VisTW: Benchmarking Vision-Language Models for Taiwanese Mandarin in Taiwan

Zhi Rui Tam **Yung-Yu Shih** **Yen-Wei Lee**
Ya-Ting Pai **Wen Yu Chang** **Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
y.v.chen@ieee.org

Abstract

Vision-Language Models (VLMs) often struggle in **Taiwanese Mandarin** environments due to region-specific orthographic and cultural context. We introduce **VisTW**, a comprehensive benchmark featuring (i) multiple-choice questions (3,795 academic questions) and (ii) free-form generation evaluation (141 Taiwanese-context free-form pairs). Beyond standard accuracy, we investigate **character mixing**, the unintended production of Simplified Chinese characters under Taiwanese-Mandarin-style prompts, and propose a **human-grounded purity penalty** derived from perceptual thresholds measured from users. Our evaluation reveals substantial character contamination (3%–19%) across state-of-the-art VLMs. We find that Gemini-3-Pro significantly outperforms the strongest open-weight baseline, Qwen3 235B MoE, by up to 22 percentage points on dialogue tasks once the purity penalty is applied. These results highlight orthographic consistency as a vital, yet overlooked, dimension for localized multimodal evaluation and deployment.[1]

1 Introduction

Vision-Language Models (VLMs) have achieved remarkable success in tasks such as image captioning, visual question answering (VQA), and cross-modal retrieval (Chen et al., 2022; Li et al., 2022; Faysse et al., 2024). However, current multimodal benchmarks predominantly focus on English, and when extended to Chinese, they primarily employ Simplified Chinese scripts common in mainland China (Das et al., 2024; Winata et al., 2024; Wang et al., 2024). Traditional Chinese, widely used in Taiwan and Hong Kong, remains significantly underrepresented. Consequently, VLM capabilities in Traditional Chinese contexts are largely untested and unknown.

Evaluating VLMs in Traditional Chinese (TC) poses distinct challenges for three main reasons. First, TC characters are visually and structurally more complex, and the correspondence between Traditional and Simplified forms is not always one-to-one. As a result, VLMs trained predominantly on Simplified Chinese corpora (Chen et al., 2024a) may exhibit degraded performance and increased errors when operating in TC contexts. Second, TC usage is often intertwined with region-specific linguistic and cultural variation. For example, Taiwanese Mandarin and Cantonese are distinct spoken varieties yet share the same written standard (TC). Beyond text, images from TC regions (e.g., Taiwan) frequently contain culturally grounded references as well as TC typography in signs, labels, and official documents, which can affect multimodal understanding (Hsieh et al., 2025). Third, because web-scale Chinese data is dominated by Simplified Chinese, pretrained models may produce *mixed* orthography, emitting Simplified and Traditional characters within the same response. Similar phenomena have been reported in Japanese, where models may confuse Shinjitai and Kyūjitai forms or mix Chinese-origin Kanji with Japanese variants.

In this work, we focus on VLMs performance in **Taiwanese Mandarin** under **multimodal** settings. Figure 1 presents representative examples from VisTW that illustrate the challenges VLMs face when processing Taiwan-specific content. On the left, we show **MCQ-style** items drawn from Taiwan's educational context, which require knowledge of Taiwan-centric geography and local conventions. On the right, we include **dialogue** scenarios that depend on culturally grounded understanding, such as interpreting Traditional Chinese menus and reading utility meters. Together, these examples highlight that evaluating TC capabilities cannot be reduced to a simple Simplified-to-Traditional character conversion: the inputs and ex-

[1] https://huggingface.co/datasets/miulab/vistw/

Findings of the Association for Computational Linguistics: ACL 2026, pages 36711–36756

pected outputs additionally involve Taiwan-specific terminology, formats, and cultural knowledge.

In summary, our contributions are three-fold:

1. We introduce the **first** comprehensive vision-language benchmark under Taiwanese Mandarin context, featuring 3,795 test questions across 21 subjects (VisTW-MCQ) and 141 culturally-grounded dialogue pairs (VisTW-Dialogue).

2. We construct a validated and cost-effective VLM as a judge evaluation framework with high human correlation ($\rho = 0.847$) and a calibration approach that ensures continuity in evaluations across model generations.

3. We conduct a human sensitivity study on human perception on mixing traditional-simplified Chinese characters. And apply a Simplified Chinese character penalty score to all benchmarks.

4. We provide empirical evidence from 53 VLMs, revealing systematic gaps between Simplified- and Traditional-Chinese settings, and analyze trends in scaling behavior and image-resolution sensitivity.

2 Related Work

2.1 English-based Multimodal Benchmarks

Vision-language benchmarks like ChartQA (Masry et al., 2022), TextVQA (Singh et al., 2019), and DocVQA (Mathew et al., 2021) have advanced visual text understanding through specialized tasks requiring OCR and reasoning. More recent exam-based evaluations such as MMMU (Yue et al., 2024) test expert-level multimodal reasoning across 30 subjects, pushing beyond surface-level image understanding to complex problem-solving. For open-ended generation, benchmarks like Vibe-Eval (Padlewski et al., 2024) assess multimodal chat models with 269 visually-grounded prompts. However, these benchmarks remain English-centric with limited applicability to other languages and cultural contexts.

2.2 Traditional Chinese and Multilingual Benchmarks

For Traditional Chinese, recent benchmarks focus primarily on text-only evaluation. TMMLU (Hsu et al., 2023) provides 3,300 multiple-choice questions from Taiwanese exams, while TMMLU+ (Tam et al., 2024) expands this to 22,690 questions across 66 subjects. However, both lack multimodal components essential for vision-language evaluation. TMMBench included in the release of Breeze2 (Hsu et al., 2025) is a multiple-choice Traditional Chinese visual QA dataset that covers topics related to Taiwan, such as Taiwanese attractions, daily life, and Taiwan's university entrance exams. However, as of this writing, the details and TMMBench dataset have not been publicly released yet.

Prior multilingual benchmarks include JMMMU (Onohara et al., 2024) for Japanese (with only 30 questions per subject) and CMMU (He et al., 2024) for Simplified Chinese (with limited public access and limited cultural grounding). M3Exam (Zhang et al., 2023) spans nine languages, including Simplified Chinese, but follows generic curricula and does not incorporate Traditional Chinese script or region-specific content. TaiwanVQA (Hsieh et al., 2025) partially addresses this gap by releasing a culture-focused multiple-choice VQA dataset curated from local sources. In contrast, VisTW-MCQ targets more *general* multimodal competencies aligned with Taiwan's educational setting, such as chart interpretation and understanding electrical-circuit diagrams that reflect STEM subjects. Therefore, **TaiwanVQA** and **VisTW** are complementary: the former emphasizes cultural specificity, while the latter also include broad academic and STEM-oriented abilities under Taiwanese Mandarin.

However, there is yet a comprehensive benchmark for evaluating vision-language models under Taiwanese Mandarin context that address both structured reasoning and open-ended generation like MT-Bench (Bai et al., 2024) or AlpacaEval (Li et al., 2023) which evaluates VLM understanding in a freeform generation format. Our work on VisTW-Dialogue aims to address this specific need in the field.

3 VisTW-Dialogue: Visual Free Form Dialogue Benchmark

VisTW-Dialogue is designed to bridge the gap between real-world user interactions and conventional model-evaluation protocols. We manually curate images and their corresponding dialogue prompts from everyday scenarios in Taiwan, such as interpreting metro route maps or computing shared bills from restaurant menus. To minimize data leakage, each image is **captured in-house** and selected to be

Figure 1: VisTW consists of two subsets: (1) MCQ - a collection of multiple-choice questions from 21 academic subjects; and (2) Dialogue - real-life images with corresponding questions requiring understanding of Traditional Chinese and Taiwan-specific cultural context.

unlikely to appear in public web corpora used for VLM training. The photographers are also tasked with authoring challenging questions grounded in the captured content, targeting both **spatial reasoning** and **Taiwan-specific cultural knowledge**.

To enable efficient data collection across diverse regions, we set up a dedicated Discord server integrated with a custom data-collection bot. We selected Discord for practical reasons: many participants already used the app, it supports direct photo uploads from mobile devices, and its threaded conversations simplify discussion and iteration. Contributors therefore only needed to join the server to submit examples. Our bot provides (i) submission acknowledgements, (ii) monitoring of emoji-based annotations for quality control, and (iii) a sandbox for testing different rating prompts within submission threads. Additional labeling details are provided in Appendix A.

Following each submission, we apply a multi-stage quality-control (QC) pipeline:

1. Answer provision. The contributors who submitted an image are required to provide a reference answer to their question in the corresponding submission thread.

2. Peer review. Other team members review the question–answer pair and revised the reference answer when necessary.

3. Community moderation. We implement lightweight quality control via Discord emoji reactions: members could flag unsuitable submissions using downvotes, which were automatically logged by our bot.

VLM Judge	With Img	W/o Img
Claude-3.5-Sonnet	**.834**	.828
Gemini-2.0-Flash	**.824**	.819
GPT-4o	.799	**.828**
Gemini-2.0-Pro	.820	**.832**
Qwen2.5-VL-72B	**.750**	.734
GPT-4o-mini	.735	**.736**
Ensemble	.846	**.847**

Table 1: Spearman correlations between human-assigned scores and VLM-generated scores. *With Img* includes the image in the evaluation prompt; *W/o Img* uses only text. Higher values indicate better alignment with human judgment.

4. Filtering. During dataset cleaning, we remove submissions that received more than two downvotes (typically from a participant and a moderator).

Data collection ran from **July 2024** to **December 2025**, yielding **141** unique image–question pairs with reference answers. Image resolution and aspect-ratio statistics for **VisTW-Dialogue** are reported in Appendix C.3.

3.1 Automatic Evaluation : VLM as a Judge

To reduce the barrier of quick evaluations, we use a VLM as an automated judge to score responses based on questions, images, and ground truth answers. Using LLMs to evaluate free-form generation has become standard practice, exemplified by benchmarks such as AlpacaEval (Li et al., 2023), MT-Bench (Zheng et al., 2023), Vibe-Eval(Padlewski et al., 2024). Following prior works, we craft an evaluation prompt that explicitly includes scoring criteria, the original question, the assistant's response and specifically the ground truth of the question. The LLM judge is instructed to first provide a detailed explanation followed by the final numerical score from 0 to 10, the detail breakdown of score criteria can be found in Appendix I.1.

To validate the reliability of LLM-generated scores, we also conduct human evaluations across four selected models (gpt-4o, claude-3-5-sonnet, gpt-4o-mini, claude-3-7-sonnet). Each response is independently scored by three human annotators, and the average rating is calculated to establish a human evaluation baseline. We evaluate several prominent VLMs including gemini-2.0-pro, gemini-2.0-flash, Qwen 2.5 VL 72B, claude-3-5-sonnet-20241022, gpt-4o-2024-08-06 and gpt-4o-mini-2024-07-18. Table 1 reports the Spearman correlations between LLM-generated scores and human evaluations.

Our analysis reveals that incorporating images into the evaluation generally improves model-human score correlations, contradicting previous findings by Padlewski et al. (2024), which indicate minimal improvements in multimodal evaluation. Notably, the `gemini-2.0-flash` achieves performance comparable to `claude-3-5-sonnet-20241022`, despite costing approximately 36 times less. When we averages the scores from all VLM judges into an ensemble approach, we find that it exhibits the strongest correlation with human scores, reaching 0.8466 without images and 0.8463 with images, suggesting that combining diverse model judgments provides more robust evaluation than any single model alone.

All reported correlations are statistically significant with $p < 0.001$. The strongest correlation is observed with `claude-3-5-sonnet-20241022` ($\rho = 0.8336$, $p < 0.001$) when images are included. Given these results and future cost considerations, we select `gemini-2.0-flash-001` as our primary LLM judge. The complete evaluation of the performance of one model on VisTW-Dialogue cost merely $0.05 USD when using `gemini-2.0-flash-001` as the automated judge. Previous work (Dubois et al., 2024) found using LLMs as judge is prone to prefer longer response than shorter ones, we find no significant correlation between response length from automated judge scores with a Spearman's $\rho = 0.26789$, $p < 0.001$, indicating that our evaluation framework does not exhibit substantial length bias. Additional details can be found in Appendix I.1.

Dataset	Lang.	Subj.	Test / Val
MMMU	EN	30	10,500 / 900
JMMMU	JA	28	1,320 / –
CMMU	ZH-CN	7	1,803 / 1,800
TaiwanVQA	ZH-TW	13	2,000 / –
VisTW-MCQ	**ZH-TW**	**21**	**3,795 / 869**

Table 2: Comparison with similar datasets: MMMU (Yue et al., 2024), JMMMU (Onohara et al., 2024), CMMU (He et al., 2024), and Taiwan-VQA (Hsieh et al., 2025).

