24.1%	22.6%
Late	4	0.3068	0.1950	67.95	43.19, 18.6%	0.3235	0.1907	5.40	3.35, 22.1%	27.9%	19.3%	20.8%
Attention (Global Weights)	2	0.3469	0.1977	76.83	43.78, 18.9%	0.3635	0.2034	6.18	3.63, 23.9%	26.9%	21.5%	24.3%
Attention (Global Weights)	4	0.3013	0.1819	66.73	40.27, 17.4%	0.3120	0.1692	5.26	3.01, 19.8%	23.2%	19.1%	18.2%
Attention (Cross-Modal)	2	0.4214	0.2452	93.31	54.31, 23.4%	0.4662	0.2719	7.91	4.83, 31.8%	37.7%	31.4%	28.0%
Attention (Cross-Modal)	4	0.4715	0.2710	104.42	60.03, 25.9%	0.4873	0.2914	8.58	5.24, 34.6%	37.4%	36.1%	30.8%
Gated (Per-Feature Weights)	2	0.3153	0.1870	69.82	41.41, 17.9%	0.3461	0.1965	5.87	3.54, 23.3%	25.2%	24.3%	20.9%
Gated (Per-Feature Weights)	4	0.3310	0.2004	73.30	44.37, 19.1%	0.4023	0.2267	6.93	4.06, 26.8%	29.3%	25.7%	26.2%
Gated (Scalar Weight)	2	0.3639	0.2317	80.60	51.31, 22.1%	0.3757	0.2288	6.65	4.15, 27.4%	27.4%	28.3%	26.4%
Gated (Scalar Weight)	4	0.2628	0.1590	58.19	35.20, 15.2%	0.2876	0.1591	4.73	2.81, 18.5%	23.1%	19.1%	14.4%

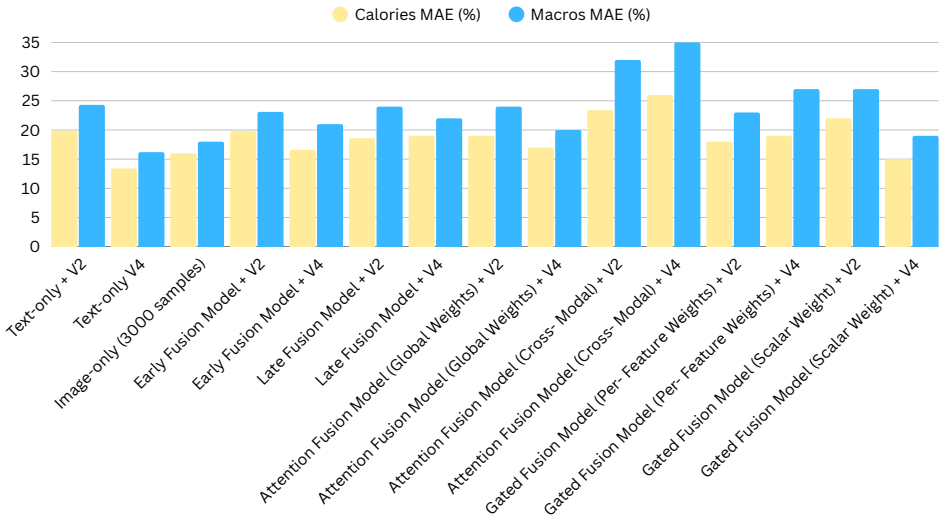

Figure 2: Performance Comparison of Models for Calorie and Macronutrient Estimation (MAE %)

6.1 Evaluation of the Text-Based Approach

6.1.1 Impact of Ingredient Ordering and Per-Ingredient Mass

Across all models, text variant 4 consistently achieves the best performance, showing strong advantages over other variants. In particular, the model trained with the full fine-tuning strategy achieves a 13.4% error in calories MAE and a 16.2% error in combined macros when using text variant 4. This result supports Hua et al.'s study (Hua et al. 2025), highlighting the importance of precise quantitative data in the text-only approach.

