FoodieQA: A Multimodal Dataset for Fine-Grained Understanding of Chinese Food Culture

Anonymous ACL submission

Abstract

Food is a rich and varied dimension of cultural heritage, crucial to both individuals and social groups. To bridge the gap in the literature on the often-overlooked regional diversity in this domain, we introduce FoodieQA, a manually curated, fine-grained image-text dataset capturing the intricate features of food cultures across various regions in China. We evaluate vision–language Models (VLMs) and large language models (LLMs) on newly collected, unseen food images and corresponding questions. FoodieQA comprises three multiple-choice question-answering tasks where models need to answer questions based on multiple images, a single image, and text-only descriptions, respectively. While LLMs excel at text-based question answering, surpassing human accuracy, the open-sourced VLMs still fall short by 41% on multi-image and 21% on single-image VQA tasks, although closed-weights models perform closer to human levels (within 10%). Our findings highlight that understanding food and its cultural implications remains a challenging and under-explored direction.

1 Introduction

One of the most popular dishes in China is *hotpot*, which comes in many varieties, as shown in Figure 1: Beijing is renowned for its mutton hotpot served with a traditional copper pot (铜锅涮羊肉 tóng guō shuàn yáng ròu). Guangdong province is home to a famous porridge-based hotpot (粥底火锅 zhōu dǐ huǒ guō), while its coastal region of Chaoshan is known for beef hotpot (潮汕牛肉火锅 cháo shàn niú ròu huǒ guō). The hotpot varieties from Sichuan and Chongqing are celebrated for their flavorful broths, with chili peppers and Sichuan peppercorns that create a unique numbing-spicy sensation. The variation among regional cultures within a country highlights the challenges that language models face in understanding cultural knowledge and context-specific information in the food domain.

Figure 1: An example of regional food differences in referring to *hotpot* in China. The depicted soups and dishware visually reflect the ingredients, flavors, and traditions of these regions: Beijing in the north, Sichuan in the southwest, and Guangdong in the south coast.

Existing datasets and models that focus on food and culinary practices primarily concentrate on tasks such as food recognition, recipe generation, food knowledge probing or recipe-related question answering (Chen et al., 2017; Cao et al., 2024a; Zhou et al., 2024; Yagcioglu et al., 2018). However, they often take a coarse view, conflating country, culture and language. Important regional cultural differences remain under-studied (Palta and Rudinger, 2023).

We introduce **FoodieQA**, a manually curated set of multimodal test questions designed to probe fine-grained cultural awareness with a focus on the food domain. Our dataset targets two under-explored directions: regional cultural diversity within a country and challenging fine-grained vision-language understanding in the culinary domain.

To build a regionally diverse dataset, we gather dishes and images selected by native Chinese speakers from various regions, covering 14 dis-

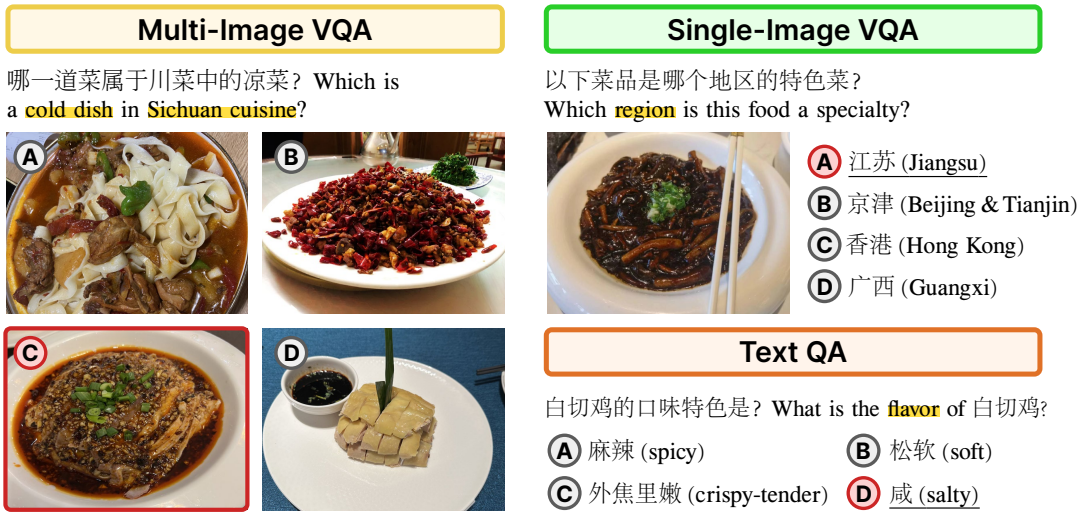


Figure 2: The tasks in FoodieQA evaluate food culture understanding from three perspectives. *Multi-image VQA* requires the ability to compare multiple images, similar to how humans browse a restaurant menu. *Single-image VQA* assesses whether models can use visual information to better understand food culture. *Text-based* questions probe model performance without multimodal data. Fine-grained attributes that the questions focus on are highlighted.

tinct cuisine types across China. To ensure the images used for benchmarking are fresh and have no chance of leaking into the pretraining data of VLMs, we collect images uploaded by local people, which are not publicly available online. We then define multiple attributes associated with the dishes and have native Chinese annotators create multiple-choice questions based on their expertise. Our dataset includes both text-based question answering and vision-based question answering tasks, as illustrated in Figure 2.

We benchmark a series of state-of-the-art models, including seven LLMs and eight VLMs, on the Foodie dataset using zero-shot evaluation. By comparing their performance to human accuracy, we highlight the gap between open-weight and closed-weight models and demonstrate their limitations in understanding Chinese regional food culture. Additionally, we compare the performance of bilingual models trained on both Chinese and English datasets to English-focused models, revealing biases in their understanding of region-specific food culture and the language of the questions. Finally, our analysis shows that visual information improves the performance of VLMs compared to text-only inputs, although some models struggle with identifying dishes from images.

2 Related Work

Multilingual Multimodal Datasets Multimodal systems are typically evaluated on English due to the widespread availability of English-language datasets. However, there are some examples of research on training and evaluating models beyond English for image captioning (Elliott et al., 2016), image–sentence retrieval (Srinivasan et al., 2021), visual reasoning (Liu et al., 2021), and question-answering (Pfeiffer et al., 2022). This paper focuses on Chinese visual question answering, with fine-grained attributes in the food domain.