4 VisTW-MCQ : Visual Multi-Choice Question Answering

Multiple-choice question answering (MCQA) is widely used to assess the knowledge and reasoning capabilities of large language models (LLMs). In this section, we introduce **VisTW-MCQ**, a visual MCQA benchmark built from past examination questions spanning diverse educational levels in Taiwan, from primary and secondary school to specialized undergraduate courses (e.g., veterinary medicine).

Our benchmark dataset is constructed using real-world examination papers collected from publicly available sources spanning 2013 to 2024. We select subjects specifically requiring visual comprehension, such as medical diagnostics (e.g., interpreting X-ray and ultrasound images), geometry, electronic circuit design, and chemistry. The curation process involves manual extraction of questions and images from official exam archives and online PDFs. Each pair of questions and images is independently reviewed by a second annotator to verify content accuracy, clarity, and completeness, ensuring the absence of contextual errors, missing information, or invalid answer choices. To maintain consistent baseline guessing performance, the answer choices for each question are randomly shuffled, ensuring a uniform probability (approximately 25%) of selecting the correct answer by chance.

The final **VisTW-MCQ** dataset includes 21 distinct subjects covering diverse areas such as medical diagnosis, nautical science with cartographic analysis and technical disciplines. Each subject contains a minimum of 100 test questions, along with 5 few-shot learning examples as a development set, and 10 questions dedicated to validation purposes. For more details on image resolutions and aspect ratio, we include this in Appendix C.2. Compared with existing multi-choice benchmarks in Table 2, existing visual multi-choice benchmarks, such as CMMU (He et al., 2024), JMMMU (Onohara et al., 2024), and ALM-Bench (Vayani et al., 2024), **VisTW-MCQ** stands out with broader subject diversity.

4.1 Evaluation Method

To evaluate the visual reasoning capabilities of Vision-Language Models (VLMs) across different subjects, we employ a zero-shot Chain-of-Thought prompting (Kojima et al., 2022). Each VLM is instructed to explicitly provide reasoning steps followed by by an explicit final option. However, we observe that certain VLMs occasionally struggle to comply with formatting instructions, resulting in invalid responses. To mitigate evaluation bias due to parsing errors, we subsequently employ an auxiliary LLM-based parser (`gpt-4o-mini`) to extract and validate the final answer choice from each VLM response. We report the average scores for all 21 subjects as our final score.

However, one limitation of the MCQA benchmark format is that it does not reflect the real world use of VLM which mostly in free-form question answering format, which we have address this limitation in Section 3.

5 Benchmark Results

We evaluated 53 vision-language models (VLMs) spanning both closed and open-weight categories; however, due to space limitations, we report only the top 10 overall models and the best open-weight models under 20B parameters in Table 8. The comprehensive evaluation results for all models are detailed in Appendix D. Detailed model specifications are provided in Appendix E. Our analysis reveals a strong correlation between VisTW-Dialogue and VisTW-MCQ performance (Spearman's $\rho = 0.7746$, $p < 0.001$; Kendall's $\tau = 0.6086$, $p < 0.001$). In the overall leaderboard, the Gemini series maintains its lead, with consistent improvements from version 2.5 to the latest 3 series. While OpenAI models show notable gains from o3 to GPT-5.2, GPT-5.2 still falls short of Gemini 2.5-flash. Among open-weight models, the Qwen3 235B MoE model achieves the best performance.

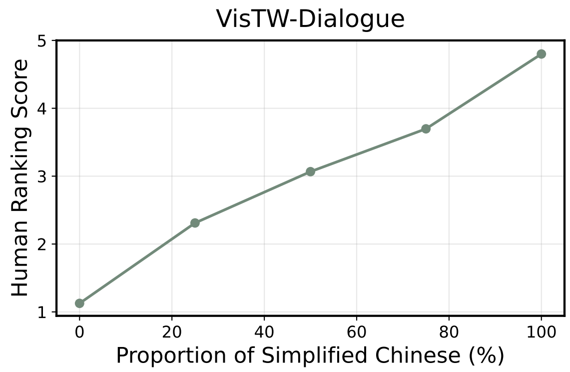


Figure 2: Human preference rankings across varying proportions of Simplified Chinese characters. Three annotators ranked five variants per question; a lower rank indicates higher preference (1 = best, 5 = worst).

Among Traditional Chinese-specialized VLMs, we evaluated three models: Breeze2-3B, Breeze2-8B, and llama3.2-ffm-11b-v-32k-chat. The latter is a fine-tuned version of Llama 3.2 11B Vision Instruct, post-trained on Taiwanese Mandarin corpus by the Taiwanese cloud provider twcloud. Despite these localization efforts, the Qwen3 series achieves the best results among open-weight models, leading both the open-weight and overall leaderboards.

Examining the performance of smaller VLMs reveals interesting efficiency-capability tradeoffs. Models with fewer parameters show consistent degradation in VisTW-Dialogue scores: Breeze2-3B scores just 2.84 compared to 4.06 for its 8B counterpart, a pattern also observed in the Qwen3 and InternVL2.5 series. Interestingly, for VisTW-MCQ, Breeze2-3B (0.2947) performs similarly with Breeze2-8B (0.2894), this is observed in Qwen3 series (8B and 4B) as well. This suggests that the dialogue task, which requires more nuanced multimodal understanding, may be more sensitive to model scale than the multiple-choice format.

6 Analysis

6.1 Penalty Based on Human Preference

When inspecting VLM outputs, we find that some responses mix Simplified and Traditional Chinese characters, while others are written entirely in Simplified Chinese. This observation motivates a key question: *to what extent do Traditional Chinese readers perceive such mixing as distracting or unacceptable?*

Model	VisTW-MCQ		VisTW-Dialogue		Avg
	Accuracy	Rank	Score 0-10	Rank	Rank
Overall Top 10					
gemini-3-pro-preview	0.9290	1	8.32	1	1.0
gemini-3-flash-preview	0.8938	2	8.15	2	2.0
gemini-2.5-flash	0.7779	5	7.38	5	5.0
gpt-5.2	0.8278	4	6.93	6	5.0
o3-2025-04-16	0.7695	6	6.81	7	6.5
gemini-2.5-pro-preview-03-25	0.6035	12	7.91	3	7.5
o4-mini-2025-04-16	0.6977	7	6.47	9	8.0
gemini-2.0-flash-001	0.6515	9	6.49	8	8.5
Qwen3-vl-235B-A22B-Instruct	0.6947	8	6.46	10	9.0
gpt-4.1	0.6433	10	6.31	11	10.5
Open-Weight Sub-20B Models					
Qwen3-VL-8B-Instruct	0.5427	18	5.59	18	18.0
Qwen3-VL-4B-Instruct	0.5046	19	5.19	23	21.0
gemma-3-12b-it	0.4841	21	4.39	28	24.5
llama3.2-ffm-11b-v-32k-chat	0.3113	41	3.55	39	40.0
Qwen2.5-VL-7B-Instruct	0.2764	50	4.16	31	40.5
Breeze2-8B-Instruct	0.2894	48	4.06	34	41.0
Qwen2-VL-7B-Instruct	0.2987	43	3.41	42	42.5
Llama-3.2-11B-Vision-Instruct	0.3234	39	3.03	47	43.0
Ministral-3-14B-Instruct-2512	0.3489	36	2.39	53	44.5
InternVL2-5-8B	0.2688	53	3.91	37	45.0
Nemotron-Nano-12B-v2-VL	0.3029	42	2.93	49	45.5
InternVL2-8B-MPO	0.2449	54	3.61	38	46.0
gemma-3-4b-it	0.2984	44	2.96	48	46.0
InternVL2-5-4B	0.2440	55	3.54	40	47.5
Breeze2-3B-Instruct	0.2947	46	2.84	50	48.0

Table 3: For VisTW-MCQ we prompt our model in zero shot chain of thought fashion and report the average score from all 21 subjects. Models are sorted by average rank from both subsets.

To quantify human sensitivity to Simplified Chinese character intrusion, we conduct a controlled survey based on **VisTW-Dialogue**. For each prompt, we start from a response generated by **Gemini-2.5-Pro** and create five variants by systematically controlling the Simplified-to-Traditional character ratio (**0%**, **25%**, **50%**, **75%**, **100%**). Participants then ranked the variants by overall preference.

Figure 2 shows the resulting preference curve. Responses written entirely in Traditional Chinese (**0%**) receive the highest preference, and preference decreases monotonically as the proportion of Simplified characters increases. Given the ordinal nature of rankings, we summarize preferences by the **mean rank** for each proportion. To assess reliability, we measure inter-annotator agreement with rank-based metrics. Annotators exhibit strong consistency, with Kendall's $\tau = 0.7277$ ($p < 0.001$) and Spearman's $\rho = 0.8133$ ($p < 0.001$). These results indicate a clear human aversion to character mixing in Traditional Chinese settings, motivating a **Simplified-character penalty** in our evaluation protocol to better align automatic scores with human preferences. The exact penalty formulation is provided in Appendix G.

6.2 Impact of Penalty on Model Ranking

As shown in Table 4, the implementation of the penalty mechanism reveals a clear divergence in ranking sensitivity between model tiers. **Top-tier models** (e.g., Gemini 3 series and GPT-5.2) remain stable in their rankings ($\Delta = 0$). In contrast, we observe that the penalty primarily affects open-

Model	R_{orig}	R_{pen}	Δ
Closed-Source SOTA Models			
gemini-3-pro-preview	1.0	1.0	**0**
gemini-3-flash-preview	2.0	2.0	**0**
gemini-2.5-flash	5.0	5.0	**0**
gpt-5.2	5.0	5.0	**0**
o3-2025-04-16	6.5	6.5	**0**
Open-Weight Sub-20B Models			
Qwen3-VL-8B-Instruct	19.5	18.0	**+1.5**
Qwen3-VL-4B-Instruct	23.5	21.0	**+2.5**
Gemma-3-12b-it	27.0	24.5	**+2.5**
llama3.2-ffm-11b-v-32k-chat	47.0	40.0	**+7.0**
Qwen2.5-VL-7B-Instruct	33.5	40.5	**-7.0**
Breeze2-8B-Instruct	46.0	41.0	**+5.0**

Table 4: Comparison of model average rankings before and after applying the preference penalty. R_{orig} denotes the ranking based on raw performance scores; R_{pen} represents the ranking after applying the penalty; Δ indicates the rank shift.

weight sub-20B models. For instance, *Qwen2.5-VL-7B-Instruct* saw a significant drop ($\Delta = -7.0$), while models like *Breeze2-8B* improved their relative standing ($\Delta = +5.0$) due to better adherence to language constraints. This behavior can be attributed to the fact that smaller open-weight models are more prone to generating responses in Simplified Chinese or mixed scripts, thereby triggering the penalty more frequently. As a result, the ranking adjusting mainly reflects differences in language control and output robustness, rather than changes in core reasoning ability. The leaderboard results evaluated **without penalty** are provided in Appendix F.

6.3 Sensitivity of Image Resolution

We investigate the effect of image resolution on model performance via systematic scaling experiments. For VisTW-MCQ, we evaluate both directions: **upscaling** images to $2\times$ and $4\times$ the original resolution using SwinIR (Liang et al., 2021), and **downscaling** images to $1/2$, $1/4$, and $1/8$ of the original size. For VisTW-Dialogue, because images are already high-resolution, we only perform downscaling at $1/2$, $1/4$, and $1/8$.

Figure 3 illustrates our findings on resolution sensitivity. For VisTW-MCQ, performance remained consistent across resolution variations, suggesting that even lower resolutions contained sufficient information for multiple-choice tasks. In contrast, for VisTW-Dialogue, model performance stabilized at approximately half the original resolution. We observed model-specific variations: `gemini-2.0-flash-lite` showed slight

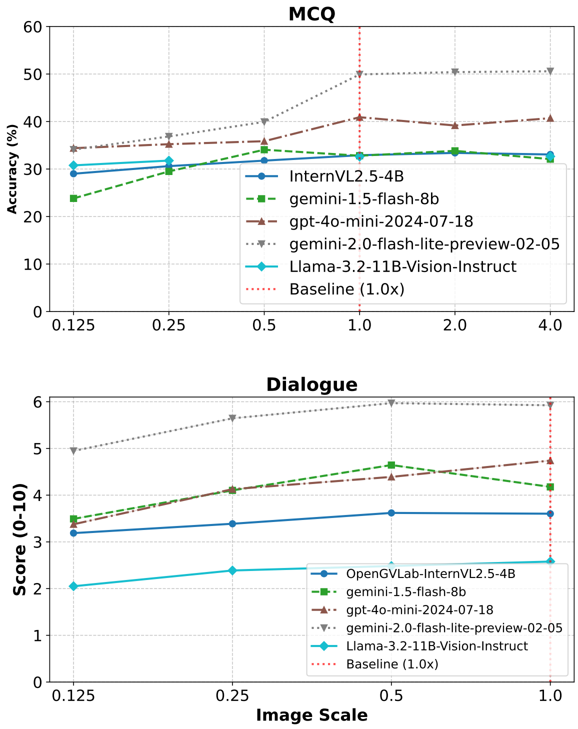


Figure 3: Effect of image resolution scaling on model performance. Left: Performance on MCQ with both upscaled ($2\times$, $4\times$) and downscaled ($1/2$, $1/4$, $1/8$) images. Right: Performance on Dialogue with downscaled images.

improvements with increased resolution, while `gemini-1.5-flash-8B` exhibited minor performance declines, indicating different architectural sensitivities to input resolution.

