myFoodQA: A Multimodal Dataset for Evaluating Cultural and Visual Reasoning in Myanmar Gastronomy

Shin Thant Phyo
University of Information Technology
Yangon, Myanmar
stphyo@uit.edu.mm

Pyae Linn
University of Information Technology
Yangon, Myanmar
pyaelinn@uit.edu.mm

Lynn Myat Bhone
University of Computer Studies
Yangon, Myanmar
lynnmyatbhone@ucsy.edu.mm

Thet Hmue Khin
University of Information Technology
Yangon, Myanmar
thethmuekhin@uit.edu.mm

Eaint Kay Khaing Kyaw
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
eaintkay.kyaw@kmitl.ac.th

Ye Kyaw Thu
Language and Semantic Technology Research Team, NECTEC
Bangkok, Thailand
yekyaw.thu@nectec.or.th

***Abstract*—This paper introduces myFoodQA (Myanmar (Burmese) Food Question Answering), the first multimodal benchmark designed to evaluate AI models on Myanmar's rich gastronomic culture. A core contribution of this work is the thorough construction of the benchmark itself, which involved curating a diverse set of food images for 20 distinct dishes and, crucially, generating a novel corpus of 2,485 question-answer pairs. The benchmark features tasks for single-image, multi-image, and text-only reasoning, specifically designed to evaluate model understanding of ingredient recognition, cultural context, preparation methods, and comparative logic. To ensure authenticity, data was sourced and collected from personal photography and web-crawling, with all annotations, prompts and questions validated by native Burmese speakers. Leading vision-language models were evaluated in a zero-shot condition and revealed a large performance disparity. While models perform well on text-based tasks, the performance significantly deficit on image-based reasoning, which needs specific image understanding and extensive cultural knowledge. These findings reveal the limitations of current Large Language and Vision Models (LLMs and VLMs) regarding the Myanmar gastronomic domain. Consequently, this work establishes myFoodQA as a foundational resource for advancing multimodal AI in culturally relevant and low-resource settings.**

***Index Terms*—Visual Question Answering, Myanmar Food, Multimodal Benchmark, Cultural Reasoning, Low-Resource AI, Vision-Language Models**

I. Introduction

The intersection of food, culture, and artificial intelligence presents a formidable challenge for modern AI, demanding a shift from simple object recognition to cultural comprehension. While numerous food-centric datasets exist, they primarily focus on semantic tasks like dish classification but often lack the rich contextual and ideological knowledge required to understand the process of a dish's religious or traditional preparation. This gap is particularly critical for low-resource languages like Burmese, where culturally-grounded multimodal benchmarks are virtually non-existent.

This paper addresses this deficiency by introducing myFoodQA, the first comprehensive multimodal benchmark for Myanmar's traditional food culture. The primary contribution of this research is the benchmark itself, including not only a curated dataset of images and cultural facts about Myanmar gastronomy, but also a novel, structured set of question-answer pairs designed to test deep understanding. Inspired by prior work like FoodieQA, we developed distinct tasks for single-image, multi-image, and text-only reasoning to evaluate the model knowledge about Myanmar food. High cultural fidelity was ensured through a meticulous creation process where native Burmese speakers curated all data, questions, and annotations.

Our second contribution is a rigorous zero-shot evaluation of state-of-the-art models on this benchmark. The results reveal a critical limitation: modern VLMs and LLMs handle general knowledge capably but fail on tasks requiring deep cultural reasoning tied to visual cues. myFoodQA thus serves as a vital tool for diagnosing these weaknesses and paving the way for more inclusive and culturally-intelligent AI.

II. Related Works

Research in multimodal culinary AI has evolved from large-scale classification to nuanced cultural reasoning.

979-8-3315-0217-1/25/$31.00 © 2025 IEEE

Early, large-scale benchmarks linked images and text, such as Recipe1M+ [1], which provided over one million image-recipe pairs focused on Western cuisines and semantic matching. Other foundational work, like ETH Food-101 [2], established robust classification standards.

Subsequent work in Visual Question Answering (VQA), such as VQA-Food [3], advanced beyond static recognition to include procedural questions (e.g., ingredients, methods). However, these datasets often lack the deeper cultural context required to probe implicit knowledge, such as a dish's ritualistic or traditional significance.

