TaiwanVQA:
A Benchmark for Visual Question Answering for Taiwanese Daily Life

Hsin-Yi Hsieh[†] [1] Shang-Wei Liu[†] [1] Chih-Chang Meng[‡] [2] Chien-Hua Chen[‡] [2]
Shuo-Yueh Lin[§3] Hung-Ju Lin[⋆4] Hen-Hsen Huang[¶5] I Chen Wu[‡] [2]

[†]National Center for High-performance Computing, Taiwan [§]National Central University, Taiwan
[‡]National Yang Ming Chiao Tung University, Taiwan [⋆]National Taiwan University, Taiwan
[¶]Institute of Information Science, Academia Sinica, Taiwan

[1]{2403055,2403056}@narlabs.org.tw [2]{mcc.cs11,chchen.cs12,icwu}@nycu.edu.tw
[3]johnnylin@g.ncu.edu.tw [4]r11922147@csie.ntu.edu.tw [5]hhhuang@iis.sinica.edu.tw

Abstract

We introduce **TaiwanVQA**, a novel visual question answering benchmark designed to evaluate vision language models' (VLMs) ability to recognize and reason about Taiwan-specific multimodal content. TaiwanVQA comprises 2,000 image-question pairs covering diverse topics relevant to Taiwanese culture and daily life. We categorize the questions into recognition and reasoning tasks, further subclassifying reasoning questions based on the level of external knowledge required. We conduct extensive experiments on state-of-the-art VLMs, including GPT-4o, Llama-3.2, LLaVA, Qwen2-VL, and InternVL2 models. Our findings reveal significant limitations in current VLMs when handling culturally specific content. The performance gap widens between recognition tasks (top score 73.60%) and reasoning tasks (top score 49.80%), indicating challenges in cultural inference and contextual understanding. These results highlight the need for more culturally diverse training data and improved model architectures that can better integrate visual and textual information within specific cultural contexts. By providing TaiwanVQA, we aim to contribute to the development of more inclusive and culturally aware AI models, facilitating their deployment in diverse real-world settings. TaiwanVQA can be accessed on our GitHub page.

1 Introduction

Multimodal vision-language models (VLMs) have achieved remarkable success in integrating visual and textual information, enabling applications ranging from image captioning to visual question answering (Li et al., 2023; Dai et al., 2023). Despite these advances, most existing benchmarks focus on general-domain knowledge and widely spoken languages, often overlooking the challenges posed by culturally specific content and underrepresented languages (Yue et al., 2024a,b; Fu et al., 2024).

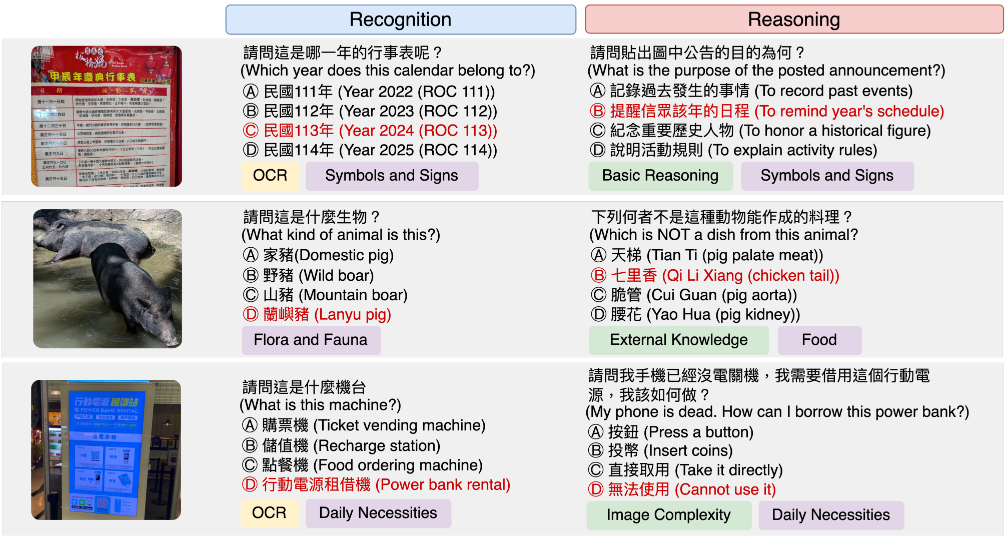
Understanding and reasoning about culturally nuanced content is crucial for deploying AI systems in diverse real-world settings (Nayak et al., 2024). For instance, accurately interpreting traditional symbols, local customs, or region-specific artifacts requires models to possess not only visual recognition capabilities but also contextual and cultural knowledge (Hershcovich et al., 2022).

To address this gap, we introduce **TaiwanVQA**, a visual question answering benchmark specifically designed to evaluate VLMs' abilities to recognize and reason about Taiwan-specific content. TaiwanVQA comprises 1,000 images paired with 2,000 questions covering a diverse range of topics relevant to Taiwanese daily life and culture, such as traditional cuisine, local festivals, historical landmarks, and public signage. Our contributions are threefold:

- We introduce TaiwanVQA, the first VQA benchmark specifically designed for Taiwanese cultural content, with data categorized based on aspects of daily life
- We propose a taxonomy of culture-specific visual questions into recognition and reasoning types, with reasoning questions sub-classified based on required external knowledge levels
- We provide comprehensive experiments on state-of-the-art VLMs including GPT-4 (OpenAI, 2023), revealing their limitations in handling culture-specific content.

Our findings indicate that while models perform reasonably well on recognition tasks, their performance significantly drops on reasoning tasks that require deeper cultural understanding. This underscores the need for more culturally diverse training data and enhanced model architectures capable of integrating visual and textual information within specific cultural contexts.

Proceedings of the First Workshop of Evaluation of Multi-Modal Generation, pages 57–75
January 20, 2025. ©2025 Association for Computational Linguistics



請問這是哪一年的行事表呢？
(Which year does this calendar belong to?)
Ⓐ 民國111年 (Year 2022 (ROC 111))
Ⓑ 民國112年 (Year 2023 (ROC 112))
Ⓒ 民國113年 (Year 2024 (ROC 113))
Ⓓ 民國114年 (Year 2025 (ROC 114))

請問貼出圖中公告的目的為何？
(What is the purpose of the posted announcement?)
Ⓐ 記錄過去發生的事情 (To record past events)
Ⓑ 提醒信眾該年的日程 (To remind year's schedule)
Ⓒ 紀念重要歷史人物 (To honor a historical figure)
Ⓓ 說明活動規則 (To explain activity rules)

請問這是什麼生物？
(What kind of animal is this?)
Ⓐ 家豬(Domestic pig)
Ⓑ 野豬 (Wild boar)
Ⓒ 山豬 (Mountain boar)
Ⓓ 蘭嶼豬 (Lanyu pig)

下列何者不是這種動物能作成的料理？
(Which is NOT a dish from this animal?
Ⓐ 天梯 (Tian Ti (pig palate meat))
Ⓑ 七里香 (Qi Li Xiang (chicken tail))
Ⓒ 脆管 (Cui Guan (pig aorta))
Ⓓ 腰花 (Yao Hua (pig kidney))

請問這是什麼機台
(What is this machine?)
Ⓐ 購票機 (Ticket vending machine)
Ⓑ 儲值機 (Recharge station)
Ⓒ 點餐機 (Food ordering machine)
Ⓓ 行動電源租借機 (Power bank rental)

請問我手機已經沒電關機，我需要借用這個行動電源，我該如何做？
(My phone is dead. How can I borrow this power bank?)
Ⓐ 按鈕 (Press a button)
Ⓑ 投幣 (Insert coins)
Ⓒ 直接取用 (Take it directly)
Ⓓ 無法使用 (Cannot use it)

Figure 1: An Illustration of the TaiwanVQA Benchmark. Each row shows an image paired with two questions: a recognition question (left) and a reasoning question (right), both in multiple-choice format with the correct answers highlighted in red. Below each question, topic categories are labeled in purple (e.g., "Symbols and Signs", "Daily Necessities"), with additional labels in yellow for OCR requirements in recognition questions and in green for knowledge types in reasoning questions.

By providing TaiwanVQA, we aim to contribute to the development of more inclusive and culturally aware AI models, facilitating their deployment in diverse real-world scenarios and promoting research in underrepresented languages and cultures.