Food Datasets In recent years, most food datasets have been designed for food image classification (Chen et al., 2017), food captioning (Ma et al., 2023), and recipe-focused generation and question answering (Yagcioglu et al., 2018; Min et al., 2018; Liu et al., 2022). For culture knowledge probing in the food domain, some of the recent datasets span multiple countries and include broad cultural or regional metadata (Min et al., 2018; Ma et al., 2023). However, they often use country as a proxy for culture, such as the country of origin for the food. For example, Palta and Rudinger (2023) introduced a test set to probe culinary cultural biases by considering US and non-US traditions, and Cao et al. (2024a) focuses on recipe transfer between Chinese and English. Investigating cultural differences within a country remains an under-explored area (Palta and Rudinger, 2023).

Fine-grained vision-language understanding Bugliarello et al. (2023) quantified the fine-grained vision-language understanding capabilities in exist-

Figure 3: Geographical distribution of cuisine types.[1]

Figure 4: Meta-info annotation for local specialty.

ing models, focusing on aspects within the general domain. Later works focus on the culture understanding in VL models (Liu et al., 2023; Cao et al., 2024b). However, current fine-grained VL datasets (Zhang et al., 2021; Parcalabescu et al., 2022; Thrush et al., 2022; Hendricks and Nematzadeh, 2021) are often framed as binary classification tasks, which limits their difficulty. Our multi-choice vision question answering dataset aims to advance the boundaries of fine-grained understanding in the context of food and culture.

3 FoodieQA: Dataset Annotation

China, with its expansive territory and long history, has cultivated rich and diverse food culture and traditions. Focusing on regional food culture differences, our dataset collection contains five distinct phases. 1) selection of cuisine types inside China; 2) collection of private images; 3) individual dish annotation; 4) visual question formulation; 5) text question formulation.

3.1 Selection of Cuisine Types

The well-recognized "eight major cuisines" in China are Sichuan (川菜), Guangdong (i.e., Cantonese, 粤菜), Shandong (鲁菜), Jiangsu (苏菜), Zhejiang (浙菜), Fujian (闽菜), Hunan (湘菜), Anhui (徽菜) cuisines (Zhang and Ma, 2020). This categorization is based on historical, cultural, and geographical factors that have influenced the development of distinct cooking styles and flavors in different regions of the country. For a better geographical coverage, we extend the eight cuisine types to additionally include Northwest (西北菜), Northeast (东北菜), Xinjiang (新疆菜), Jiangxi (赣菜) and, Mongolian cuisines (内蒙古菜) in this study. This results in 14 types (Figure 3) in total, for which we collect dish images and annotations.

3.2 Collection of Images

To ensure that the images are not used in the pre-training of existing models and contaminating evaluation, we designed and distributed a survey for Chinese locals to upload their own dish images (Figure 11).[2] We provide detailed guidelines for image uploading, specifying that: (1) the image should be clear, with a single dish as the focal point in the center; (2) participants should select the cuisine type of the dish from our list or specify it if it is not listed; (3) participants should provide the specific name of the dish, e.g., “mapo tofu (麻婆豆腐)" instead of “tofu (豆腐)"; (4) participants should indicate where the dish was served in their image, choosing from options such as cooked at home, restaurant, canteen, or delivery; (5) participants need to grant us permission to use the image for research purposes and confirm the image is not publicly available online, i.e., it has neither been downloaded from nor uploaded to the web or social media. In other words, the images we collected only existed on their phones or cameras. The uploaded images genuinely represent the locals’ daily diet and culinary experiences, showcasing dishes that are currently popular.

We manually filter out 102 images that are blurry, have the dish off-center, or show a mismatch between the dish and the image.

3.3 Local Specialty Annotation

We also gather text annotations of representative local specialties for each cuisine type on our list. Annotators are asked to collect meta information

[1] We omit the Islands of the South China Sea in the figure for visualization simplicity.

[2] The survey is distributed through WeChat and Douban.

for representative local dishes for each cuisine type, based on their life experience and knowledge obtained from the web. These meta-fields provide information beyond recipes, offering insights into how the food looks and tastes when people are eating it. An example is provided in Figure 4. The annotation is done by eight native Chinese speaker which include five PhD students and three postdoctoral researchers from different provinces in China.

The 17 meta-info fields cover the looks, taste, and culinary attributes of a dish. They include the food category, dish name, alternative names, main ingredient, characteristics of the main ingredient, three other key ingredients, dish flavor, presentation style, dish color, serving temperature (cold or warm), dishware used, region and province of origin, cuisine type, three primary cooking techniques, eating habits (if any), and reference links.

3.4 Visual Question Answering Annotation

One major consideration for vision-language understanding is that models can rely on language priors, consequently neglecting visual information (Goyal et al., 2017; Zhang et al., 2016). This underscores the importance of formulating visual questions in such a way that they can only be answered by examining visual features, rather than relying on text priors. Based on the number of images used as inputs, we formulate both multi-image VQA questions and single-image VQA questions.

3.4.1 Multi-image VQA

Multi-image VQA requires the ability to compare detailed visual features from multiple images, similar to how humans browse a restaurant menu.

Question formulation We ask the annotators to write challenging questions that require: (1) looking at the dish images to answer, (2) thinking beyond merely recognizing the dish and questions that may require multi-hop reasoning, (3) asking diverse questions that belong to a diverse set of question types such as food type, flavor, color, expense, amount, and etc., (4) only one image is the correct answer to the question. The multi-image VQA questions are written by five native speakers from five different regions in China.

We organize the collected images into 28 groups based on cuisine types and food categories, as outlined in Section 3.2. This allows annotators to write questions sequentially for related images extracted from the same group. Each annotator is asked to write two–three questions, given a four-image group. We note that in order to avoid the bias from language priors, dish names corresponding to the images are not presented. The user interface that we use for annotation is shown in Figure 12.