6.4 Comparison between With and Without Vision Cues

In the previous section, we show score degradation with reduced image resolution. Here, we examine performance when image context is entirely omitted. For MCQ tasks, this analysis establishes a baseline that measures how much linguistic knowledge alone contributes to task performance before visual information becomes beneficial. Similarly, for dialogue tasks, images-free score serve as baseline measurements, indicating how significantly performance drops when visual context is absent. We conduct this experiment using the same models evaluated in the previous section. Table 5 presents the results, which show consistent score reductions across all four models in both datasets when images are withheld. The performance degradation is expected, as it indicates that VLMs struggle when visual information is not provided.

Model	MCQA		Dialogue	
	w/	w/o	w/	w/o
Gemini-2.0-Flash-Lite	49.92	32.37	5.92	2.35
GPT-4o-mini	40.91	32.76	4.74	1.95
Gemini-1.5-Flash-8B	32.80	25.62	4.18	1.85
InternVL2.5-4B	32.91	28.64	3.60	1.29

Table 5: Performance comparison of various multimodal models across different evaluation tasks in VisTW. MCQA scores represent accuracy (%), while dialogue scores represent quality ratings. Higher scores indicate better performance.

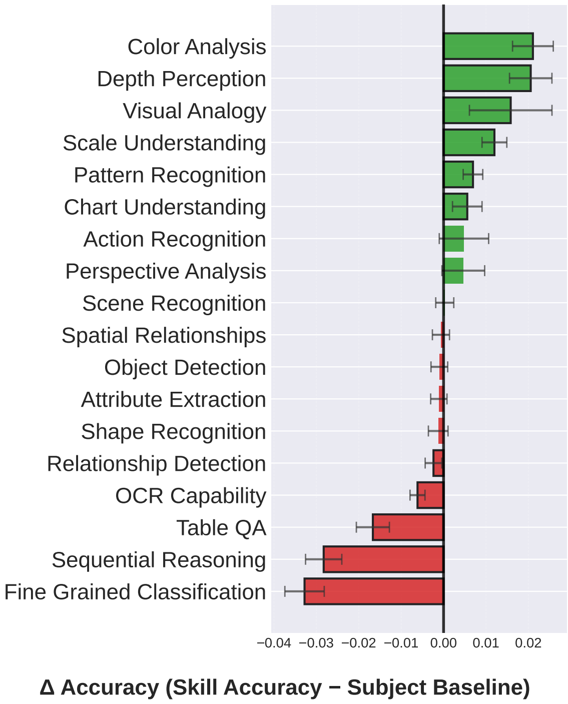


Figure 4: Skill Difficulty Relative to Subject Baseline Across 30 Models, black borders highlight statistically significant differences.

6.5 Skills breakdown

To characterize what current VLMs handle well—and where they fail—in Traditional Chinese settings, we annotate each VisTW-MCQ item with a set of image-grounded skills. Concretely, we define 18 vision skills required to solve the full benchmark, including chart understanding and OCR for recognizing Chinese characters. Our analysis shows that **table understanding** remains challenging, as do (i) **multi-step reasoning** that links sequential concepts within an image and (ii) **fine-grained classification** under an extremely long-tailed label distribution.

6.6 Correlations with Other Benchmarks

As shown in Figure 5, we compare our VisTW-MCQ results with exam style benchmarks: MMMU (Yue et al., 2024) and CMMU (He et al., 2024) on a selected set of models spanning different scales. Despite the relatively small subset (due to computational and financial constraints), we observe a broadly consistent trend: models that perform well on MMMU or CMMU also tend to score higher on VisTW-MCQ. However, we also note that some points deviate from the main diagonal, suggesting that the three benchmarks are not fully interchangeable. The models show stronger relative performance on MMMU and CMMU compare to VisTW-MCQ scores.

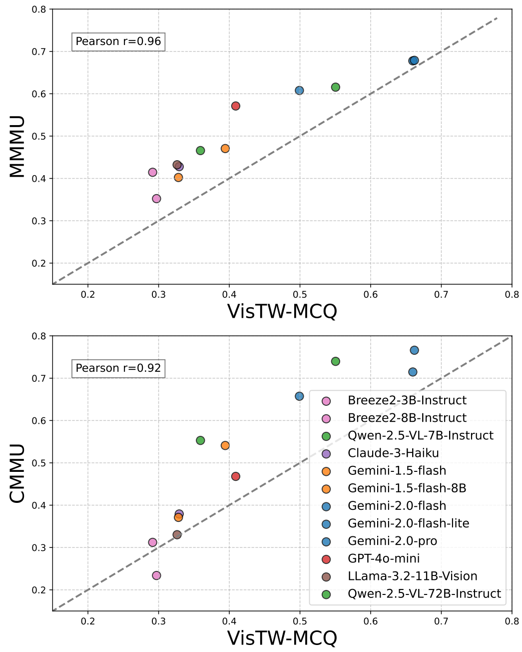


Figure 5: Comparison of VisTW-MCQ scores with MMMU (top) and CMMU (bottom) on a selected subset of models of varying scales. We observe a rough correlation across the three benchmarks, though some deviations suggest differences in the specific knowledge or reasoning skills each test emphasizes.

7 Conclusion

In this work, we introduce **VisTW-MCQ** and **VisTW-Dialogue**, the first comprehensive benchmarks for evaluating vision–language models in **Traditional Chinese** under **Taiwanese Mandarin** contexts. Our results show that even VLMs trained specifically for Traditional Chinese can substantially underperform their Simplified-Chinese counterparts on Taiwan-centric multimodal inputs. We also observe a strong correlation between performance on structured (MCQ) and free-form dialogue tasks, suggesting that capabilities transfer across evaluation formats while capturing complementary aspects of VLM behavior. Finally, our analyses of image resolution and model scaling offer practical insights for deployment. Looking forward, we encourage work on improving robustness to Traditional Chinese orthography and on expanding VisTW to cover broader cultural contexts within Taiwan.

Acknowledgments

This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 112-2223-E002-012-MY5. We thank the National Center for High-performance Computing of National Institutes of Applied Research (NIAR) in Taiwan for providing computational and storage resources.

Limitations

All images in VisTW-Dialogue were collected by our researchers specifically for this benchmark, with contributors explicitly waiving ownership rights and consenting to public use of their images. We implemented strict guidelines to ensure no personally identifiable information (PII) was included—images containing names, faces of non-public figures, Wi-Fi passwords, or residential addresses were prohibited and filtered through community moderation. For VisTW-MCQ, we used only publicly available exam questions from official sources spanning 2013-2024, maintaining academic integrity while respecting intellectual property.

Our work directly addresses the underrepresentation of Traditional Chinese in vision-language benchmarks, promoting greater linguistic and cultural inclusion in AI evaluation. By creating resources specifically for Traditional Chinese, we help ensure that NLP advancements benefit communities beyond those using dominant languages like English or Simplified Chinese. Our benchmark incorporates diverse geographical sampling across Taiwan through contributions from nine individuals in different regions, though we acknowledge a concentration of samples from Taipei.

References

Mistral AI. 2025. Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1.

Anthropic. 2024a. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Anthropic. 2024b. Claude 3.5 sonnet model card addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and 1 others. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

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. *Preprint*, arXiv:2502.13923.

Kaibing Chen, Dong Shen, Hanwen Zhong, Huasong Zhong, Kui Xia, Di Xu, Wei Yuan, Yifei Hu, Bin Wen, Tianke Zhang, and 1 others. 2024a. Evlm: An efficient vision-language model for visual understanding. *arXiv preprint arXiv:2407.14177*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, and 16 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Preprint*, arXiv:2404.16821.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Google DeepMind Gemma Team. 2025. Gemma 3 technical report. https://storage.googleapis.com/deepmind-media/gemma/Gemma3Report.pdf.

Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. 2024. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning. *arXiv preprint arXiv:2401.14011*.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, and 6 others. 2024. Cogvlm2: Visual language models for image and video understanding. *Preprint*, arXiv:2408.16500.

Hsin-Yi Hsieh, Shang-Wei Liu, Chang-Chih Meng, Chien-Hua Chen, Shuo-Yueh Lin, Hung-Ju Lin, Hen-Hsen Huang, and I-Chen Wu. 2025. Taiwanvqa: Benchmarking and enhancing cultural understanding in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da-shan Shiu. 2023. Advancing the evaluation of traditional chinese language models: Towards a comprehensive benchmark suite. *arXiv preprint arXiv:2309.08448*.

Chan-Jan Hsu, Chia-Sheng Liu, Meng-Hsi Chen, Muxi Chen, Po-Chun Hsu, Yi-Chang Chen, and Da-Shan Shiu. 2025. The breeze 2 herd of models: Traditional chinese llms based on llama with vision-aware and function-calling capabilities. *arXiv preprint arXiv:2501.13921*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2024. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. *arXiv preprint arXiv:2410.17250*.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, and 1 others. 2024. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *arXiv preprint arXiv:2405.02287*.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Preprint*, arXiv:2204.05149.

MediaTek Research, :, Chan-Jan Hsu, Chia-Sheng Liu, Meng-Hsi Chen, Muxi Chen, Po-Chun Hsu, Yi-Chang Chen, and Da-Shan Shiu. 2025. The breeze 2 herd of models: Traditional chinese llms based on llama with vision-aware and function-calling capabilities. *Preprint*, arXiv:2501.13921.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Zhi Rui Tam, Ya Ting Pai, Yen-Wei Lee, Hong-Han Shuai, Jun-Da Chen, Wei Min Chu, and Sega Cheng. 2024. Tmmlu+: An improved traditional chinese evaluation suite for foundation models. In *First Conference on Language Modeling*.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Yuxuan Wang, Yijun Liu, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, and Wanxiang Che. 2024. Cvlue: A new benchmark dataset for chinese vision-language understanding evaluation. *arXiv preprint arXiv:2407.01081*.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, and 1 others. 2024. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. *arXiv preprint arXiv:2410.12705*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. 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.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Creating VisTW-Dialogue on Discord

Guidelines: The collected question-image pairs were designed to be challenging for Vision Language Models (VLMs), focusing on realistic scenarios where users might seek assistance (e.g., for visual impairment). All contributors were instructed to ensure their submissions contained no personally identifiable information (PII), including names or faces of non-public figures, Wi-Fi passwords, or any information that could reveal residential addresses. The dataset comprises question-image pairs contributed by 9 different individuals from various locations across Taiwan. While the geographic distribution covers multiple regions, there is a higher concentration of samples from Taipei (the capital of Taiwan) as the majority of our contributors resided there.

A.1 Human Annotation Process

As illustrated in Figure 6, steps to label a problem are:

1. Upload an image, use @ + hanji to submit it to the bot on discord, and include the question and ground_truth <standard answer>

2. The bot will create a discussion thread, or people can wait until the thread appears before adding the standard answer.

3. Format for updating the standard answer: as long as the text mentions "ground_truth:", the system will automatically capture the text that follows as the standard answer.

4. The standard answer can be updated repeatedly using step 3.

5. People can type "validate_result" to get score evaluations for responses from two models.

Note: When people upload images, they waive ownership rights and agree to make the photos public.

