KWordinaryVQA: A Keyword-Driven Generative Visual Question Answering System for Culinary Exploration

Huy Trieu[1,2], Thanh Thai Nguyen[1,2], Thanh Nghia Vo[1,2],
Thinh Vuong Vo[1,2], Thanh Tu Dang[1,2], Tung Le[1,2,*]
[1]Faculty of Information Technology, University of Science, Ho Chi Minh city, Vietnam
[2]Vietnam National University, Ho Chi Minh city, Vietnam
{tghuy22,ntthai22,vtnghia22,vtvuong22,dttu22}@clc.fitus.edu.vn
[*]Corresponding author: lttung@fit.hcmus.edu.vn

Abstract

Visual Question Answering (VQA) has seen significant progress on general images, yet food imagery presents unique challenges requiring domain-specific understanding. This paper presents KWordinaryVQA[1], an end-to-end automated pipeline to construct large-scale food VQA datasets. Starting from raw images, we employ advanced Large Language Models (LLMs) to generate detailed descriptions and synthesize diverse question-answer pairs, followed by targeted manual validation to ensure high-quality evaluation data. We then benchmark multiple approaches–including LLMs under zero-shot and few-shot settings, a traditional retrieval baseline, and representative fine-tuned vision-language models–evaluating them on accuracy, human judgment, and inference efficiency. Our workflow mirrors a standard data science process of data collection, exploration, evaluation, and model building, providing a systematic framework for domain-specific VQA.

1 Introduction

Visual Question Answering (VQA) aims to develop systems that can answer natural language questions based on the visual content of an image. This inherently multimodal task requires a sophisticated integration of computer vision, natural language understanding, and often, commonsense reasoning. While significant strides have been made with general-purpose VQA benchmarks like VQA v2.0 (Goyal et al., 2017) and OK-VQA (Marino et al., 2019), largely fueled by advances in transformer-based multimodal architectures, these models often struggle when applied to specialized domains. The food domain, in particular, presents unique challenges and opportunities, yet remains relatively underexplored despite its profound practical relevance.

Addressing VQA in the food domain extends beyond academic inquiry, unlocking numerous real-world applications. Such systems could revolutionize dietary tracking by identifying ingredients and portion sizes, enhance cooking education with interactive guidance, and offer crucial assistive technologies for individuals with visual impairments or dietary restrictions. For example, the ability to accurately detect allergens or forbidden food items is crucial for users with health concerns, such as food allergies and diabetes, or those following religious dietary laws, underscoring the need for ingredient-level understanding. Furthermore, visual cues like color and texture are vital indicators of food condition–whether an item is raw, perfectly cooked, or fresh–making an effective food VQA system invaluable for culinary evaluation, food safety, and quality control in both domestic and industrial settings such as restaurants and food processing plants. Spatial reasoning also plays a crucial role; understanding the arrangement of food components on a plate can inform analyses of food presentation, a key factor in professional culinary arts and brand consistency across F&B chains. AI systems could thereby assess presentation uniformity, portion control, and adherence to plating standards, directly impacting customer experience and brand perception.

Despite these compelling applications and the increasing capabilities of general VQA models, our preliminary analysis and a review of existing literature indicate that current state-of-the-art approaches, including large language models (LLMs) with visual grounding, classification-based methods, and statistical models, often falter on domain-specific food-related queries (as detailed in Section 2). This performance gap stems from several factors: the inherent visual complexity of food items (e.g., fine-grained differences between ingredients, varied cooking states), the need for specialized, implicit domain knowledge (e.g., culinary

[1] The final dataset is available on Kaggle.

techniques, cultural nuances), and critically, the scarcity of large-scale, high-quality, open-ended VQA datasets tailored specifically for the food domain.

To address the lack of domain-adapted benchmarks in visual question answering, we introduce KWordinaryVQA, an end-to-end automated pipeline for constructing VQA datasets from food imagery. Beginning with raw images from the public culinary platform Allrecipes (Allrecipes contributors, 2025), the pipeline first employs Gemini 2.0 Flash (Google DeepMind, 2025) to perform image captioning and then extract key phrases from these captions. Subsequently, question-answer pairs are generated from the captions and key phrases using DeepSeek-V3 (Liu et al., 2024). This automated workflow enables the construction of scalable datasets with minimal human effort, promoting deeper reasoning through open-ended formats rather than restrictive multiple-choice or classification tasks.

Our contributions are fourfold. First, we design and implement a fully automated pipeline for generating domain-specific VQA data from food imagery. Second, we construct the KWordinaryVQA dataset, consisting of 43,455 QA pairs across 8,693 images, including manually validated test and validation splits. Third, we benchmark a diverse set of VQA approaches–ranging from zero-shot LLMs and TF-IDF retrieval to fine-tuned vision-language transformers. Finally, we conduct in-depth dataset analysis, covering question types, linguistic features, and performance breakdowns, offering insight into domain-specific VQA challenges.

To facilitate further research, we publicly release the KWordinaryVQA dataset, generation pipeline, and evaluation code, aiming to advance domain-adapted VQA in food and other specialized domains.

2 Related Works

The growing interest in AI for the food domain has led to the development of several datasets for food-related tasks. These datasets, while valuable, can be broadly categorized by their focus on either cultural depth or task-specific evaluation, each with inherent limitations.

Several benchmarks offer deep insights into specific culinary traditions. For instance, FoodieQA (Li et al., 2024) provides a manually annotated, multimodal benchmark for Chinese cuisine, while IndiFoodVQA (Agarwal et al., 2024) uses a knowledge-graph-enhanced pipeline to assess reasoning in the Indian food domain. Although these datasets are rich in specialized knowledge, their cultural specificity and, in the case of FoodieQA, reliance on manual annotation, can limit their scalability and general applicability.

In parallel, other large-scale efforts often concentrate on more constrained task formulations. WorldCuisines (Winata et al., 2024), despite its impressive scale and multilingual support, primarily targets dish and origin identification rather than complex reasoning about visual attributes. Similarly, Food-VQA-Benchmark (Cheng et al., 2024) evaluates a suite of tasks but largely relies on closed-set formats or structured outputs, which may not fully capture the complexities of truly open-ended VQA where models must generate free-form answers.