Question verification Once the questions and answers for the multi-image multiple-choice questions are collected, we verify the questions by asking the annotators (who did not create the questions) to answer them. If a question does not meet our defined criteria, annotators are instructed to flag it as a "bad question." Through this process, 87 questions were discarded. Additionally, when answering the questions, annotators are required to provide the rationale they use to reach the answer, as well as judge whether the question requires multi-hop reasoning. The user interface that we use for verification is shown in Figure 13. Each question is verified by two annotators, and we exclude the questions that do not have full agreement.

3.4.2 Single-Image VQA

Besides using images as multiple-choice answer options, we also ask diverse fine-grained questions about various aspects of a dish based on its meta-information. We identify dishes that have both meta-information annotations and collected images, and then create questions based on the meta-information. As shown in the example in Figure 2, the dish name is intentionally omitted from the questions to ensure they can only be answered by examining the visual features.

Question formulation We adopt a template-based approach, where a question about the same meta-field is asked multiple times, varying factors like the image of the dish, while the answer options are carefully selected from the wrong candidates in the meta-field to ensure that only one answer is correct. The single-image VQA questions are generated using a rule-based method, followed by thorough human verification and filtering through that is similar to the multi-image VQA verification process. Please see details in the Appendix A.

Question verification Similar to verification for the multi-image VQA questions, annotators are asked to answer the question given the text question and the corresponding image, and raise a “bad-question” flag to filter out questions that does not satisfy the criteria. 88 questions were discarded as bad. Note that the name of the dish is not revealed

in the text question so that the question needs to be answered based on visual information. Annotators are asked to write “I don’t know" in the rationale and randomly guess an answer if they think the question is beyond their knowledge.

3.5 Text Question Answering Annotation

We formulate the text-based questions by combining human annotations and rule-based generation. Similar to the single-image VQA approach described in Section 3.4.2, we generate questions and multiple-choice answer options based on the meta-information fields. However, instead of using the dish image, we included the dish name directly in the question. The questions are formulated using templates, where only the dish names and meta-fields are varied. A same human verification process to single-image question answering is included. 135 bad questions are discarded. Notice that annotators are asked to answer the questions based on their knowledge without using search engines, this makes the task challenging as it would be hard for one to answer questions about unfamiliar foods and regions without any other available information besides names of the food.

4 Dataset Statistics

4.1 Human Validation

In Table 1, we calculate human accuracy and inter-annotator agreement scores based on human-verified questions, excluding those identified as bad questions. For the single-image VQA and text QA questions, given the diverse cultural backgrounds of the human annotators, some questions can be challenging if the required food culture knowledge falls outside an annotator’s cultural experience. Annotators are instructed to indicate "I don’t know" when they lack the cultural knowledge to answer a question. These questions are classified as out-of-domain. For out-of-domain questions, the answer is randomly selected from the provided choices when calculating human accuracy and Cohen’s Kappa scores. We also report Cohen’s Kappa (κ), and human accuracy for in-domain questions. The human validation process involves eight native Chinese speakers from seven different provinces across China[3], including three postdoctoral researchers and five PhD students. Each question is verified and answered by two annotators.

[3]The annotators are from Sichuan, Shaanxi, Guangdong, Jiangsu, Jiangxi, Shandong, and Chongqing.

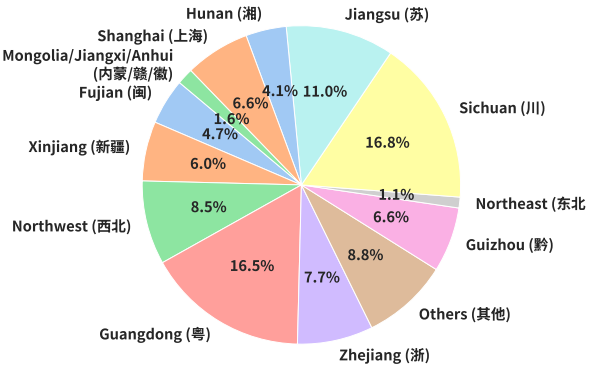


Figure 5: Region distribution of collected food images.

Task	Questions	κ	Accuracy
Multi-image VQA	403	.834	.916
Single-image VQA	256	.556	.744
- In-domain	168	.674	.818
Text QA	705	.470	.562
- In-domain	307	.808	.857

Table 1: Question Statistics per task in FoodieQA.

	Multi-image	Single-image	TextQA
Avg. length	12.9	17.0	14.9
Question types	14	6	7
Multi-hop (%)	25.3	73.4	1.6
Unique Images	403	103	-

Table 2: Question Statistics.

4.2 Image and Question Distribution

Image statistics We collected 502 images but discarded 113 due to quality control issues. The final dataset of 389 images are distributed across regions in China as shown in Figure 5. All 389 images are used for multi-image VQA; a subset of 103 images are used for single-image VQA.

Question statistics After human verification, we obtain 403 multi-image VQA questions, where each question needs to be answered with a set of four provided images. Single-image VQA tasks consists of 256 question in total, and text QA consists of 705 questions in total (Table 1). We report the key statistics of the questions in Table 2. Please see more details in Appendix B.

5 Baselines: How Much of a Foodie are the LLMs/VLMs?

We evaluate open-weight and API-based state-of-the-art LLMs and VLMs to probe their culture

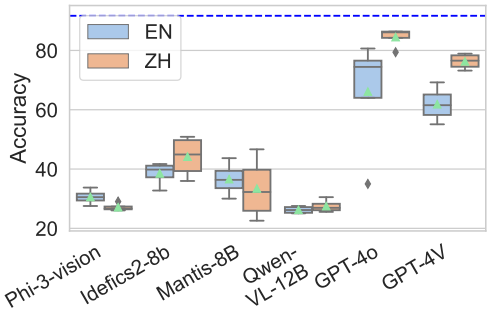


Figure 6: Accuracy of multi-image VQA tasks across four different prompts compared to a 91.96% human accuracy in Chinese. Although Idefics2 and Mantis have higher accuracy than other models, they show greater variation across different prompts.