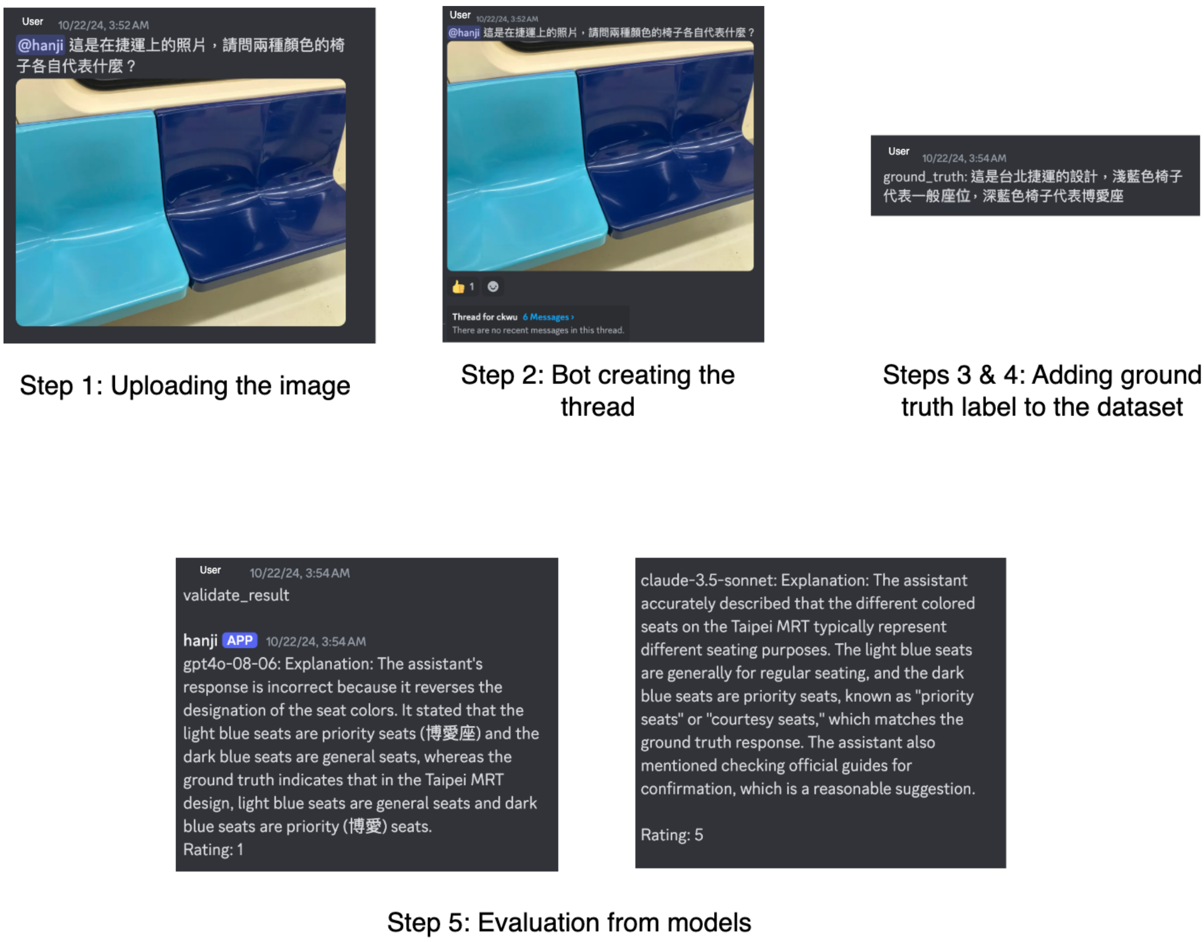


Figure 6: Workflow of adding images, labeling, and evaluating

B Contingency Plan When the Gemini-2.0-Flash Is Retired

Due to our choice of using a close weight VLM : Gemini 2.0 Flash as our judge; this means it will eventually become unavailable when newer versions become available. Consequently, scores assigned by a retired model, such as *gemini-2.0-flash* will become difficult to compare with those of the newer VLMs. Many LLM-as-judge works overlook this scenario, typically re-running an entire benchmark using the new model as judge. However, large-scale evaluations make such an approach infeasible.

To address this, we propose a calibration-based contingency plan using *Qwen2.5-VL 72B*. Figure 7 (top) shows that the raw scores for Qwen2.5-VL are systematically higher than those from gemini-2.0-flash. We therefore fit a linear mapping (based on five-vote distributions) to align Qwen2.5-VL scores with gemini-2.0-flash. After calibration, instances that gemini-2.0-flash scores as zero now cluster near zero under Qwen2.5-VL, however the number of perfect scores is reduced (Figure 7, bottom). After calibration, the spearman correlation with human score and ensemble of Qwen2.5-VL scores improve from 0.7993 to 0.8070.

Table 6 compares raw vs. calibrated Qwen2.5-VL scores to Gemini-2.0-flash on *VisTW-Dialogue*, demonstrating how calibration improves alignment. Our result offers a practical strategy to maintain continuity of evaluation even if the original judge model is no longer available.

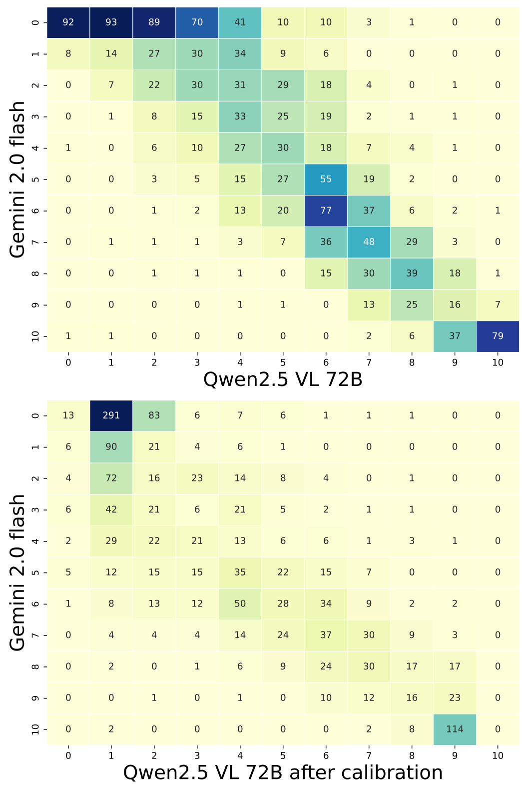


Figure 7: Comparison of Gemini 2.0 flash and Qwen 2.5 VL 72B score distribution rounded to the nearest integer (top) and Qwen 2.5 VL 72B scores after calibrated against Gemini 2.0 flash (bottom).

Eval Model	Gemini	Qwen	Qwen (calibrated)	Δ (Cal - Gem)
Gemini-2.0-flash-001	6.15	6.65	5.73	-0.42
Qwen2.5-vl-72b-instruct	4.87	6.18	5.12	+0.25
Llama-Breeze2-8B-Instruct	3.14	4.12	3.16	+0.02
Llama-Breeze2-3B-Instruct	2.90	4.07	3.06	+0.16

Table 6: Comparison of scores assigned by different judge models (Gemini, Qwen, and calibrated Qwen) across various evaluated models. The Δ column shows the difference between calibrated Qwen VL scores and Gemini scores, highlighting the effectiveness of the calibration. Values close to zero indicate better alignment.

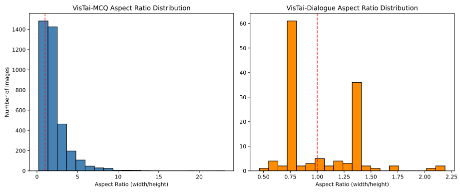


Figure 8: Histograms comparing the aspect ratio distributions (width/height) of the VisTW-MCQ dataset (left) and the VisTW-Dialogue dataset (right). The vertical dashed line on the right indicates an aspect ratio of 1 (i.e., square). Most MCQ images fall below an aspect ratio of 1 (portrait), while the Dialogue dataset shows a broader spread, with many images close to square.

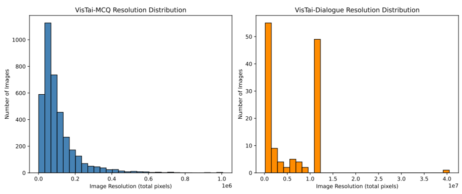


Figure 9: Comparison of image resolution distributions (in total pixels) for VisTW-MCQ (left) and VisTW-Dialogue (right). The x-axis shows the total number of pixels in each image, and the y-axis indicates how many images fall within each resolution range. The MCQ dataset tends to contain lower-resolution images, whereas the Dialogue dataset spans a broader range of resolutions.

C Additional Details for VisTW-MCQ and VisTW-Dialogue

C.1 Details of subjects for VisTW-MCQ

Table 7 shows each of the subjects found in VisTW-MCQ with their broad category, as well as total number of questions.

C.2 VisTW-MCQ

Figure 8 (left) shows the aspect ratio of mostly long rectangle, most of the width is longer than the height. In the Figure 9 (left) for the MCQ dataset most of the image pixel is below 1M pixels.

C.3 VisTW-Dialogue

Figure 8 (right) shows the aspect ratio of mostly long rectangle, most of the width is longer than the height. In the Figure 9 (right) for the MCQ dataset most of the image pixel is above 1M pixels.

Subject Name	**Chinese Name**	**Test**	**Val**	**Dev**
Accounting	會計學	100	29	5
Arts	藝術	385	91	5
Biology	生物學	150	8	5
Chemistry	化學	172	38	5
Chinese Literature	中國文學	100	45	5
Dentistry	牙醫學	220	49	5
Electronic Circuits	電子電路	388	91	5
Fundamentals of Physical Therapy	物理治療基礎	150	23	5
Geography	地理學	100	39	5
Mathematics	數學	240	54	5
Mechanics	力學	217	49	5
Medical	醫學	172	37	5
Music	音樂	100	6	5
Natural Science	自然科學	303	70	5
Navigation	航海學	100	16	5
Pharmaceutical Chemistry	藥物化學	100	45	5
Physics	物理學	100	23	5
Sociology	社會學	348	81	5
Statistics	統計學	100	45	5
Structural Engineering	結構工程	100	19	5
Veterinary Medicine	獸醫學	150	11	5

Table 7: Overview of subjects with VisTW-MCQ splits

D Detailed Experimental Results

In this appendix, we provide the comprehensive evaluation results for all VLMs tested on our VisTW benchmark. As discussed in Table 8, while the main text highlights the top 10 models, this section presents the full performance breakdown across both `VisTW-Dialogue` and `VisTW-MCQ` subsets.

Model	VisTW-MCQ		VisTW-Dialogue		Avg
	Accuracy	Rank	Score 0-10	Rank	Rank
gemini-3-pro-preview	0.9290	1	8.32	2	1.5
gemini-3-flash-preview	0.8938	2	8.15	3	2.5
gpt-5.2	0.8278	4	7.73	4	4.0
o3-2025-04-16	0.7695	6	6.64	6	6.0
qwen3-vl-235b-a22b-instruct	0.6947	8	6.60	7	7.5
o4-mini-2025-04-16	0.6977	7	6.47	9	8.0
gemini-2.5-pro-preview-03-25	0.6035	13	7.31	5	9.0
gpt-4.1	0.6433	11	6.54	8	9.5
gemini-2.0-flash-001	0.6515	10	6.47	10	10.0
claude-sonnet-4-5	0.6645	9	6.05	13	11.0
gpt-4o-2024-11-20	0.5700	17	6.09	12	14.5
Qwen3-VL-30B-A3B-Instruct	0.6242	12	5.74	18	15.0
claude-3-5-sonnet-20241022	0.5955	14	5.82	16	15.0
gpt-4.1-mini	0.5754	16	5.88	15	15.5
gemini-2.5-pro-exp-03-25	0.3727	34	8.33	1	17.5
gemini-2.0-flash-lite-preview-02-05	0.4947	20	5.79	17	18.5
qwen2.5-vl-72b-instruct	0.5437	18	4.58	24	21.0
claude-haiku-4.5	0.5766	15	4.36	28	21.5
qwen2.5-vl-32b-instruct	0.4630	23	5.09	21	22.0
gemini-2.0-flash-thinking-exp-1219	0.3676	35	6.44	11	23.0
gemini-1.5-pro	0.4368	26	5.00	22	24.0
gpt-4o-2024-08-06	0.4000	31	5.59	19	25.0
gpt-4o-mini-2024-07-18	0.4066	29	4.66	23	26.0
mistral-small-3.1-24b-instruct-2503	0.4517	25	4.40	27	26.0
gemma-3-27b-it	0.4332	27	4.54	26	26.5
gemini-2.5-pro	0.8415	3	2.53	51	27.0
gemini-2.5-flash	0.7779	5	2.39	52	28.5
Llama-3.2-90B-Vision-Instruct	0.4078	28	4.17	30	29.0
gemini-1.5-flash	0.3887	33	4.23	29	31.0
llama-4-maverick	0.4561	24	3.48	41	32.5
gpt-4.1-nano	0.3938	32	3.99	36	34.0
qvq-72b-preview	0.4066	30	3.76	39	34.5
gemini-1.5-flash-8b	0.3251	39	4.11	32	35.5
claude-3-haiku-20240307	0.3266	38	4.07	33	35.5
qwen2-vl-72b-instruct	0.0000	53	4.55	25	39
llama-4-scout	0.3518	36	3.22	43	39.5
gemma-3n-e4b-it	0.3153	41	2.80	50	45.5
nova-lite-v1	0.2976	46	3.17	45	45.5
Open-Weight Sub-20B Models					
Qwen3-VL-8B-Instruct	0.5427	19	5.93	14	16.5
Qwen3-VL-4B-Instruct	0.4901	21	5.19	20	20.5
gemma-3-12b-it	0.4841	22	3.90	37	29.5
Qwen-Qwen2.5-VL-7B-Instruct	0.2764	49	4.16	31	40.0
MediaTek-Research-Llama-Breeze2-8B-Instruct	0.2894	48	4.06	34	41.0
OpenGVLab-InternVL2-5-8B	0.2688	50	4.06	35	42.5
llama3.2-ffm-11b-v-32k-chat	0.3113	42	3.21	44	43.0
Qwen-Qwen2-VL-7B-Instruct	0.2987	44	3.41	42	43.0
meta-llama-Llama-3.2-11B-Vision-Instruct-Turbo	0.3234	40	3.03	47	43.5
Ministral-3-14B-Instruct-2512	0.3489	37	2.17	53	45.0
OpenGVLab-InternVL2-5-4B	0.2440	52	3.80	38	45.0
NVIDIA-Nemotron-Nano-12B-v2-VL-BF16	0.3029	43	2.95	48	45.5
gemma-3-4b-it	0.2984	45	3.09	46	45.5
OpenGVLab-InternVL2-8B-MPO	0.2449	51	3.61	40	45.5
Breeze2-3B-Instruct	0.2947	47	2.83	49	48.0

Table 8: Detailed results on the VisTW benchmark. Models are sorted by average rank from both subsets.