In addition, text variants 2 and 3 follow text variant 4, showing similar performance to each other across all models. These results reveal a key finding of this study, which is that textual structure (ingredient ordering) is as important as quantitative data for text-only

nutrition estimation. When quantitative data is unavailable, ingredient ordering serves as a strong alternative to improve accuracy. When both are available, combining clear textual structure with quantitative data can further improve model performance.

6.1.2 Impact of Fine-Tuning Strategy

Among the four fine-tuning configurations, full fine-tuning consistently achieves the best performance across all text variants. In contrast, the fixed feature configuration with all DistilBERT layers frozen achieves the worst performance. This indicates that DistilBERT's original training data is very different from the data used in this research. Therefore, fine-tuning is crucial to enable DistilBERT to learn relevant features from the domain-specific data to achieve strong results.

6.1.3 Performance Across Macronutrient Metrics

The ranking of model versions and text input variants remains consistent across all evaluated metrics, including calories, fat, carbohydrates, and protein. Among these, calorie is the easiest metric to predict, followed by carbohydrates, then protein, with fat being the most difficult to predict. This pattern is likely due to differences in the complexity of these macronutrients within text descriptions and the level of correlation of these macronutrients to common ingredients.

Calories and carbohydrates are generally easier to predict because their amounts are often more directly related to the common ingredients listed. For example, ingredients like rice and bread directly contribute to carbohydrates. However, fat is more difficult to predict because only a small number of ingredients have a strong correlation with fat, such as olive oil. Many ingredients have variable fat content, and it is difficult to fully capture these relationships using only the current data. The difficulties in fat prediction suggest that it is a crucial metric to focus on when assessing model performance. Since fat contains more calories per gram than carbohydrates and protein (9 calories per gram vs. 4 calories per gram), improving fat prediction is likely to significantly improve calorie prediction as well.

6.1.4 Relationship Between RMSE and MAE

For the full fine-tuning model, RMSE values are consistently larger than MAE values for both calories and combined macros. For example, with text variant 4, the model's standardised MAE for calories is 0.1444, while its RMSE is 0.6419. Since RMSE penalises larger errors more heavily than MAE, this pattern indicates that this model performs well on most samples but is affected by a few difficult-to-predict cases.

In contrast, underperforming models (e.g., the model with all training layers frozen) tend to have a smaller gap between RMSE and MAE. This suggests that their errors are more evenly spread, and the MAE reflects their overall performance more directly.

6.2 Evaluation of the Image-Only Approach

6.2.1 Impact of Dataset Size

Model performance generally improves with more data, with two exceptions in the fixed-features model, which shows a slight increase in error when the data size grows from 600

to 1,000 samples and from 2,000 to 3,000 samples. This unusual pattern further confirms that models with all training layers frozen struggle to adapt to domain-specific tasks.

The difference in performance between the datasets of 3,000 and 3,490 samples is very small across all models. However, it is worth noting that the models with full fine-tuning and with only the last layer unfrozen both achieve their best results using the 3,000-sample subset rather than the full 3,490 samples. Meanwhile, RMSE for calories and combined macros increases as the data subset grows from 3000 to 3490. This result indicates that the 3,000-sample subset achieves the best balance between diversity and noise, and there are outlier samples in the additional data.

This finding suggests that, to reduce noise and maintain consistency, all pipelines should be evaluated using the 3,000-sample subset instead of the full dataset. This approach ensures that the results can better reflect the models' true capacity.

6.2.2 Impact of Fine-Tuning Strategy

A similar pattern to the text-based approach is present in the image-only approach, with full fine-tuning of ResNet-18 consistently achieving the best performance across all data subset sizes. For example, with a data subset of 3,000 samples, full fine-tuning results in a 24.8% error in calories MAE and a 36.4% error in combined macros, achieving the best performance across all configurations. This pattern supports the idea discussed in Section 6.1.2, suggesting that fine-tuning is crucial for enabling pretrained models to achieve strong performance in domain-specific tasks.