Evaluation	Multi-image VQA		Single-image VQA	
	ZH	EN	ZH	EN
Human	91.69	77.22†	74.41	46.53†
Phi-3-vision-4.2B	29.03	33.75	42.58	44.53
Idefics2-8B	**50.87**	41.69	46.87	**52.73**
Mantis-8B	46.65	**43.67**	41.80	47.66
Qwen-VL-12B	32.26	27.54	48.83	42.97
Yi-VL-6B	-	-	**49.61**	41.41
Yi-VL-34B	-	-	52.73	48.05
GPT-4V	78.92	69.23	63.67	60.16
GPT-4o	**86.35**	**80.64**	**72.66**	**67.97**

Table 3: Comparison of Multi-image and Single-image VQA Performance in Chinese and English. We report the best accuracy from four prompts. †: see Footnote 4.

knowledge in the food domain. We evaluate the models in both Chinese and English[4]. For VQA questions are translated to English using the DeepL free API[5] and validated by two PhD students.

5.1 Multi-Image VQA is Difficult

We evaluate the multi-image VQA task using open-weight models that are capable of handling multiple image inputs, including Phi-3-vision-128k-instruct (Abdin et al., 2024), Idefics2-8B (Laurençon et al., 2024), Mantis-8B-Idefics2 (Jiang et al., 2024), and English-Chinese bilingual Qwen-VL-12B (Bai et al., 2023), and Yi-VL 6B and 34B models (AI et al., 2024), as well as API-based models such as GPT-4V and GPT-4o.

We experimented with four different prompts that utilized lists of images and texts or interleaved image-text inputs. Details can be found in Appendix C. As shown in Figure 6, when compared to the human accuracy of 91.69% in Chinese, the best-performing open-weight model, Idefics2-8B, achieves an accuracy of 50.87%, which is still significantly lower than human performance. This indicates that current state-of-the-art models are still weak at distinguishing differences among food from visual input. This underscores that multi-image understanding, especially in contexts requiring cultural knowledge in the food domain, remains a challenging problem. When evaluating on the translated English questions, model performance decreases for all models except Phi-3-vision.

[4]We also include an estimate, calculated over 100 random samples, of Human performance on the English Multi-Image and Single-Image VQA from one native speaker with **no** specialized knowledge of Chinese food culture.

[5]https://www.deepl.com/en/translator

5.2 Single-Image VQA Results

Besides the four open sourced models that we used for multi-image VQA, we also evaluate the bilingually trained (Chinese and English) Yi models (AI et al., 2024) for the single-image VQA task.

The evaluation accuracy is reported in Table 3. Almost every open-weight model performs better on Single-image VQA than Multi-image VQA. We can observe that, for the bilingually trained models, i.e., Qwen-VL and Yi-VL, their performance is better when evaluated in Chinese. However, for the multilingual models, i.e. Phi-3, Idefics2, and Mantis-8B, their performance is better when evaluated in English. The best performing models are the API-based models from OpenAI.

5.3 Models are Strong at Text QA

We evaluate text question answering with a series of open-weight models, including Phi-3-medium-4k-instruct (Abdin et al., 2024), Llama3-8B-Chinese (Wang and Zheng, 2024), Mistral-7B-Instruct-v0.3 (Wang and Zheng, 2024), Yi-6B and 34B models (AI et al., 2024), and Qwen2-7B-instruct (qwe, 2024), as well as API-based model GPT-4.

Given that dish names translation is challenging and would likely introduce additional information and unfair comparison, we only evaluate the text questions in Chinese. For example, a famous Sichuan dish "夫妻肺片" can be translated to "couple's lung slices" if translate word by word, however it would be translated as "Sliced Beef and Ox Tongue in Chilli Sauce" by meaning. While the literal translation makes no sense, translation by meaning would hint the flavor and ingredients that are not included in its original Chinese name.

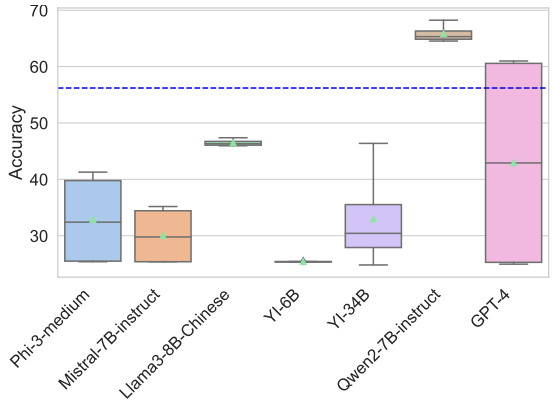


Figure 7: Accuracy of text QA across four different prompts. The blue dashed line indicates human accuracy (56.2%).

Input	prompt1	prompt2	prompt3	prompt4
Dish name only	28.52	27.73	36.72	37.11
+ dish image	40.23	41.41	40.62	42.19

Table 4: Accuracy on two variants of Single-image VQA task, showing that visual information of food images is crucial for Idefics2 to correctly answer the questions.

From Figure 7, we see that the Qwen2-7B-instruct model surpasses human performance on the text QA task, where the questions are formulated based on the local specialty annotations in Section 3. Since the local specialty annotations are collected and summarized from public resources such as Baidu-Baike by local representatives, we suspect that the high performance could be due to the inclusion of domain-specific training data.

6 Analysis

Visual information helps. In Single-image VQA, the default setting is to use only dish image without specifying the dish name. We now examine whether the visual information is beneficial using the Idefics2-8B model.[6] Results are shown in Table 4, where we investigate two variants of Single-image VQA: providing the model with dish name only versus both the dish name and image. We observe that the Idefics2 model consistently performs better when dish images are available as additional information. Please see comparison examples in Appendix E.2.

Dish names could be helpful clues for some of the models. As discussed in Section 4.2, over 73.4% of single-image questions require multi-hop reasoning, which typically involves identifying the dish and then leveraging related knowledge to answer the questions. To determine whether the identification of the food image and the utilization of visual information are bottlenecks for the models, we compare their performance on single-image VQA when provided with the dish name in the question.

[6]We selected this model because it supports text-only inputs, unlike some other models such as the Yi-VL series.

