

# TaiwanVQA: Benchmarking and Enhancing Cultural Understanding in Vision-Language Models

---

Hsin-Yi Hsieh<sup>1</sup> Shang-Wei Liu<sup>1</sup> Chang-Chih Meng<sup>2</sup> Chien-Hua Chen<sup>2</sup>

Shuo-Yueh Lin<sup>3</sup> Hung-Ju Lin<sup>4</sup> Hen-Hsen Huang<sup>5</sup> I-Chen Wu<sup>2</sup>

<sup>1</sup>National Center for High-performance Computing <sup>2</sup>National Yang Ming Chiao Tung University

<sup>3</sup>National Central University <sup>4</sup>National Taiwan University <sup>5</sup>Institute of Information Science, Academia Sinica



# Motivation

- Most VLM benchmarks focus on dominant languages and cultures
- Limited evaluation of **localized or Taiwan-specific** content
- VLMs often recognize objects but miss cultural meaning
- → Need benchmarks that capture **cultural reasoning and adaptation**

Global / Common Food



**Hamburger**

vs.

Local / Cultural Food



**挫冰 (Taiwanese shaved ice)**

traditional summer dessert made of finely shaved ice topped with boba, beans, jellies, or condensed milk.

---

# What We Do in This Paper

- Build **TaiwanVQA**, the first benchmark for evaluating Taiwanese cultural understanding in VLMs
- Provide a **generalizable taxonomy** separating recognition and reasoning tasks
- Define **structured annotation rules** for consistent, culturally rich data
- Conduct **comprehensive evaluation** on 12 SOTA VLMs (e.g., Gemini-2.5, InternVL3)
- Demonstrate **culture-aware fine-tuning** that substantially narrows reasoning gaps

Dataset

Taxonomy

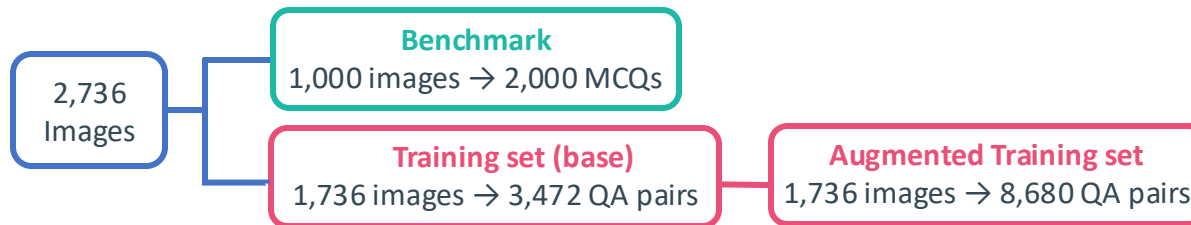
Evaluation

Fine-tuning



# Dataset Overview

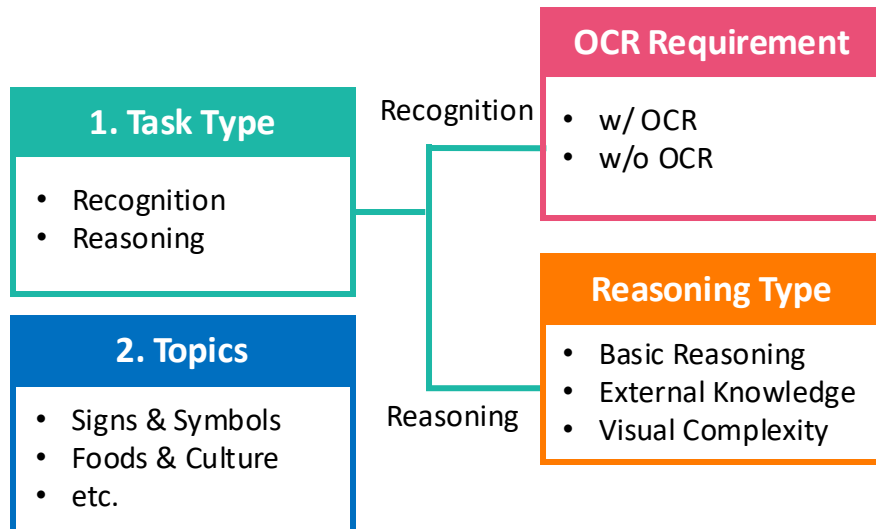
- Each image includes **two human-annotated MCQs** — one for **Recognition** (object or OCR understanding) and one for **Reasoning** (cultural meaning or symbolism)
  - Benchmark:** 1,000 images, 2,000 MCQs — used only for evaluation.
  - Training set:** 1,736 images → 3,472 human QA pairs, later augmented to 8,680 QA pairs.



