Multimodal Dietary Knowledge Graph-Driven Visual Language Model for Food Question Answering

Anonymous ACL submission

Abstract

Food analysis is crucial for personalized nutrition guidance and disease management. However, existing Visual Language Models (VLMs) have limitations in understanding deep, multi-dimensional food knowledge, such as nutritional composition, cultural background, and health impacts. Current food datasets and knowledge graphs often focus on textual knowledge, lacking visual information or failing to integrate cross-domain knowledge. To address these challenges, we constructed DietKG-VQA—the first large-scale food analysis benchmark (3404 images, 10219 question-answering pairs) that fuses multi-domain (nutrition, culture, health) structured knowledge with visual information. We also propose a novel method for enhancing VLMs based on a Multimodal Dietary Knowledge Graph (MDKG): by constructing an MDKG that incorporates visual information, and combining visual similarity retrieval, knowledge graph querying, and our proposed VLM-guided Knowledge Pruning & Selection (V-KPS) mechanism, we precisely extract core knowledge to enhance VLM reasoning, especially for uncommon food items. Experimental results on the DietKG-VQA benchmark show that the proposed method significantly outperforms baseline VLMs; for example, gpt_4o_mini's comprehensive average score increased substantially from a baseline of 34.81% to 76.02%. The DietKG-VQA benchmark and related code will be publicly released.

1 Introduction

Dietary habits are closely related to human health and are key influencing factors for many chronic diseases such as obesity, diabetes, and cardiovascular diseases (Amiri et al., 2019; Min et al., 2019). Therefore, accurate dietary assessment, nutritional monitoring, and personalized dietary management play an increasingly important role in modern healthcare and daily life (Rollo et al., 2016). For example, diabetic patients need to accurately estimate dietary carbohydrate content to determine insulin dosage, as incorrect estimations can lead to severe health complications (Contreras et al., 2023; Buck et al., 2022).

In recent years, Visual Language Models have shown great potential in general image understanding and question answering. However, when applied to the complex field of food analysis, a significant gap exists between their current capabilities and application needs. Firstly, existing VLMs often lack deep, domain-specific food knowledge. They might be able to identify basic food categories in images, but perform poorly when answering complex questions requiring specialized knowledge, such as precise quantification of nutritional components, cultural traceability, or reasoning about the health impacts of specific foods. Secondly, existing food-related datasets have limitations. Many datasets either only contain images (Ma et al., 2023; Liang and Li, 2017; Tai et al., 2023), consist of tabular nutritional data lacking natural language descriptions (USDA, 2019; Nutritionix, 2018), or focus solely on a single information dimension (like recipes or nutrition), making it difficult to support complex analysis tasks that require integrating multi-source information (Hezarjaribi et al., 2017). While some work has begun to focus on using Large Language Models to estimate nutrition from textual descriptions (Hua et al., 2024), they overlook the important modality of visual information.

Knowledge Graphs, as an effective representation of structured knowledge, have been used in some food-related tasks, such as food recommendation (Haussmann et al., 2019) or text-based question answering (Chen et al., 2021). KGs can integrate heterogeneous data, promote knowledge discovery, and are valuable for food safety, nutritional assessment, and diet-disease association studies (Min et al., 2022). However, a key pain point of

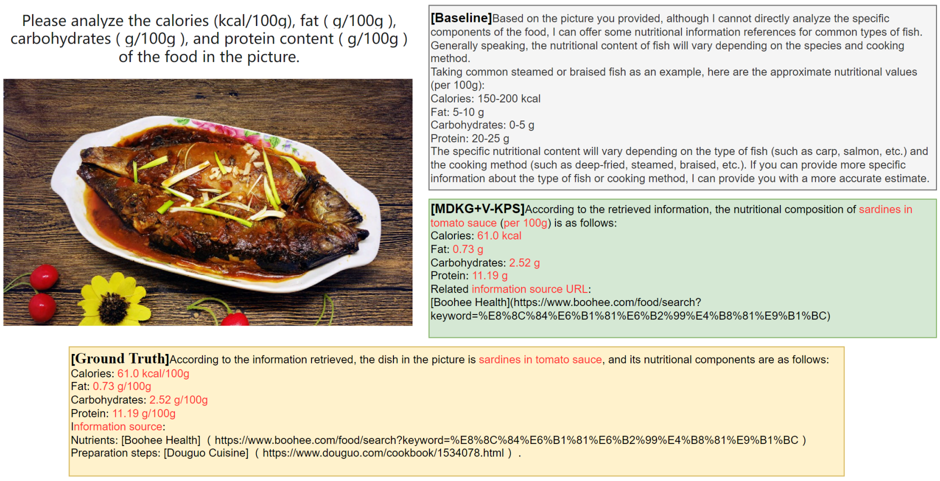

Figure 1: GPT-4o-mini answering nutrition-related questions from DietKG-VQA using different prompting strategies.

existing food KGs is that they mostly focus on organizing textual information and generally lack structured integration of visual knowledge (such as food images and their features) (Min et al., 2022), which directly limits their application potential in visual tasks (especially VQA). Currently, research on systematically utilizing multimodal dietary KGs to enhance VLMs for solving complex visual question answering tasks requiring deep, multi-dimensional knowledge reasoning (such as image-based nutrition quantification, food-disease association reasoning) is still nascent. Furthermore, existing methods often struggle to provide effective information when faced with visually unique or very uncommon dishes.

To bridge the aforementioned gaps, this study aims to answer a core question: How can structured multimodal knowledge be effectively utilized to enhance the accuracy and depth of VLMs in complex food visual question answering tasks? We are committed to exploring the use of multimodal dietary KGs to enhance the food question answering capabilities of VLMs. Our main contributions include:

1) **Proposing DietKG-VQA**: a novel benchmark specifically designed to evaluate the deep food understanding capabilities of VLMs, being the first to fuse visual information with multi-domain structured knowledge (nutrition, culture, health impacts, etc.).

2) **Designing and implementing a framework for enhancing VLMs with a multimodal dietary KG**: This framework constructs an MDKG that integrates visual information and innovatively combines visual similarity retrieval with KG querying, initiating a VLM-guided Knowledge Pruning & Selection process we propose, to precisely extract and format external knowledge. A highlight of this framework is its ability to handle uncommon food items by bridging to known food knowledge through visual similarity for reasoning, enhancing robustness.