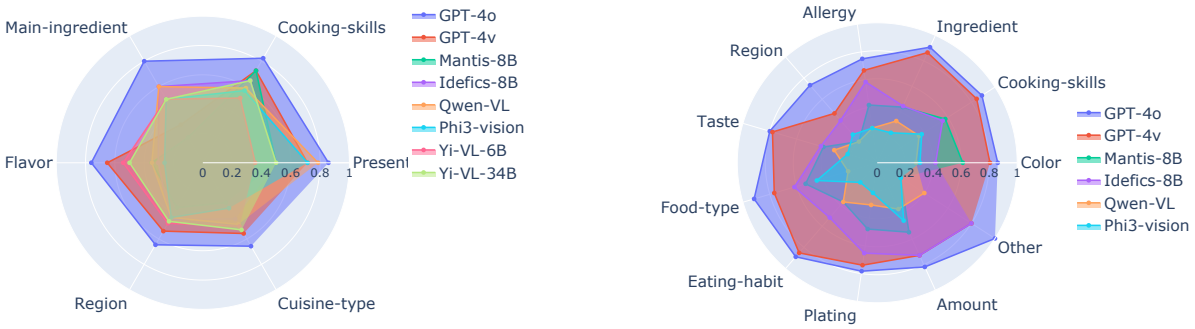
Model	Condition	p1	p2	p3	p4
Yi-VL-6B	Image-only	**49.61**	48.05	47.66	46.09
	+ dish name	73.83	74.61	**76.17**	62.50
Yi-VL-34B	Image-only	50.39	**52.73**	50.78	48.83
	+ dish name	75.39	78.13	**79.30**	75.39
Idefics2-8B	Image-only	44.53	43.75	46.09	**46.87**
	+ dish name	40.23	41.41	40.62	**42.19**

Table 5: Accuracy in the Single-image VQA task when dish name is revealed in the questions along with the image or not. While the Yi models benefit greatly from the additional information of the dish name, Idefics2 does not. "p1–4" indicates four different prompt templates.

The results in Table 5 indicate that while the Yi models significantly benefit from being given both the images and names of the dishes, the Idefics2-8B model does not show the same improvement from this additional information. This indicates that recognizing the dishes could be a possible bottleneck for the Yi series models.

Models are foodies who know cooking better than taste. Figure 8a shows the model performance under fine-grained questions attributes on Single- and Multi-image VQA. We observe that all models generally excel at answering questions related to cooking skills and ingredients. The Yi models, in particular, demonstrate a stronger ability to identify the flavors of dishes. Conversely, the Qwen-VL and Phi3-vision models perform well in observing the presentation of food when served but struggle with flavor-related questions. When answering questions based on multiple images, it also holds true that models are generally good at questions regarding cooking skills and the amount of food (Figure 8b). However, these models are weak at answering questions related to the region and taste of the dish. Idefics-8B stands out, excelling in most of the fine-grained features we evaluated.

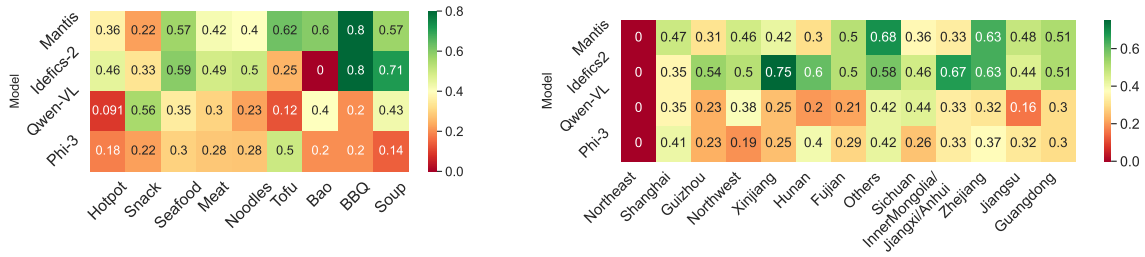
Favorite food of the models. In Figure 9, we compare model performance on multi-image VQA tasks for questions grouped by food categories and cuisine types. This analysis provides insight into how well the models can compare features from



(a) Single-image VQA.

(b) Multi-image VQA.

Figure 8: Model accuracy on fine-grained question attributes.



(a) Accuracy by food categories.

(b) Accuracy by cuisine types.

Figure 9: Model accuracy on questions categorized by food categories and cuisine types.

images within the same group. The overall best performing model on multi-image VQA tasks excels at questions about BBQ and Xinjiang cuisines, but weak at questions about Shanghai dishes. Another interesting finding is that, despite Sichuan food being one of the most popular cuisines in China, and presumably having more available images and resources online, none of the models excel at answering questions related to this cuisine type.

7 Conclusion

We introduce FoodieQA, a multimodal dataset designed to evaluate fine-grained understanding of Chinese food culture through multi-image, single-image, and text-only multiple-choice questions.

Our experiments, which focus on regional cultural differences and detailed visual features, reveal that understanding food and its cultural context remains a complex and under-explored task. We find that comparing food across multiple images—similar to the common scenario of people browsing menus—is particularly challenging. All open-source models underperform human accuracy by more than 40% in this task. This suggests that our dataset offers a more accurate assessment of the suitability of state-of-the-art models for real-world applications in the food domain.

Our analysis of language and prompt templates indicates that models can be sensitive to the language in which questions are asked—bilingually trained Chinese–English models perform better in Chinese, while other multilingual models are stronger in English. We also demonstrate the effectiveness of incorporating visual features compared to text-only settings in this context.

Improved models or methods for understanding food culture may be essential for future progress in the FoodieQA challenge. Looking ahead, we aim to expand the dataset to include dishes from other countries and regions. We make all of our data collection, annotation, and verification tools freely available for re-use, and encourage the community to create Foodie datasets for their own language.[7]

8 Limitations

The size of the FoodieQA dataset is limited by the challenge of collecting unseen images from individuals, as it requires them to voluntarily upload images from their phones or cameras. Although we

[7]We will release our dataset as a benchmark on Codabench.

have distributed the survey on two popular Chinese social media platforms, we anticipate that increased social media exposure or collaboration with food industry professionals could facilitate the collection of more images, and contribute to a training dataset for advancing this direction.

Translating Chinese dish names into other languages poses another challenge, as some dish names do not directly relate to their ingredients or cooking methods. Introducing translated dish names could potentially introduce additional information, leading to unfair comparisons among the models. Consequently, we have chosen to experiment solely with Chinese questions for the text-based queries.