2 Related Work

The evaluation of VLMs has progressed from general visual recognition to understanding culturally specific content. Early datasets like DOLLAR STREET (Rojas et al., 2022) and GLDV2 (Weyand et al., 2020) provided extensive collections of images from diverse regions but focused primarily on recognition tasks without delving into cultural nuances.

Recent benchmarks have aimed to directly assess cultural understanding in VLMs. Burda-Lassen et al. (2024) introduced MOSAIC-1.5K, a culture-specific captioning dataset that includes images from various regions to test models' cultural awareness in captioning tasks. Similarly, Bhatia et al. (2024) proposed GLOBALRG, evaluating retrieval and grounding capabilities across 15 countries, emphasizing local concepts within a global context. Nayak et al. (2024) introduced the CULTUREVQA dataset, a benchmark designed to evaluate VLMs on cultural understanding across multiple countries and cultures. CULTUREVQA comprises 2,378 image-question pairs from 11 countries spanning 5 continents, with questions focusing on traditions, rituals, and cultural artifacts. While this dataset advances the evaluation of cultural understanding in VLMs, it allocates a smaller proportion of its dataset to traditions and rituals compared to our benchmark and uses a multiple-choice evaluation format, which may not fully capture the depth of models' cultural reasoning capabilities.

Other efforts target more specific cultural domains. Li et al. (2024b) introduced FOODIEQA, which examines fine-grained understanding of Chinese food culture through multiple-choice tasks. Although it addresses a culturally rich dimension (food), current VLMs still lag behind human-level performance, especially on image-based tasks. Meanwhile, Liu et al. (2021) proposed MARVL, focusing on visually grounded reasoning across multiple languages and cultures, but it does not explicitly assess rich cultural common sense related to traditions and also utilizes a true/false format.

Our work differs by focusing specifically on the Taiwanese cultural context, providing an in-depth evaluation of VLMs' abilities to understand and reason about Taiwan-specific content. TaiwanVQA includes 2,000 image-question pairs with a significant emphasis on traditions, rituals, and daily life. We adopt a multiple-choice format, and ensure diverse and carefully designed distractors to challenge the models' cultural understanding. By categorizing questions into recognition and reasoning tasks, and further sub-classifying reasoning

	w/ OCR	w/o OCR	All
Recognition	339	661	1,000

	Basic	External Knowledge	Image Complexity	All
Reasoning	246	674	80	1,000

Table 1: Statistics of Recognition and Reasoning Questions by Types

questions based on the level of external knowledge required, our benchmark offers a comprehensive assessment of VLMs' cultural understanding within a specific regional context. This structured approach enables a more detailed analysis of models' capabilities and limitations in handling culturally rich content.

3 TaiwanVQA

3.1 Tasks

In constructing TaiwanVQA, we were inspired by two recent VLM evaluation benchmarks: MME(Fu et al., 2024) and TransportationGames(Zhang et al., 2024). MME's division of questions into perception and cognition guided our approach, as understanding Taiwan-related visual content requires both basic recognition and deeper reasoning. Thus, we structured TaiwanVQA by assigning two questions to each image to fully assess models' understanding of Taiwanese culture and knowledge:

- **Recognition Questions** – These questions evaluate models' ability to accurately identify Taiwan-specific visual elements, including local cuisine, transportation facilities, native ecology, and folk activities.
- **Reasoning Questions** – These questions test models' advanced analytical abilities, requiring them to not only identify visual elements but also understand relationships between them (such as spatial relations, usage contexts, and cultural implications), integrating local Taiwanese knowledge to reach accurate conclusions.

Within *recognition questions*, we specifically marked those requiring Optical Character Recognition (OCR) capabilities. These questions assess models' ability to recognize Traditional Chinese text in images, crucial for understanding Taiwan's visual elements such as public signs and notices.

Additionally, to better evaluate models' reasoning capabilities, we further categorize *reasoning questions* into three types:

- **Basic Reasoning Required** - Questions that can be answered through straightforward inference from the image content, requiring no external knowledge.
- **External Knowledge Required** - Questions that cannot be answered through image content alone, requiring specific knowledge about Taiwanese culture, customs, or context for accurate responses.
- **Image Complexity Required** - Images contain multiple visual elements or complex spatial relationships, requiring deep visual analysis for accurate judgment.

A detailed annotation process for both task types can be found in Appendix A, and Table 1 shows the statistical distribution across different types.

3.2 Data Collection

To construct the TaiwanVQA dataset, we selected 1,000 representative images of Taiwan, each paired with one identification and reasoning question, generating 2,000 questions in total. Due to licensing concerns, all images and questions were manually designed. We recruited 9 annotators from diverse backgrounds (varying in residence location, ethnic identity, gender, and academic fields), who underwent a week-long training before formal annotation. Detailed annotation guidelines can be found in Appendix A.

Beyond the task type classification in subsection 3.1, to ensure comprehensive coverage of Taiwan's daily life and cultural aspects, we established a question classification framework comprising 13 topics and 27 subtopics. We employed GPT-4o to perform the classification tasks to ensure consistency throughout the dataset. As shown in Figure 2, our questions primarily focus on signs and food culture, as these elements are most closely related to Taiwanese daily life. The remaining questions are evenly distributed across other categories, demonstrating the diversity of our data. Detailed classification criteria and prompts used can be found in Appendix B.

3.3 Data Quality

To validate the quality of TaiwanVQA benchmark, evaluation was performed by annotators on 10% randomly sampled data across three aspects:

- **Question Type Correctness** - compliance

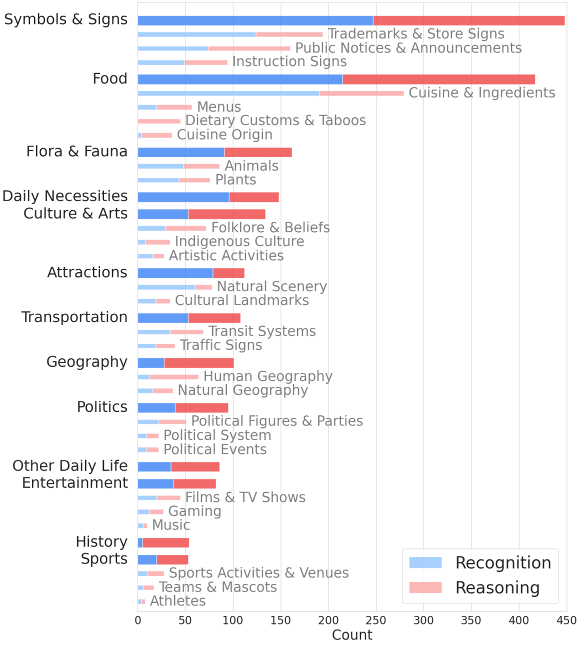


Figure 2: Distribution of Question Categories. The blue and red bars stand for the recognition and reasoning questions respectively. The darker bars represent the total number of questions in the topic, and the lighter bars represent the number of questions in the sub topics under the topic. If there are no shallower bars, it means that the topic has no sub topics, such as Daily Necessities.

	Q1: Recog. Compl.	Q2: Reas. Compl.	Q3: Topic Approp.	Q4: Subtopic Approp.	Q5: Question Clarity	Q6: Img Clarity	Q7: Img Need
A1	91	98	89.7	86.1	100	98	99.5
A2	88	99	88.1	85.1	99.5	100	100
A3	94	93	94.8	92.3	99	98.5	98
A4	93	91	96.4	90.2	98.5	94	98.5
Avg.	91.5	95.3	92.3	88.4	99.3	97.6	99
N	100	100	194	194	200	200	200

Table 2: Results of quality assessment, reported in Accuracy (%). Four annotators (A1-A4) evaluated sampled data across three aspects: question compliance (Recognition Q1 and Reasoning Q2), topic appropriateness (Topic Q3 and Subtopic Q4), and clarity (Question Q5, Image Q6, and Image Need Q7). N indicates the number of samples evaluated for each question.

with recognition and reasoning question design guidelines

- **Topic Classification Appropriateness** - compliance with topic and subtopic classification definitions
- **Content Clarity** - question comprehensibility, image clarity, and the necessity of the image for answering the question

As shown in Table 2, all criteria achieved over 85% agreement rate from annotators, demonstrating high consistency in question design and content presentation.

	Overall		Recognition		Reasoning	
Model	w/	w/o	w/	w/o	w/	w/o
GPT-4o	61.7	12.5	73.6	7.7	49.8	17.2
Llama-3.2-90B	51.6	11.1	61.8	7.0	41.4	15.2
InternVL2-76B	64.3	21.9	75.9	17.2	52.6	26.5
Qwen2-VL-72B	75.0	24.8	83.7	18.3	66.2	31.3

Table 3: Performances of VLMs in Normal and Text-only Conditions. The Accuracies (%) evaluated with images (w/) and without images (w/o) are reported.