More recently, benchmarks have begun to address this cultural gap. FoodieQA [4], for instance, introduced a benchmark for Chinese cuisine with culturally-grounded questions that revealed significant limitations in state-of-the-art models.

Our work, myFoodQA, builds directly on this progression. We extend the cultural reasoning framework of FoodieQA to the context of Myanmar, a domain that is distinct for its combination of challenges: 1) a low-resource language with limited pre-training corpora, 2) high cultural diversity from indigenous and regional influences, and 3) significant ethnolinguistic variation in its culinary traditions.

III. Data Preparation and Benchmark Construction

The construction of myFoodQA was a multi-stage process centered on authenticity and cultural relevance. We began by selecting 20 popular traditional Myanmar foods, categorized into five groups: soups, snacks, beverages, meals, and salads.

A. Data Collection and Curation

1) Image Collection: The visual dataset was constructed from two main sources: crowd-sourced personal photographs and publicly available internet images. Personal images were collected via a Google Form and through direct field photography by team members to ensure authentic representation. From an initial 988 submissions, 713 images passed quality filtering and were combined with 1,750 internet-sourced images. The final dataset consists of 2,463 images with a 70/30 online–personal split, as summarized in Table I.

TABLE I
Distribution of online and personal images after filtering.

Images	Count	Percentage
Online	1750	71.1%
Personal	713	28.9%
Total	**2463**	**100%**

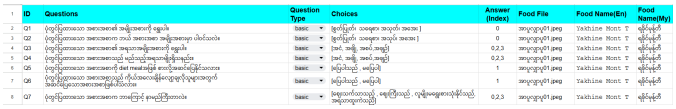
2) Cultural Information Gathering: To support reasoning beyond visual recognition, detailed cultural metadata were compiled for each dish using Myanmar cookbooks, historical texts, and native researcher expertise. For each food item, information was recorded on local and official names, regional origins, preparation methods, common ingredients, and cultural or festive relevance. This textual layer enables the benchmark to assess reasoning grounded in cultural context, such as identifying dishes associated with specific traditions or ceremonies.

B. Question Generation and Annotation

A structured question–answer benchmark was developed to evaluate multimodal reasoning in relation to Myanmar cuisine, focusing on understanding ingredients, preparation, and cultural significance. A template-based pipeline was used to ensure systematic coverage, linguistic diversity, and consistency across three major task types.

- *Single-Image VQA:* Contains 1,100 question–image pairs derived from 25 validated templates expanded across 600 images.
- *Text-Only QA:* Comprises 1,100 questions converted from the Single-Image VQA task by removing image references and explicitly adding dish names, isolating text-based reasoning.
- *Multi-Image VQA:* Includes 300 questions, each paired with four images (one correct target and three distractors) to assess fine-grained cross-modal understanding.

Table II illustrates examples of Burmese–English question formats for the VQA and text-only tasks, while Figure 1 presents representative question samples.



(a) A sample question from the Single-Image VQA task.

(b) A sample question from the Multi-Image VQA task.

Fig. 1. Question formats for the two VQA tasks in the myFoodQA dataset.

This structured generation process involved creating all question–answer pairs in a spreadsheet (Excel) to streamline authoring and management. The finalized collection was then converted into JSON format to enable efficient data processing and automated evaluation. In total, the process yielded a benchmark of 2,485 question–answer pairs designed to comprehensively assess multimodal reasoning abilities, encompassing visual, contextual, and cultural understanding of Myanmar's diverse food heritage.

TABLE II
Translation of question formats in Burmese and English.

Single-Image VQA

Original Burmese	English Translation
ပုံတွင်ပြထားသော အစားအစာ၏ အမျိုးအစားကို ရွေးပါ။	Choose the type of food shown in the picture.
Category: အခြေခံ	**Category:** basic
Options: [စွတ်ပြုတ်၊ သရေစာ၊ အသုတ်၊ အအေး]	**Options:** [Soup, Snack, Salad, Cold Drink]
Correct Answer Index: 0	**Correct Answer Index:** 0
Image File Name: အာပူလျှာပူ01.jpeg	**Image File Name:** အာပူလျှာပူ01.jpeg
Food Name (Phonetic): Yakhine Mont T	**Food Name (Phonetic):** Yakhine Mont T
Food Name (Burmese): ရခိုင်မုန့်တီ	**Food Name (English):** Rakhine Mont Ti