We have benchmarked fifteen popular models using our dataset. However, due to the rapid advancements in the field, it is impossible to benchmark all trending models continuously. We hope our dataset will inspire future researchers to develop similar Foodie datasets for their own regions and languages, thereby guiding LLMs and VLMs towards a better understanding of regional food cultures.

References

2024. Qwen2 technical report.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring Progress in Fine-grained Vision-and-Language Understanding. ArXiv:2305.07558 [cs].

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024a. Cultural Adaptation of Recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Yong Cao, Wenyan Li, Jiaang Li, Yifei Yuan, Antonia Karamolegkou, and Daniel Hershcovich. 2024b. Exploring visual culture awareness in gpt-4v: A comprehensive probing. *ArXiv*, abs/2402.06015.

Xin Chen, Hua Zhou, Yu Zhu, and Liang Diao. 2017. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. arXiv2405.01483.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models?

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiao Liu, Yansong Feng, Jizhi Tang, Chengang Hu, and Dongyan Zhao. 2022. Counterfactual recipe generation: Exploring compositional generalization in a realistic scenario. *arXiv preprint arXiv:2210.11431*.

Zhixuan Liu, You Won Shin, Beverley-Claire Okogwu, Youngsik Yun, Lia Coleman, Peter Schaldenbrand, Jihie Kim, and Jean Oh. 2023. Towards equitable representation in text-to-image synthesis models with the cross-cultural understanding benchmark (ccub) dataset. *ArXiv*, abs/2301.12073.

Zheng Ma, Mianzhi Pan, Wenhan Wu, Kanzhi Cheng, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2023. Food-500 Cap: A Fine-Grained Food Caption Benchmark for Evaluating Vision-Language Models. ArXiv:2308.03151 [cs].

W. Min, B. K. Bao, S. Mei, Y. Zhu, Y. Rui, and S. Jiang. 2018. You are what you eat: Exploring rich recipe information for cross-region food analysis. *IEEE Transactions on Multimedia*, 20(4):950–964.

Shramay Palta and Rachel Rudinger. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5238–5248.

Shenzhi Wang and Yaowei Zheng. 2024. Llama3-8b-chinese-chat (revision 6622a23).

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. " my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes.

Na Zhang and Guansheng Ma. 2020. Nutritional characteristics and health effects of regional cuisines in china. *Journal of Ethnic Foods*, 7.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022.

Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. 2021. Vsr: a unified framework for document layout analysis combining vision, semantics and relations. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pages 115–130. Springer.

Li Zhou, Taelin Karidi, Nicolas Garneau, Yong Cao, Wanlong Liu, Wenyu Chen, and Daniel Hershcovich. 2024. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. *arXiv preprint arXiv:2404.06833*.

A Rule-based question formulation

For text-based question answering we develop a rule-based question formulation method. For each of the question type, we have the meta information from the local specialty annotation (Section 3.3). Then we design three to four templates for each of the question type. For example, for questions that ask about cuisine type, our templates include

- <dish>是哪个地区的特色菜? (What region is <dish> a specialty dish of?)
- <dish>是哪个地区的特色美食? (In which region that <dish> is a local specialty?)
- 去哪个地方游玩时应该品尝当地的特色美食<dish>? Which place should you visit to taste the local specialty food <dish>?

Then, we randomly select cuisine types that are not the correct answer to serve as the alternative options. By utilizing different meta fields, we can generate multiple questions for each dish.

For single-image VQA, we associate the questions related to the dish with the corresponding dish image in our collection. We exclude questions of the warm-cold type—those that inquire whether a

dish is served hot or cold—since these questions involve different dishes as options and are not suitable for the single-image scenario.

B Question type and answer distribution

In Table 6, 7, and 8, we show concrete statistics about distribution of question types in each task. In Figure 10, we plot the distribution answer distribution for each of the question type in the tasks.

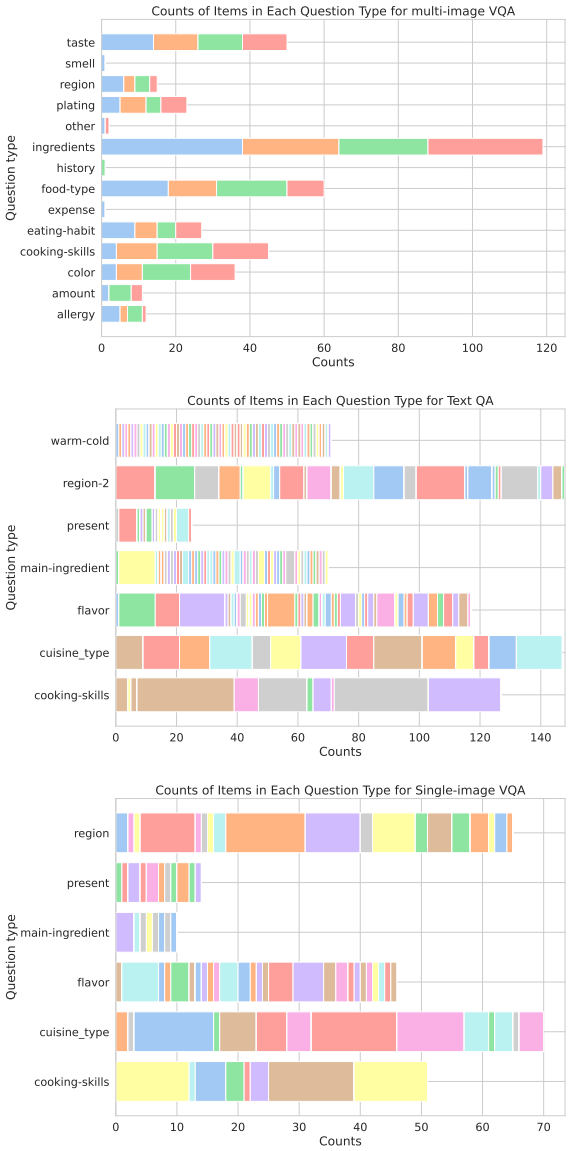


Figure 10: Answer distribution for each of the tasks.