Furthermore, to validate the necessity of visual information, in addition to the previously mentioned manual inspection of image dependency (Table 2, Q7), we compared four major VLMs' performance with and without images. Table 3 shows that all models performed significantly worse in text-only (w/o) conditions, confirming that our benchmark requires visual reasoning capabilities for accurate answers.

4 Experiments

4.1 Evaluation Strategy

Prompting Approach In our experiments, we design a standardized prompt structure to ensure consistent model evaluation. Figure 3 presents our prompt template used during the evaluation process. To directly assess models' intrinsic instruction-following capabilities, we conduct our evaluation in a zero-shot setting.

Scoring Method To obtain model predictions, we select the option token ("A", "B", "C", or "D") that receives the highest probability among the 20 most probable tokens in the model's output distribution. If none of the option tokens appear in these 20 tokens, the prediction is marked as null and counted as incorrect. We evaluate performance using accuracy as our primary metric, calculated as Accuracy $= \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100\%$, where N_{correct} represents the number of correctly answered questions, and N_{total} represents the total number of questions in our benchmark.

Robust Evaluation Recognizing the sensitivity of language models to the ordering of options in multiple-choice questions (Pezeshkpour and Hruschka, 2023), we adopt the **CircularEval** strategy proposed by (Liu et al., 2024). Details of this approach are provided in Appendix C. This strategy evaluates model responses across four iterations, each applying a circular shift to the answer choices.

[Question content]
有以下幾個選項： (Here are the following options:)
A. <Option A>
B. <Option B>
C. <Option C>
D. <Option D>

請直接使用所提供的選項字母作為答案回答。 (Please answer directly with the option letter provided.)

Figure 3: The Prompt Template for the Zero-shot Setting

A question is considered correctly answered only if the model provides the accurate answer in all iterations, ensuring robustness against option positioning.

4.2 Experimental Setup

Models We evaluate our benchmark using a diverse set of vision-language models, including both open-source and proprietary models. For open-source models, we include: (1) leading multilingual VLMs; and (2) Chinese-based VLMs, which are VLMs that integrate large language models developed in countries where Chinese is the native language. We also include different versions from a proprietary model series. A comprehensive list of the evaluated models and their specifications is provided in Table 9(Appendix D).

Implementation Details Proprietary models are evaluated through OpenAI's API, while open-source models are deployed in containers using the vLLM framework (Kwon et al., 2023). This setup maintains API consistency across all evaluations, facilitating fair comparisons. Due to API constraints, we can only access the 20 most probable tokens from the model's output distribution. All open-source models are hosted on DGX-1 V100 GPUs. Our evaluation pipeline is built upon lmms-eval[1] with modifications to accommodate our experimental requirements. Detailed implementation information, including chat completion parameters and model deployment configurations, is provided in Appendix D.

4.3 Results

We evaluate eleven VLMs and present their performance in three aspects. Table 4 shows the overall performance and results on two question types: Recognition and Reasoning. We further examine model performance across different topics for Recognition (Table 5) and Reasoning (Table 7). For more detailed analysis, we break down the performance by subtopics; complete results are in Appendix E.

Model	Overall	Recognition	Reasoning
Phi3.5-Vision-Instruct	29.95	33.80	26.10
Llama-3.2-11B	33.10	46.80	19.40
Llama-3.2-90B	51.60	61.80	41.40
LLaVA-v1.6-mistral-7B	28.90	33.50	24.30
LLaVA-v1.6-34B	49.50	57.80	41.20
InternVL2-8B	60.45	71.80	49.10
InternVL2-76B	64.25	75.90	52.60
Qwen2-VL-7B	65.35	79.40	51.30
Qwen2-VL-72B	**74.95**	**83.70**	**66.20**
GPT-4o	61.70	73.60	49.80
GPT-4o-mini	50.05	59.80	40.30

Table 4: Performance (in Accuracy, %) Comparison on Overall Performance and Two Question Types: Recognition and Reasoning

5 Analysis

5.1 Recognition and Reasoning Performance

Table 4 shows the performance variations across models in recognition and reasoning tasks related to Taiwan. Among the evaluated models, Qwen2-VL-72B demonstrates the highest overall score (74.95), significantly outperforming other models in both recognition (83.70) and reasoning (66.20). This indicates its robust capability to handle diverse knowledge-intensive tasks. Conversely, smaller models, such as LLaVA-v1.6-mistral-7B and Phi3.5-Vision-Instruct, exhibit lower scores in both categories, suggesting that model size and architectural sophistication are critical for domain-specific generalization.

Generally, model performance tends to scale with size, with larger models typically outperforming smaller ones. However, the results reveal an exception to this trend: Qwen2-VL-7B and InternVL2-8B both outperform larger models such as LLaVA-v1.6-34B and Llama-3.2-90B in both recognition and reasoning tasks. This suggests that, within our benchmark, InternVL2 and Qwen exhibit superior capabilities in both cognitive tasks and Taiwan-specific reasoning, demonstrating a clear advantage over Llama and LLaVA despite their smaller scale.

5.2 Recognition Questions

Recognition questions in the Taiwan Vision Benchmark test models on identifying Taiwan-specific

[1] https://github.com/EvolvingLMMs-Lab/lmms-eval

Model	S&S	Att	Food	Trans	C&A	Pol	Geo	Spo	F&F	His	Ent	DN	ODL
Phi3.5-Vision-Instruct	34.82	22.78	32.56	37.74	35.85	20.00	35.71	45.00	26.37	20.00	31.58	47.92	42.86
Llama-3.2-11B	55.87	35.44	42.79	49.06	33.96	37.50	35.71	65.00	34.07	40.00	55.26	64.58	34.29
Llama-3.2-90B	68.83	46.84	62.79	62.26	62.26	50.00	60.71	65.00	46.15	20.00	71.05	73.96	54.29
LLaVA-v1.6-mistral-7B	35.63	20.25	31.16	50.94	26.42	25.00	21.43	30.00	30.77	20.00	42.11	47.92	28.57
LLaVA-v1.6-34B	57.49	55.70	57.21	67.92	54.72	45.00	60.71	65.00	48.35	20.00	57.89	72.92	54.29
InternVL2-8B	82.59	60.76	65.12	71.70	77.36	67.50	75.00	75.00	62.64	60.00	71.05	72.92	77.14
InternVL2-76B	82.59	68.35	73.49	71.70	84.91	60.00	67.86	**90.00**	63.74	60.00	**84.21**	82.29	77.14
Qwen2-VL-7B	87.45	68.35	76.28	77.36	75.47	82.50	**89.29**	**90.00**	67.03	**80.00**	81.58	79.17	**88.57**
Qwen2-VL-72B	**89.88**	**78.48**	**82.79**	**73.58**	**88.68**	**90.00**	**89.29**	**90.00**	**68.13**	**80.00**	**84.21**	**84.38**	**88.57**
GPT-4o	76.52	72.15	77.21	67.92	73.58	62.50	85.71	85.00	59.34	**80.00**	76.32	77.08	62.86
GPT-4o-mini	68.42	53.16	58.60	60.38	49.06	57.50	60.71	70.00	48.35	60.00	57.89	65.62	48.57

Table 5: Performances of **recognition questions** across different models and topics, including Symbols and Signs (S&S), Attractions (Att), Food, Transportation (Trans), Culture and Arts (C&A), Politics (Pol), Geography (Geo), Sports (Spo), Flora and Fauna (F&F), History (His), Entertainment (Ent), Daily Necessities (DN), and Other Daily Life (ODL). All results are reported in Accuracy (%).

visual elements like local cuisine, transportation, native ecology, and cultural artifacts. These tasks focus on precise object detection without requiring advanced contextual reasoning.

General Patterns and High-Performing Topics Models performed best in visually distinct and simpler categories like *Transportation*, *Symbols and Signs*, and *Sports*. Qwen2-VL-72B excelled, achieving over 89% in *Geography* and *Symbols and Signs*, while InternVL2-76B also performed well, particularly in *Symbols and Signs* (82.59%) and *Daily Necessities* (82.29%). Table 5 highlights Qwen2-VL-72B's dominance and InternVL's strength. Categories like *Food* and *Daily Necessities* further show models' effectiveness in recognizing familiar objects. The high accuracy of Qwen2-VL and InternVL models reflects their robust architectures and multilingual training, enabling strong performance with Traditional Chinese text.