3) **Systematically verifying the superiority of this framework through experiments**: We conducted baseline evaluations on DietKG-VQA for various mainstream VLMs (including open-source and closed-source models) and demonstrated that our proposed MDKG method can significantly improve model performance on complex food knowledge question answering tasks.

2 Related Work

Existing work has covered food image recognition (Knez and Šajn, 2020; Klasson et al., 2020; Liu et al., 2021), portion/calorie estimation (Yunus et al., 2018; Keller et al., 2024), ingredient recognition (Chen et al., 2020), and recipe generation (Shirai et al., 2020). Visual Question Answering (VQA) (Yin et al., 2023) has also begun to be applied in the food domain, but is often limited to basic questions, primarily relying on the model's internal knowledge, and struggles with complex analyses requiring external professional knowledge. Although some studies have utilized KGs to enhance visual food recognition (Lu et al., 2020b,a), their goal was to improve recognition performance, rather than, as in this study, using a multimodal KG containing visual information for complex, multi-domain knowledge-based visual question answering.

To integrate structured food information, researchers have constructed various food KGs. These efforts often begin with the development of food ontologies, such as Taaable focusing on cooking, PIPS and FOODS on nutrition and health, and the more comprehensive FoodOn (Min et al., 2022). Based on these ontologies or directly from data, KGs like FoodKG (Haussmann et al., 2019) were created, integrating recipes, ingredients, and nutritional data, mainly for recommendation systems or text-based question answering. Other works may focus on specific regions, themes, or industrial applications (e.g., internal KGs built by Uber Eats, Edamam) (Hamad et al., 2018; Çelik, 2015). These KGs have shown value in organizing textual information, supporting semantic search (Huang et al., 2019), text-based QA (Qin et al., 2019), and discovering diet-disease associations (Afshin et al., 2019; Zhao et al., 2020). However, the vast majority of these KGs only contain textual knowledge and generally lack direct, structured association with food images (Lei et al., 2021). This limits their direct support for tasks requiring visual input (such as VQA). Our work directly addresses this key limitation by constructing an MDKG that explicitly includes image information.

Nutrition estimation is a core task in food analysis. Traditional methods rely on querying tabular databases (USDA, 2019; Nutritionix, 2018), but this often requires exact matches and is cumbersome for multi-ingredient meals (Hezarjaribi et al., 2017). Image-based methods (Keller et al., 2024; Yunus et al., 2018) are susceptible to visual factors and often lack interpretability. Recent work has started to utilize LLMs to estimate nutrition from natural language descriptions (Hua et al., 2024). This study, however, focuses on starting from visual input and combining structured multi-domain knowledge provided by an MDKG for more comprehensive and reliable visual question answering.

To overcome the knowledge limitations of (V)LMs (such as hallucinations, outdatedness), methods like Retrieval Augmented Generation (RAG) (Lewis et al., 2020) have been proposed, which enhance generation by retrieving information from external knowledge bases (databases, KGs, etc.). This approach has proven effective in both general and specific domains (Wu et al., 2024). Research on using KGs as external knowledge sources to enhance models is also ongoing (Wang et al., 2024). However, filtering useful knowledge from vast retrieval results and organizing it in a way that VLMs can easily understand remains a challenge. The VLM-guided Knowledge Pruning & Selection mechanism proposed in this study aims to address this issue, ensuring that the injected knowledge is both relevant and refined.

Compared to existing work, our core distinctions are: 1) Constructing the first multimodal benchmark, DietKG-VQA, specifically designed for evaluating deep food understanding, which integrates multi-domain knowledge with visual information; 2) Proposing and validating a new method that utilizes an MDKG and employs an advanced, VLM-driven knowledge pruning and selection strategy to enhance models for complex food visual question answering.

3 DietKG-VQA Benchmark

To systematically evaluate and advance the capabilities of multimodal models in complex food analysis tasks, we constructed the DietKG-VQA benchmark. Its core objective is to provide an evaluation platform that contains rich, multi-dimensional, structured knowledge closely associated with visual information, specifically for measuring the deep food understanding capabilities of models.

3.1 Data Sources and Integration

DietKG-VQA integrates the following six authoritative or widely used Chinese and English databases, covering key information dimensions:

- Boohee Food Database (Boohee, 2025): Provides nutritional information (calories, carbohydrates, fat, protein, in kcal/100g or g/100g) and main ingredients for common Chinese dishes.

- Douguo Recipe Database (Douguo, 2025): Offers a large number of structured recipes for Chinese dishes, including detailed ingredient lists, quantities, cooking steps, and classification information.

- Nutritionix (Nutritionix, 2018): Provides an extensive nutritional database covering common American foods (packaged foods, restaurant dishes), including macro and micronutrient information (standardized per 100g).

- Baidu Baike (Baidu Baike, 2025) & Wikipedia (Wikipedia, 2025): Used to extract historical origins, cultural backgrounds, flavor characteristics, and regional popularity of foods (especially dishes) (e.g., attribution to China's eight major cuisines, popular dishes in specific US regions).

- Baidu Health Dictionary (Baidu Health, 2025): Provides expert-reviewed association information between food and diseases, encoding the potential beneficial or harmful effects of food on approximately 1,849 common diseases.

3.2 Data Preprocessing

To ensure consistency and quality across databases, we performed the following preprocessing steps:

Unit Normalization: All nutrient contents were uniformly converted to a "per 100g" standard (e.g., kcal/100g, g/100g). Regular expressions were used to handle and eliminate heterogeneity in unit expressions.

Entity Alignment: Using the paraphrase-multilingual-MiniLM-L12-v2 model based on cosine similarity, we aligned identical food entities from different databases (e.g., "potato" (tǔdòu) and "potato" (mǎlíngshǔ), Chinese and English names for "scrambled eggs with tomatoes" (xīhóngshì chǎo dàn)).

3.3 Benchmark Construction Methodology

From the images associated with the above databases, we carefully selected 3404 representative images that were not subsequently used in the construction of the knowledge graph. Image selection followed principles of clarity, typicality, and visual diversity.