C Prompts used for evaluation

Following Durmus et al. (2023) and Wang et al. (2024), we design four prompts for each of the tasks and extract the option letter from the model response. For multi-image VQA, we specifically include prompts that feature both interleaved image and text inputs as well as separate lists of images and texts. Please see examples of the prompts in Table 9 and Table 10.

Task	Count
Cuisine Type	147
Cooking Skills	127
Main Ingredient	70
Region	148
Flavor	117
Present	25
Warm-Cold	71

Table 6: Distribution of text QA question types

Task	Count
Cuisine Type	70
Flavor	46
Region	65
Present	14
Cooking Skills	51
Main Ingredient	10

Table 7: Distribution of single-image VQA question types

D Interface of image collection, annotation and verification tool

In Figure 11, we display the survey that we used to collect images. In Figure 12 and Figure 13 show the user interface that annotators use to create questions and verify the questions.

E More examples

E.1 Examples of the questions in the dataset

E.2 Examples of comparing whether the visual information is available

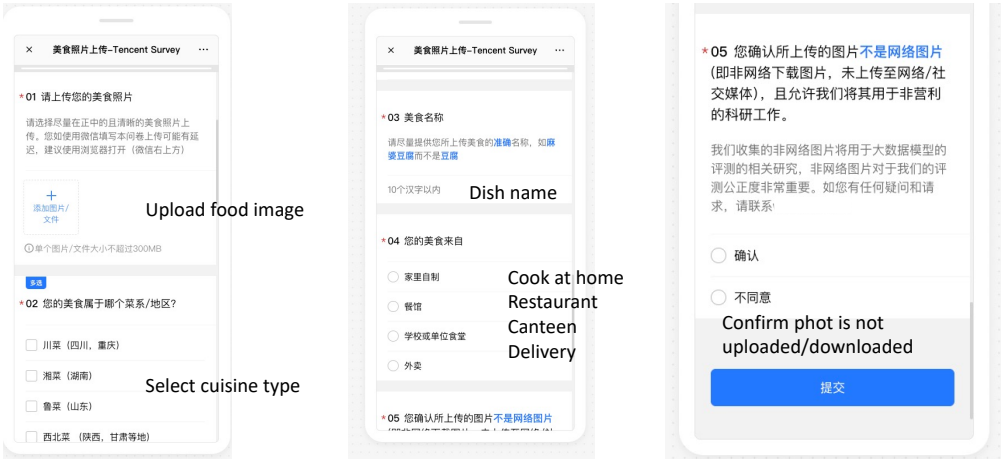


Figure 11: Survey interface of image collection

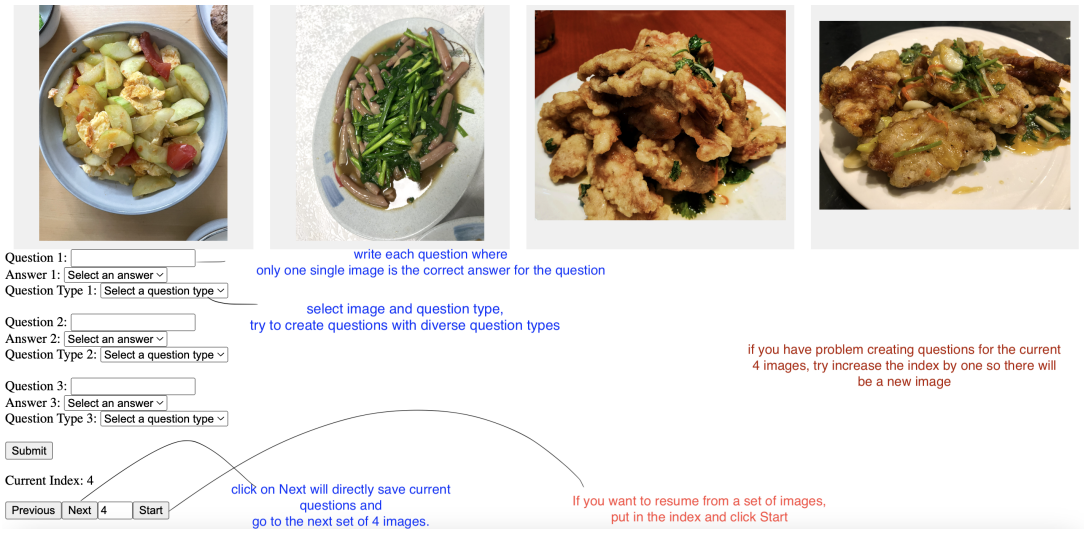


Figure 12: Annotation interface of writing questions when presented multiple images.

Multi-image VQA

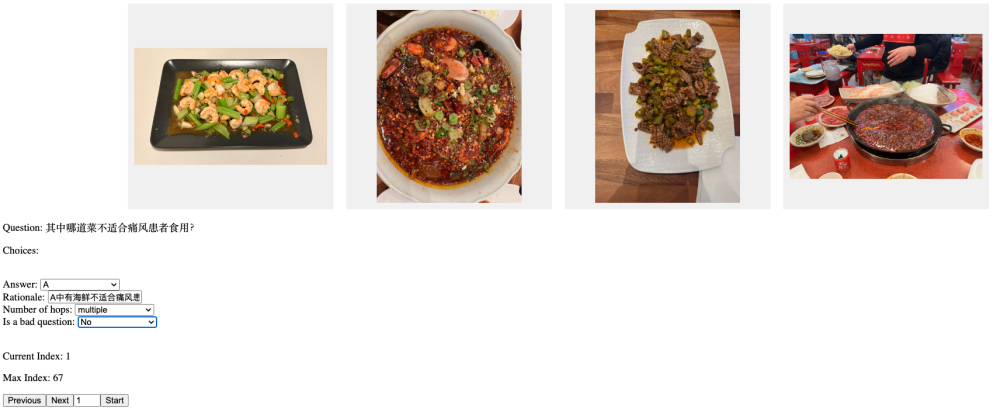
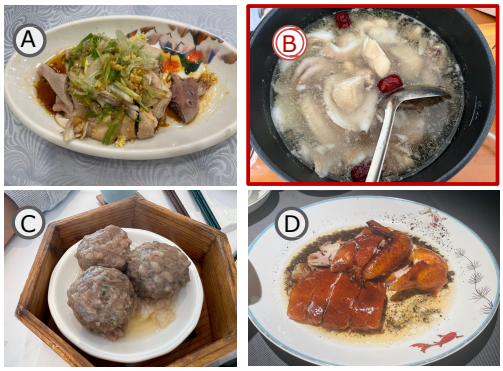


Figure 13: Annotation interface of verifying the multi-image multiple-choice questions.