Challenging Topics Across Models Despite overall progress in recognition tasks, certain topics posed significant challenges, particularly those requiring nuanced cultural understanding or visual differentiation. Categories such as *Politics*, *Flora and Fauna*, and *History* consistently recorded lower accuracy, with models like Phi3.5-Vision-Instruct scoring as low as 20% in *Politics*. Table 5 shows a pronounced dip in performance for smaller and less advanced models like LLaVA-v1.6-mistral-7B across complex topics. Text-heavy categories, such as *Politics* and *Culture and Arts*, were particularly difficult for models without some cultural knowledge of Taiwanese culture. These findings emphasize the need for enriched cultural datasets and improved linguistic understanding to enhance performance in these challenging areas.

Comparison of Models The Qwen2-VL models outperformed others in recognition tasks, with the 72B model excelling in Politics (90.00%), Geography (89.29%), and Culture and Arts (88.68%). The smaller 7B version also performed well in visually distinct areas like Symbols and Signs (87.45%). InternVL models were balanced, with the 76B model strong in Symbols and Signs (82.59%) and Daily Necessities (82.29%) but slightly behind Qwen2 in nuanced tasks. GPT models excelled in reasoning-heavy areas like History (80.00%) and Sports (85.00%) but struggled in visual categories, especially smaller versions. LLaVA models, even the larger 34B version, lagged in nuanced areas like Politics (45.00%). Overall, Qwen2-VL led in accuracy, highlighting the importance of model size and training depth.

Model	w/ OCR	w/o OCR
Phi3.5-Vision-Instruct	31.56	34.95
Llama-3.2-11B	47.20	46.60
Llama-3.2-90B	59.59	62.93
LLaVA-v1.6-mistral-7B	23.89	38.43
LLaVA-v1.6-34B	49.26	62.18
InternVL2-8B	84.96	65.05
InternVL2-76B	83.19	72.16
Qwen2-VL-7B	92.63	72.62
Qwen2-VL-72B	93.51	78.67
GPT-4o	75.81	72.47
GPT-4o-mini	63.72	57.79

Table 6: Performances of Recognition Task with and without OCR, reported in Accuracy (%)

OCR and Text Recognition As shown in Table 6 the OCR capabilities of the Phi, Llama, and GPT series models are similar to their performance in general QA tasks, showing no significant differentiation. In contrast, the LLaVA series struggles noticeably with OCR-related questions. Notably,

Model	S&S	Att	Food	Trans	C&A	Pol	Geo	Spo	F&F	His	Ent	DN	ODL
Phi3.5-Vision-Instruct	31.34	24.24	20.79	25.45	30.86	23.64	23.29	39.39	25.35	20.41	27.27	30.77	19.61
Llama-3.2-11B	26.37	15.15	17.33	20.00	20.99	7.27	10.96	36.36	11.27	14.29	34.09	23.08	13.73
Llama-3.2-90B	54.23	30.30	38.61	40.00	39.51	27.27	20.55	51.52	40.85	40.82	45.45	53.85	37.25
LLaVA-v1.6-mistral-7B	23.88	33.33	21.78	25.45	25.93	21.82	17.81	42.42	29.58	22.45	27.27	26.92	15.69
LLaVA-v1.6-34B	48.76	36.36	42.08	43.64	40.74	36.36	26.03	48.48	43.66	36.73	36.36	46.15	31.37
InternVL2-8B	58.21	45.45	45.05	60.00	54.32	45.45	32.88	60.61	33.80	42.86	54.55	57.69	45.10
InternVL2-76B	62.19	33.33	45.05	61.82	58.02	49.09	39.73	54.55	49.30	48.98	63.64	61.54	49.02
Qwen2-VL-7B	65.67	36.36	45.54	56.36	51.85	47.27	39.73	60.61	42.25	42.86	56.82	63.46	39.22
Qwen2-VL-72B	**76.12**	**57.58**	**65.84**	**63.64**	**69.14**	**70.91**	**52.05**	**72.73**	**50.70**	**67.35**	**65.91**	**73.08**	**56.86**
GPT-4o	60.20	42.42	45.54	47.27	53.09	52.73	41.10	51.52	40.85	42.86	59.09	57.69	39.22
GPT-4o-mini	52.74	30.30	37.13	47.27	48.15	30.91	30.14	45.45	29.58	30.61	43.18	42.31	31.37

Table 7: Performances of **reasoning questions** across different models and topics, including Symbols and Signs (S&S), Attractions (Att), Food, Transportation (Trans), Culture and Arts (C&A), Politics (Pol), Geography (Geo), Sports (Spo), Flora and Fauna (F&F), History (His), Entertainment (Ent), Daily Necessities (DN), and Other Daily Life (ODL). All results are reported in Accuracy (%).

the InternVL2 and Qwen models perform better on OCR tasks than on general QA, suggesting a strong specialization. Given that our benchmark primarily consists of Traditional Chinese OCR tasks, we speculate that InternVL2 and Qwen were trained with more extensive Traditional Chinese OCR data compared to other models.

5.3 Reasoning Questions

Reasoning questions required models to interpret visual elements and apply external knowledge, such as culture or history, to answer questions beyond the image content. Unlike Recognition tasks, these questions tested deeper, abstract understanding, posing unique challenges for VLMs.

General Patterns and High-Performing Topics Reasoning tasks revealed significant variation in model performance. Categories like Transportation, Symbols and Signs, and Daily Necessities were strengths for larger models. Qwen2-VL-72B led across the board, achieving top scores in Symbols and Signs (76.12%), Politics (70.91%), and Daily Necessities (73.08%). InternVL2-76B also performed well, excelling in Transportation (61.82%) and Culture and Arts (58.02%). Other models like GPT-4o showed strength in reasoning-intensive topics such as Politics (52.73%), but struggled in more visually complex tasks. Table 7 highlights Qwen2-VL-72B's dominance across reasoning tasks.

Challenging Topics Across Models Topics requiring cultural or linguistic reasoning, such as Politics, Flora and Fauna, and History, were difficult for most models. Smaller models like Phi3.5-Vision-Instruct and Llama-3.2-11B scored poorly in these areas, with accuracy as low as 14.29% in History and 7.24% in Politics, respectively. Even intermediate models like LLaVA-v1.6-34B struggled in nuanced reasoning, achieving only 36.36% in Politics, emphasizing a need for better Taiwanese linguistic and cultural training.

Comparison of Models Qwen2-VL-72B outperformed all others, achieving exceptional accuracy in reasoning categories like Culture and Arts (69.14%), Geography (52.05%), and Politics (70.91%). Its smaller version, Qwen2-VL-7B, maintained competitive scores in areas like Daily Necessities (63.46%) and Symbols and Signs (65.67%). InternVL2-76B offered balanced results across most tasks, while GPT-4o excelled in text-heavy reasoning but fell short in visual topics. Smaller models like LLaVA consistently underperformed, demonstrating the importance of scale and training diversity.

Analysis of Types of Reasoning Questions Model size generally correlates strongly with reasoning ability, a trend also observed within the same model series in Figure 4. However, InternVL2-8B and Qwen2-VL-7B, despite being smaller models, outperform larger models such as LLaVA-34B and Llama-90B in reasoning tasks, an unexpected result. Across our types of reasoning questions, Qwen2-VL-72B consistently demonstrate a deeper understanding of Taiwan-specific content compared to other models.