6.2.3 Performance Across Macronutrient Metrics

For the image-based approach, the pattern across macronutrients is similar to the pattern in the text-based approach. Calories and carbohydrates are the easiest to predict, with fat being the most difficult. For example, the full fine-tuning model trained on the 3,000-sample subset shows errors of 24.8% for calories MAE, 30.5% for carbohydrates MAE, 35.4% for protein MAE, and 39.7% for fat MAE. This result supports the finding from the text-based approach, that a common challenge in nutritional estimation is accurately predicting the fat value.

6.3 Evaluation of the Multimodal Approach

6.3.1 Best Performing Multimodal Fusion Strategies

Gated fusion with scalar weight using text variant 4 achieves the best results among all configurations, with a 15.2% error in calories MAE and an 18.5% error in combined macros. This is likely due to its ability to effectively balance and selectively integrate text features with visual information. The runner-up is early fusion with text variant 4, achieving a 16.6% error in calories MAE and a 21% error in combined macros. This result suggests that using low-level feature combination to merge raw features before deeper processing can still be effective for leveraging both textual and visual data in nutritional estimation tasks.

When leveraging text variant 2, gated fusion with per-feature weights achieves the best result, with a 17.9% error in calories MAE and a 23.3% error in combined macros. This result suggests that when precise pre-ingredient mass data is not available, the image modality can successfully guide portion estimation in this architecture. However,

the less impressive result of gated fusion with per-feature weights with text variant 4, with a 19.1% error in calories MAE and a 26.8% error in combined macros, also suggests that precise text mass values may introduce noise that harms fusion performance in this architecture.

6.3.2 Impact of Text Variants in the Multimodal Setting

Unlike the consistent advantage observed for text variant 4 in the text-based models, text variant 4 does not consistently outperform text variant 2 across all models in the multimodal setting. In this research, four multimodal architectures, including early fusion, late fusion, attention fusion with global weights, and gated fusion with scalar weight, achieve better results using text variant 4. Meanwhile, the attention fusion model with global weighting and the gated fusion model with per-feature weighting perform better when using text variant 2. These mixed findings suggest that while pre-ingredient mass data offer additional valuable information for some multimodal models, including them may introduce noise or the overfitting issue in certain specific architectures. When leveraging both textual and visual data in the multimodal setting, it is important to select a text variant choice appropriate to the specific multimodal architecture.

6.3.3 Performance of Late Fusion with Different Text Variants

Late fusion performs similarly across all evaluated metrics when using both text variant 2 and variant 4, achieving 18.5% and 18.6% error in calories MAE, respectively. These results further support the observation made in Section 6.3.1, which is that low-level feature combination methods can be effective to a certain degree in nutritional estimation tasks. Moreover, the minimal difference in performance between text variant 2 and variant 4 suggests that when text features are integrated too late in the pipeline, after the visual feature extraction stage, the added value of precise pre-ingredient mass data is limited.

6.3.4 Underperformance of Attention with Cross-Modal Fusion

The attention with cross-modal fusion model results in the worst performance among all configurations when using both text variant 2 and text variant 4. This model employs multiple attention heads to match individual text tokens to corresponding image patches, which makes it a theoretically powerful model. However, the results suggest that this approach overcomplicates the problem in the context of nutrition estimation. In addition, the dataset of 3,000 samples may be insufficient for this architecture to learn comprehensively.