Multi-Image VQA

如果你想要喝汤，以下食物你会选择哪一道？If you want soup, which dish would you choose?



Single-Image VQA

以下菜品是哪个地区的特色菜？
Which region is this food a specialty?

A. 宁波 (Ningbo)
B. 福建 (Fujian)
C. 广东 (Guangdong)
D. 安徽 (Anhui)

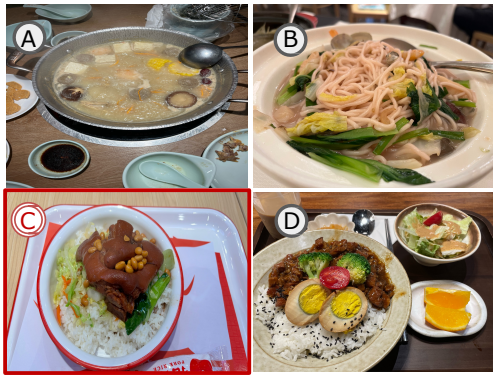
Text QA

阳澄湖大闸蟹是什么口味？
What is the flavor of 阳澄湖大闸蟹?

A 软香 (Soft & fragrant)　B 甜 (Sweet)
C 肉香 (Meaty aroma)　D 鲜美 (Fresh & tasty)

Multi-Image VQA

哪一道菜适合喜欢吃肥肉的人？Which dish is good for people who like fatty foods?



Single-Image VQA

以下菜品是哪个地区的特色菜？
Which region is this food a specialty?

A. 川渝 (Sichuan & Chongqing)
B. 西宁 (Xining)
C. 嘉兴 (Jiaxing)
D. 南疆 (South Xinjiang)

Text QA

阳澄湖大闸蟹是哪个菜系的经典菜？
In which regional cuisine is 阳澄湖大闸蟹 a specialty?

A 川菜 (Sichuan cuisine)　B 苏菜 (Jiangsu cuisine)
C 家常菜 (home-style cuisine)　D 鲁菜 (Shandong cuisine)

Multi-Image VQA

哪一道菜的口味最辣？Which dish is the spiciest?

Single-Image VQA

以下菜品是哪个地区的特色菜？
Which region is this food a specialty?

A. 陕西 (Shaanxi)
B. 东北 (Northeast of China)
C. 扬州 (Yangzhou)
D. 徽州 (Huizhou)

Text QA

鱼丸粉是哪个菜系的经典菜？
In which regional cuisine is 鱼丸粉 a specialty?

A 粤菜 (Cantonese cuisine)　B 苏菜 (Jiangsu cuisine)
C 新疆菜 (Xinjiang cuisine)　D 赣菜 (Jiangxi cuisine)

Figure 14: More examples in FoodieQA evaluate food culture understanding from three perspectives.

Task	Count
Ingredients	119
Food Type	60
Color	36
Taste	50
Cooking Skills	45
Plating	23
Eating Habit	27
Allergy	12
Region	15
Expense	1
Other	2
Amount	11
Smell	1
History	1

Table 8: Distribution of multi-image VQA question types

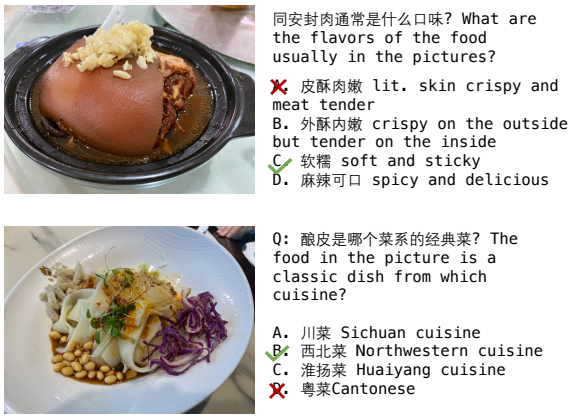


Figure 15: Examples where the Idefics-2-8B model correctly answers the question when the image is available but failed when it is not.

Prompt 1	<img1>, <img2>, <img3>, <img4> 根据以上四张图回答问题，他们分别为图A, 图B, 图C, 图D, 请从给定选项ABCD中选择一个最合适的答案。问题：<question>, 答案为：图
Prompt 2	<img1>, <img2>, <img3>, <img4> 根据以上四张图回答问题, 请从给定选项ABCD中选择一个最合适的答案。问题：<question>, 答案为：图
Prompt 3	根据以下四张图回答问题, 请从给定选项ABCD中选择一个最合适的答案。 <img1>图A <img2>图B <img3>图C <img4>图D 问题：<question>, 答案为：图
Prompt 4	Human: 问题<question>，选项有: 图A<img1> 图B<img2> 图C<img3> 图D<img4> Assistant: 如果从给定选项ABCD中选择一个最合适的答案， 答案为：图

Table 9: Chinese prompts for zero-shot evaluation for multi-image VQA.

Prompt	Content
Prompt 0	<img1><img2><img3><img4> Answer the following question according to the provided four images, they correspond to Option (A), Option (B), Option (C), Option (D). Choose one best answer from the given options. Question: , your answer is: Option (
Prompt 1	Answer the following question according to the provided four images which correspond to Option (A), Option (B), Option (C), Option (D). Choose one best answer from the given options. The options are: <img1>Option (A) <img2>Option (B) <img3>Option (C) <img4>Option (D) Question: <question>, your answer is: Option (
Prompt 2	Answer the following question according to the provided four images, and choose one best answer from the given options. The options are: <img1>Option (A) <img2>Option (B) <img3>Option (C) <img4>Option (D) Question: <question>, your answer is: Option (
Prompt 3	Human: Question <question> The options are: Option (A)<img1> Option (B)<img2> Option (C)<img3> Option (D)<img4> Assistant: If I have to choose one best answer from the given options, the answer is: Option (

Table 10: Prompts for zero-shot evaluation