For each image and its associated metadata, we defined four types of questions:

1. *Nutritional Analysis*: "Please analyze the calorie (kcal/100g), fat (g/100g), carbohydrate (g/100g), and protein content (g/100g) of the food in the image."

2. *Regional Popularity*: "In which regions is the food in the image most popular?"

3. *Pathological Association*: "What diseases might the food in the image be beneficial for, and what diseases might it be harmful for?"

4. *Basic Information*: "What food is in the image? Please provide some information about this food."

To ensure the high quality and consistency of the "Ground Truth," we prioritize the use of unified, standardized prompts rather than pursuing diversity. This approach is designed to guarantee the precise correspondence between answers and metadata, and to enhance the efficiency of expert review. The diversity of the dataset is already provided by 3,404 visually rich images and four core question types.

The generation of benchmark answers relies on the **GPT-4o** model. The process involves: first, constructing a multimodal knowledge graph from the original metadata; second, employing visual similarity retrieval, knowledge graph querying, and the V-KPS (VLM-guided Knowledge Pruning & Selection) mechanism to precisely extract core knowledge highly relevant to the input image from the graph; finally, this refined knowledge, along with the corresponding image, is jointly input into the **GPT-4o** model (parameters: `max_new_tokens=2048`, `temperature=0.3`, `top_p=0.75`) to generate preliminary answers.

To guarantee the quality and accuracy of the final answers, we assembled a professional review team. All team members are nutrition experts who have passed the Chinese Registered Dietitian examination and have accumulated over 5 years of experience in the field of public nutrition. The expert team is responsible for meticulously reviewing the generated preliminary answers, ensuring their complete fidelity to the original metadata. They make final adjudications, modifications, and confirmations independently of the V-KPS mechanism, and eliminate any model hallucinations or content inconsistent with the metadata. Through this process, the **DietKG-VQA** benchmark, comprising 10,219 high-quality image-question-answer pairs, was ultimately established.

4 Methodology

We propose a novel method for enhancing VLMs for food question answering based on a multimodal dietary knowledge graph. The core idea of this method is to use external, structured food knowledge containing visual information to compensate for the model's deficiencies in domain depth and knowledge accuracy. The overall framework consists of three main parts: construction of the multimodal food knowledge graph, knowledge retrieval and summarization based on the MDKG, and the knowledge-enhanced visual question answering pipeline.

4.1 Multimodal Food Knowledge Graph Construction

The knowledge graph we constructed aims to integrate multi-dimensional food information and explicitly incorporate visual information into the KG structure.

Data Integration and Schema Design: Based on the six data sources described in Section 3.1, we designed a knowledge graph schema, defining core entity types and relation types.

Category	Count
Disease	1849
Food	6058
Ingredient	4149
Region	1484
Classification	2306
Not_Recommend_Food	954
Recommend_Food	802
Relation	**Count**
Belongs_To	7299
Contains_Ingredient	23029
Not_Recommend_To_Eat	5749
Popular_In	8898
Recommend_To_Eat	6277

Table 1: Statistics of Knowledge Graph Entity and Relation Types.

Data Processing and Ingestion: After the data preprocessing described in Section 3.2 (unit normalization, entity alignment), the structured data was converted into knowledge graph triples (Head Entity, Relation, Tail Entity) and entity attributes (Entity, Attribute, Value), and stored in a graph database (Neo4j).

Visual Information Embedding: To associate Food entities in the KG with corresponding images, we linked nodes to image files by storing the local image URL (local_image_url attribute). We pre-computed DINOv2-large high-dimensional feature vectors for 5108 representative images associated with Food entities in the MDKG. These vectors were stored in an efficient vector index library (FAISS), and the mapping between image vectors and their corresponding Food nodes in the KG was preserved. This step is key to achieving efficient visual similarity retrieval.

Note: The current knowledge graph was constructed to validate the effectiveness of the contributions and does not include all food and disease information. Nevertheless, its coverage is sufficient to support the construction of the DietKG-VQA benchmark and to preliminarily validate the effectiveness of the proposed method. Further enrichment of the graph entities will be carried out in subsequent work.

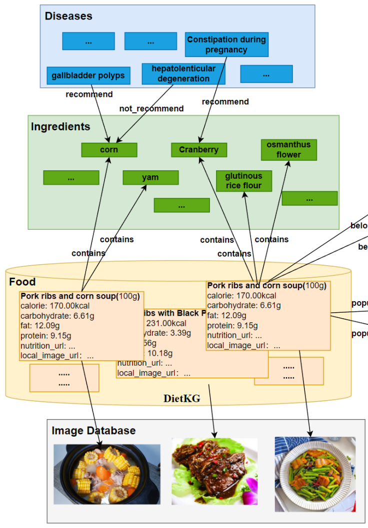

Figure 2: Multimodal Knowledge Graph Schema Diagram.

4.2 Visual Question Answering Based on Multimodal Dietary Knowledge Graph

When given a user-input food image (Query Image) and a related question, our method utilizes the knowledge graph to enhance the VLM's answering ability through the following three key stages.

4.2.1 Visual Similarity Retrieval and Raw Knowledge Extraction

Image Feature Extraction and Knowledge Retrieval: Extract the feature vector of the Query Image using DINOv2-large.In the pre-built FAISS vector index library, perform a K-Nearest Neighbors (KNN) search using this feature vector to retrieve the food images from the knowledge graph that are visually most similar to the Query Image, along with their corresponding Food nodes.