E Benchmarked VLM Details

For open weight models, we utilize local GPU resources (3090) for models under 11B parameters, while larger models are accessed through the Open Router API service. For closed weight models, we access them directly through each organization's official API.

Model	Organization	Size	Arch
Closed Source Models			
claude-3-haiku-20240307 (Anthropic, 2024a)	Anthropic	-	-
claude-3.5-sonnet-20241022 (Anthropic, 2024b)	Anthropic	-	-
gemini-1.5-flash (Team, 2024)	Google	-	Dense
gemini-1.5-flash-8b (Team, 2024)	Google	8B	Dense
gemini-1.5-pro (Team, 2024)	Google	-	MoE
gemini-2.0-pro-exp-02-05 (Team, 2024)	Google	-	-
gemini-2.0-flash-thinking-exp-1219 (Team, 2024)	Google	-	-
gemini-2.0-flash-001 (Team, 2024)	Google	-	-
gemini-2.0-flash-lite-preview-02-05 (Team, 2024)	Google	-	-
gpt-4o-2024-11-20 (OpenAI, 2024)	OpenAI	-	-
gpt-4o-2024-08-06 (OpenAI, 2024)	OpenAI	-	-
gpt-4o-mini-2024-07-18 (OpenAI, 2024)	OpenAI	-	-
nova-lite-v1	Amazon	-	-
Open Weights Models			
Mistral 24B Instruct (2503)(AI, 2025)	Google	24B	Dense
Gemma3 27B Instruct (Gemma Team, 2025)	Google	27.4B	Dense
Qwen3-238B-A22B-Instruct (Bai et al., 2025)	Alibaba	238B	MoE
Qwen3-30B-A3B-Instruct (Bai et al., 2025)	Alibaba	30B	MoE
Qwen2.5-VL-72b-instruct (Bai et al., 2025)	Alibaba	72B	Dense
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	Alibaba	7B	Dense
Qwen2-VL-72b-instruct (Bai et al., 2023)	Alibaba	72B	Dense
Qwen2-VL-7B-Instruct (Bai et al., 2023)	Alibaba	7B	Dense
InternVL2.5-8B (Chen et al., 2025)	OpenGVLab	8B	Dense
InternVL2-8B-MPO (Chen et al., 2024b)	OpenGVLab	8B	Dense
InternVL2.5-4B (Chen et al., 2025)	OpenGVLab	4B	Dense
InternVL2-8B (Chen et al., 2024b)	OpenGVLab	8B	Dense
InternVL2-4B (Chen et al., 2024b)	OpenGVLab	4B	Dense
InternVL2-2B (Chen et al., 2024b)	OpenGVLab	2B	Dense
InternVL2-1B (Chen et al., 2024b)	OpenGVLab	1B	Dense
Llama-3.2-90B-Vision-Instruct (Patterson et al., 2022)	Meta	90B	Dense
Llama-3.2-11B-Vision-Instruct (Patterson et al., 2022)	Meta	11B	Dense
CogVLM2-llama3-chinese (Hong et al., 2024)	THUDM	19B	Dense
Breeze2-8B-Instruct (Research et al., 2025)	MediaTek	8B	Dense
Breeze2-3B-Instruct (Research et al., 2025)	MediaTek	3B	Dense
deepseek-vl2-tiny (Wu et al., 2024)	DeepSeek	-	MoE
deepseek-vl2-small (Wu et al., 2024)	DeepSeek	-	MoE

Table 9: Overview of evaluated models. For closed source models, sizes are marked with '-' where not publicly disclosed. Dense stands for Decoder only Dense Transformer architecture, MoE stands for Mixture of Experts architecture.

F Leaderboard Results without Penalty

As shown in Table 10, we presents the complete leaderboard results evaluated **without applying language preference penalty**. These results are provided to facilitate direct comparison with the penalized setting discussed in the main text and to isolate the impact of the penalty mechanism on model ranking.

Model	VisTW-MCQ		VisTW-Dialogue		Avg
	Accuracy	**Rank**	**Score 0-10**	**Rank**	**Rank**
Overall Top 10					
gemini-3-pro-preview	0.9327	1	8.50	1	1.0
gemini-3-flash-preview	0.9019	2	8.35	2	2.0
gemini-2.5-flash	0.7862	5	7.55	5	5.0
gpt-5.2	0.8363	4	7.04	6	5.0
o3-2025-04-16	0.7769	6	6.99	7	6.5
o4-mini-2025-04-16	0.7364	7	6.78	8	7.5
gemini-2.5-pro-preview-03-25	0.6072	13	8.11	3	8.0
Qwen3-vl-235b-A22b-instruct	0.7016	8	6.64	10	9.0
gemini-2.0-flash-001	0.6596	9	6.65	9	9.0
gpt-4.1	0.6503	11	6.48	11	11.0
Open-Weight Sub-20B Models					
Qwen3-VL-8B-Instruct	0.5479	19	5.69	20	19.5
Qwen3-VL-4B-Instruct	0.4639	24	5.31	23	23.5
Gemma-3-12b-it	0.4863	23	4.46	31	27.0
Qwen2.5-VL-7B-Instruct	0.3592	38	4.54	29	33.5
Ministral-3-14B-Instruct-2512	0.4431	26	3.40	47	36.5
Ministral-3-8B-Instruct-2512	0.4233	29	2.97	52	40.5
OpenGVLab-InternVL2-5-8B	0.3447	40	4.05	41	40.5
OpenGVLab-InternVL2-8B-MPO	0.3533	39	3.68	43	41.0
Qwen2-VL-7B-Instruct	0.3004	53	4.21	33	43.0
OpenGVLab-InternVL2-8B	0.3431	41	3.45	46	43.5
OpenGVLab-InternVL2-5-4B	0.3291	46	3.73	42	44.0
Breeze2-8B-Instruct	0.2915	55	4.12	37	46.0
llama3.2-ffm-11b-v-32k-chat	0.3119	50	3.65	44	47.0
Nemotron-Nano-12B-v2-VL	0.3330	44	3.02	51	47.5
Llama-3.2-11B-Vision-Instruct	0.3262	48	3.12	49	48.5

Table 10: Performance results on VisTW-MCQ using zero-shot chain-of-thought prompting **without character penalty** and report the average score from all 21 subjects. Models are sorted by average rank from both subsets.

G Simplified Chinese Penalty Mechanism

We compute the ratio of Simplified Chinese usage in the output text to discourage excessive use of Simplified characters. Only Chinese characters are considered, while non-Chinese characters are ignored. For each character, if the original form matches its Simplified variant but not its Traditional variant, it is treated as a *strong Simplified signal*. Conversely, characters that match the Traditional form but not the Simplified form are treated as *strong Traditional signals*. The Simplified ratio (ratio) is defined as the proportion of strong Simplified characters among all disambiguated Chinese characters.

A penalty is applied to the raw score using the following formula:

$$\text{penalty\_factor} = \max(0,\ 1 - \lambda \cdot \text{ratio})$$

where λ denotes the penalty coefficient (set to 1 in our experiments). When ratio $= 0$, indicating no Simplified bias, the score remains unchanged. When ratio $= 1$, indicating fully Simplified output, the final score is reduced by half. This mechanism introduces a smooth penalty while leaving non-Chinese and Traditional Chinese text unaffected.

H Direct Answer vs CoT in VisTW-MCQ

To analyze the impact of reasoning on performance, we compare two prompting strategies: Zero-Shot Chain of Thought (CoT) and Direct Answer prompting. The latter asks models to provide answers without intermediate reasoning steps. Table 11 presents the comparative results.

Interestingly, CoT does not universally benefit all vision-language models (VLMs). The performance gain from reasoning is predominantly observed in models that already demonstrate strong direct answer capabilities. For instance, Gemini-2.0-Flash-001, despite being smaller than Gemini-2.0-Pro (as evidenced by its lower direct answer performance), exhibits a substantially larger improvement margin (13.9%) when employing reasoning steps.

Our analysis reveals that high-performing models generally gain 2-10% improvement through CoT prompting. Conversely, lower-performing VLMs show degraded performance with CoT, as evidenced by the negative differences in the lower portions of Table 11. We hypothesize that inferior vision processing capabilities in these models may introduce hallucinations when given extended reasoning space, ultimately diminishing their performance.

Model Name	**CoT**	**Direct Answer**	**Difference**
gemini-2.0-flash-001	0.660	0.521	0.139
gemini-2.0-pro-exp-02-05	0.662	0.569	0.093
gpt-4o-2024-11-20	0.576	0.486	0.090
claude-3-5-sonnet-20241022	0.602	0.519	0.083
qwen2.5-vl-72b-instruct	0.502	0.433	0.069
gpt-4o-mini-2024-07-18	0.409	0.350	0.059
Llama-Breeze2-3B-Instruct	0.297	0.267	0.030
gemini-1.5-flash	0.394	0.371	0.023
gemini-1.5-pro	0.442	0.420	0.022
gemini-1.5-flash-8b	0.328	0.309	0.019
deepseek-ai-deepseek-vl2-tiny	0.278	0.259	0.019
OpenGVLab-InternVL2-8B-MPO	0.353	0.338	0.015
Qwen-Qwen2-VL-7B-Instruct	0.300	0.287	0.013
Qwen2.5-VL-7B-Instruct	0.359	0.346	0.013
Llama-3.2-90B-Vision-Instruct-Turbo	0.412	0.404	0.008
THUDM-cogvlm2-19B	0.278	0.277	0.001
claude-3-haiku-20240307	0.329	0.330	-0.001
OpenGVLab-InternVL2-8B	0.343	0.347	-0.004
OpenGVLab-InternVL2-2B	0.289	0.294	-0.005
OpenGVLab-InternVL2_5-8B	0.345	0.352	-0.007
OpenGVLab-InternVL2-4B	0.308	0.315	-0.007
Llama-Breeze2-8B-Instruct	0.292	0.299	-0.007
Llama-3.2-11B-Vision-Instruct	0.326	0.335	-0.009
deepseek-ai-deepseek-vl2-small	0.318	0.327	-0.009
OpenGVLab-InternVL2-1B	0.269	0.282	-0.013
OpenGVLab-InternVL2_5-4B	0.329	0.345	-0.016

Table 11: Model Performance Comparison: CoT vs Direct Answer (Sorted by Difference) **without character penalty**

I Prompts

I.1 VisTW-Dialogue Judgment Prompt

Apart from providing judgement prompts Figure10, we utilize a 10 point scale for evaluation, with the score decreasing from 10. If there are any omissions or mistakes in the model explanations, appropriate deductions will be made. Besides, in Scoring Guidelines Figure11 we also explain the range definition represented by each score, so as to better evaluate the model's capabilities. In the Figure14 example, according to the Scoring Method and Scoring Guidelines, the model explanations accurately aligned with the Ground Truth and addressed all answers with clear and structured language. This evaluation got a perfect score of 10.

I.2 VisTW-MCQ Prompt

In VisTW-MCQ, our prompts can be found in Figure12 and Figure13 . We divide the evaluation into two prompts: Zero-Shot COT and Direct-Answer to examine VLM model capabilities. In these examples of Figure15 and Figure16, we use red color to mark the model's responses and COT processes. For COT prompts, we specifically use "think step by step" to guide the model's reasoning further.