Collectively, while these datasets have advanced the field, a clear gap persists. They are often constrained by the laborious nature of manual annotation, which impacts scale and diversity, or their task formulations favor identification and structured information over fostering a broad spectrum of open-ended inquiries that demand nuanced visual reasoning. This highlights a pressing need for a large-scale, open-ended VQA dataset for general food imagery, developed through a scalable and adaptable pipeline–a need that our work, KWordinaryVQA, directly aims to address.

3 Dataset Acquisition and Preprocessing

To construct the KWordinaryVQA dataset, we initially crawled 49,332 structured data entries from the public recipe source Allrecipes (Allrecipes contributors, 2025), covering a broad diversity of cuisines, food types, and presentations, encompassing both professionally staged studio shots and user-submitted home-cooking photographs. Each entry included a food image accompanied by metadata such as food names, descriptions, and ingredients. Given the scale of this raw dataset and computational constraints, we designed a multi-step preprocessing pipeline to curate a high-quality, representative, and manageable subset suitable for robust model training and evaluation.

3.1 Initial Data Filtering

The initial preprocessing involved two main stages. First, to normalize food names, a fuzzy string matching technique using the RapidFuzz (Bach-

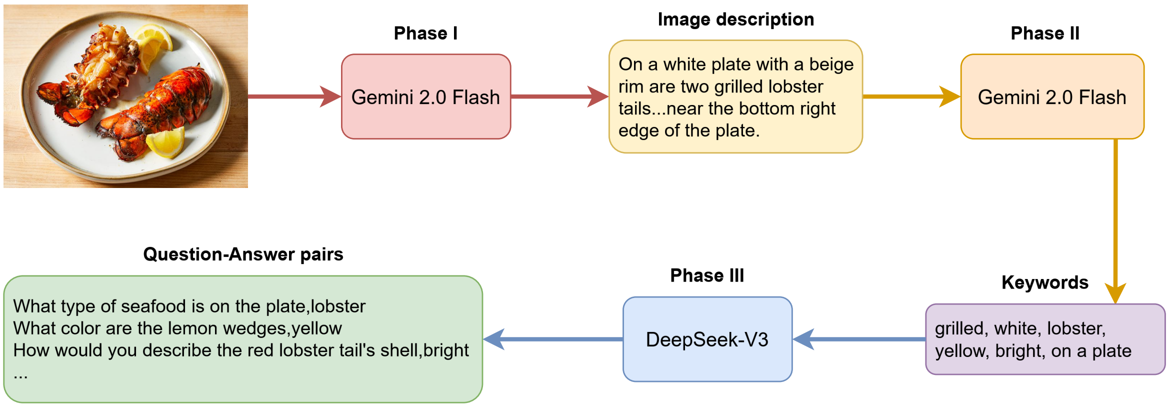

Phase I
Gemini 2.0 Flash
Image description
On a white plate with a beige rim are two grilled lobster tails...near the bottom right edge of the plate.
Phase II
Gemini 2.0 Flash
Keywords
grilled, white, lobster, yellow, bright, on a plate
Phase III
DeepSeek-V3
Question-Answer pairs
What type of seafood is on the plate,lobster
What color are the lemon wedges,yellow
How would you describe the red lobster tail's shell,bright
...

Figure 1: The dataset creation pipeline consists of three stages: (I) generating image descriptions, (II) extracting relevant keywords, and (III) creating question-answer pairs.

mann, 2024) library was applied with a token set ratio scorer and an 80% similarity threshold. Within each group of near-duplicate names, the shortest variant was selected as the canonical form to improve naming consistency. This step resulted in 27,270 entries. Subsequently, we removed all samples with incomplete metadata to ensure data integrity for downstream tasks. Specifically, 498 entries lacked calorie data and 9 were missing image dimensions. By retaining only complete entries, we obtained a final dataset of 21,370 samples.

3.2 Outlier Analysis and Retention

Outliers were analyzed using the interquartile range (IQR) method, identifying caloric values outside an 800-calorie range as potential outliers. In the context of KWordinaryVQA, high-calorie dishes (e.g., energy-rich foods) were deemed valuable for calorie-related questions. Thus, we retained these outliers to preserve informative samples, ensuring the dataset remained representative of diverse nutritional profiles.

3.3 Undersampling for Class Balance

The dataset exhibited significant class imbalances. To address this, we implemented a two-tiered undersampling strategy. First, to mitigate the dominance of common dishes (e.g., chicken, cake), we capped their sample count at 100-150 per dish. Conversely, extremely rare dishes (1-4 samples) were removed, as their low representation was insufficient for effective model learning. This process reduced the dataset to 16,817 entries.

Second, to address the severe imbalance between vegan and non-vegan dishes, we capped the number of samples for non-vegan dishes with more than 20 instances at 20, while all vegan and rare non-vegan dishes were kept unchanged. This strategy yielded a more reasonable class ratio (approximately 1:4) while preserving dataset diversity, resulting in a raw set of 9,191 unique food images.

3.4 Scope-based Filtering

Finally, a manual review was conducted to ensure linguistic and cultural consistency. To create a robust benchmark, we narrowed the dataset's scope to focus on food items commonly understood within general English culinary discourse. Entries for dishes requiring specific, non-English cultural context for identification (e.g., "Cao Lau," a Vietnamese specialty) were excluded. This deliberate choice, while limiting cultural breadth, was crucial for enhancing the dataset's internal consistency and suitability for our defined VQA task. This step resulted in the final pool of 8,693 images used for generation.

4 Dataset Creation

Following the preprocessing pipeline, a final collection of 8,693 unique food images served as the visual foundation for the KWordinaryVQA dataset. This curated set captures a wide variety of cuisines, food types, and presentation styles. The subsequent dataset creation process involved three main automated stages: description generation, keyword extraction, and question-answer synthesis.