6.4 Discussion

When precise mass data is available, the text-only approach alone is sufficient for nutrition estimation, with DitilBERT combined with full fine-tuning and text variant 4 achieving the best performance across all pipelines and configurations. This is because the ingredient list and individual mass values together enable the model to effectively learn the linear relationships between ingredient content and nutritional values. Even simple fusion methods, such as early fusion with text variant 4, underperform compared to the text-only approach with text variant 4, suggesting that adding additional images in this setting introduces unnecessary noise. This is likely because textual information

is so precise that images have little to contribute additionally. However, as discussed in Section 1. Introduction, acquiring precise per-ingredient mass data in real life is not practical. Therefore, research should focus on how to compensate for missing quantitative data to achieve better overall performance, with strategies such as leveraging images.

Although the image-only approach underperforms, which is expected as previously discussed in Section 3. Research Methodology, images play an important role when text lacks pre-ingredient quantitative data in the multimodal setting. When the best performance with text variant 2 in the text-based approach achieves a 19.9 % error in calories MAE and a 24.3% error in combined macros, gated fusion with per-feature weights achieves a 17.9 % error in calories MAE and a 23.3% error in combined macros when leveraging text variant 2 and images, presenting a 2.0% MAE improvement. This result reveals that images are able to compensate for missing masses in text in nutritional estimation by adopting per-feature gated fusion to uniquely leverage visual portion cues. This finding can be further examined by testing text data that follows a different format or structure, with and without pre-ingredient mass, just like in this research.

The underperformance of certain models and architectures reveals the nature of the nutritional estimation task. The failure of cross-modal attention fusion suggests that the approach to match individual text tokens to corresponding image patches is powerful theoretically, but overkill in practice in the current setting. Whether the reason for the underperformance is this approach overcomplicating the problem in the context of nutrition estimation, or the lack of sufficient data to enable proper learning, it will require a larger dataset to be fully investigated. Therefore, future work should investigate the performance of the cross-modal attention fusion model when trained on a larger dataset.

Across all pipelines, the performance across macronutrients presents a consistent pattern. Calories and carbohydrates are the easiest to predict, with fat being the most challenging to predict. This reveals that a key area to focus on in nutritional estimation is fat estimation. Although fat estimation remains consistent with the improvement in calories and combined macros estimation, how to address this specific metric remains an area to further explore in future work.

7 Conclusion and Future Work

This study focuses on investigating how images compensate for missing quantitative data in text in nutritional estimation within a multimodal setting. As previous studies have revealed the strengths and limitations of the image-only approach, text-only approach, LLMs, and MLLMs in nutritional estimation, as well as the advantages of DistilBERT and ResNet for domain-specific text and image processing, this study implements three parallel pipelines to systematically review and compare the text-based, image-based, and multimodal approaches in this domain by leveraging both textual and visual data. In addition, this study systematically evaluates how different multimodal fusion techniques and architectures impact model performance in scenarios where textual data lacks precise pre-ingredient mass information and therefore requires additional visual cues for more accurate estimation.

This study confirms one key finding from previous studies, which is that the text-only approach requires precise quantitative data to provide accurate nutritional estimation. Furthermore, this study finds that when precise quantitative data is available, the text-only approach alone is sufficient for nutrition estimation, with DistilBERT combined

with full fine-tuning achieving the best results across all pipelines and models in this study. In scenarios where textual data lacks precise mass information, images can provide visual portion cues in the multimodal setting effectively, resulting in 2.0% improvement in calories MAE error compared to using textual data alone.

Furthermore, this study finds that precise text mass values do not always contribute to more accurate estimations in the multimodal setting. Whether the pre-ingredient quantitative text data brings valuable information or introduces unnecessary noise depends on the specific multimodal architecture. In addition, the results of this study suggest that while gated fusion with scalar weight and gated fusion with per-feature weights achieve the best results when using text variant 2 and text variant 4 respectively, low-level feature combination methods such as early fusion can also effectively leverage different data modalities in nutritional estimation tasks.