Knowledge Graph Querying and Raw Knowledge Extraction: Based on the Food node identifiers associated with the retrieved most similar images, query the MDKG. Using a graph query language (e.g., Cypher), starting from this Food

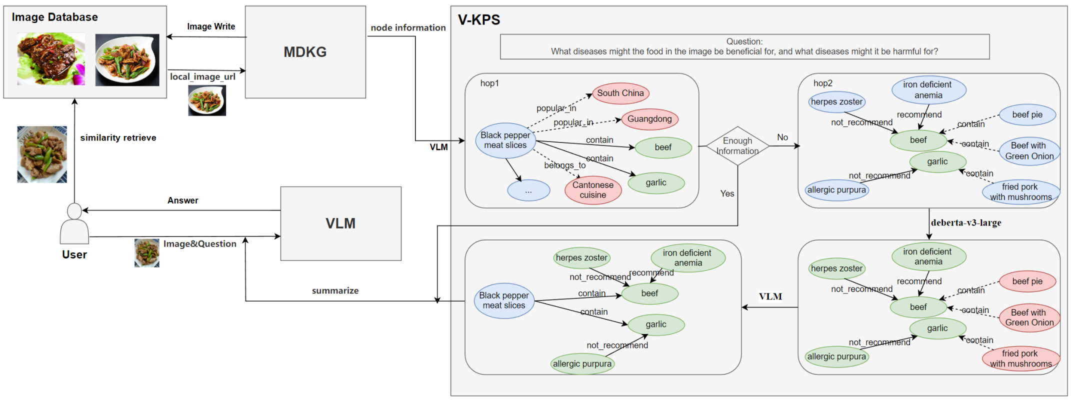

Figure 3: MDKG-based Visual Question Answering Method Flowchart.

node, query its one-hop and two-hop neighboring nodes, and then filter the knowledge snippets using the VLM-guided Knowledge Pruning & Selection method described below.

```
// Query disease association
MATCH p=(n:Food {local_image_url:
$image_path})-[*1..2]-(m:Disease)
WHERE length(p) = 2
AND type(relationships(p)[1])
= 'not_recommend_to_eat'
RETURN m {.*} AS Disease,
       [node IN nodes(p) WHERE node <> m
       | node.name] AS RelatedNames
```

4.2.2 VLM-guided Knowledge Pruning & Selection (V-KPS))

After identifying initial food entities in the MDKG through visual similarity retrieval, the goal of this stage is to precisely extract and filter knowledge snippets most relevant to the user's question from the knowledge graph.

1. One-hop Knowledge Filtering: Based on the previously queried knowledge snippets, filter its one-hop neighbor knowledge S_{hop1_raw}. Using a VLM combined with the question Q and query image I_Q, determine which knowledge in S_{hop1_raw} (e.g., key ingredients) is most critical for answering the question or serving as an intermediate node, filtering out the Top-N_1 to get $S_{hop1\_selected}$.

2. One-hop Information Sufficiency Judgment: Use the model to evaluate whether $S_{hop1\_selected}$ is already sufficient to answer Q. If the model judges "yes," it may directly format the output or skip two-hop retrieval; if "no" (e.g., only ingredients are known, and further information about their association with diseases is needed), proceed to two-hop retrieval.

3. Two-hop Knowledge Retrieval and Filtering: If two-hop retrieval is needed, expand the entities in $S_{hop1\_selected}$ (e.g., "beef," "onion") in the knowledge graph to obtain raw two-hop knowledge S_{hop2_raw} (e.g., (Iron-deficiency Anemia, recommended_to_eat, Beef)).

   Contextual Relevance Coarse Filtering: Use the deberta-v3-large model to calculate the semantic relevance of each knowledge snippet in S_{hop2_raw} with the query context (Q, I_Q), filtering the Top-W knowledge snippets with the highest scores to form $S_{hop2\_candidate}$.

   VLM-assisted Fine-grained Selection:Using a model (with P_{select_final} prompt) combined with Q and I_Q, perform fine-grained selection on the knowledge snippets in $S_{hop2\_candidate}$ to choose the final Top-K knowledge set S_{final}.

4. Formatted Injection: Format the final selected knowledge set S_{final} (or the one-hop filtered result) into concise natural language text, which is then injected as knowledge enhancement into the subsequent module.

4.2.3 Knowledge-Enhanced Prompt Construction and VLM Inference

Combine the knowledge snippets from the previous step with the original user question to construct a more informative enhanced prompt. Input the Query Image, the question, and the constructed enhanced prompt together into a pre-trained VLM. The model utilizes its own image understanding and language generation capabilities, combined with the external knowledge provided in the prompt, to generate the final answer.

5 Experiments

The experiments aim to validate the effectiveness of our proposed method for enhancing VLMs based on an MDKG, with the following specific objectives:

5.1 Experimental Setup

We use the DietKG-VQA dataset described in Section 3. Evaluation focuses on its four main food knowledge question categories: Nutritional Analysis, Regional Popularity, Pathological Association, and Basic Information. As shown in Table 2, we comprehensively evaluated 10 state-of-the-art VLMs (parameters uniformly set to: max_new_tokens=2048, temperature=0.3, top_p=0.75, top_k=50).

Experimental Groups:

- **Baseline Group**: Directly use the original model to process images and questions from DietKG-VQA to generate answers. Input is (Image, Question).
- **MDKG+V-KPS Enhanced Group**: Apply our proposed method. As detailed in Section 4, through visual similarity retrieval and KG querying, and utilizing the VLM-guided Knowledge Pruning & Selection mechanism for pruning and selection, the VLM's answer generation is enhanced. Input is (Image, V-KPS enhanced Prompt).

Across all four question categories and their respective evaluation metrics, the enhanced models demonstrated significantly superior performance compared to their baseline counterparts.

Nutritional Analysis: Enhanced models consistently achieved significantly lower Mean Absolute Error. For instance, *gpt_4o_mini* reduced the MAE for calories from 47.86 to 4.40.

Regional Popularity: The F1 score for *gpt_4o_mini* increased from 40.28 to 84.24, while for *DeepSeek-VL2-16B*, it rose from 35.73 to 79.29.

Pathological Association: The baseline F1 score of 10.17 for gpt_4o_mini, for instance, reflects that general-purpose vision-language models struggle with accurate fine-grained food-disease association reasoning due to a lack of specialized knowledge. As illustrated in Figure 5, the baseline model tends to rely more on general common-sense judgments (e.g., vaguely stating that "high-salt foods might be detrimental to hypertension") and seldom actively cites or links to professional medical information or verified food-disease knowledge entries. This precisely underscores the critical role that the structured, specialized knowledge provided by the knowledge graph in our study plays in enhancing performance on such complex reasoning tasks. Moving forward, we also plan to optimize the baseline, for example, by employing guiding prompts to encourage the model to retrieve and reason about associations from its internal knowledge.