4.1 Description Generation

For each image, we first generated a detailed textual description using Gemini 2.0 Flash (Google DeepMind, 2025). Given a food image, the model was prompted (see Appendix A.1) to produce a concise, standalone paragraph of up to 200 words, describ-

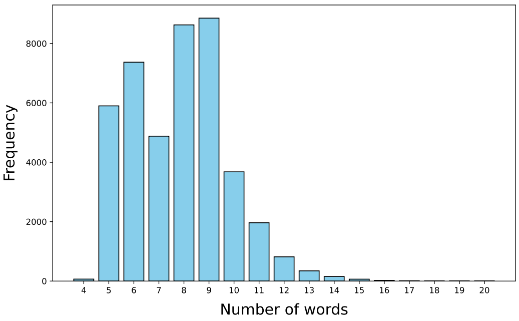


Figure 2: Distribution of question lengths.

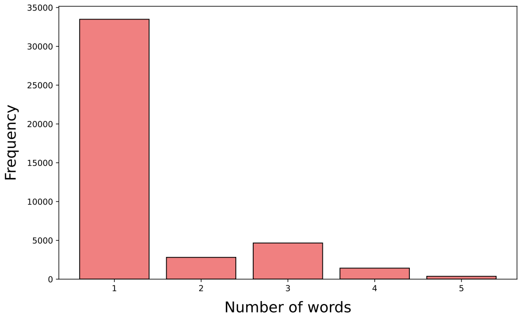


Figure 3: Distribution of answer lengths.

ing its salient contents, including ingredients, dish type, and serving context. For example, for an image of ramen, the model might output: "A bowl of noodle soup with sliced pork, half an egg, green onions, and seaweed on top, on a wooden table." A detailed description is foundational, as it provides the rich, factual grounding required for generating diverse and meaningful questions. While the process was largely automated, minor manual corrections were occasionally applied to address clear factual inaccuracies (e.g., misidentifying an ingredient).

4.2 Keyword Extraction

To guide the question synthesis process and establish ground-truth answers, we then extracted key terms from each image description. This step, again utilizing Gemini 2.0 Flash, was designed to produce a set of target answers for the subsequent QA generation. A specific prompt (see Appendix A.2) instructed the model to identify six diverse keywords–including at least one verb, adjective, noun, color, and preposition–to ensure a variety of question types. For instance, from the ramen description above, the extracted keywords, which would later serve as target answers, might be: *slice, fresh, pork, green, on the table, soup*.

4.3 Question-Answer Synthesis

Using the generated descriptions and extracted keywords, we synthesized QA pairs with DeepSeek-V3 (Liu et al., 2024). For each image, the model received its description and keywords, tasked with generating one question per keyword. The prompt (Appendix A.3) enforced several critical constraints: questions had to be grounded in the description, the answer for each question had to be the corresponding keyword, and "Where" questions were specifically generated for prepositional keywords. This automated workflow allowed for the rapid, large-scale synthesis of question-answer pairs, averaging approximately five pairs per image.

4.4 Data Splitting and Validation

The dataset was partitioned at the image level to prevent content leakage between splits: 90% of the images and their corresponding QA pairs were allocated to the training set (39,051 pairs), and the remaining 10% were reserved for the test set (4,342 pairs). A subset of the training data was subsequently held out for validation.

To ensure the high quality of our evaluation benchmark, the initial test set underwent a rigorous manual validation process for factual accuracy, visual relevance, and clarity. This involved removing or rephrasing samples that: (1) referenced details not visually apparent (i.e., hallucinations); (2) were vague or inadequately specified; or (3) required subjective inference or external knowledge. For example, speculative queries about non-visible attributes (e.g., spiciness) were discarded. This meticulous process resulted in a high-confidence test set of 3,699 QA pairs (a 15.0% reduction). The training and validation sets were not subjected to this manual filtering. Illustrative examples from the resulting dataset can be found in Appendix D.

5 Dataset Analysis

To characterize the newly constructed KWordinaryVQA dataset and assess the output of our generation pipeline, we conducted an exploratory data analysis. This analysis provided insights into properties such as question diversity and potential biases, informing our understanding of the dataset's characteristics and suitability for evaluating VQA models.

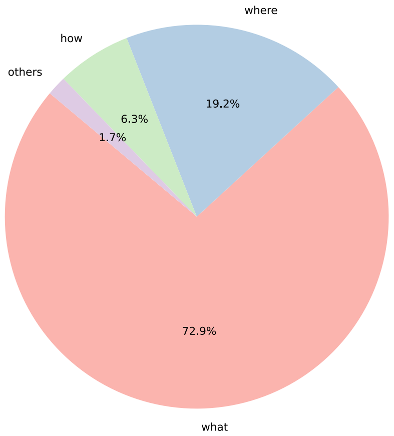


Figure 4: Question types distribution.

5.1 Length Distributions

We first analyzed the length of questions and answers in terms of word count. Questions in KWordinaryVQA tend to be concise, with a mean length of 7.79 and a median of 8 words. As shown in Figure 2, the distribution peaks around 8-9 words, with a long tail extending to about 15 words, corresponding to more complex inquiries. Answers are typically even shorter; over 80% are single or two-word responses. The average answer length is 1.41 words, with a heavy concentration on one-word answers (Figure 3). This finding has direct implications for evaluation design, suggesting that metrics should accommodate brief answers and that exact-match criteria may be overly stringent for some phrasal responses.

5.2 Question Types

We categorized the questions by their starting words to identify common patterns. As illustrated in Figure 4, the dataset is overwhelmingly dominated by “What” questions, which typically inquire about ingredients, objects, or dish names. “Where” questions focusing on spatial relations are the next most frequent category, followed by “How” questions, which encompass both counting and descriptive inquiries.

Notably, other question types such as polar and causal “Why” questions are extremely rare. This distribution highlights a potential bias in our automated generation pipeline, which favors descriptive inquiries grounded directly in visual evidence over more abstract or inferential reasoning.

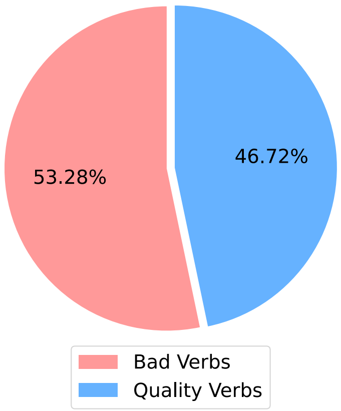


Figure 5: Analysis of verb quality in test set answers. The chart illustrates the proportion of quality verbs among all verb-based answers. Of the answers identified as verbs, 46.72% were deemed to be meaningful (quality verbs).