Text-Only QA

Original Burmese	English Translation
Question: ရခိုင်မုန့်တီက ဘယ် အစားအစာ အမျိုးအစားမှာ ပါဝင်သလဲ။	**Question:** Which food category does Rakhine Mont Ti belong to?
Choices: [စွတ်ပြုတ်၊ သရေစာ၊ အသုပ်၊ အအေး]	**Choices:** [Soup, Snack, Salad, Cold Drink]
LLM's Answer: 0,1	**LLM's Answer:** 0,1
Correct Answer: 0	**Correct Answer:** 0
Is Correct: false	**Is Correct:** false
Partially Correct: true	**Partially Correct:** true
Score: 0.5	**Score:** 0.5

C. Data Cleaning and Validation

Each data instance was subjected to a dual independent review for quality assurance. Reviewers labeled entries as `good`, `bad`, or `requires revision`. Low-quality items were removed and ambiguous cases were jointly refined to ensure factual accuracy and linguistic clarity. The verified data set, including all reference and metadata images, was serialized in JSON format for reproducibility and standardized evaluation.

D. Prompt Templates

Three prompt templates were designed for the food evaluation task to test different aspects of model performance:

- *Template 0 (Detailed English Instruction):* Provides precise guidance in English for both single and multi-image tasks, outlining expected responses and optional hints.

TABLE III
Summary of myFoodQA Question Types

Question Type	Number of Questions
Single Image	1098
Text Only	1100
Multi-Image	300
Total	**2498**

TABLE IV
Distribution of All Question Types per Food Item

Food Type	Single Image	Multi Image
ကတ်ကြေးကိုက် (Kat Kyay Kite)	55	15
မုန့်လင်မယား (Mont Lin Ma Yarr)	54	15
နန်းကြီးသုပ် (Nan G Thote)	55	15
အုန်းနို့ခေါက်ဆွဲ (Coconut Noodle)	55	15
ခေါက်ဆွဲသုပ် (Noodle Salad)	55	15
ကောက်ညှင်းပေါင်း (Kout Nyin Paung)	54	15
ထမနဲ (Hta Ma Nae)	55	15
ရွှေရင်အေး (Shwe Yin Aye)	55	15
ဝက်သားတုတ်ထိုး (Wat Thar Tote Htoe)	55	15
မုန့်ဖက်ထုပ် (Mont Phat Htote)	55	15
မုန့်လက်ဆောင်း (Mont Let Saung)	55	15
သာကူ (Thar Ku)	55	15
ရခိုင်မုန့်တီ (R Pu Shar Pu)	55	15
ထမင်းပေါင်း (Htamin Paung)	55	15
ကြေးအိုးဆီချက် (Kyay Ohh See Chat)	55	15
တိုဟူးနွေး (Tofu Nwe)	54	15
ကြာဇံချက် (Kyar San Chat)	55	15
မုန့်ဟင်းခါး (Mohinga)	55	15
လက်ဖက်သုပ် (Lat Phat Thote)	56	15
ရှမ်းခေါက်ဆွဲ (Shan Noodle)	55	15
Total	**1098**	**300**

- *Template 1 (Myanmar Language Instruction):* Presents equivalent instructions in the Myanmar language, ensuring accessibility for native users.
- *Template 2 (Step-by-Step Analytical Approach):* Guides reasoning in sequential steps—identification, analysis, and decision—to evaluate structured thinking.

Together, these templates establish a multilingual and reasoning-oriented framework suitable for evaluating both simple recognition and advanced analytical capabilities in vision–language models.

Prompt 0 (Multiple Images)

Please carefully examine ALL the provided photos and identify the different foods shown in each image. Consider all images together to answer the question below.Choose the best answer from options {choice_range}.