Basic Information: The BERTScore F1 for *gpt_4o_mini* increased from 32.53 to 55.1, and for *Qwen2.5-VL-72B-Instruct*, it improved from 30.35 to 56.93.

This improvement was consistent across all evaluated models, with even smaller models like *LLaVA 1.6 Mistral 7B* demonstrating significant gains. This highlights the generalizability and effectiveness of our MDKG+V-KPS methodology. The targeted, factual information supplied by the knowledge graph enables models to overcome their inherent knowledge limitations, reduce hallucinations, and provide more accurate, detailed, and contextually-aware answers to complex food-related queries.

5.2 Ablation Study

To further validate the effectiveness of the key modules (especially V-KPS) in our proposed MDKG enhancement framework, we conducted a series of ablation studies. We compared the performance of the following three settings on DietKG-VQA, using the same comprehensive average score (avg) calculation method defined in the main experiment as the evaluation metric.

- **Baseline**: The original model, without any external knowledge enhancement, directly answering questions.
- **Simple KG Enhancement (MDKG)**: The model is combined with the MDKG, but only performs basic KG querying without VLM-guided multi-stage knowledge priority pruning. Retrieved knowledge is provided to the model in a relatively raw form.
- **MDKG+V-KPS Enhanced Group**: The complete MDKG enhancement framework

Baseline	Nutrition MAE calorie	fat	carbohydrate	protein	Region – F1	Disease – F1	Base BERTScore F1	avg
gpt_4o_mini	**47.86**	**3.71**	**7.22**	**3.55**	**40.28**	**10.17**	**32.53**	**34.81**
DeepSeek-VL2-**16B**	104.36	6.37	12.97	5.80	35.73	8.36	25.88	27.69
DeepSeek-VL2-**27B**	129.96	6.72	14.85	5.20	34.53	8.46	25.37	22.24
Llama-3.2-11B-Vision	95.08	6.68	12.74	5.12	33.34	6.54	19.79	25.25
LLaVA 1.6 **Vicuna 7B**	159.14	7.33	21.18	6.96	20.57	8.64	14.2	12.44
LLaVA 1.6 **Vicuna 13B**	158.28	7.42	26.30	6.15	26.93	6.25	16.8	17.34
LLaVA 1.6 **Mistral 7B**	163.74	6.89	17.28	5.70	22.2	5.67	6.23	13.18
Pixtral **12B**	83.89	5.88	10.76	5.03	28.59	9.22	25.16	26.73
Qwen2.5-VL-7B	113.38	6.19	13.92	4.67	32.99	7.36	24.82	23.99
Qwen2.5-VL-72B	71.00	5.25	8.72	3.36	36.02	8.26	30.35	31.83
MDKG+V-KPS								
gpt_4o_mini	**4.40**	**0.41**	**0.75**	**0.18**	**84.24**	**93.9**	55.1	**82.68**
DeepSeek-VL2-**16B**	5.03	1.16	1.46	0.68	79.29	88.58	50.23	78.29
DeepSeek-VL2-**27B**	5.54	0.50	0.97	0.23	78.69	87.32	44.57	76.81
Llama-3.2-11B-Vision	5.55	0.43	0.87	0.23	76.85	86.06	43.96	75.96
LLaVA 1.6 **Vicuna 7B**	12.63	3.40	4.26	0.38	78.33	81.56	38.47	71.20
LLaVA 1.6 **Vicuna 13B**	28.19	2.72	6.31	1.68	77.97	87.1	37.39	69.16
LLaVA 1.6 **Mistral 7B**	5.88	0.60	1.00	0.24	76.2	55.22	31.61	64.69
Pixtral **12B**	5.21	0.47	0.84	0.31	80.86	92.56	56.94	81.73
Qwen2.5-VL-7B	59.44	2.96	7.77	2.27	74.41	67.47	38.53	60.11
Qwen2.5-VL-72B	7.55	0.57	0.83	0.46	76.41	93.73	**56.93**	80.18

Table 2: Experimental results for baseline VLMs and their V-KPS enhanced counterparts. Best results are in **bold**. The metric for Nutritional Analysis questions is Mean Absolute Error (MAE). Metrics for Regional Popularity and Pathological Association questions are F1 scores, derived from confusion matrix-based classification evaluation. The metric for Basic Information questions is BERTScore F1. The last column "avg" is a composite average score reflecting overall model performance, calculated as: avg = 0.25 * (1 - Normalized MAE for Nutrition) + 0.25 * F1 Regional Popularity + 0.25 * F1 Pathological Association + 0.25 * BERTScore F1 Basic Info. In this formula, the MAE for nutrition analysis has been normalized.

proposed in this paper, including the VLM-guided Knowledge Pruning & Selection mechanism.

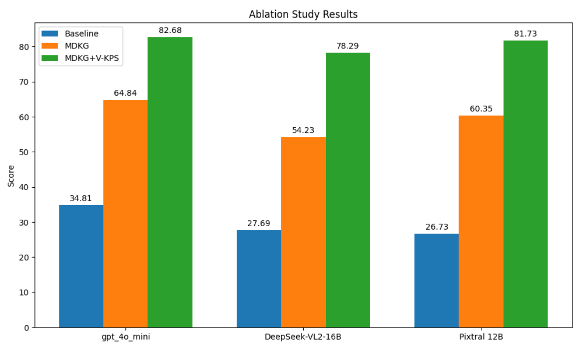

Figure 4: Ablation Study Results. (This figure would typically be a bar chart showing the ‘avg‘ score for Baseline, Simple KG Enhancement, and MDKG+V-KPS for several representative VLMs.)

Ablation studies demonstrate that even simple Knowledge Graph enhancement significantly improves baseline model performance (e.g., boosting the F1 score of `gpt_4o_mini` from 34.81% to 64.84%). However, our proposed Vision-Language Model -guided knowledge pruning and selection framework exhibits even better results, achieving state-of-the-art performance across all evaluated models. This framework excels by precisely filtering key relevant knowledge and reducing interference, thereby enabling models to utilize information more intelligently and accurately. Consequently, it achieves superior performance in complex food visual question answering tasks.