5.3 Quality Estimation

While the training sets were not manually examined, we aimed to quantitatively estimate the incidence of semantically poor labels (i.e., label noise). To this end, we conducted a comparative analysis of verb quality[2] between the unfiltered test set and its manually validated counterpart. We argue that this analysis is representative of the entire dataset, given that all splits were generated from the same pipeline and exhibit consistent linguistic distributions, as detailed in Appendix E and F.

Our methodology leveraged the pre-existing filtered test set as ground truth. First, we employed a BERT-uncased part-of-speech tagging model (Blagojevic, 2024) to programmatically identify all verb-based answers in the initial, unfiltered test set, resulting in a total of 137 such answers. We then applied the same process to the manually validated test set, which yielded only 64 verb-based answers.

By this definition, the 64 verbs that survived the manual validation process were considered "high-quality" (e.g., "eat", "contain"), while the 73 verbs that were filtered out were considered "low-quality" or generic (e.g., “do”, “be”). As illustrated in Figure 5, this comparison yields a verb quality rate of 46.72%. However, given that verb-based answers constitute a small fraction of the dataset, accounting for just 3.2% of all questions in the unfiltered test set. Therefore, while this analysis highlights

[2] For this analysis, a “verb” is defined as a word tagged as VERB, from which we excluded words ending in “-ed” to filter out potential passive voice forms.

a qualitative weakness in our pipeline, its overall quantitative impact on the integrity of the automatically generated training and validation sets is likely limited.

5.4 Data Summary

Statistic	Number
Size of dataset	42,750
Unique questions	29,822
Unique answers	4,413
Number of images	8,693
Average question length	7.79
Average answer length	1.42

Table 1: Dataset statistics.

Table 1 summarizes the key properties of the KWordinaryVQA dataset. Crucially, our analysis confirms that key linguistic characteristics are consistently maintained across the training, validation, and test splits (see Appendix E and F). This consistency ensures that our benchmark provides a fair and representative basis for evaluating model performance.

6 Experimental Setup

6.1 Evaluation Metrics

To provide a holistic assessment of model performance on the KWordinaryVQA test set, we employed a comprehensive suite of metrics targeting various dimensions of answer quality. We began by measuring Accuracy, defined as the proportion of predictions that exactly matched the ground-truth answers. This was followed by the computation of standard token-level Precision, Recall, and F1-score (Goutte and Gaussier, 2005) using normalized text. For lexical overlap, we used BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004). Lastly, to evaluate semantic relevance beyond surface-level matching, we computed GPTScore (Fu et al., 2024) using the GPT-4o model via the OpenAI API.

6.2 Traditional Baselines

We first established performance using two traditional methods that treat VQA as a retrieval or classification task over a fixed answer set.

6.2.1 Lexical Retrieval

As a non-parametric benchmark, we adopt a TF-IDF retrieval method (Sparck Jones, 1972) to capture lexical patterns independent of visual input. In this setup, each question is represented as a TF-IDF weighted bag-of-words vector after stopword removal and stemming. For any given test question, the answer of the most lexically similar training question–determined by cosine similarity–is adopted as the prediction. This approach, while computationally efficient, entirely disregards visual input.

6.2.2 Fine-Tuned Classifiers

To evaluate supervised, classification-based VQA, we fine-tuned two models representing distinct architectural paradigms: BEiT-3 (Wang et al., 2023), a state-of-the-art unified model with an end-to-end fused architecture, and LXMERT (Tan and Bansal, 2019), a canonical two-stream model that operates on pre-extracted visual region features.

Both models were fine-tuned on the KWordinaryVQA training set with a classification objective, using a cross-entropy loss over a predefined answer set. To prepare the data for this task, answer labels in the training and validation sets underwent a consistent normalization process, including lowercasing and punctuation removal. Reflecting its two-stream design, LXMERT also required a separate feature extraction step, for which we employed a pretrained Faster R-CNN model (Ren et al., 2015) with a ResNet-50 backbone (He et al., 2016) and Feature Pyramid Network (Lin et al., 2017). Further details on the fine-tuning hyperparameters are provided in Appendix C.

6.3 Generative Model Baselines

Next, we evaluated the performance of several state-of-the-art multimodal Large Language Models (LLMs) under different prompting conditions.

6.3.1 Zero-Shot Evaluation

In the primary evaluation setting, four models were tested under strict zero-shot setting: Llama 3.2-Vision 11B Instruct (Meta AI, 2024), MiniCPM-o 2.6 (Yao et al., 2024), Qwen2.5-VL 7B Instruct (Bai et al., 2025), and [3]Gemini 2.0 Flash (Google DeepMind, 2025). Each model received only the test image (resized to 480 × 480 pixels) and a question, without any in-context examples.

Initial tests with both 'Raw' (unconstrained) and 'Instructed' (concise format) prompting yielded low

[3] Questions were synthesized by DeepSeek-V3, mitigating potential self-enhancement bias in Gemini's evaluation.

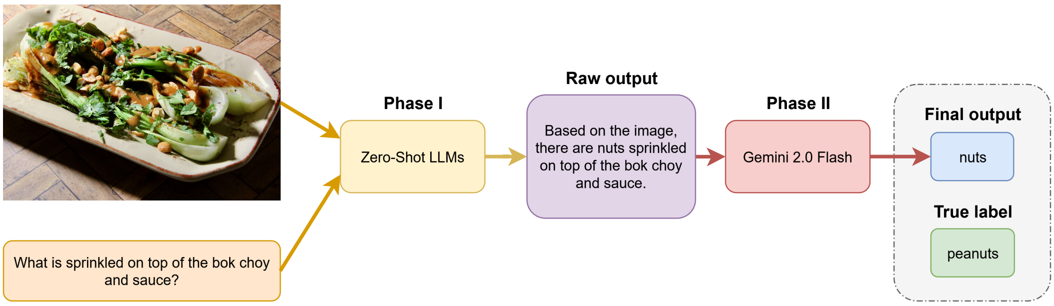

What is sprinkled on top of the bok choy and sauce?
Phase I
Zero-Shot LLMs
Raw output
Based on the image, there are nuts sprinkled on top of the bok choy and sauce.
Phase II
Gemini 2.0 Flash
Final output
nuts
True label
peanuts

Figure 6: The two-phase post-processing pipeline applied to our Zero-Shot VLM baselines. (I) A LLM initially generates a raw, often verbose, answer (e.g., 'The top of the enchiladas is golden brown.') to the visual question. (II) This raw output is then post-processed by Gemini 2.0 Flash to extract a concise final answer (e.g., 'golden brown'), aligning it with the desired concise format.