5.4 Model Analysis and Insights

Analysis of Chinese-based Model In Figure 5, we analyze base models, where "O" represents Chinese-based models and "X" represents non-Chinese-based models. The choice of base model has a significant impact on our TaiwanVQA benchmark. Chinese-based models excel in recognition

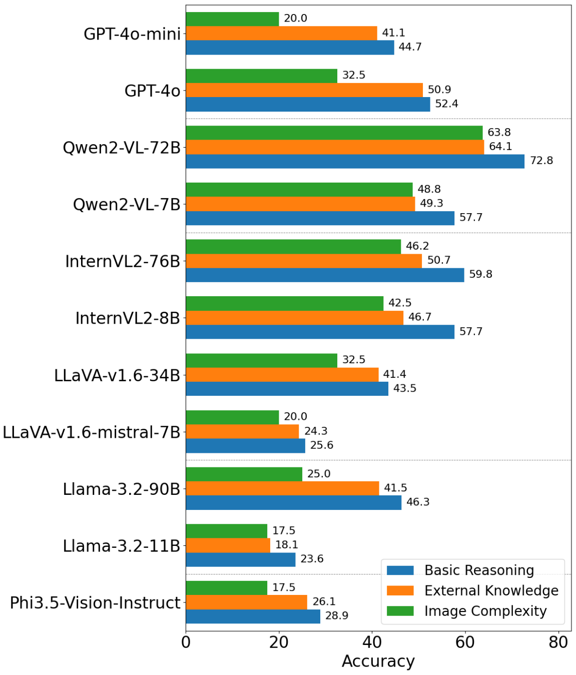


Figure 4: Analysis of Types of Reasoning Questions

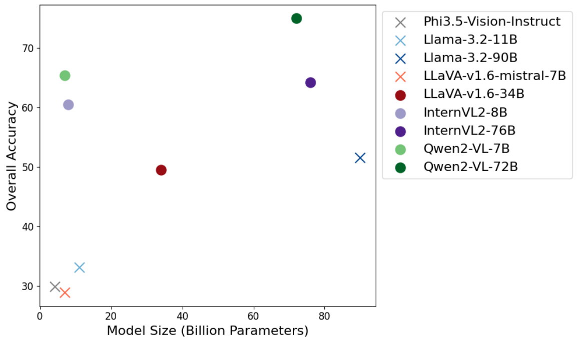


Figure 5: The Impact of Base Model Selection: Comparing Chinese-Based ("O") and Non-Chinese-Based ("X") Models

tasks, while also outperforming in reasoning tasks due to their optimization for Chinese semantic understanding and content-specific pretraining. Notably, InternVL2-8B and Qwen2-VL-7B achieve higher overall scores than Llama-90B, despite their smaller size.

Impact of Model Size on Accuracy The relationship between model size and overall accuracy underscores the significant impact of scale on performance. Larger models, such as Qwen2-VL-72B and InternVL2-76B, consistently achieved the highest overall accuracy, exceeding 70% and 60% accuracy, respectively. In contrast, smaller models like Phi3.5-Vision-Instruct and LLaVA-v1.6-mistral-7B struggled to surpass 30% accuracy, demonstrating a clear limitation in their ability to handle complex tasks. Notably, mid-sized models such as LLaVA-v1.6-34B showed moderate improvements in accuracy (around 50%), indicating that scaling up provides diminishing but still significant returns in accuracy. This trend emphasizes the importance of large-scale architectures and extensive training datasets for achieving state-of-the-art performance in multimodal recognition and reasoning tasks, though some smaller models still demonstrate reasonable accuracy.

6 Conclusion

In this paper, we introduced TaiwanVQA, a novel visual question answering benchmark specifically designed to evaluate the capabilities of VLMs in understanding and reasoning about Taiwan-specific content. TaiwanVQA consists of 1,000 images and 2,000 questions covering a diverse range of topics relevant to Taiwanese daily life and culture, including local cuisine, public signage, tourist attractions, and local flora and fauna. We categorized the questions into recognition and reasoning tasks, further sub-classifying the reasoning questions based on the level of external knowledge required.

Our extensive experiments with state-of-the-art models, including GPT-4 (OpenAI, 2023), revealed significant limitations in current VLMs when dealing with culturally specific content. The results demonstrated that while models perform reasonably well on recognition tasks, their performance on reasoning tasks that require deeper cultural understanding is substantially lower. This highlights the need for more culturally diverse training data and improved model architectures that can better integrate visual and textual information in culturally nuanced contexts.

By providing the first VQA benchmark that focuses on culturally rich content specific to Taiwan, TaiwanVQA fills a critical gap in the evaluation of VLMs. We believe this benchmark will contribute to the development of more inclusive and culturally aware AI models, ultimately facilitating their deployment in diverse real-world scenarios (Nayak et al., 2024).

7 Limitations

While TaiwanVQA makes significant strides in evaluating VLMs on culturally specific content, several limitations exist in our current work. First, due to technical challenges during the experimentation phase, we were unable to successfully infer and evaluate some models. These models are marked with an asterisk (*) or dagger (†) in our experimental settings and results (see Appendix D and E). The inability to include these models may affect the comprehensiveness of our evaluation. In future work, we plan to resolve these technical issues and include a broader range of models in our analysis.

Second, the dataset, though diverse, may not cover all aspects of Taiwanese culture and daily life. Certain niche or less visually represented cultural elements might be underrepresented, potentially limiting the assessment of models' understanding in those areas.

Third, the dataset primarily focuses on visual content accompanied by textual questions in Traditional Chinese. This language-specific focus might make it challenging to generalize the findings to other underrepresented languages and cultures without additional adaptation.

Finally, our current evaluation is conducted in a zero-shot setting without fine-tuning on Taiwan-specific data. While this approach highlights inherent model capabilities, it does not account for improvements that might be achieved through targeted training or domain-specific adaptation (Li et al., 2024a).

Acknowledgments

This work was supported by the National Center for High-performance Computing (NCHC), National Applied Research Laboratories (NARLabs), and the National Science and Technology Council (NSTC) under the project "Taiwan's 113th year endeavoring in the promotion of a trustworthy generative AI large language model and the cultivation of literacy capabilities (Trustworthy AI Dialog Engine, TAIDE)".

References

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *arXiv preprint arXiv:2407.00263*.

Olena Burda-Lassen, Aman Chadha, Shashank Goswami, and Vinija Jain. 2024. How culturally aware are vision-language models? *arXiv preprint arXiv:2405.17475*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Wenyan Li, Xinyu Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, et al. 2024b. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. *arXiv preprint arXiv:2406.11030*.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

OpenAI. 2023. Gpt-4 technical report.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *Preprint*, arXiv:2308.11483.

William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

T. Weyand, A. Araujo, B. Cao, and J. Sim. 2020. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Xue Zhang, Xiangyu Shi, Xinyue Lou, Rui Qi, Yufeng Chen, Jinan Xu, and Wenjuan Han. 2024. Transportationgames: Benchmarking transportation knowledge of (multimodal) large language models. *Preprint*, arXiv:2401.04471.

A Annotation Guideline

In this section we demonstrate the detail annotation guideline we asked annotator to do. There are three steps in our annotation step. First, we give annotators an general guideline and asked them to take a picture with Taiwan information. Second, we asked annotator to generate a recognition question. Final, we asked annotator to generate a reasoning question.

A.1 General Guideline

Before the annotators begin annotating data, we first provided them with a general guideline. This guideline asked the annotator follow the rules to write the recognition question and choices, including:

- The primary purpose of data collection: to collect images and questions featuring elements specific to Taiwan.
- Ensuring that the language used in questions reflects common terms and expressions used in Taiwan.
- Ensuring that annotators do not violate any legal issues, such as those related to privacy or copyright.

After reading the overall guideline, the annotator should upload an image containing a Taiwan-specific object.

A.2 Recognition Question

Next, we asked them to generate a recognition question and corresponding multiple-choice answers. To help annotators understand the guidelines, we provide clear examples and detailed explanations, ensuring both the questions and answer choices meet the required conditions. This guideline introduces key concepts of writing a recognition question, including:

- The definition of a recognition question: questions that assess whether the model can identify and name the object in an image without requiring analysis or inference.
- Emphasize that the question should be answerable solely based on all visible text or clearly identifiable objects in the image, and that the designed options do not include these visible texts or identifiable objects as possible answers.
- Ensure that questions cannot be answered without actually viewing the image.
- If there are multiple objects in the image, specify exactly which person or object to identify to avoid overly simplistic questions.
- Include misleading choices to make it harder for the model to select the correct answer, increasing the challenge.
- No length limit for questions and options.

Additionally, we asked annotators to classify whether the recognition question required ORC capability or not.

Once the question is written, annotators are required to categorize the question's topic. The topics definition is shown in Table 8. This helps in further analyzing the questions and ensuring data quality.

A.3 Reasoning Question

After writing a recognition question, annotator should write a reasoning question with the guideline. This guideline introduces key concepts of writing a reasoning question, including:

- The definition of a reasoning question: questions that require not only identifying the object but also understanding additional information, such as quantity, use, location, relative position, physical properties, or price, to provide an answer.
- Ensure that questions cannot be answered without actually viewing the image.
- No length limit for questions and options.

Once the reasoning question is written, we also asked the annotator to classify the question topic, similar to the recognition question. Additionally, we asked them to further label the question by identifying the capabilities required to answer it. The annotator should also indicate whether the question requires information about current events.