6 Conclusion

Addressing the limitations of Visual Language Models in deep food understanding, this research proposes an enhancement method based on a multimodal dietary knowledge graph. We constructed the *DietKG-VQA* benchmark and a multimodal dietary knowledge graph, and designed a VLM-guided knowledge pruning and selection mechanism. This mechanism integrates visual similarity with knowledge graph queries, enabling it to handle uncommon foods and complex reasoning. Experimental results demonstrate that our approach significantly improves model performance on *DietKG-VQA*, particularly in deep knowledge question answering tasks such as nutritional quantification and pathological association, thereby underscoring the substantial potential of multimodal dietary KGs in empowering intelligent dietary analysis.

Limitations

First, the current Multimodal Dietary Knowledge Graph predominantly covers Chinese and Western recipes, lacking comprehensive coverage of other regional cuisines. The number of food entities

(6,058) and disease entities (1,849) it encompasses, while substantial, still offers room for expansion when compared to the complexity of the real world. Although the visual similarity retrieval mechanism (as described in Section 4.2.1) offers some inferential capability for handling uncommon food items not directly included in the MDKG, the breadth and depth of the underlying knowledge base remain critical.

Second, while the VLM-guided Knowledge Pruning & Selection process is effective, it necessitates multiple calls to the Vision-Language Model. Depending on the depth of knowledge retrieval and the complexity of the query, this typically involves 2-3 VLM calls for filtering and assessment, potentially increasing inference latency and computational overhead. Future work could explore more lightweight pruning models or optimize model interaction prompts to enhance both efficiency and robustness.

Ethical Considerations

This research is dedicated to the responsible advancement of dietary analysis technology. Our benchmark construction utilizes publicly available data, which is expert-reviewed to ensure accuracy and reduce bias. The outputs generated by our model are intended as supplementary information only and cannot substitute for professional medical advice; users are advised to exercise discretion.

We acknowledge the current limitations in the coverage of our knowledge graph and are committed to continuously enhancing its fairness, generalizability, and transparency through open benchmarks, code, and other publicly available resources. Our ultimate commitment is to ensure that technological progress serves human well-being.

References

Ashkan Afshin, Patrick John Sur, Kairsten A Fay, Leslie Cornaby, Giannina Ferrara, Joseph S Salama, Erin C Mullany, Kalkidan Hassen Abate, Cristiana Abbafati, Zegeye Abebe, and 1 others. 2019. Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The lancet*, 393(10184):1958–1972.

Hadi Amiri, Andrew Beam, and Isaac S Kohane. 2019. Learning to estimate nutrition facts from food descriptions. In *AMIA*.

Baidu Baike. 2025. Baidu Baike. https://baike.baidu.com.

Baidu Health. 2025. Baidu Health Dictionary. https://jiankang.baidu.com/widescreen/entitylist.

Boohee. 2025. Boohee Food Database. https://www.boohee.com/food.

Sina Buck, Collin Krauss, Delia Waldenmaier, Christina Liebing, Nina Jendrike, Josef Högel, Boris M Pfeiffer, Cornelia Haug, and Guido Freckmann. 2022. Evaluation of meal carbohydrate counting errors in patients with type 1 diabetes. *Experimental and Clinical Endocrinology & Diabetes*, 130(07):475–483.

Duygu Çelik. 2015. Foodwiki: Ontology-driven mobile safe food consumption system. *The scientific World journal*, 2015(1):475410.

Jingjing Chen, Liangming Pan, Zhipeng Wei, Xiang Wang, Chong-Wah Ngo, and Tat-Seng Chua. 2020. Zero-shot ingredient recognition by multi-relational graph convolutional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10542–10550.

Yu Chen, Ananya Subburathinam, Ching-Hua Chen, and Mohammed J Zaki. 2021. Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. In *Proceedings of the 14th ACM international conference on web Search and data mining*, pages 544–552.

Ivan Contreras, Marti Guso, Aleix Beneyto, and Josep Vehi. 2023. Photo-based carbohydrates counting using pre-trained transformer models. *IFAC-PapersOnLine*, 56(2):11533–11538.

Douguo. 2025. Douguo Recipe Database. https://www.douguo.com.

Ferras Hamad, Isaac Liu, and X Zhang. 2018. Food discovery with uber eats: Building a query understanding engine. *Uber Engineering*.

Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L McGuinness, and Mohammed J Zaki. 2019. Foodkg: a semantics-driven knowledge graph for food recommendation. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18*, pages 146–162. Springer.

Niloofar Hezarjaribi, Sepideh Mazrouee, and Hassan Ghasemzadeh. 2017. Speech2health: a mobile framework for monitoring dietary composition from spoken data. *IEEE journal of biomedical and health informatics*, 22(1):252–264.

Andong Hua, Mehak Preet Dhaliwal, Ryan Burke, Laya Pullela, and Yao Qin. 2024. Nutribench: A dataset for evaluating large language models on nutrition estimation from meal descriptions. *arXiv preprint arXiv:2407.12843*.

Lan Huang, Congcong Yu, Yang Chi, Xiaohui Qi, and Hao Xu. 2019. Towards smart healthcare management based on knowledge graph technology. In *Proceedings of the 2019 8th International Conference on Software and Computer Applications*, pages 330–337.

Matthew Keller, Chi-en Amy Tai, Yuhao Chen, Pengcheng Xi, and Alexander Wong. 2024. Nutritionverse-direct: exploring deep neural networks for multitask nutrition prediction from food images. *arXiv preprint arXiv:2405.07814*.

Marcus Klasson, Cheng Zhang, and Hedvig Kjellström. 2020. Using variational multi-view learning for classification of grocery items. *Patterns*, 1(8).

Simon Knez and Luka Šajn. 2020. Food object recognition using a mobile device: evaluation of currently implemented systems. *Trends in Food Science & Technology*, 99:460–471.

Zhenfeng Lei, Anwar Ul Haq, Adnan Zeb, Md Suzauddola, and Defu Zhang. 2021. Is the suggested food your desired?: Multi-modal recipe recommendation with demand-based knowledge graph. *Expert Systems with Applications*, 186:115708.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yanchao Liang and Jianhua Li. 2017. Computer vision-based food calorie estimation: dataset, method, and experiment. *arXiv preprint arXiv:1705.07632*.