Model	Accuracy	Precision	Recall	F1-score	BLEU	ROUGE-L	GPTScore
TF-IDF	0.2163	0.2756	0.2762	0.2728	0.2209	0.2981	0.4285
BEiT-3 (fine-tuned)	**0.4804**	0.5543	**0.5436**	**0.5455**	**0.4806**	**0.5780**	**0.6864**
LXMERT (fine-tuned)	0.3574	**0.6203**	0.3574	0.3017	0.3549	0.4457	0.5686

Table 2: Evaluation of classification-based VQA models. Fine-tuned BEiT-3 achieved the best overall performance among non-generative models, while LXMERT showed high precision but lower recall. The TF-IDF baseline, though simple, performed reasonably well in repetitive query scenarios.

scores due to the models' tendency to generate verbose answers. Consequently, we implemented a crucial post-processing step, where an LLM was used to extract a concise, relevant answer from the raw output (see Appendix B). This refinement was applied to all zero-shot results to ensure fair comparison.

6.3.2 Few-Shot Evaluation

To further assess in-context learning capabilities, we conducted few-shot evaluations on two of the models: Llama 3.2-Vision 11B Instruct and MiniCPM-o 2.6. The evaluation was performed under two distinct conditions, utilizing both 5 and 10 demonstration exemplars. These exemplars, carefully curated from the training set, consisted of image-question-answer triplets. This contextual prefix was prepended to each test instance to prime the models towards the appropriate response format. Unlike the zero-shot setting, no post-processing was applied to the few-shot outputs.

7 Results and Analysis

The empirical results of our benchmarking experiments are presented below. These evaluations use a range of current models to probe the difficulty and characteristics of the KWordinaryVQA dataset.

7.1 Traditional and Fine-Tuned Baselines

The performance of the fine-tuned models and the lexical retrieval baseline is presented in Table 2. A substantial performance gap was observed between the fine-tuned models and the TF-IDF retrieval baseline. BEiT-3, the strongest performer in this category, achieved a modest accuracy of 0.4804. This result indicates that the dataset is not easily solved even by powerful, domain-adapted vision-language models, suggesting that successful task completion requires a level of reasoning beyond simple pattern recognition.

7.2 Generative Model Baselines in Zero-Shot Settings

The performance of large generative models in a zero-shot setting further highlights the challenge of KWordinaryVQA (Table 3). Without intervention, all evaluated LLMs performed poorly, primarily due to a systemic failure to produce answers in the required concise format. This outcome demonstrates that the benchmark effectively tests a model's ability to adhere to precise output constraints in addition to its content understanding. While a post-processing pipeline substantially improved the scores–with Gemini 2.0 Flash achieving the highest GPTScore–the necessity of this external step underscores a fundamental difficulty that the dataset exposes in current generative models.

Model	Version	Accuracy	Precision	Recall	F1	BLEU	ROUGE-L	GPTScore
LLaMA 3.2 Vision	Raw	0.0224	0.1045	**0.6439**	0.159	0.0419	0.1591	0.3476
	Instructed	0.2160	0.3489	0.5723	0.3923	0.2437	0.4193	0.6245
	Post-Processed	**0.2958**	**0.4355**	0.5346	**0.4620**	**0.3151**	**0.5066**	**0.6782**
MiniCPM-o 2.6	Raw	*negligible*	0.0699	**0.6665**	0.1160	0.0172	0.1148	0.3787
	Instructed	0.2690	0.4143	0.5533	0.4502	0.2978	0.4908	0.6599
	Post-Processed	**0.3160**	**0.4568**	0.5287	**0.4724**	**0.3275**	**0.5228**	**0.6936**
Qwen2.5-VL	Raw	*negligible*	0.0539	**0.6494**	0.0953	0.0119	0.0910	0.3503
	Instructed	0.0657	0.2443	0.5890	0.3163	0.1062	0.3442	0.6599
	Post-Processed	**0.2609**	**0.4111**	0.5083	**0.4356**	**0.2756**	**0.4832**	**0.6668**
Gemini 2.0 Flash	Raw	0.0370	0.1476	**0.6633**	0.2150	0.0100	0.2490	0.5607
	Instructed	0.1703	0.3153	0.3690	0.3239	0.1815	0.3710	0.6113
	Post-Processed	**0.3533**	**0.4974**	0.5694	**0.5130**	**0.3627**	**0.5597**	**0.7239**

Table 3: Zero-shot performance of generative vision-language models on the KWordinaryVQA benchmark under three evaluation settings: **Raw** denotes direct model output from image-question input without prompt engineering; **Instructed** augments input with manually designed prompts to guide answer generation; and **Post-Processed** applies a secondary model to refine raw outputs for improved alignment with reference answers. Bold values indicate the best scores within each model.

Model	Version	Accuracy	Precision	Recall	F1	BLEU	ROUGE-L	GPTScore
LLaMA 3.2 Vision	5 samples	0.3060	0.4196	0.4866	0.4360	0.3198	0.4755	0.6108
	10 samples	**0.3466**	**0.4559**	**0.5226**	**0.4694**	**0.3568**	**0.5096**	**0.6541**
MiniCPM-o 2.6	5 samples	0.2374	0.3971	0.5472	0.4363	0.2613	0.4810	0.6528
	10 samples	**0.3225**	**0.4546**	**0.5394**	**0.4761**	**0.3400**	**0.5178**	**0.6702**

Table 4: Performance of generative models with few-shot setting. Both LLaMA 3.2 Vision and MiniCPM-o 2.6 struggled to match the concise answer format of the dataset, yielding moderate scores across all metrics. This highlights the importance of output refinement for domain-specific VQA.