Topic	**Subtopic**	**Definition**
Symbols and Signs		Recognition and understanding of symbols, like priority seating, restrooms, no smoking, etc.
	Trademarks and Store Signs	Registered trademarks and store signs, such as FamilyMart, Louisa Coffee, YongChing Real Estate, Hua Nan Bank, etc.
	Public Notices and Announcements	Images or text providing information, such as advertisements, banners, usage instructions, and rules.
	Instruction Signs	Signs indicating rules or directions, like no smoking, emergency exit, restrooms, priority seating, parking, turn off devices, etc.
Attractions		Including Taiwan's natural and cultural landscapes.
	Natural Scenery	Includes Taiwan's mountains, coastlines, lakes, etc., such as Alishan, Taroko National Park, etc.
	Cultural Landmarks	Covers Taiwan's historical sites, architectural landmarks, and other non-natural tourist spots, such as Anping Fort in Tainan, Chiang Kai-shek Memorial Hall in Taipei, National Palace Museum, Jiufen Old Street.
Food		Including content related to Taiwan's culinary culture.
	Cuisine and Ingredients	Names of dishes and their ingredients, including distinctive foods, components, and garnishes on plates.
	Dietary Customs and Taboos	Features of Taiwan's daily dietary habits and customs, including combinations and taboos, like breakfast culture, adding cilantro, etc.
	Menus	Judging information based on menu or price list content; images only show text, no actual dishes.
	Cuisine Origin	Judging a dish's origin by time or location, or associating it with the culture that originated it.
Transportation		Including content related to Taiwan's transportation.
	Transit Systems	Includes Taiwan's metro, train, and bus systems, their operations and features.
	Traffic Signs	Covers Taiwan's traffic lights, violation checks, driving tests, etc.
Culture and Arts		Including content related to Taiwan's culture and arts.
	Folklore and Beliefs	All things related to culture and religion, including Taiwan's festivals, customs, and taboos like the Mid-Autumn Festival, Dragon Boat Festival, marriage and funeral traditions, religious buildings and decorations, gods, religious practices, temple culture, folk beliefs like Mazu worship.
	Indigenous Culture	Taiwan's indigenous customs, languages, and arts, such as those of the Amis and Atayal tribes.
	Artistic Activities	Activities like art exhibitions, cultural artifacts, musical instruments, operas, etc.
Politics		Including content related to Taiwan's politics.

Topic	**Subtopic**	**Definition**
	Political System	Taiwan's political system and electoral system, such as central and local government bodies, legislative election systems, etc.
	Political Events	Activities like elections and social movements.
	Political Figures and Parties	Contemporary Taiwanese political figures or parties, such as Lai Ching-te, Chu Li-lun, Taiwan People's Party.
Geography		Including content related to Taiwan's geography.
	Natural Geography	Taiwan's landforms and natural features, such as the Central Mountain Range and the eastern coast.
	Human Geography	Taiwan's administrative divisions, place name origins, population distribution, industry distribution, etc.
Sports		Including content related to Taiwan's sports and athletics.
	Sports	Types of sports and sports venues, such as tennis, badminton, baseball fields.
	Athletes	Taiwanese athletes, such as Chuang Chih-yuan, Tai Tzu-ying, Wang Chien-ming.
	Teams and Mascots	Taiwan's professional or amateur teams and mascots, such as the Uni Lions, Rakuten Monkeys, Monkeys Kids, Ryan.
Flora and Fauna		Including Taiwan's common flora and fauna.
	Animals	Common animal species in Taiwan, such as the Taiwan blue magpie and the Formosan landlocked salmon.
	Plants	Common plant species in Taiwan, such as the blackboard tree and large flower impatiens.
History		Covers historical events (e.g., the February 28 Incident, Kaohsiung Incident) and figures who impacted Taiwanese history, such as Chiang Ching-kuo, Lee Teng-hui.
Entertainment		Including content related to Taiwan's entertainment.
	Films and TV Shows	Movies, TV series, related events, and venues.
	Music Industry	Music genres, important music events, music works, and related venues.
	Gaming Industry	Games and industry development.
Daily Necessities		Common items or tools with specific purposes in daily life, requiring identification of the items and their possible uses or purposes.
Other Daily Life		Other content related to the daily lifestyle and habits of Taiwanese people.

Table 8: Definition of Each Topic

B Topic Definition and Classification Prompt

In this section, we show the detail of the definition of the topics and the analysis of it.

B.1 Definition

We classify the questions into 13 topics and 27 subtopics. The definition of the topics and subtopics is shown in Table 8.

B.2 Classification Prompt

In this section, we present the detailed prompt used to instruct GPT-4o to classify question topics.

The system prompt is shown in Figure 6. It includes a role-play request, asking GPT-4o to act as an assistant with a deep understanding of Taiwanese culture. Furthermore, we instruct GPT-4o to respond in a specific format, which includes both the topic and subtopic of the question. Additionally, we emphasize that GPT-4o should avoid selecting subtopics that do not align with the chosen topic.

你是一個專業的主題分類助理，且十分理解台灣的日常生活文化。請根據以下的分類標準，為每個問題選擇最適合的主題類別和子類別。

<Topics Definition>

分類標準：

- 若選擇 "交通標示" 作為子類別，主題必須是 "交通"。
- 若選擇 "文宣與告示" 作為子類別，主題必須是 "標誌標示"。

注意：
1. 部分主題沒有子類別，這種情況下只需提供主題即可。
2. 子類別必須屬於其主題，例如：
主題：[主題名稱]
子類別：[子類別名稱]（若該主題沒有子類別則此行可省略）。
回答格式：
- 主題：[主題名稱]。
- 子類別：[子類別名稱]。

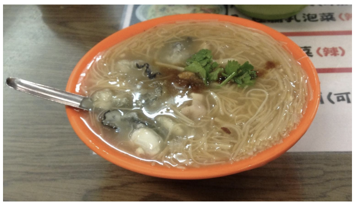
Figure 6: System Prompt for Classifying Question Topics

The user prompt is shown in Figure 7. This section directly includes the question, the options, and the correct answer.

問題：<question>
選項：
A. <option A>
B. <option B>
C. <option C>
D. <option D>
答案：<correct option>

Figure 7: User Prompt for Classifying Question Topics

Original Question:



請問照片拍攝的是以下哪種台灣小吃？ (Which Taiwanese snack is shown in the photo?)
A. 蚵仔煎 (Oyster Omelette)
B. 地瓜球 (Sweet Potato Balls)
C. 牛肉湯 (Beef Soup)
D. 蚵仔麵線 (Oyster Vermicelli)
Answer: D

Four Iterations with Circular Shifts:
1: A. 蚵仔煎 B. 地瓜球 C. 牛肉湯 D. 蚵仔麵線 → D
2: A. 地瓜球 B. 牛肉湯 C. 蚵仔麵線 D. 蚵仔煎 → C
3: A. 牛肉湯 B. 蚵仔麵線 C. 蚵仔煎 D. 地瓜球 → B
4: A. 蚵仔麵線 B. 蚵仔煎 C. 地瓜球 D. 牛肉湯 → A

Figure 8: CircularEval example. A model must correctly track the target answer (Oyster Vermicelli) through all shifted positions to be considered successful.

C Evaluation Strategy

C.1 Robust Evaluation

To ensure robust evaluation of model performance on multiple-choice questions, we implement the CircularEval strategy as illustrated in Figure 8. This approach addresses potential biases in model responses due to option positioning.

Consider an example where the model is asked to identify a Taiwanese snack from an image. The original question is presented with four options (A: Oyster Omelette, B: Sweet Potato Balls, C: Beef Soup, D: Oyster Vermicelli), where the correct answer is "Oyster Vermicelli" (Option D). CircularEval then creates four iterations by circularly shifting these options:

- Original: The correct answer "Oyster Vermicelli" is at position D
- First shift: The answer moves to position C
- Second shift: The answer moves to position B
- Third shift: The answer moves to position A

For a model's prediction to be considered correct, it must accurately track the answer through all

Model	Language Model	Vision Encoder	Size (B)
Phi3.5-Vision-Instruct	Phi-3.5-mini-instruct	CLIP ViT-L/14	4.2
Llama-3.2-11B	Llama-3.1-8B	ViT–H/14	11
Llama-3.2-90B	Llama-3.1-70B	ViT–H/14	90
LLaVA-v1.6-mistral-7B	Mistral-7B	CLIP ViT-L/14	7
LLaVA-v1.6-34B	Nous-Hermes-2-Yi-34B	CLIP ViT-L/14	34
InternVL2-8B	InternLM2.5-7B-Chat	InternViT-300M	8
InternVL2-76B	Hermes-2-Theta-Llama-3-70B	InternViT-6B	76
Qwen2-VL-7B	Qwen2-7B	CLIP ViT-L/14	7
Qwen2-VL-72B	Qwen2-72B	CLIP ViT-L/14	72
GPT-4o	–	–	–
GPT-4o-mini	–	–	–

Table 9: Model specifications of evaluated VLMs. Size is measured in billions of parameters (B).

four positions (D→C→B→A). This methodology ensures that the model's performance is based on genuine understanding rather than position-based biases or patterns.