Yao Liu, Hongbin Pu, and Da-Wen Sun. 2021. Efficient extraction of deep image features using convolutional neural network (cnn) for applications in detecting and analysing complex food matrices. *Trends in Food Science & Technology*, 113:193–204.

Ya Lu, Thomai Stathopoulou, Maria F Vasiloglou, Stergios Christodoulidis, Zeno Stanga, and Stavroula Mougiakakou. 2020a. An artificial intelligence-based system to assess nutrient intake for hospitalised patients. *IEEE transactions on multimedia*, 23:1136–1147.

Ya Lu, Thomai Stathopoulou, Maria F Vasiloglou, Lillian F Pinault, Colleen Kiley, Elias K Spanakis, and Stavroula Mougiakakou. 2020b. gofoodtm: an artificial intelligence system for dietary assessment. *Sensors*, 20(15):4283.

Peihua Ma, Yixin Wu, Ning Yu, Yang Zhang, Michael Backes, Qin Wang, and Cheng-I Wei. 2023. Vision-language models boost food composition compilation. *CoRR*.

Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. 2019. A survey on food computing. *Acm Computing Surveys (CSUR)*, 52(5):1–36.

Weiqing Min, Chunlin Liu, Leyi Xu, and Shuqiang Jiang. 2022. Applications of knowledge graphs for food science and industry. *Patterns*, 3(5).

Nutritionix. 2018. Api nutrition. Retrieved March 26, 2018.

Li Qin, Zhigang Hao, and Liang Zhao. 2019. Food safety knowledge graph and question answering system. In *Proceedings of the 2019 7th International Conference on Information Technology: IoT and smart city*, pages 559–564.

Megan E Rollo, Rebecca L Williams, Tracy Burrows, Sharon I Kirkpatrick, Tamara Bucher, and Clare E Collins. 2016. What are they really eating? a review on new approaches to dietary intake assessment and validation. *Current nutrition reports*, 5:307–314.

Sola Shirai, Oshani Seneviratne, Minor Gordon, Ching-Hua Chen, and Deborah L McGuinness. 2020. Semantics-driven ingredient substitution in the foodkg.

Chi-en Amy Tai, Saeejith Nair, Olivia Markham, Matthew Keller, Yifan Wu, Yuhao Chen, and Alexander Wong. 2023. Nutritionverse-real: An open access manually collected 2d food scene dataset for dietary intake estimation. *arXiv preprint arXiv:2401.08598*.

USDA. 2019. Fooddata central. Accessed: 2024-09-30.

Zhunheng Wang, Xiaoyi Liu, Mengting Hu, Rui Ying, Ming Jiang, Jianfeng Wu, Yalan Xie, Hang Gao, and Renhong Cheng. 2024. Ecok: Emotional commonsense knowledge graph for mining emotional gold. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8055–8074.

Wikipedia. 2025. Wikipedia, The Free Encyclopedia. `https://en.wikipedia.org`.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. 2023. Foodlmm: A versatile food assistant using large multi-modal model. *arXiv preprint arXiv:2312.14991*.

Raza Yunus, Omar Arif, Hammad Afzal, Muhammad Faisal Amjad, Haider Abbas, Hira Noor Bokhari, Syeda Tazeen Haider, Nauman Zafar, and Raheel Nawaz. 2018. A framework to estimate the nutritional value of food in real time using deep learning techniques. *IEEE Access*, 7:2643–2652.

Zhiyun Zhao, Mian Li, Chao Li, Tiange Wang, Yu Xu, Zhizheng Zhan, Weishan Dong, Zhiyong Shen, Min Xu, Jieli Lu, and 1 others. 2020. Dietary preferences and diabetic risk in china: A large-scale nationwide internet data-based study.

A More DietKG-VQA Q&A Examples

What diseases might the food in the image
be beneficial for, and what diseases might it be harmful for?

[Baseline]This dish looks like a hot soup containing noodles and vegetables, possibly some kind of Asian-style dish. Here are some potential benefits and drawbacks:
Benefits:
Rich in nutrients: If the soup contains various vegetables and protein (such as meat or tofu), it can provide abundant vitamins, minerals, and fiber, helping to boost immunity and improve digestion.
Warms the body: Hot soup can help soothe the throat and alleviate cold or flu symptoms.
Hydration: Soups help replenish fluids and maintain the body's hydration balance.
Low in calories: If low-fat ingredients are used, it may be suitable for individuals aiming for weight loss or weight management.
Drawbacks:
High sodium content: If a large amount of soy sauce or seasonings is used in the soup, it may lead to excessive sodium intake, which is detrimental for individuals with hypertension.
Allergic reactions: Certain ingredients (like seafood, nuts, etc.) may trigger allergic reactions, requiring careful selection.
Excessive carbohydrates: If the proportion of noodles is high, it may be unfavorable for individuals with diabetes, who need to control their intake.
Fat content: If the soup contains a lot of oil/fat, it may negatively impact cardiovascular health.
In summary, the benefits and drawbacks of food often depend on the specific ingredients and an individual's health condition. It is advisable to consume it in moderation according to one's own circumstances.

[MDKG+V-KPS]Suanlafen (Hot and Sour Noodles) may be beneficial for the following conditions:
Motor Neuron Disease (MND): The high-calorie and high-protein characteristics of Suanlafen may, to some extent, align with the dietary recommendations for this disease.
Disease description URL: [Motor Neuron Disease](https://www.baidu.com/bh/dict/ydxx_8178830962759247390?from=dicta&sf_ref=med_pc&sf_ch=ch_med_pc)