7.3 Few-Shot vs. Post-Processed Zero-Shot

Our final set of experiments confirmed the dataset's robustness against common learning strategies. As shown in Table 4, the effectiveness of providing in-context examples was inconsistent. While a 10-shot configuration surpassed the F1-score of the post-processed zero-shot baseline, a 5-shot configuration proved insufficient to achieve the same, indicating that adapting to the dataset's diversity necessitates a substantial number of exemplars. Crucially, a persistent trade-off was observed across both settings: the few-shot approach improved exact-match accuracy but resulted in lower semantic relevance (GPTScore) compared to the post-processed counterpart. This finding demonstrates that the core challenges of KWordinaryVQA, particularly the dual demand for fine-grained reasoning and strict output formatting, are not easily circumvented by simple prompting strategies.

7.4 Inference Cost

Our evaluation reveals critical trade-offs between performance, cost, and computational requirements, as detailed in Table 7. Fine-tuned models, particularly BEiT-3, offer the highest efficiency, delivering moderate accuracy with minimal inference time and no financial cost.

In contrast, large generative models present a more complex cost-benefit profile. Notably, a zero-shot approach combined with our optional post-processing step achieves superior performance to 10-shot prompting but at a fraction of the computational cost and time. This result suggests that for tasks like KWordinaryVQA, refining the output of a cost-effective zero-shot model can be a more pragmatic and effective strategy than computationally expensive few-shot prompting.

7.5 Overall Evaluation

Our collective results reveal that no single modeling paradigm excels across all dimensions of performance, efficiency, and cost on the KWordinaryVQA benchmark. Instead, the findings highlight a series of critical trade-offs that present a nuanced decision for practical applications.

Fine-tuned models, exemplified by BEiT-3, achieve the highest exact-match accuracy and inference efficiency, but require a significant upfront investment in training. Conversely, large generative models offer flexibility and eliminate training costs, with a post-processed zero-shot approach–using Gemini 2.0 Flash–delivering the best semantic relevance. However, this approach's reliance on an external refinement step and the general failure of expensive few-shot prompting underscore the inherent challenges these models face with the dataset's specific constraints.

Ultimately, these complex trade-offs solidify KWordinaryVQA's value as a multifaceted benchmark. It effectively probes models on distinct capabilities–from precise classification to semantic understanding and adherence to formatting–demonstrating that a truly robust system for food-domain VQA must balance these competing demands.

8 Conclusion

In this paper, we introduced KWordinaryVQA, a large-scale, automatically generated dataset for visual question answering in the food domain. Our primary contribution is a novel, challenging benchmark designed to probe the reasoning capabilities of modern vision-language models. Our comprehensive evaluations demonstrate that while no single modeling paradigm excels across all metrics, a series of critical trade-offs exist between accuracy, semantic relevance, and computational efficiency.

The empirical results underscore the dataset's difficulty. We found that even powerful, fine-tuned models like BEiT-3 achieve only modest accuracy, while large generative models struggle with the dataset's concise formatting requirements, necessitating external post-processing steps. Furthermore, our experiments revealed that for deployment, refining the output of a cost-effective zero-shot model can be a more pragmatic and effective strategy than computationally expensive few-shot prompting.

Our work has two main limitations that open avenues for future research. First, the automated generation pipeline introduces a degree of semantic noise. Future work should focus on developing methods for automated noise filtering to further enhance dataset integrity. Second, our post-processing technique relies on a static answer-length threshold, limiting its applicability in dynamic, real-world systems. We believe that developing adaptive output refinement strategies is a crucial next step.

By publicly releasing the KWordinaryVQA dataset, our generation pipeline, and evaluation code, we aim to facilitate further research into developing more robust and accurate food-centric AI systems and to provide a valuable resource for benchmarking in this specialized domain.

Limitations

A foundational limitation of our study is the integrity of the dataset itself, which is constrained by semantic noise from the automated generation process. This noise manifests as factually incorrect ground-truth answers, where the generated text fails to align with the visual evidence (detailed in Section 5.3). The implications of this label noise are twofold. First, it corrupts the learning signal during training, potentially forcing the model to form erroneous associations rather than robust, generalizable knowledge. Second, it complicates evaluation, as a model providing a visually faithful answer may be marked incorrect, leading to an underestimation of its actual reasoning abilities. Consequently, our reported results should be viewed as a conservative baseline, acknowledging that model performance is likely suppressed by these data artifacts.

Beyond the data itself, a second limitation lies in the practical applicability of our post-processing technique. The method relies on a pre-determined threshold for answer length–for instance, instructing a model to generate 'no more than 5 words'. Setting an optimal threshold requires analyzing the entire dataset in advance to understand its global statistics. This assumption of having full, prior knowledge of the corpus is feasible for static, offline benchmarks but is unrealistic for real-world, dynamic systems where data arrives sequentially. This dependency thus restricts the direct deployment of this specific method and underscores the need for more adaptive post-processing strategies in future work.

Acknowledgments

Huy Trieu is supported by the research funding from the Faculty of Information Technology, University of Science, VNU-HCM, Vietnam. This work was also supported in part by the Air Force Office of Scientific Research under award number FA2386-24-1-4034 granted to Tung Le.

References

Pulkit Agarwal, Settaluri Sravanthi, and Pushpak Bhattacharyya. 2024. Indifoodvqa: Advancing visual question answering and reasoning with a knowledge-infused synthetic data generation pipeline. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1158–1176.

Allrecipes contributors. 2025. Allrecipes. Accessed: 2025-03-20.