Question: {question}
Choices: {choice_text}
Answer:

Prompt 0 (Single Image)

Please carefully look at the provided photo and identify the food shown. Based on this, answer the question below. Note:There may be multiple correct answers. Select all that apply. Choose from options {choice_range}. Question: {question} Choices: {choice_text} Answer:

Prompt 1 (Multiple Images)

ပုံများအားလုံးကိုသေချာစွာကြည့်ပြီးမေးခွန်းကိုဖြေပါ။ ရွေးချယ်မှုများထဲမှ အကောင်းဆုံးအဖြေကိုရွေးပါ။ Question: {question} Choices: {choice_text} Answer:

Prompt 1 (Single Image)

ပုံကိုကြည့်ပြီးမေးခွန်းကိုဖြေပါ။
မှတ်ချက်–အဖြေများစွာမှန်ကန်နိုင်ပါတယ်။
သင့်လျော်သမျှရွေးပါ။ ရွေးချယ်မှုများထဲမှရွေးပါ။
Question: {question} Choices: {choice_text} Answer:

Prompt 2 (Multiple Images)

Step 1: Look at each image and identify what food is shown. Step 2: Consider the relationship between all foods shown in the images. Step 3: Based on your analysis, answer the question. Step 4: Choose the best answer from options {choice_range}. Question: {question} Available Options: {choice_text} Your Answer:

Prompt 2 (Single Image)

Step 1: Carefully examine the image and identify the food. Step 2: Consider the food's characteristics, ingredients, or preparation method. Step 3: Based on your analysis, answer the question. Note: Multiple answers may be correct. Select all that apply. Step 4: Choose from options {choice_range}. Question: {question} Available Options: {choice_text} Your Answer:

IV. Evaluation Results

In this study, we evaluate the performance of Large Language Models (LLMs) and Vision-Language Models (VLMs) in understanding Myanmar food through both visual and textual inputs. The experiments cover single-image and multi-image evaluations, a comparison between the two settings, text-only model assessments, and cross-model analysis between LLMs and VLMs.

A. Comparison of Single vs. Multi-Image

We compared model performance between single-image and multi-image settings to assess how additional visual context affects recognition. Gemma3:12b showed a minor gain from 43.38% to 44.89%, while Pixtral-12b-2409 dropped from 53.37% to 37.58%. Qwen2.5VL:7b nearly doubled its accuracy from 30.74% to 65.30%, indicating strong multi-image reasoning. Conversely, Gemini 1.5 Flash fell from 64.27% to 36.76%. These findings reveal that multi-image inputs benefit some models but hinder others, underscoring the importance of robust visual aggregation mechanisms in VLMs.

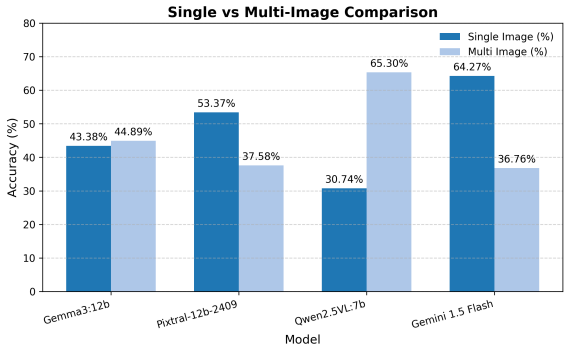


Fig. 2. Model accuracy on the Single-Image VQA task compared to the Multi-Image VQA task.

B. Text-Only Models

Text-only LLMs were evaluated to determine their knowledge of Myanmar food without visual context. LLaMA3.2:latest (3B) achieved 37.77%, and Mistral:latest (7B) reached 43.04%. Although performance was lower than that of VLMs, these results suggest partial retention of cultural and contextual knowledge in text-based settings. Figure 3 compares the two models.

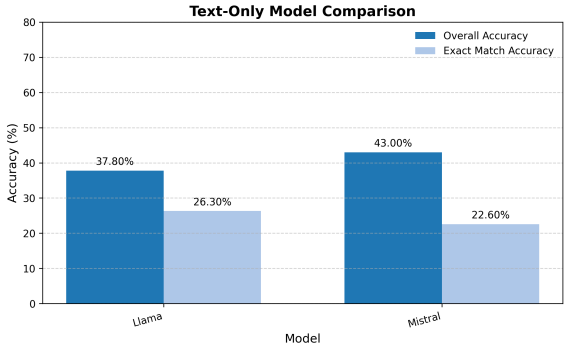


Fig. 3. LLaMA vs Mistral Text-Only Model Comparison

C. Comparison of Text-Only and Vision-Language Models

We compared text-only LLMs with Vision-Language Models (VLMs) to evaluate the impact of visual grounding. On average, VLMs outperform text-only models (47.9% vs. 40.4%), demonstrating the added value of visual features. However, performance varies notably across VLMs, with Gemini 1.5 Flash achieving 64.27% and Qwen2.5VL:7b only 30.74%, overlapping with text-only results such as Mistral:latest at 43.04%. These findings highlight that multimodal effectiveness depends on model architecture and training rather than the mere inclusion of images.