Judgement Prompt

請根據使用者詢問的問題 [Question] 與正確答案 [Ground Truth], 去評價助手的回覆 [Assistant Response] 的評分，評分依照下方的評價指導手冊去評分。
Please evaluate the assistant's response based on the user's question [Question] and the correct answer [Ground Truth], and assess the score of the assistant's response [Assistant Response] according to the evaluation guidelines provided below.
[問題] [Question]
{question}
[評價助手的回覆] [Assistant Response]
{response}
[正確答案] [Ground Truth]
{ground_truth}

評分標註指南 # Scoring Guidelines
{Scoring Guidelines}

評分方式 ## Scoring Method

1. 首先將回答與標準答案比較 1. Responses were compared with standard answers.
2. 評估以下幾點：2. Evaluate the following aspects:
   - 是否準確？ - Whether the responses were accurate?
   - 是否回答了問題的所有部分 - Whether all parts of the question were addressed?
   - 是否清晰且有條理？ - Whether the responses were clear and well-structured?
   - 是否提供有幫助的補充說明？ - Whether helpful explanations were provided?
3. 從10分開始扣分：3. Points were deducted for errors from an initial 10-point score.
   - 每個事實錯誤（-1至-2分） - Each factual error deducted 1–2 points.
   - 遺漏資訊（-1至-2分） - Missing information deducted 1–2 points.
   - 語言組織不佳（-1分） - Incoherent language structure deducted 1–2 points.
   - 補充說明不當或錯誤（-1分） - Inappropriate or incorrect explanations deducted 1 point.
4. 請簡短說明評分理由，包含：Scoring rationale includes concise explanations of:
   - 做得好的地方- Well-addressed components or exceptional performance.
   - 缺少或錯誤的部分- Missing or incorrect information
   - 為何給予此分數- Why this score was awarded?

你的回覆格式應該是如下: Your response should be in the format:

[解釋]: (你的解釋) [Explanations]: (Your explanations)

[評分]: (int分數) [Scoring]: (Int score)

Figure 10: Instruction prompt used in scoring the sample.

Scoring Guidelines

評分範圍（0-10分）Score Range (0 - 10 points)

10分：完美 10 points: Perfect

- 完全準確無誤 - Fully accurate and error-free
- 回答問題的所有部分 - Address all parts of the questions
- 清晰且條理分明 - Clear and well-structured
- 提供有幫助的補充說明 - Helpful explanations

8 - 9分：非常好 8-9 points: Excellent

- 有些微錯誤或遺漏 - Minor errors or omissions
- 主要重點都有涵蓋 - Key points are covered
- 組織良好 - Clear and well-structured
- 提供有用的細節說明 - Helpful explanations

6 - 7分：良好 6-7 points: Good

- 有一些小錯誤 - Some minor errors
- 大部分重點都有提到 - Most key points are addressed
- 組織尚可 - Adequate structure
- 有一些相關說明 - Some relevant explanations

4 - 5分：普通 4-5 points: Average

- 有數個錯誤 - Several errors
- 遺漏一些重點 - Some key points are omitted
- 基本的組織結構 - Basic structure
- 說明不夠完整 - Incomplete explanations

2 - 3分：不佳 2-3 points: Poor

- 有許多錯誤 - Numerous errors
- 遺漏重要資訊 - Critical information is omitted
- 組織不清楚 - Disorganized structure
- 說明不正確或不足 - Incorrect or insufficient explanations

0 - 1分：不及格 0-1 points: Failing

- 大部分錯誤 - Mostly incorrect
- 未回答問題重點 - The core question remains unanswered
- 沒有明確的組織 - Absence of structure
- 說明不相關或有誤導 - Irrelevant or misleading explanations

Figure 11: Scoring guidelines outline the criteria for each point

VisTW-MCQ - Zero-Shot COT Prompt

回答以下的多選題問題。並且在回覆的最後記得講格式： 答案: $字母 而字母是 ABCDEFG 的其中一個。回答前請先一步一步(think step by step)想好答案。你必須使用中文回答。
Answer the following multiple-choice questions. Additionally, remember to include the format at the end of your response: Answer: $letter, where the letter is one of A, B, C, D, E, F, G. Before answering, carefully deliberate through the answer step by step. You must respond in Chinese.

[問題]: [Question]
{question}

[COT]：[COT:]
{COT reasoning}

[答案]：[Answer:]
{your answer}

Figure 12: Instruction prompt used in VisTW-MCQ zero-shot COT

VisTW-MCQ - Direct-Answer Prompt

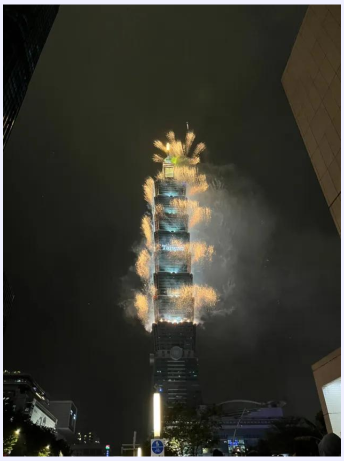
回答以下的多選題問題。並且在回覆的最後記得講格式： 答案: $字母 而字母是 ABCD 的其中一個。直接給我答案回覆，Do not think step by step。你必須使用中文回答。
Answer the following multiple-choice question. Remember to include the format at the end of your response: Answer: $letter, where the letter is one of A, B, C, or D. Provide only the answer in Chinese without any additional explanations. Do not think step by step. You must respond in Chinese.

[問題]: [Question]
{question}

[答案]：[Answer:]
{your answer}

Figure 13: Instruction prompt used in VisTW-MCQ direct answer

VisTW-Dialogue Example



請根據使用者詢問的問題 [Question] 與正確答案 [Ground Truth], 去評價助手的回覆 [Assistant Response] 的評分，評分依照下方的評價指導手冊去評分。
Please evaluate the assistant's response based on the user's question [Question] and the correct answer [Ground Truth], and assess the score of the assistant's response [Assistant Response] according to the evaluation guidelines provided below.

[問題] [Question]
這在台灣哪的建築？能猜出這是什麼時間點拍的嗎？
Where is this building located in Taiwan?
Can the time period during which this photo was taken be determined?

[評價助手的回覆] [Assistant Response]
這張照片拍攝的是台北101大樓，畫面中出現的煙火表演通常是在跨年時舉行的，因此可以推測這張照片大約是跨年夜所拍攝的。台北101在新年倒數時會舉行盛大的煙火秀，吸引了許多民眾前來觀賞。
This photograph depicts Taipei 101, with fireworks display typically held during New Year's Eve celebrations. It is inferred that this photo was taken on New Year's Eve, as Taipei 101 annually hosts a grand fireworks show during the New Year countdown, drawing large crowds of spectators.

[正確答案] [Ground Truth]
這是台灣台北著名的「台北101」，此圖是元旦跨年時101的煙火秀
This is the renowned Taipei 101 in Taipei, Taiwan. The image captures the New Year's Eve fireworks display at the building during the annual countdown celebration.

你的回覆格式應該是如下: Your response should be in the format:

[解釋]: (你的解釋) [Explanations]: (Your explanations)
[Explanations]: Completely correct

[評分]: (int分數) [Scoring]: (Int score)
[Scoring]: 10

Figure 14: VisTW-Dialogue for evaluating the response with prefect 10 points.

VisTW-MCQ - Zero-Shot COT Example

帳齡	借方餘額	損失比率
30 天以下	$489,000	0.5%
31-90 天	201,000	2%
91-180 天	106,000	5%
181-360 天	48,500	20%
361 天以上	27,500	60%
	$872,000	

回答以下的多選題問題。並且在回覆的最後記得講格式： 答案: $字母 而字母是 ABCDEFG 的其中一個。回答前請先一步一步(think step by step)想好答案。你必須使用中文回答。

Answer the following multiple-choice questions. Additionally, remember to include the format at the end of your response: Answer: $letter, where the letter is one of A, B, C, D, E, F, G. Before answering, carefully deliberate through the answer step by step. You must respond in Chinese.

[問題]:
甲公司係採用帳齡分析法估計預期信用減損損失，X8 年 12 月 31 日應收帳款帳齡分析資料如下：X8 年 5 月 12 日曾將確認無法收回之應收帳款$35,680 沖銷；X8 年 12 月 31 日調整前備抵損失為借餘$970。甲公司 X8 年度應認列之預期信用減損損失為何？

[Question]:
Company A uses the aging analysis method to estimate expected credit losses. The aging analysis data for accounts receivable as of December 31, Year X8 is as follows: On May 12, Year X8, the company wrote off accounts receivable of $35,680 confirmed to be uncollectible; the allowance for credit losses had a debit balance of $970 before adjustment on December 31, Year X8. What amount of expected credit loss should Company A recognize for the Year X8?

Ａ. $37,965 Ａ. $37,965
Ｂ. $36,935 Ｂ. $36,935
Ｃ. $36,995 Ｃ. $36,995
Ｄ. $38,935 Ｄ. $38,935

[COT]:
我將計算預期信用損失。我將每個帳齡類別的餘額乘以相應的損失率，然後將結果相加。最後，我會調整信用損失準備金中的借方餘額。 預期信用損失計算如下： 1. 30天以下：$489,000 * 0.5% = $2,445 2. 31-90天：$201,000 * 2% = $4,020 3. 91-180天：$106,000 * 5% = $5,300 4. 181-360天：$48,500 * 20% = $9,700 5. 361天以上：$27,500 * 60% = $16,500 總預期信用損失：$2,445 + $4,020 + $5,300 + $9,700 + $16,500 = $37,965
調整前信用損失準備的借方餘額：$970 因此，應確認的預期信用損失金額為：$37,965 + $970 = $38,935
[答案]: D

[COT]:
I will calculate the expected credit loss by multiplying the balance of each aging category by the corresponding loss rate and summing the results. Finally, adjustments are made for the debit balance in the credit loss reserve. The expected credit loss calculation is as follows: 1. Under 30 days: $489,000 × 0.5% = $2,445 2. 31-90 days: $201,000 × 2% = $4,020 3. 91-180 days: $106,000 × 5% = $5,300 4. 181-360 days: $48,500 × 20% = $9,700 5. Over 361 days: $27,500 × 60Total expected credit loss: $2,445 + $4,020 + $5,300 + $9,700 + $16,500 = $37,965
Debit balance of credit loss reserve before adjustment: $970 Accordingly, the recognized expected credit loss amount is: $37,965 + $970 = $38,935
[Answer]: D

Figure 15: VisTW-MCQ - Zero-Shot COT with reasoning and correct response

VisTW-MCQ - Direct Answer Example

回答以下的多選題問題。並且在回覆的最後記得講格式： 答案: $字母 而字母是 ABCD 的其中一個。直接給我答案回覆，Do not think step by step。你必須使用中文回答。
Answer the following multiple-choice question. Remember to include the format at the end of your response: Answer: $letter, where the letter is one of A, B, C, or D. Provide only the answer in Chinese without any additional explanations. Do not think step by step. You must respond in Chinese.

[問題]:
X1 年間總公司總計將成本$150,000 之商品運交分公司，轉撥價格均按成本加價 10%，且分公司不自外界進貨。總公司運交分公司之一批商品，至期末分公司仍未收到，而導致「總公司往來」與「分公司往來」金額不一致。總公司與分公司於 X1 年底調整前試算表部分資料如下：試問分公司帳上期初存貨餘額為何？
[Question]:
During the year X1, the Head Office transferred goods costing $150,000 to the Branch at a transfer price of cost plus 10%, with the Branch not sourcing inventory externally. A shipment of goods dispatched by the Head Office to the Branch remained undelivered at period-end, creating discrepancies between the balances of the Head Office Current Account and Branch Current Account. Selected pre-adjustment trial balance data from both entities as of year-end X1 are provided.
What was the Branch's beginning inventory balance?

Ａ. $33,000 Ａ. $33,000
Ｂ. $30,000 Ｂ. $30,000
Ｃ. $0 Ｃ. $0
Ｄ. $20,000 Ｄ. $20,000

[答案]：B
[Answer]: B

Figure 16: VisTW-MCQ - direct answer example with incorrect responses (the answer is A)

J Question Diversity in VLM

Both VisTW-MCQ and VisTW-Dialogue contain diverse visual contexts and challenge VLMs' skills such as content comprehension, image depth judgment, and cultural-specific analysis. Correctly answering each question requires multiple visual competencies, such as table comprehension, logical reasoning, visual content interpretation, and cultural background knowledge.

J.1 Skills set found in VisTW-Dialogue

Context Comprehension Accurate question-answering in VLM requires both visual content comprehension and subsequent reasoning skills. Visual content includes diverse types of content such as table-related information Figure 17, daily life materials Figure 18, 19, and implicit semantic content Figure 20. The comprehension of these contents involves OCR technology, semantic analysis and latent reasoning. Therefore, answering each question precisely demands VLM integrate multiple skills for comprehensive understanding and response.