D Experimental Setup

D.1 Models

We evaluate a diverse set of vision-language models in our experiments, categorized into three groups based on their primary language capabilities and model characteristics.

The first category includes leading **multilingual VLMs**:

- **Phi3.5-Vision-Instruct**: A lightweight model from Microsoft.
- **Llama-based models**: Including Llama-3.2-11B and Llama-3.2-90B.
- **LLaVA-v1.6-mistral-7B**: Designed for multilingual tasks.

The second category comprises **Chinese-based VLMs**:

- **InternVL2 series**: Consisting of InternVL2-8B and InternVL2-76B.
- **Qwen2-VL series**: Including Qwen2-VL-7B and Qwen2-VL-72B.
- **LLaVA-v1.6-34B**: Tailored for Chinese language understanding.

The third category consists of **proprietary models**:

- **GPT-4o series**: This includes GPT-4o and GPT-4o-mini, proprietary models whose architectural details are not publicly disclosed.

Parameter	Value	Description
logprobs	True	Return log prob. of output tokens
top_logprobs	20	Return top 20 likely tokens
temperature	0	Deterministic sampling

Table 10: Chat completion parameters for model inference.

Table 9 presents the specifications of all evaluated models. For open-source models, we detail their language models, vision encoders, and total parameters in billions (B). The size ranges from 4.2B (Phi3.5) to 90B (Llama-3.2-90B) parameters, offering a comprehensive evaluation across different model scales. For proprietary models in the GPT-4o series, these specifications are not publicly available and thus marked with dashes.

D.2 Implementation Details

In this subsection, we present our experimental configurations for both model inference and deployment. Table 10 shows the chat completion parameters used consistently across all evaluations. For serving open-source models, we utilize the vLLM framework (Kwon et al., 2023) to evaluate the performance and scalability of the serving infrastructure under different configurations, which are detailed in Table 11.

The evaluated models include a wide range of vision-language models such as **LLaVA**, **Qwen-**

VL, **InternVL**, among others. For each model, key configuration parameters were recorded:

- **Maximum Model Length (`max-model-len`):** The maximum sequence length supported by the model.
- **Tensor Parallel Size (`tensor-parallel-size`):** The number of GPUs allocated for parallel inference.
- **GPU Memory Utilization:** The proportion of GPU memory utilized during serving.
- **Batching Parameters:**
  - **Maximum Number of Batched Tokens:** The maximum number of tokens that can be processed in a single batch.
  - **Maximum Number of Sequences:** The maximum number of sequences processed in parallel.
- **Swap Space:** Indicates whether disk-based swap space is enabled to handle memory overflow scenarios.
- **Worker Configuration (`worker-use-ray`):** Specifies whether Ray-based worker management is employed for distributed serving.

To clarify the model status during the experiments:

- Models currently **in progress** or **pending evaluation** are marked with **‘†‘** before their names.
- Models encountering **errors** during serving are marked with **‘*‘** before their names.

The `vLLM` framework was used for all experiments. This framework is optimized for high-throughput inference with features such as:

- Token-level pipelining to maximize GPU utilization.
- Tensor-parallel support for efficient multi-GPU inference.
- Dynamic batching for reducing latency and improving throughput.

Table 11 provides a detailed summary of the experiment configurations and results. These settings can serve as a practical reference for deploying vision-language models in research or production environments.

E Experiment Results

Detailed performance results for recognition and reasoning questions across various subtopics are presented in Table 12 and Table 13.

Model	max-model-len	tensor-parallel-size	gpu-memory-utilization	max-num-batched-tokens	max-num-seqs	swap-space	worker-use-ray
Llama-3.2-11B	16384	4	0.8	16384	4	1	✓
Llama-3.2-90B	16384	8	0.9	16384	8	-	-
Qwen2-VL-7B-Instruct	16384	4	0.85	16384	8	1	✓
Qwen2-VL-72B-Instruct	16384	8	0.85	16384	8	1	✓
LLaVA-1.6-Mistral-7B	32000	4	-	8192	4	-	✓
*LLaVA-1.6-Vicuna-7B	4096	4	0.9	4096	4	1	✓
*LLaVA-1.6-Vicuna-13B	4096	4	0.9	4096	4	1	✓
LLaVA-1.6-34B	4096	8	0.9	4096	16	1	✓
†LLaVA OneVision 7B	8192	4	0.88	8192	4	1	✓
†LLaVA OneVision 72B	8192	8	0.85	8192	8	1	✓
*Pixtral 12B	8192	4	0.9	8192	4	1	✓
†Molmo-D 7B	4096	4	0.88	4096	4	1	✓
†Molmo 72B	8192	8	0.85	8192	16	1	✓
InternVL2-8B	12288	4	0.85	12288	4	1	✓
†InternVL2-26B	8192	8	0.85	8192	16	1	✓
InternVL2-Llama-3-76B	12288	8	0.85	12288	4	1	✓
†InternVL-Chat-V1-5	4096	8	0.9	4096	8	1	✓
†Mono-InternVL-2B	4096	4	0.70	4096	2	1	-
*cogvlm2-llama3-chinese-chat-19B	4096	4	0.85	4096	4	1	-
*Deepseek-vl-7b-chat	4096	4	0.9	4096	4	1	-
*BLIP-3 (XGen-MM)	4096	4	0.9	4096	8	1	✓
Phi3.5-Vision-Instruct	8192	4	0.9	8192	8	1	✓

Table 11: Configuration and Status of Vision-Language Models in vLLM Serving Framework. The table summarizes the key parameters used for serving various models, including model length, tensor parallelism, GPU utilization, and batching settings. Models marked with ‘*‘ encountered errors during the experiments, while models marked with ‘†‘ are in progress or pending evaluation.

Model	**Symbols & Signs**			**Attractions**		**Food**			**Transport**		**Culture & Arts**			**Politics**		
	T&S	PN	IS	NS	CL	C&I	Men	CO	TS	TrS	F&B	IC	AA	PS	PE	PFP
Phi3.5-Vision	33.87	29.73	44.90	15.79	25.00	33.51	30.00	0.00	47.06	21.05	34.48	25.00	43.75	33.33	44.44	4.55
Llama-3.2-11B	67.74	36.49	55.10	36.84	35.00	45.55	20.00	25.00	61.76	26.32	27.59	25.00	50.00	66.67	55.56	18.18
Llama-3.2-90B	71.77	62.16	71.43	36.84	50.00	63.87	55.00	50.00	67.65	52.63	55.17	50.00	81.25	88.89	66.67	27.27
LLaVA-v1.6-m-7B	30.65	31.08	55.10	15.79	21.67	31.41	35.00	0.00	55.88	42.11	27.59	0.00	37.50	55.56	44.44	4.55
LLaVA-v1.6-34B	52.42	58.11	69.39	57.89	55.00	60.73	35.00	0.00	73.53	57.89	51.72	25.00	75.00	55.56	55.56	36.36
InternVL2-8B	84.68	81.08	79.59	63.16	60.00	65.97	55.00	75.00	76.47	63.16	72.41	75.00	87.50	88.89	88.89	50.00
InternVL2-76B	87.90	75.68	79.59	57.89	71.67	76.44	45.00	75.00	79.41	57.89	79.31	87.50	93.75	88.89	88.89	36.36
Qwen2-VL-7B	93.55	82.43	79.59	57.89	71.67	75.39	85.00	75.00	85.29	63.16	75.86	50.00	87.50	100.00	88.89	72.73
Qwen2-VL-72B	92.74	85.14	89.80	68.42	81.67	83.77	75.00	75.00	82.35	57.89	93.10	75.00	87.50	100.00	100.00	81.82
GPT-4o	87.90	58.11	75.51	68.42	73.33	81.68	35.00	75.00	76.47	52.63	65.52	87.50	81.25	88.89	88.89	40.91
GPT-4o-mini	72.58	62.16	67.35	63.16	50.00	61.78	30.00	50.00	70.59	42.11	37.93	50.00	68.75	88.89	66.67	40.91