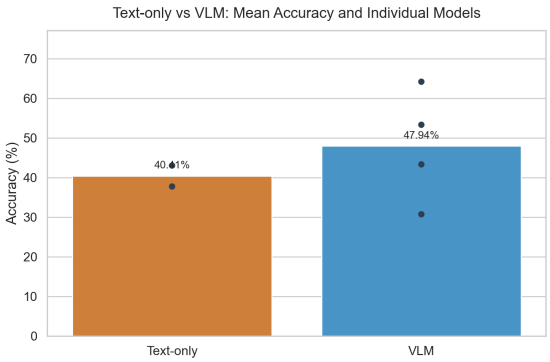


Fig. 4. Text-Only vs Vision-Language Models Comparison

V. Experimental Results and Discussion

A. Experimental Setting

The myFoodQA benchmark is introduced for multimodal evaluation of Myanmar food recognition and question answering. A range of state-of-the-art Vision-Language Models (VLMs)—Gemini 1.5 Flash [5], Pixtral-12b-2409 [6], Qwen2.5VL:7b [7], and Gemma3:12b—and text-only models—Mistral:latest (7B) [9] and LLaMA3.2:latest (3B) [8]—were benchmarked. Experiments were conducted under three primary configurations: single-image, multi-image, and text-only question answering. Results indicate that VLMs generally outperform text-only models, emphasizing the significance of visual cues, while all models exhibit limitations in culturally specific reasoning, suggesting the need for more diverse pre-training data.

B. Results

1) Single-Image Visual Question Answering (VQA): This setting assesses a model's ability to identify a dish and answer questions based on a single image.

2) Multi-Image Visual Question Answering (VQA): This setting requires comparative reasoning between two or more images, often involving subtle differences between similar dishes.

TABLE V
Performance of VLMs on the Single-Image VQA Task.

Model	Accuracy
Gemini 1.5 Flash	64.27
Pixtral-12b-2409	53.37
Gemma3:12b	43.38
Qwen2.5VL:7b	30.74

TABLE VI
Performance of VLMs on the Multi-Image VQA Task.

Model	Accuracy
Qwen2.5VL:7b	65.30
Gemma3:12b	44.89
Pixtral-12b-2409	37.58
Gemini 1.5 Flash	36.76

3) Text-Only Question Answering (QA): This baseline setting evaluates the models' internal, pre-trained knowledge base about Myanmar food without relying on visual input.

TABLE VII
Performance of Text-Only Models on the QA Task.

Model	Accuracy
Mistral:latest (7b)	43.04
Llama3.2:latest (3b)	37.77

4) Impact of Providing Dish Name Hints: An ablation study was conducted to analyze the effect of textual hints provided alongside images.

TABLE VIII
Effect of Dish Name Hints on Single-Image VQA Performance.

Model	Without Hint	With Hint	Improvement
Pixtral-12b-2409	53.37	62.00	+8.63
Qwen2.5VL:7b	30.74	37.87	+7.13
Gemma3:12b	43.38	46.78	+3.40
Gemini 1.5 Flash	64.27	66.08	+1.81

C. Discussion and Analysis

The experimental results provide several insights into the strengths and limitations of current multimodal and text-only models for the Myanmar Food Recognition task.

1) VLMs vs. Text-Only Models: VLMs outperform text-only models across all configurations. Gemini 1.5 Flash achieved the highest accuracy (64.27%) in single-image VQA, surpassing the best-performing text-only model, Mistral:latest (43.04%). This demonstrates the advantage of visual grounding for fine-grained cultural recognition, where textual corpora alone remain insufficient.

2) Model Specialization Across Tasks: Gemini 1.5 Flash exhibits superior single-image recognition, reflecting strong visual–semantic alignment. In contrast, Qwen2.5VL:7b performs best in the multi-image setting (65.30%), improving by approximately 35% over its single-image accuracy. The results suggest that model architectures influence strengths in either direct recognition or comparative reasoning.

3) Impact of Textual Hints: Providing dish name hints enhances accuracy across all models, particularly in weaker systems such as Pixtral-12b-2409 (+8.63%) and Qwen2.5VL:7b (+7.13%). This pattern indicates that textual cues compensate for limitations in visual grounding. In contrast, Gemini 1.5 Flash shows minimal improvement (+1.81%), confirming its robustness in visual understanding.