Data Split	Purpose	images	Question Type	Total Questions
<b>Benchmark</b>	Evaluation only	1,000	2 MCQs per image ( 1 Recognition + 1 Reasoning)	<b>2,000</b>
<b>Training (Augmented)</b>	Training	1,736	5 QA pairs per image	<b>8,680</b>

# Taxonomy

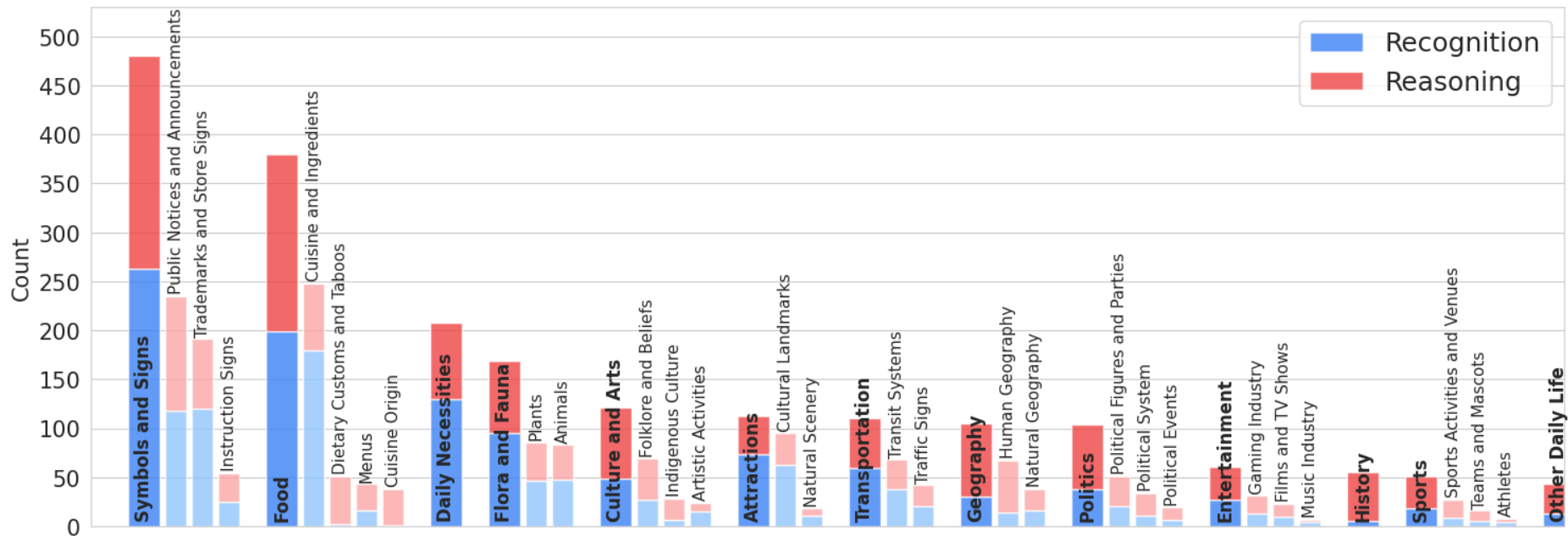
- **Task Type:** Recognition | Reasoning
- **OCR Requirements** (for Recognition): with / without Traditional Chinese text
- **Reasoning Type** (for Reasoning): Basic | External Knowledge | Visual Complexity
- **Topic Classification:** 13 cultural topics, expanded into 27 fine-grained subtopics



Recognition		Reasoning	
	<p>請問以下哪個食物沒有出現在照片裡？ (Which food is NOT in the photo?)</p> <p>Ⓐ 醃蘿蔔 (Pickled radish) Ⓑ 米飯 (Rice) Ⓒ 魚湯 (Fish soup) Ⓓ 牛肉麵 (Beef Noodle Soup)</p> <p>Food</p>	<p>請問照片上方魚湯裡的魚，大多是從哪裡獲得？ Where are the soup's fish usually sourced from?</p> <p>Ⓐ 國外進口 (Imported from abroad) Ⓑ 魚塢養殖 (Aquaculture) Ⓒ 外海捕撈 (Caught in offshore waters) Ⓓ 近岸捕撈 (Caught near the coast)</p> <p>Basic Reasoning      Food</p>	
	<p>圖片中的黃底看板上宣傳的是什麼祭典？ (What ceremony is on the yellow sign?)</p> <p>Ⓐ 大獵祭 (Mangayaw) Ⓑ 矮靈祭 (Pas-taii) Ⓒ 小米收穫祭 (Masalut / Masuvaqu) Ⓓ 豐年祭 (Malalikit / Malikoda)</p> <p>OCR      Culture and Arts</p>	<p>請問這個祭典主要是台灣哪一個原住民族的傳統祭儀？ (Which Taiwanese indigenous tribe holds this ceremony?)</p> <p>Ⓐ 賽夏族 (Saisiyat) Ⓑ 太魯閣族 (Truku) Ⓒ 排灣族 (Paiwan) Ⓓ 泰雅族 (Atayal)</p> <p>External Knowledge      Culture and Arts</p>	
	<p>請問這張照片並未出現下列哪個地區的地圖？ (Which region's map is NOT shown in this photo?)</p> <p>Ⓐ 菲律賓 (Philippines) Ⓑ 印尼 (Indonesia) Ⓒ 日本 (Japan) Ⓓ 台灣 (Taiwan)</p> <p>Geography</p>	<p>根據照片裡黑板上的資訊，這可能講述的是台灣什麼時期的情形？ (Which period of Taiwan's history is shown on the blackboard?)</p> <p>Ⓐ 20 世紀 (20th century) Ⓑ 17 世紀 (17th century) Ⓒ 15 世紀 (15th century) Ⓓ 二戰期間 (Post-WWII)</p> <p>Image Complexity      History</p>	