2D Image Depth Judgment Certain 2D imagery questions in VLMs demand both 3D object recognition and spatial reasoning abilities. These questions evaluate whether VLMs can leverage geometric cues such as perspective distortion, occlusion patterns, and relative scale to reconstruct 3D depth relationships like Figure 21.

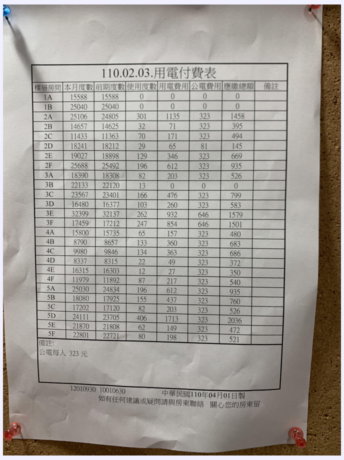
Cultural-Specific Knowledge

Cultural contextual understanding constitutes a crucial capability for VLMs. Many daily life problems require cultural knowledge for resolution. Hence, within VisTW-Dialogue like Figure 22, 23, 24, there are specific evaluations of background knowledge, especially culture and landscape.

J.2 VisTW-MCQ skills introduction

Strong Subjects knowledge The VisTW-MCQ benchmark evaluates the multimodal capabilities of VLMs by integrating visual comprehension and semantic reasoning. Rather than merely assessing basic visual processing skills such as OCR and visual content analysis, VLMs are required to demonstrate deductive reasoning to choose correct answers for multiple-choice questions. These questions appraise the reasoning capabilities of VLMs across 21 diverse subjects, , including a sample of disciplines such as Figure 25 Arts, Figure 26 Chemistry, Figure 27 Chinese Literature, Figure 28 Music, Figure 29 Sociology, and Figure 30 Geography.

OCR - Understanding Table-Related Question



110.02.03.用電付費表

樓層房間	本月度數	前期度數	使用度數	用電費用	公電費用	應繳總額	備註
1A	15588	15588	0	0	0	0	
1B	25040	25040	0	0	0	0	
2A	25106	24805	301	1135	323	1458	
2B	14657	14625	32	71	323	395	
2C	11433	11363	70	171	323	494	
2D	18241	18212	29	65	81	145	
2E	19027	18898	129	346	323	669	
2F	25688	25492	196	612	323	935	
3A	18390	18308	82	203	323	526	
3B	22133	22120	13	0	0	0	
3C	23567	23401	166	476	323	799	
3D	16480	16377	103	260	323	583	
3E	32399	32137	262	932	646	1579	
3F	17459	17212	247	854	646	1501	
4A	15800	15735	65	157	323	480	
4B	8790	8657	133	360	323	683	
4C	9980	9846	134	363	323	686	
4D	8337	8315	22	49	323	372	
4E	16315	16303	12	27	323	350	
4F	11979	11892	87	217	323	540	
5A	25030	24834	196	612	323	935	
5B	18080	17925	155	437	323	760	
5C	17202	17120	82	203	323	526	
5D	24111	23705	406	1713	323	2036	
5E	21870	21808	62	149	323	472	
5F	22801	22721	80	198	323	521	

備註:
公電每人 323 元

12010930 10010630 中華民國110年04月01日製
如有任何建議或疑問請與房東聯絡 關心您的房東留

[問題] [Question]

圖片中的電費單中，哪一樓層用電最多？
In the electricity bill presented in the image, which floor exhibits the highest electricity consumption?

[參考答案]:
我們從每一樓層的電費總額來計算：
1樓： 1A, 1B 用電總額為 0。
2樓： 2A 是 1458， 2B 是 395，2C 是 494， 2D 是 145，2E 是 669，2F 是 935；因此總電額度為 4,096。
3樓：3A 是 526， 3B 是 0，3C 是 799， 3D 是 583，3E 是 1579，3F 是 1501 ；因此總電額度為 4,988 元。
4樓：4A 是480， 4B 是 683，4C 是 686， 4D 是 372，4E 是 350，4F 是 540；因此總電額度為 3,111 元。
5樓：5A 是935， 5B 是 760，5C 是 526， 5D 是 2036，5E 是 472，5F 是 521；因此總電額度為 5,250 元。
綜上所述，用電額度最高的是5樓。
[Reference]:
let's add up the electricity costs for each floor：
For the 1st floor, units 1A and 1B used zero, so that's easy.
On the 2nd floor, we've got 1458 for 2A, 395 for 2B, 494 for 2C, 145 for 2D, 669 for 2E, and 935 for 2F. So, that floor comes out to 4,096.
The 3rd floor: 526 for 3A, nothing for 3B, 799 for 3C, 583 for 3D, 1579 for 3E, and 1501 for 3F. That totals 4,988.
Moving to the 4th floor: 4A is 480, 4B is 683, 4C is 686, 4D is 372, 4E is 350, and 4F is 540. So the total there is 3,111.
Finally, the 5th floor: 935 for 5A, 760 for 5B, 526 for 5C, a big one at 2036 for 5D, 472 for 5E, and 521 for 5F. That brings their total to 5,250.
So, looking at all that, the 5th floor had the highest electricity bill.

Figure 17: OCR - Floor Power: Table-Related Insights

Daily Life Content Comprehension Question - 1

[問題] [Question]

如果我來台灣5天最划算的配套是什麼？

What is the most cost-effective package for a 5-day trip to Taiwan?

[參考答案]:

選 4G Unlimited data 的5天配套，只需要300元即可。

[Reference]:

You can get 5 days of unlimited 4G data for just 300.

Figure 18: Daily Life Comprehension: Results via Reasoning

Daily Life Content Comprehension Question - 2

[問題] [Question]

假設該期統一發票的中獎號碼是94172788，這張發票中了多少錢？

Assuming the winning number of the uniform invoice for this period is 94172788, what is the prize amount for this invoice?

[參考答案]:

統一發票於每單月之二十五日，就前期之統一發票，開出特獎一至三組及其他各獎三至十組之中獎號碼，並視財政狀況增開特別獎一組，其獎別及獎金如下：

一、特別獎：統一發票八位數號碼與中獎號碼完全相同者，獎金新臺幣一千萬元。

二、特獎：統一發票八位數號碼與中獎號碼完全相同者，獎金新臺幣二百萬元。

三、其他各獎：

（一）頭獎：統一發票八位數號碼與中獎號碼完全相同者，獎金新臺幣二十萬元。

（二）二獎：統一發票末七位數號碼與中獎號碼之末七位完全相同者，獎金新臺幣四萬元。

（三）三獎：統一發票末六位數號碼與中獎號碼之末六位完全相同者，獎金新臺幣一萬元。

（四）四獎：統一發票末五位數號碼與中獎號碼之末五位完全相同者，獎金新臺幣四千元。

（五）五獎：統一發票末四位數號碼與中獎號碼之末四位完全相同者，獎金新臺幣一千元。

（六）六獎：統一發票末三位數號碼與中獎號碼之末三位完全相同者，獎金新臺幣二百元。

您的發票號碼最後一位數字和中獎號碼的最後一位數字不同，故沒有中獎。

[Reference]: Main Prizes:

1. There's a huge Grand Special Prize of NT$10 million.

2. Special Prize of NT$2 million. For these, your entire 8-digit receipt number must exactly match the winning number.

3. Then there's a First Prize of NT$200,000, also for an exact 8-digit match.

4. Smaller Prizes :

For these, you need to match the last few digits of your receipt number with the last few digits of one of the First Prize winning numbers.

The last digit of your invoice number is different from the last digit(s) of the applicable winning numbers (e.g., for the Sixth Prize), so you have not won a prize.

Figure 19: Daily Life Comprehension: Local Cultural-Specific Daily Question

Implicit Contextual Understanding Question

[問題] [Question]

為什麼我同事週一傳了這張圖給我？

Why did my colleague send me this image on Monday?

[參考答案]:

既然提到時間是週一，而這隻狗看起來有點困惑或不安的樣子,站在雜亂的環境中。也許你的同事想表達週一開始工作的心情,感覺像這隻狗一樣被雜務和工作"包圍"

[Reference]:

Seeing as it's Monday and the dog looks kind of dazed or stressed out in that messy spot, your colleague probably means that's how they feel starting the work week – completely swamped with work and chores, just like the dog is hemmed in by all that stuff.

Figure 20: Implicit Contextual Understanding: Beyond Direct Visual Reasoning

2D Image Depth Judgment Question

[問題] [Question]

圖中紅色標點數字，幫我按從拍攝視角的距離近到遠排序

Please sort the red-labeled numbers in the figure according to their distance from the camera perspective, from near to far.

[參考答案]:

圖片中距離攝影師最近到最遠的分別是：5, 3, 1, 2, 4

[Reference]:

From closest to farthest from the photographer, the order of the items is: 5, then 3, then 1, then 2, and finally 4.

Figure 21: 2D Image Depth Judgment: Red-Labeled Number Distance Sorting

Cultural-Specific Question - 1

[問題] [Question]

圖片位於台灣哪個風景區？

In which scenic area of Taiwan is the image located?

[參考答案]:

這位於台灣高雄的「蓮池潭風景區」，蓮池潭在清代又譽為「鳳山八景」之一。

而圖片中出現的是著名的「龍虎塔」，是高雄市最具傳統宗教色彩的風景區之一。

[Reference]:

This is the Lotus Pond Scenic Area in Kaohsiung, Taiwan. Tiger Pagodas are part of one of Kaohsiung City's most traditional and religiously significant scenic areas.

Figure 22: Cultural-Specific: Taiwan Scenic Area Identification

Cultural-Specific Question - 2

[問題] [Question]

這具有宗教色彩的景點位於台灣何處？ 圖中出現了什麼特色元素？

Where is this religiously significant attraction located in Taiwan? What distinctive elements appear in the image?

[參考答案]:

這是位於台灣基隆的「中正公園」。在公園的最高點處，有一座22.5公尺高的白色觀音大佛像，像內分五層，可登上鳥瞰基隆全景，在佛像旁有兩頭巨大的金獅坐鎮，十分祥和莊嚴。「中正公園」是基隆市最具代表性的地標。

[Reference]:

This is Zhongzheng Park in Keelung, Taiwan. At the park's highest point, there stands a 22.5-meter-tall (approximately 73.8 feet) white statue of the Bodhisattva Guanyin. The statue's interior is divided into five levels, which visitors can ascend to get a bird's-eye view of the entire Keelung cityscape. Flanking the Guanyin statue are two enormous golden lions, creating a serene and majestic atmosphere. Zhongzheng Park is considered Keelung City's most iconic landmark.

Figure 23: Cultural-Specific: Religiously Significant Local Sites

Cultural-Specific Question - 3

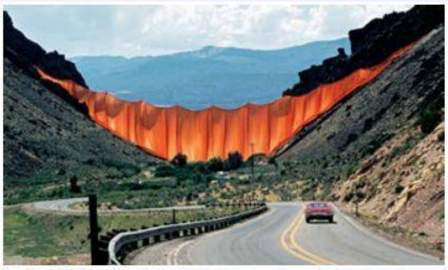
[問題] [Question]

這是什麼經典建築？圖中有多少個砲臺？
What classic building is this? How many gun emplacements are there in the figure?

[參考答案]:
這是位於台灣淡水「紅毛城」。早期在西班牙人侵略淡水時期在此建「聖多明哥城」為基地，後來荷蘭人驅逐了西人，在舊址重建該城，當時稱荷蘭人為「紅毛」，因此「紅毛城」沿用至今。 共有「5座」砲臺。
[Reference]:
We're looking at Hongmao Cheng, also known internationally as Fort San Domingo, in Tamsui, Taiwan. Its story begins with the Spanish, who established 'Fort San Domingo' on this spot as their stronghold during their time in Tamsui. Subsequently, the Dutch drove out the Spanish and reconstructed the fort on the same foundations. Because the local population called the Dutch 'hóngmáo', meaning 'red hairs,' the fort became popularly known as Hongmao Cheng, a name that has persisted to this day. It is also equipped with five artillery batteries.

Figure 24: Cultural-Specific: Local Cultural-Historical Heritage Sites

VisTW-MCQ: Arts Question



[問題] [Question]

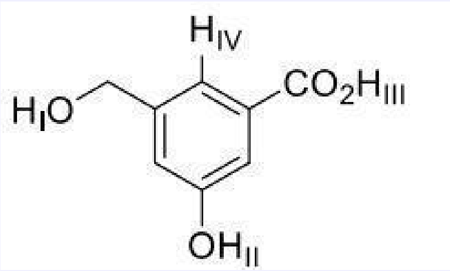
對於美國科羅拉多州於1970年設立的《山谷簾幕》，這是一個引發關注和環境議題討論的巨大人造景觀，請問這是誰的作品？
Regarding the "Valley Curtain" installed in Colorado, USA, in 1970—a large-scale artificial landscape that sparked significant attention and discussions on environmental issues—who is the artist behind this work?