Max Bachmann. 2024. Rapidfuzz. Version 3.12.1.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Vladimir Blagojevic. 2024. Bert english uncased fine-tuned pos. Accessed: 2025-04-15.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Google DeepMind. 2025. Gemini 2.0 flash. Accessed: 2025-03-25.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in Information Retrieval*, pages 345–359, Berlin, Heidelberg. Springer Berlin Heidelberg.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Meta AI. 2024. Llama 3.2 11b vision instruct. Released under the Llama 3.2 Community License. Accessed: 2025-04-20.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong

Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Ching Lam Cheng, Daud Abolade, Emmanuele Chersoni, and 32 others. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *CoRR*, abs/2410.12705.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Appendices

A Prompts for Data Generation

A.1 Image Description Generation

```
The following food items are present in
    this image: {food_name}.
Describe the color and relative
    location of each food item in
    details.

Instruction:
- Only print the final description, do
    not print anything else like
    headers or human-like response!
- The summary has maximum 200 words
- Do not break the line!
```

A.2 Keyword Extraction

```
Given a summary of an image for the VQA
    task, extract the top 6 important
    keywords, ensuring diversity in
    word types, including at least one
    verb, one adjective, one noun, one
    color, and one preposition (if
    present).

Summary: {summary}
Instruction:
- Only print the extracted keywords as
    a comma-separated list.
- If there is a preposition, include
    the full phrase containing it
    (e.g., 'on the table' instead of
    just 'on').
- Do not print anything else like
    headers or human-like responses!
```

A.3 Question-Answer Generation

```
Based on these keywords: {keywords} And
    the food image description:
    {summary}
Please generate one simple question per
    keyword where:
1. Each question is based on the food
    image description.
2. The answer to each question must
    exactly match its corresponding
    keyword
3. The number of questions must be
    equal to the number of keywords
4. For keywords that are prepositions
    (e.g., on the table), ask a Where
    question.
5. The order of the questions must
    match the order of the keywords.
Instruction:
- Generate only the questions as a
    comma-separated list.
- Do not include headers, explanations,
    or human-like responses.
```

B Prompt for Post-Processing

```
You are a helpful VQA assistant.
Your task is to extract a single, most
    relevant answer to the question,
    based on the given prediction text.
The answer must:
- Directly address the question's intent
- Contain no more than 5 words
- Be written entirely in lowercase
    letters
- Not include any commas, lists, or
    explanations
- Be concise and natural, like a
    typical VQA answer (e.g., 'red
    shirt', 'top left', 'enchiladas',
    'golden-brown', etc.)
- If multiple candidates appear in the
    prediction, select the one most
    relevant to the question
- Respond with only the final answer
    and nothing else

Question: {question}
Predict: {prediction}
```

C Training Configuration

Hyperparameter	Value
Optimizer	Adam
Learning rate	5×10^{-5}
Batch size	8
GPU	NVIDIA Tesla P100

Table 5: Training configuration for BEiT-3 and LXMERT.

For the BEiT-3 model, we fine-tuned only the classification head and the text embedding layers over 4 epochs, which took approximately 6 hours, while keeping the remaining parameters frozen. The LXMERT model was fine-tuned for 6 epochs, requiring about 1 hour.

D Illustrative Examples

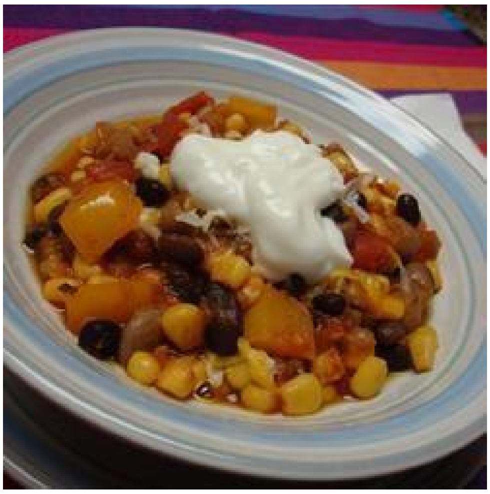
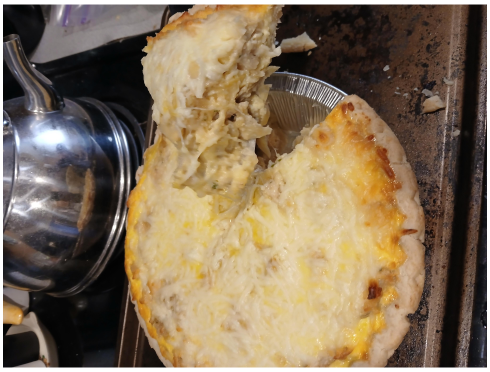
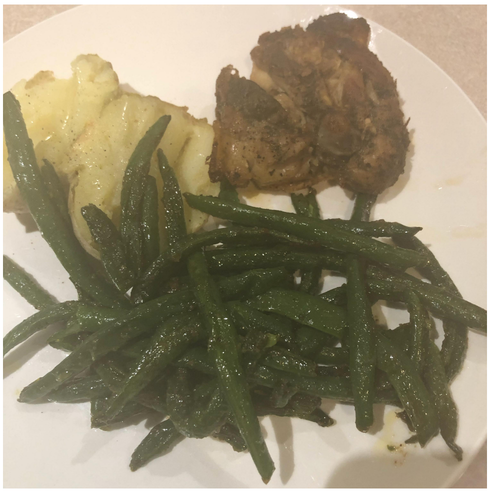
Image	**Question & Answer**
	**Question:** What color are the corn kernels in the stew? **Answer:** yellow **Question:** Where is the dollop of sour cream located? **Answer:** atop the center
	**Question:** What is the state of the cheese layer on top of the pie? **Answer:** melted **Question:** What type of food is primarily located on the right side of the image? **Answer:** pie **Question:** What color is the crust of the pie? **Answer:** tan **Question:** What color is the baking tin the pie is sitting in? **Answer:** silver
	**Question:** What color are the string beans? **Answer:** green **Question:** How are the string beans arranged in the image? **Answer:** tangled **Question:** How are the string beans arranged in the image? **Answer:** tangled **Question:** Do the string beans overlap any other food in the image? **Answer:** yes **Question:** Where are the string beans located in the image? **Answer:** in the bottom center