One limitation of this study is that it does not further investigate the underperformance of the cross-modal attention fusion model, which is theoretically powerful because it can match individual text tokens to corresponding image patches. This study proposes two possible underlying causes, both of which require a larger dataset to be further investigated. In addition, this study does not provide additional solutions specifically for addressing the difficulty in fat estimation, which could be a focus of future work. Last but not least, another area to focus on in future work is to further examine the impact of gated fusion with per-feature weights on text data that hs a different format or structure from the text data generated in this study.

References

Amandeep & Suresh, S. (2025), 'Transforming fake news detection: Leveraging distilbert models for enhanced accuracy', *Procedia Computer Science* **260**, 283–290.

Anisuzzaman, D. M., Malins, J. G., Friedman, P. A. & Attia, Z. I. (2025), 'Fine-tuning large language models for specialized use cases', *Mayo Clinic Proceedings: Digital Health* **3**(1), 100184.

Basiri, R., Seidu, B. & Cheskin, L. J. (2023), 'Key nutrients for optimal blood glucose control and mental health in individuals with diabetes: A review of the evidence', *Nutrients* **15**(18), 3929.

Burke, L. E., Wang, J. & Sevick, M. A. (2011), 'Self-monitoring in weight loss: A systematic review of the literature', *Journal of the American Dietetic Association* **111**(1), 92–102.

Hua, A., Dhaliwal, M. P., Pullela, L., Burke, R. & Qin, Y. (2025), Nutribench: A dataset for evaluating large language models on nutrition estimation from meal descriptions, *in* 'Proceedings of the International Conference on Learning Representations (ICLR)'.

Khan, H. U., Naz, A., Alarfaj, F. K. & Almusallam, N. (2025), 'Identifying artificial intelligence-generated content using the distilbert transformer and nlp techniques', *Scientific Reports* **15**(1), 20366.

Małkiński, M., Pawlonka, S. & Mańdziuk, J. (2025), 'Reasoning limitations of multimodal large language models: A case study of bongard problems'. Submitted to ICLR 2025. **URL:** *https://openreview.net/forum?id=BTk1hNuIPq*

O'Hara, C., Kent, G., Flynn, A. C., Gibney, E. R. & Timon, C. M. (2025), 'An evaluation of chatgpt for nutrient content estimation from meal photographs', *Nutrients* **17**(4), 607.

Qi, S., Cao, Z., Rao, J., Wang, L., Xiao, J. & Wang, X. (2023), 'What is the limitation of multimodal llms? a deeper look into multimodal llms through prompt probing', *Information Processing & Management* **60**(6).

Rohanian, O., Nouriborji, M., Jauncey, H., Kouchaki, S., Nooralahzadeh, F., Clifton, L., Merson, L., Clifton, D. A. & Group, I. C. C. (2024), 'Lightweight transformers for clinical natural language processing', *Natural Language Engineering* **30**(5), 887–914.

Sanh, V., Debut, L., Chaumond, J. & Wolf, T. (2019), Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *in* 'Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing'. arXiv:1910.01108.

Swan, W. I., Vivanti, A., Hakel-Smith, N. A., Hotson, B., Orrevall, Y., Trostler, N., Howarter, K. B. & Papoutsakis, C. (2017), 'Nutrition care process and model update: Toward realizing people-centered care and outcomes management', *Journal of the Academy of Nutrition and Dietetics* **117**(12), 2003–2014.

Tabassum, I. & Nunavath, V. (2024), 'A hybrid deep learning approach for multi-class cyberbullying classification using multi-modal social media data', *Applied Sciences* **14**(24), 12007.

Tanabe, H. & Yanai, K. (2025), 'Reasoning-driven food energy estimation via multimodal large language models', *Nutrients* **17**(7), 1128.

Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T. & Sim, J. (2021), Nutrition5k: Towards automatic nutritional understanding of generic food, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)', IEEE, pp. 8903–8911.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T. & Chen, E. (2024), 'A survey on multimodal large language models', *National Science Review* **11**(12), nwae403.
URL: *https://doi.org/10.1093/nsr/nwae403*