# Topic Distribution

- Dataset covers **13 topics** and **27 subtopics**, reflecting Taiwan's cultural, social, and linguistic diversity.
- Recognition** and **Reasoning** questions are balanced across domains, allowing fair evaluation.



---

# Data Collection & Quality

- **Annotation Process**

- **Annotators:** 9 contributors with diverse educational and regional backgrounds.
- **Calibration:** 1-week training using shared examples to unify question style and cultural tone.
- **Peer Review:** Each QA pair reviewed by a second annotator.
- **Lead Adjudication:** Final quality checks by senior reviewers.

- **Quality Assurance**

- **Agreement:** > 95% inter-annotator agreement on answers and tags.
- **Audit:** 10% random rechecks.
- **Scope:** Only public, culture-related images — no PII or sensitive content.

# Training Data Augmentation Pipeline

- **Step 1. Image Captioning:** Qwen2-VL generates captions for each image.
- **Step 2. Dialogue Generation:** GPT-4o creates 3 QA types per image:
  - Visual Conversation: overall visual context
  - Attribute Recognition: key attributes of the object
  - Contextual Inference: situational or functional reasoning

- **Outcome:** From 2 → 5 questions per image (8,680 total)

Data Source	Total	Source
Seed QA pairs	3,472	Human
Visual Conversation	1,736	Generated
Attribute Recognition	1,736	Generated
Contextual Inference	1,736	Generated
<b>Total</b>	<b>8,680</b>	<b>Mixed</b>



# Experimental Setup & Results

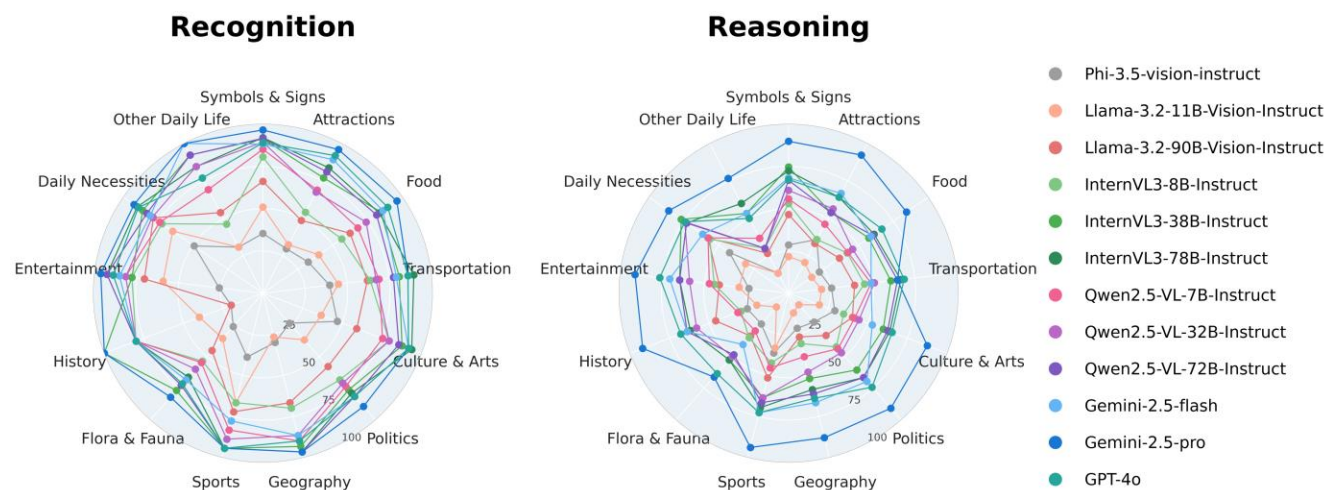
## MCQ Evaluation

### Evaluation Setup


- **Format:** Zero-shot MCQ with **CircularEval**<sup>1</sup> over shuffled options.
- **Models:** 12 VLMs, open (e.g., InternVL3, Qwen2.5-VL, LLaMA-3.2-Vision) and proprietary (Gemini-2.5, GPT-4o).