Model	**Geography**		**Sports**			**F&F**		**His**	**Entertainment**			**DN**	**ODL**
	NG	HG	SAV	Ath	T&M	Ani	Pla	His	FTS	Mus	Gam	DN	ODL
Phi3.5-Vision	50.00	16.67	60.00	25.00	33.33	29.17	23.26	20.00	35.00	50.00	16.67	47.92	42.86
Llama-3.2-11B	37.50	33.33	50.00	75.00	83.33	33.33	34.88	40.00	55.00	66.67	50.00	64.58	34.29
Llama-3.2-90B	68.75	50.00	70.00	25.00	83.33	50.00	41.86	20.00	80.00	50.00	66.67	73.96	54.29
LLaVA-v1.6-m-7B	37.50	0.00	20.00	0.00	66.67	31.25	30.23	20.00	55.00	33.33	25.00	47.92	28.57
LLaVA-v1.6-34B	75.00	41.67	70.00	0.00	100.00	45.83	51.16	20.00	60.00	66.67	50.00	72.92	54.29
InternVL2-8B	81.25	66.67	60.00	100.00	83.33	64.58	60.47	60.00	80.00	83.33	50.00	72.92	77.14
InternVL2-76B	68.75	66.67	80.00	100.00	100.00	70.83	55.81	60.00	90.00	83.33	75.00	82.29	77.14
Qwen2-VL-7B	87.50	91.67	80.00	100.00	100.00	64.58	69.77	80.00	85.00	66.67	83.33	79.17	88.57
Qwen2-VL-72B	87.50	91.67	80.00	100.00	100.00	62.50	74.42	80.00	90.00	66.67	83.33	84.38	88.57
GPT-4o	87.50	83.33	70.00	100.00	100.00	56.25	62.79	80.00	75.00	66.67	83.33	77.08	62.86
GPT-4o-mini	62.50	58.33	50.00	75.00	100.00	47.92	48.84	60.00	50.00	66.67	66.67	65.62	48.57

Table 12: Subtopic Performance of Recognition Questions (Accuracy, %).
Subtopics: T&S=Trademarks & Store Signs, PN=Public Notices & Announcements, IS=Instruction Signs, NS=Natural Scenery, CL=Cultural Landmarks, C&I=Cuisine & Ingredients, Men=Menus, CO=Cuisine Origin, TS=Transit Systems, TrS=Traffic Signs, F&B=Folklore & Beliefs, IC=Indigenous Culture, AA=Artistic Activities, PS=Political System , PE=Political Events, PFP=Political Figures & Parties, NG=Natural Geography, HG=Human Geography, SAV=Sports Activities & Venues, Ath=Athletes, T&M=Teams & Mascots, Ani=Animals, Pla=Plants, His=History, FTS=Films & TV Shows, Mus=Music, Gam=Gaming, DN=Daily Necessities, ODL=Other Daily Life.

Model	**Symbols & Signs**			**Attractions**		**Food**				**Transport**		**Culture & Arts**		
	T&S	PN	IS	NS	CL	C&I	DCT	Men	CO	TS	TrS	F&B	IC	AA
Phi3.5-Vision	30.00	26.74	42.22	33.33	16.67	21.59	17.78	24.32	18.75	28.57	20.00	39.53	23.08	16.67
Llama-3.2-11B	24.29	18.60	44.44	20.00	11.11	14.77	24.44	16.22	15.62	25.71	10.00	27.91	15.38	8.33
Llama-3.2-90B	52.86	46.51	71.11	20.00	38.89	40.91	48.89	21.62	37.50	45.71	30.00	44.19	23.08	58.33
LLaVA-v1.6-m-7B	27.14	16.28	33.33	33.33	33.33	19.32	33.33	16.22	18.75	25.71	25.00	30.23	15.38	33.33
LLaVA-v1.6-34B	47.14	46.51	55.56	33.33	38.89	44.32	53.33	16.22	50.00	45.71	40.00	44.19	34.62	41.67
InternVL2-8B	54.29	59.30	62.22	46.67	44.44	42.05	51.11	43.24	46.88	62.86	55.00	65.12	42.31	41.67
InternVL2-76B	61.43	60.47	66.67	26.67	38.89	48.86	42.22	37.84	46.88	68.57	50.00	65.12	42.31	66.67
Qwen2-VL-7B	68.57	61.63	68.89	33.33	38.89	47.73	46.67	40.54	43.75	60.00	50.00	55.81	42.31	58.33
Qwen2-VL-72B	74.29	76.74	77.78	53.33	61.11	72.73	57.78	56.76	68.75	68.57	55.00	79.07	53.85	66.67
GPT-4o	61.43	58.14	62.22	46.67	38.89	53.41	55.56	21.62	37.50	51.43	40.00	55.81	42.31	66.67
GPT-4o-mini	55.71	44.19	64.44	26.67	33.33	39.77	46.67	21.62	34.38	42.86	55.00	51.16	38.46	58.33

Model	**Politics**			**Geography**		**Sports**			**F&F**		**His**	**Entertainment**			**DN**	**ODL**
	PS	PE	PFP	NG	HG	SA&V	Ath	T&M	Ani	Pla	His	F&TS	Mus	Gam	DN	ODL
Phi3.5-Vision	15.38	61.54	10.34	33.33	19.23	55.56	25.00	18.18	23.68	30.30	20.41	20.00	50.00	33.33	30.77	19.61
Llama-3.2-11B	7.69	23.08	0.00	9.52	11.54	50.00	25.00	18.18	15.79	6.06	14.29	24.00	50.00	46.67	23.08	13.73
Llama-3.2-90B	30.77	46.15	17.24	38.10	13.46	72.22	50.00	18.18	44.74	36.36	40.82	44.00	50.00	46.67	53.85	37.25
LLaVA-v1.6-m-7B	30.77	38.46	10.34	23.81	15.38	61.11	50.00	9.09	28.95	30.30	22.45	24.00	50.00	26.67	26.92	15.69
LLaVA-v1.6-34B	30.77	61.54	27.59	38.10	21.15	66.67	25.00	27.27	44.74	42.42	36.73	24.00	50.00	53.33	46.15	31.37
InternVL2-8B	53.85	76.92	27.59	28.57	34.62	83.33	75.00	18.18	31.58	36.36	42.86	48.00	75.00	60.00	57.69	45.10
InternVL2-76B	69.23	76.92	27.59	47.62	36.54	72.22	50.00	27.27	57.89	39.39	48.98	60.00	50.00	73.33	61.54	49.02
Qwen2-VL-7B	69.23	69.23	27.59	38.10	40.38	77.78	75.00	27.27	39.47	45.45	42.86	48.00	50.00	73.33	63.46	39.22
Qwen2-VL-72B	92.31	76.92	58.62	47.62	53.85	83.33	100.00	45.45	55.26	45.45	67.35	60.00	50.00	80.00	73.08	56.86
GPT-4o	69.23	76.92	34.48	52.38	36.54	77.78	25.00	18.18	52.63	27.27	42.86	56.00	75.00	60.00	57.69	39.22
GPT-4o-mini	53.85	61.54	6.90	23.81	32.69	55.56	50.00	27.27	34.21	24.24	30.61	36.00	50.00	53.33	42.31	31.37

Table 13: Subtopic Performance of Reasoning Questions (Accuracy, %). Subtopics: T&S=Trademarks & Store Signs, PN=Public Notices & Announcements, IS=Instruction Signs, NS=Natural Scenery, CL=Cultural Landmarks, C&I=Cuisine & Ingredients, DCT=Dietary Customs & Taboos, Men=Menus, CO=Cuisine Origin, TS=Transit Systems, TrS=Traffic Signs, F&B=Folklore & Beliefs, IC=Indigenous Culture, AA=Artistic Activities, PS=Political System, PE=Political Events, PFP=Political Figures & Parties, NG=Natural Geography, HG=Human Geography, SA&V=Sports Activities & Venues, Ath=Athletes, T&M=Teams & Mascots, Ani=Animals, Pla=Plants, His=History, F&TS=Films & TV Shows, Mus=Music, Gam=Gaming, DN=Daily Necessities, ODL=Other Daily Life.