Suanlafen may be detrimental for the following conditions:
Acute Eczema: Patients with acute eczema should avoid easily sensitizing and irritating foods. The spicy nature of Suanlafen may not be suitable for such patients.
Disease description URL: [Acute Eczema](https://www.baidu.com/bh/dict/ydxx_11283957043424045442?from=dicta&sf_ref=med_pc&sf_ch=ch_med_pc)
Renal Cell Carcinoma: Patients with renal cell carcinoma should limit salt intake. The preparation of Suanlafen may involve adding a large amount of salt, making it unsuitable for these patients.
Disease description URL: [Renal Cell Carcinoma](https://www.baidu.com/bh/dict/ydxx_8141964621220728723?from=dicta&sf_ref=med_pc&sf_ch=ch_med_pc)
Polycystic Kidney Disease (PKD): Patients with PKD should avoid increased blood pressure, and those with renal insufficiency should follow a low-salt diet. Suanlafen may not be suitable for such patients.
Disease description URL: [Polycystic Kidney Disease](https://www.baidu.com/bh/dict/ydxx_10767364304051955308?from=dicta&sf_ref=med_pc&sf_ch=ch_med_pc)
Hematuria: Patients with hematuria should abstain from all spicy and irritating foods. The spicy nature of Suanlafen may not be suitable for these patients.
Disease description URL: [Hematuria](https://www.baidu.com/bh/dict/ydzz_10913604809621716394?from=dicta&sf_ref=med_pc&sf_ch=ch_med_pc)
Sprouted Potato Poisoning: Patients who have experienced sprouted potato poisoning (or to avoid it) should avoid consuming sprouted or green/purplish potatoes. Care should be taken regarding the quality of potatoes potentially used in Suanlafen, ensuring they are not sprouted or discolored.
Disease description URL:[Sprouted Potato Poisoning](https://www.baidu.com/bh/dict/ydxx_11137272167793392119?from=dicta&sf_ref=med_pc&sf_ch=ch_med_pc)

Figure 5: GPT-4o-mini answering pathological association questions from DietKG-VQA using different prompting strategies.

In which regions is the food in the image most popular?

[Baseline]The food in the picture looks like mooncakes or similar Chinese pastries. These are highly popular in China and other East Asian regions like Hong Kong, Macau, and Taiwan, particularly during the Mid-Autumn Festival. Mooncakes are also a cherished traditional food in many Chinese communities across Southeast Asia.

[MDKG+V-KPS]Lao Po Bing (Wife Cake, also known as Sweetheart Cake) is a Cantonese-style pastry originating from Guangdong Province, China, and is therefore most popular in Guangdong Province and its surrounding regions. Additionally, with the spread of Cantonese culinary culture, Lao Po Bing is also very popular in Hong Kong, Macau, and overseas Chinese communities.
If you need more detailed information, you can refer to the following website:
Description and ingredient information for Lao Po Bing: [Baidu Baike]
(https://baike.baidu.com/item/%E8%80%81%E5%A9%86%E9%A5%BC?fromModule=lemma_search-box)

Figure 6: GPT-4o-mini answering regional popularity questions from DietKG-VQA using different prompting strategies.

What food is in the image?
Please provide some information about this food.

[Baseline]The image shows a bowl of Chinese-style soup, likely fish-based, garnished with red dates (jujubes) and some vegetables or medicinal herbs. Such soups are very common in traditional Chinese cuisine and are often believed to have nourishing and health-preserving benefits.
Food Information:
Main Ingredients:
Fish: Fresh fish is typically chosen, rich in protein and Omega-3 fatty acids.
Red Dates (Jujubes): Considered in Traditional Chinese Medicine (TCM) as an ingredient for enriching the blood and enhancing beauty, rich in vitamins and minerals.
Other Ingredients: May include ginger, scallions, carrots, etc., to enhance the soup's flavor and nutrition.
Preparation Method:
Clean the fish thoroughly and cut it into pieces.
Add water to a pot, put in the fish pieces, red dates, and other ingredients. Bring to a boil, then reduce to a low heat and simmer slowly until the fish is cooked through and the broth becomes rich.
Nutritional Value:
This type of soup is generally low in fat and high in protein, suitable for various groups of people, especially those in need of nourishment.
The addition of red dates not only adds sweetness but also helps to boost the immune system and improve sleep.
Cultural Background:
In China, soup is an indispensable part of the dining table, especially during family gatherings or festive occasions, where it is often seen as a symbol of warmth and reunion.
This soup is not only delicious but also beneficial for health, making it a very popular traditional delicacy.

[MDKG+V-KPS]The image shows Cordyceps Flower and Silkie Chicken Soup, a delicacy primarily made with Cordyceps flower and Silkie chicken.
Cordyceps Flower and Silkie Chicken Soup has the effects of benefiting the liver and kidneys, replenishing vital essence, stopping bleeding, and resolving phlegm. It is suitable for adjunctive treatment of symptoms such as dizziness and tinnitus, forgetfulness and insomnia, soreness and weakness of the lower back and knees, impotence and premature ejaculation, chronic cough, and asthenia-induced wheezing.
Every 100 grams of Cordyceps Flower and Silkie Chicken Soup contains 89.0 kcal of calories, 13.2 grams of protein, 1.36 grams of fat, and 6.18 grams of carbohydrates.
The steps to make Cordyceps Flower and Silkie Chicken Soup are as follows:
1.Prepare the ingredients.
2.Chop the Silkie chicken into small pieces, blanch to remove blood, then rinse clean with water. Wash the red dates, goji berries, and Cordyceps flowers. Slice the ginger.
3.Add the Silkie chicken, ginger slices, red dates, and an appropriate amount of water to a clay pot (or casserole/stew pot).
4.Start the "Soup" function (e.g., on an electric pressure cooker or rice cooker); the default time is 2 hours.
5.After simmering for 1 hour, add the Cordyceps flowers and continue to stew.
6.Add the goji berries during the last 15 minutes.
7.Before serving, season with salt to taste.
Source URLs for related information:
Baidu Baike: [Cordyceps Flower and Silkie Chicken Soup]
(https://baike.baidu.com/item/%E8%99%AB%E8%8D%89%E8%8A%B1%E4%B9%8C%E9%B8%A1%E6%B1%A4?fromModule=lemma_search-box)
Recipe Source: [Douguo Meishi](https://www.douguo.com/cookbook/1728430.html)
Nutrition Information Query: [Boohee.com](https://www.boohee.com/food/search?
keyword=%E8%99%AB%E8%8D%89%E8%8A%B1%E4%B9%8C%E9%B8%A1%E6%B1%A4)

Figure 7: GPT-4o-mini answering basic information questions from DietKG-VQA using different prompting strategies.