### Key Findings

- **Overall trend:** Proprietary models lead; all show **Recognition > Reasoning**.
- OCR items are usually easier than non-OCR reasoning items.
- **Takeaway:** Cultural reasoning remains the key bottleneck in vision-language understanding.



**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 → Answer: D  
2: A. 地瓜球 B. 牛肉湯 C. 蚵仔麵線 D. 蚵仔煎 → Answer: C  
3: A. 牛肉湯 B. 蚵仔麵線 C. 蚵仔煎 D. 地瓜球 → Answer: B  
4: A. 蚵仔麵線 B. 蚵仔煎 C. 地瓜球 D. 牛肉湯 → Answer: A

CircularEval example

<sup>1</sup> Liu et al. MMBench: Is Your Multi-modal Model an All-around Player?

# Experimental Setup & Results

## Open-Ended QA Evaluation

### Evaluation Setup

- **Format:** Open-ended answers in Traditional Chinese; no multiple-choice options.
- **Scoring:** GPT-4.1 as judge model to evaluate semantic similarity between model output and reference answer.

### Key Findings

- **Performance gap:** Accuracy drops **10–20 points** compared with MCQ.
- **Where it hurts:** Reasoning questions with external knowledge or symbolic meaning show largest declines.
- **Takeaway:** Open-QA exposes hidden weaknesses in knowledge retrieval and cultural grounding.

Model	MCQ			Open-QA					
	All	Recog.	Reason.	All	$\Delta$	Recog.	$\Delta$	Reason.	$\Delta$
Phi-3.5-Vision	31.05	35.20	26.90	10.20	-20.85	12.70	-22.50	7.70	-19.20
Llama-3.2-11B	32.35	45.60	19.10	31.60	-0.75	39.00	-6.60	24.20	+5.10
Llama-3.2-90B	51.50	62.70	40.30	40.70	-10.80	49.80	-12.90	31.60	-8.70
InternVL3-8B	55.15	67.60	42.70	43.55	-11.60	55.30	-12.30	31.80	-10.90
InternVL3-38B-Instruct	74.10	85.30	62.90	51.80	-22.30	65.50	-19.80	38.10	-24.80
InternVL3-78B-Instruct	75.80	86.50	65.10	53.10	-22.70	65.90	-20.60	40.30	-24.80
Qwen2.5-VL-7B	59.75	74.10	45.40	50.70	-9.05	64.30	-9.80	37.10	-8.30
Qwen2.5-VL-32B-Instruct	65.65	77.70	53.60	55.85	-9.80	66.80	-10.90	44.90	-8.70
Qwen2.5-VL-72B-Instruct	73.35	84.60	62.10	58.35	-15.00	70.00	-14.60	46.70	-15.40
Gemini-2.5-flash	72.80	84.20	61.40	66.80	-5.80	76.40	-7.80	57.20	-24.70
Gemini-2.5-pro	89.35	93.40	85.30	71.90	-17.45	79.50	-13.90	64.30	-21.00
GPT-4o	77.40	87.30	67.50	67.40	-10.00	77.50	-9.80	57.30	-10.20

# Fine-Tuning Results

## Training Setup

- Base Model: Llama-3.2-11B-Vision-Instruct
- Data: 8,680 TaiwanVQA QA pairs (training split)
- Variants:
  - Base: Original, non-fine-tuned model
  - Human: Fine-tuned on *human-annotated* seed data (3,472)
  - Mixed: Fine-tuned on both *human* and *augmented* data
- Benchmarks: TaiwanVQA (cultural) + MMMU

## Key Findings

- **Mixed fine-tuning** yields the best cultural gains (+15–20 pts).
- Maintains general performance.

		Llama-3.2-11B		
		base	human	mix
TaiwanVQA	Recognition	45.6	51.6	61.0
	Reasoning	19.1	27.0	36.4
MMMU	Valid	37.7	43.7	42.8
	Pro-standard	28.0	30.4	31.7
	Pro-vision	5.6	11.2	13.0

---

# Key Takeaways

- **TaiwanVQA** offers a systematic, scalable framework for evaluating cultural understanding in VLMs.
- **Cultural reasoning** remains the primary bottleneck.
- **Open-QA evaluation** exposes hidden failures in knowledge retrieval and grounding that MCQs overlook.
- Lightweight fine-tuning on culture-specific data significantly improves reasoning while preserving general performance.
- Enables low-resource, reproducible cultural adaptation across domains.