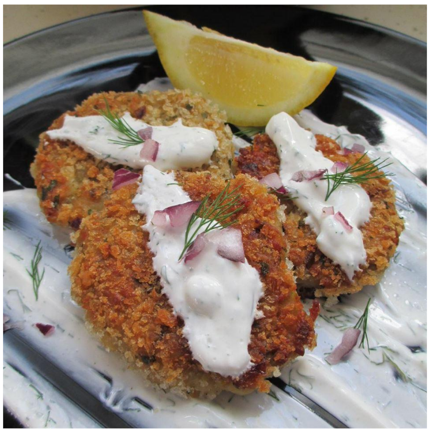
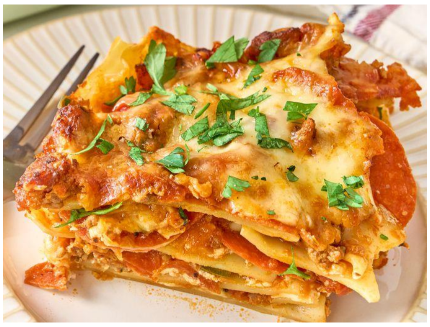
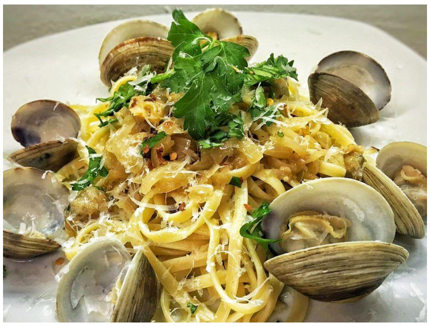
Image	Question & Answer
	**Question:** What color are the falafels? **Answer:** golden-brown **Question:** What color is the lemon? **Answer:** yellow **Question:** What type of food is shown in the image? **Answer:** falafels
	**Question:** What is the predominant color of the lasagna noodles? **Answer:** yellow **Question:** What is the name of the dish described? **Answer:** lasagna **Question:** Where are the fresh green parsley leaves placed? **Answer:** on top
	**Question:** What type of pasta is in the center of the dish? **Answer:** linguine **Question:** What color is the shredded cheese? **Answer:** white

Table 6: Illustrative examples from the KWordinaryVQA dataset.

E Length Distribution

E.1 Training Set

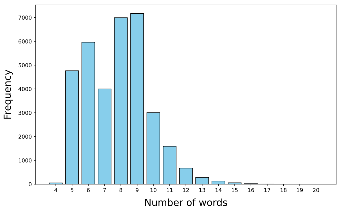


Figure 7: Question lengths distribution in training set.

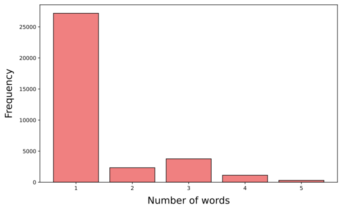


Figure 8: Answer lengths distribution in training set.

E.2 Validation Set

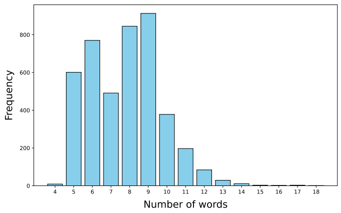


Figure 9: Question lengths distribution in validation set.

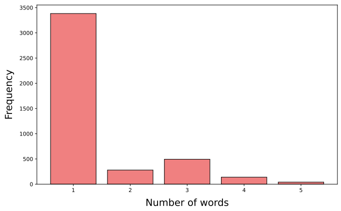


Figure 10: Answer lengths distribution in validation set.

E.3 Test Set

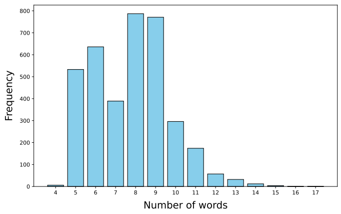


Figure 11: Question lengths distribution in test set.

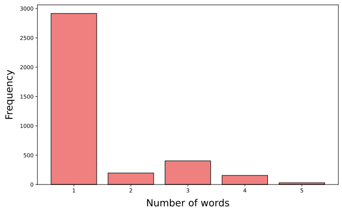


Figure 12: Answer lengths distribution in test set.

F Question Types Distribution

F.1 Training Set

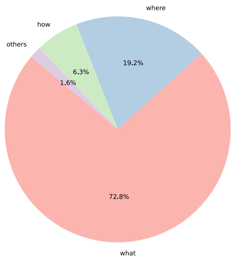


Figure 13: Distribution of question types in training set.

F.2 Validation Set

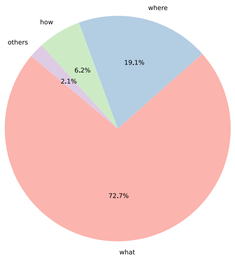


Figure 14: Distribution of question types in validation set.

F.3 Test Set

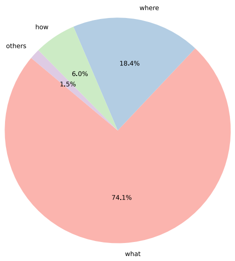


Figure 15: Distribution of question types in test set.

Model	Type	Time (hours)	Cost (VND)	GPU	Overall Evaluation
TF-IDF (Sparck Jones, 1972)	computed	**< 0.01**	**Free**	**None (CPU)**	Very Low Accuracy - Free
BEiT-3 (Wang et al., 2023)	fine-tuned	**0.06**	**Free**	P100	**Moderate Accuracy - Free**
LXMERT (Tan and Bansal, 2019)	fine-tuned	0.134	**Free**	P100	Low Accuracy - Free
Llama 3.2 Vision (Meta AI, 2024)	zero-shot	1.5	8,000	P40	Low Accuracy - Moderate Cost
Llama 3.2 Vision (Meta AI, 2024)	few-shot	4.5	100,000	A100 PCIE	Low Accuracy - Very High Cost
MiniCPM-o 2.6 (Yao et al., 2024)	zero-shot	1.5	8,000	P40	Low Accuracy - Moderate Cost
MiniCPM-o 2.6 (Yao et al., 2024)	few-shot	2.725	30,000	RTX 5090	Low Accuracy - Very High Cost
Qwen2.5-VL (Bai et al., 2025)	zero-shot	1.5	8,000	RTX 3090	Low Accuracy - Moderate Cost
Gemini 2.0 Flash (Google DeepMind, 2025)	zero-shot	1.5	**Free**	**None (API)**	Very Low Accuracy - Free

Table 7: Inference time, estimated cost (VND), GPU type, and qualitative evaluation for each model on the KWordinaryVQA test set. Zero-shot results are reported without including any post-processing time; optionally applying a post-processing step would add approximately 1.5 hours to the total inference time. For few-shot models, we only report results for the setting with 10 in-context examples. Reported inference time includes only model execution and excludes any additional refinement steps.