4) Cultural Reasoning and Knowledge Gaps: Although leading VLMs achieved strong results, top accuracies remained below 70%. A qualitative error analysis revealed two main error types: (1) visual misidentification, where models confused similar dishes such as မုန့်လင်မယား (Mont Lin Ma Yarr) and မုန့်ဖက်ထုပ် (Mont Phat Htote); and (2) cultural knowledge gaps, where models recognized the dish but failed to answer questions about its cultural significance or preparation, such as the association of ထမနဲ (Hta Ma Nae) with specific festivals. Gemini 1.5 Flash frequently exhibited the latter error type, while Pixtral-12b-2409 demonstrated slightly better, though still limited, cultural reasoning. These results emphasize persistent challenges in fine-grained cultural understanding and the importance of developing region-specific multimodal datasets.

a) Ethical Considerations.: All images and cultural information in myFoodQA were collected with consent and validated by native Burmese speakers to ensure respectful representation. Care was taken to avoid stereotyping or misrepresentation of dishes from diverse Myanmar communities. Future expansions will continue to prioritize transparency, culturally sensitive annotation practices, and community engagement to minimize bias.

VI. Conclusion and Future Work

This study presented myFoodQA, the first Visual Question Answering (VQA) benchmark centered on Myanmar cuisine, developed to evaluate multimodal reasoning that combines visual and cultural understanding across 20 representative dishes. Unlike general-purpose benchmarks, myFoodQA emphasizes culturally contextualized question–answer pairs that require reasoning beyond surface-level image recognition.

Experimental findings indicate that existing vision–language models (VLMs) demonstrate varying levels of competence in interpreting visual cues but encounter challenges in addressing culturally nuanced reasoning, revealing limitations in multimodal understanding of localized knowledge domains.

The release of myFoodQA is intended to facilitate further research on culturally aware multimodal AI and to provide a transparent, interpretable benchmark for assessing future models. To promote reproducibility and continued exploration, the complete dataset—including images, question–answer pairs, and evaluation scripts—is publicly available on Zenodo at: https://doi.org/10.5281/zenodo.17531977.

For future work, we plan to:

- *Dataset Expansion:* Broaden the scope of myFoodQA by including more regional cuisines, subcategories, and preparation variants to enhance cultural and visual diversity.
- *Multilingual Extension:* Extend the benchmark to support both Myanmar and English question–answer pairs, enabling evaluation of cross-lingual transfer and multilingual reasoning. Human assessment will be used to verify the fluency and accuracy of model responses to culturally nuanced questions.
- *Model Fine-Tuning:* Beyond zero-shot evaluation, future work will involve fine-tuning high-performing vision–language models (e.g., Gemini, Qwen-VL, Pixtral) on myFoodQA and related cultural datasets. This will assess domain adaptation, transfer learning, and bias mitigation using metrics such as cultural reasoning accuracy, visual–semantic alignment, and interpretability.

References

[1] J. Marin, A. Das, F. Ofli, N. Hynes, A. Salvador, M. Rohrbach, and I. Aytar, "Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 187–203, Jan. 2021.

[2] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101: Mining Discriminative Components with Random Forests," in *Proc. European Conf. on Computer Vision (ECCV)*, Zurich, Switzerland, 2014, pp. 446–461.

[3] Z. Peng, S. Wu, J. Xu, Z. Zheng, and G. Zhao, "VQA-Food: A Super-Fine-Grained VQA Dataset for Food," in *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV) Workshops*, Paris, France, 2023, pp. 1957–1965.

[4] A. Jacovi, D. Chen, P. Yin, L. Hou, and H. Ji, "FoodieQA: A Multimodal Dataset for Chinese Food Culture Question Answering," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023, pp. 14757–14771.

[5] Gemini Team (Google), "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context," *arXiv preprint arXiv:2403.05530*, 2024.

[6] Mistral AI, "Pixtral 12B," *arXiv preprint arXiv:2410.07073*, 2024.

[7] Alibaba Group, "Qwen2.5-VL Technical Report," *arXiv preprint arXiv:2502.13923*, 2025.

[8] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.

[9] A. Q. Jiang *et al.*, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.