A. 高沃爾思 Andy Goldsworthy
B. 席格爾 George Segal
C. 克里斯托和珍妮・克勞德 Christo and Jeanne-Claude
D. 史密斯遜 Robert Smithson

[答案]: C
Answer: C

Figure 25: VisTW-MCQ: Arts Question Example

VisTW-MCQ: Chemistry Question



[問題] [Question]

下列化合物中標示的質子酸性由高至低排列的順序何者正確？

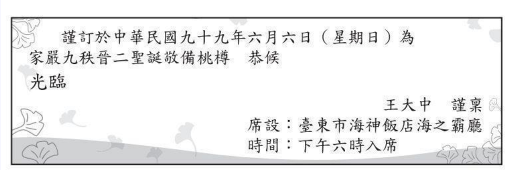
Regarding the order of proton acidity from highest to lowest for the indicated protons in the following compounds, which is correct?

Ａ. III > IV > II > I　III > IV > II > I
Ｂ. III > II > I > IV　III > II > I > IV
Ｃ. IV > I > II > III　IV > I > II > III
Ｄ. II > I > IV > III　II > I > IV > III

[答案]: B
Answer: B

Figure 26: VisTW-MCQ: Chemistry Question Example

VisTW-MCQ: Chinese Literature Question


謹訂於中華民國九十九年六月六日（星期日）為
家嚴九秩晉二聖誕敬備桃樽　恭候
光臨
王大中　謹稟
席設：臺東市海神飯店海之霸廳
時間：下午六時入席

[問題] [Question]

關於這則請柬，下列敘述何者正確？

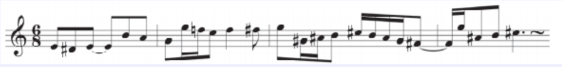
Regarding this invitation, which of the following statements is correct?

Ａ.「謹稟」應改為「叩首」
「謹稟」should be replaced with 「叩首」
Ｂ.「聖誕」應改為「壽誕」
「聖誕」should be replaced with 「壽誕」
Ｃ.「家嚴」應改為「令尊」
「家嚴」should be replaced with「令尊」
Ｄ. 「桃樽」應改為「桃符」
「桃樽」」should be replaced with「桃符」

[答案]: B
Answer: B

Figure 27: VisTW-MCQ: Chinese Literature Question Example

VisTW-MCQ: Music Question



[問題] [Question]

這個曲調的調性為？

What is the tonality of this melody?

Ａ. E 大調轉至 b 小調 The piece modulates from E major to b minor

Ｂ. e 小調轉至 B 大調 The piece modulates from e minor to B major

Ｃ. B 大調轉至升 g 小調 The piece modulates from B major to g sharp minor

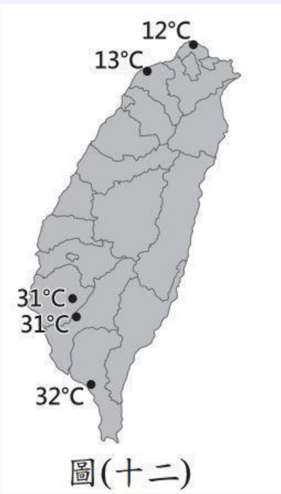
Ｄ. 升 c 小調轉至 B 大調 The piece modulates from c sharp minor to B major

[答案]: B

Answer: B

Figure 28: VisTW-MCQ: Music Question Example

VisTW-MCQ: Sociology Question



[問題] [Question]

圖(十二)為小芬在網路上看到的某日下午一點臺灣部分測站的氣温資料，當時臺灣氣温空間分布的差異與下列何者關係最為密切？

Pic(12) shows temperature data from selected weather stations in Taiwan at 1:00 PM on a certain afternoon. The spatial variation in temperature across Taiwan at that time is most closely related to which of the following factors?

Ａ. 滯留鋒面持續向北移 The stationary front continues to move northward

Ｂ. 蒸發旺盛形成熱對流 Intense evaporation generates thermal convection

Ｃ. 熱帶低氣壓逐漸增强 The tropical depression gradually intensifies

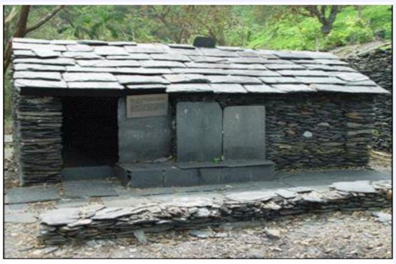
Ｄ. 大陸冷氣團逐漸南下 The continental cold air mass gradually advances southward

[答案]: D

Answer: D

Figure 29: VisTW-MCQ: Sociology Question Example

VisTW-MCQ: Geography Question



照片 2

[問題] [Question]

臺灣傳統民宅的建材，常因不同的族群配合環境特性就地取材，而有竹管屋、土堈厝、石板屋、茅草屋等形式。照片 2 是某族群的傳統民宅，該種形式的民宅最可能出現在下列何處?

Building materials for traditional Taiwanese dwellings often vary by ethnic group, utilizing local resources adapted to environmental conditions, resulting in architectural forms such as bamboo tube houses, adobe homes, stone slab houses, and thatched cottages. Photo 2 depicts a traditional residence of a specific ethnic group. Where is this type of dwelling most likely to be found?

Ａ. 埔里盆地 Puli Basin
Ｂ. 澎湖群島 Penghu Archipelago
Ｃ. 花東縱谷北部 Northern Hualien-Taitung Valley
Ｄ. 中央山脈南段 Southern Section of the Central Mountain Range

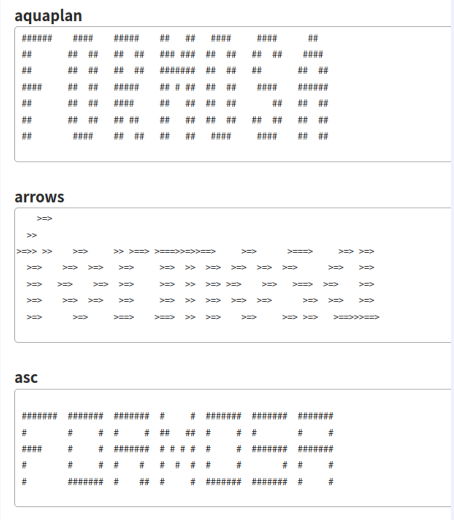
[答案]: D
Answer: D

Figure 30: VisTW-MCQ: Geography Question Example

K Failure case study found in VisTW-Dialogue

In this section, we present several failure cases from low-scoring examples identified within the VisTW-Dialogue dataset. These examples highlight specific scenarios where state-of-the-art multimodal models still encounter difficulties, including interpreting ASCII art (Figure 31), accurately reading numeric data from complex electricity meter displays (Figure 32), and reasoning errors arising from Traditional Chinese OCR on electricity bills (Figure 33). Each case underscores existing challenges in multimodal understanding and detailed visual reasoning, suggesting areas for further model development and improvement.

Failure case #1 - ASCII art reading



[Question]

圖中三個 ascii 藝術字是寫什麼字？

What are the three ASCII art characters in the image?

Claude-3.5-sonnet-20241022 這張圖片顯示了三種不同的 ASCII 藝術字體樣式:
1. aquaplan - 使用 "#" 符號排列成的字型
2. arrows - 使用箭頭符號 ">" 和 "=" 組成的字型
3. asc - 同樣使用 "#" 符號排列的另一種字型
每個樣式都展示了相同的排列方式但使用不同的符號和格式來呈現。

Gemini 2.0 flash ensemble judge score: 2.5 **Gemini 2.0 Pro** 圖片中三個 ASCII 藝術字分別是：
* **aquaplan:** 這個字很明顯，就是 "aquaplan"。
* **arrows:** 這個字也是直接顯示，為 "arrows"。
* **asc:** 這個字稍微需要辨識一下，但可以看出是 "asc"。
所以答案是：
1. aquaplan
2. arrows
3. asc

Gemini 2.0 flash ensemble judge score: 0.0

Figure 31: The best-performing models—Claude-3.5-Sonnet (which achieved the highest score on this question) and Gemini 2.0 Pro (the top model in VisTW-Dialogue)—still failed to correctly recognize that the ASCII art depicted the word "FORMOSA."

Failure case #2 - Reading electricity meter

[Question]

請問電表上的度數是幾號 Could you please tell me the reading on the electricity meter?

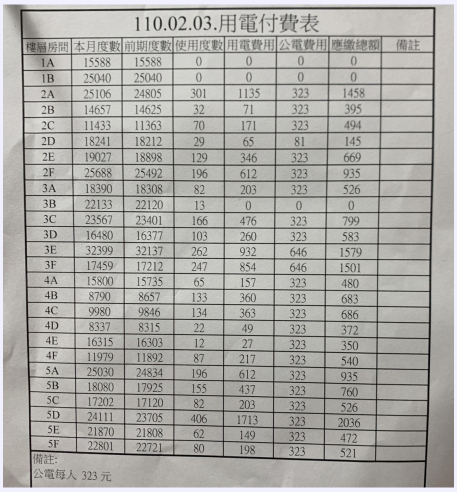
Gemini 2.0 Flash 電表上的度數是 06765.

Gemini 2.0 flash ensemble judge score: 2.0 **Gemini 2.0 Pro**根據圖片，電表上的度數讀數為：
* 第一個錶盤（最左邊）：指針在 6 和 7 之間，讀數為 6。
* 第二個錶盤：指針接近 3，讀數為 2。
* 第三個錶盤：指針接近 2，讀數為 1。
* 第四個錶盤：指針在 7 和 8 之間，讀數為 7。
* 第五個錶盤（最右邊）：指針接近 6，讀數為 5。
所以，總度數是 62175 度。

Gemini 2.0 flash ensemble judge score: 0.4

Figure 32: The best-performing model (Gemini-2.0-flash) successfully read the second meter from the left, correctly interpreting the value '6', despite numbers 0-9 being arranged clockwise and the pointer moving counter-clockwise. However, it incorrectly read the middle meter, mistaking a '9' for a '7'. Additionally, it misread the third meter from the left, interpreting an '8' as a '6', and the leftmost meter, identifying a '7' as a '5'.

Failure case #3 - Traditional Chinese OCR with Reasoning



110.02.03.用電付費表

樓層房間	本月度數	前期度數	使用度數	用電費用	公電費用	應繳總額	備註
1A	15588	15588	0	0	0	0	
1B	25040	25040	0	0	0	0	
2A	25106	24805	301	1135	323	1458	
2B	14657	14625	32	71	323	395	
2C	11433	11363	70	171	323	494	
2D	18241	18212	29	65	81	145	
2E	19027	18898	129	346	323	669	
2F	25688	25492	196	612	323	935	
3A	18390	18308	82	203	323	526	
3B	22133	22120	13	0	0	0	
3C	23567	23401	166	476	323	799	
3D	16480	16377	103	260	323	583	
3E	32399	32137	262	932	646	1579	
3F	17459	17212	247	854	646	1501	
4A	15800	15735	65	157	323	480	
4B	8790	8657	133	360	323	683	
4C	9980	9846	134	363	323	686	
4D	8337	8315	22	49	323	372	
4E	16315	16303	12	27	323	350	
4F	11979	11892	87	217	323	540	
5A	25030	24834	196	612	323	935	
5B	18080	17925	155	437	323	760	
5C	17202	17120	82	203	323	526	
5D	24111	23705	406	1713	323	2036	
5E	21870	21808	62	149	323	472	
5F	22801	22721	80	198	323	521	

備註:
公電每人 323 元

[Question]

圖片中的電費單中，哪一樓層用電最多？

In the electricity bill shown in the image, which floor consumed the most electricity?

Gemini 2.0 Pro 從圖片中的電費單來看，「使用度數」欄位中數值最高的為5D，用電度數為406度，因此5D用電最多。

Gemini 2.0 flash ensemble judge score: 2.0

Gemini 2.0 Flash Thinking 圖片中的電費單中，我們來計算度數最高的

步驟:

1. **找到「使用度數」欄位:** 在電費單表格中，找到標題為「使用度數」的欄位。這個欄位顯示了每個房間的用電量。

2. **計算每層樓的總用電量:** 將同一樓層的所有房間的「使用度數」加總起來。

* **1樓:** 1A (0) + 1B (0) = 0
* **2樓:** 2A (301) + 2B (32) + 2C (70) + 2D (29) + 2E (129) + 2F (196) = 757
* **3樓:** 3A (82) + 3B (13) + 3C (166) + 3D (103) + 3E (262) + 3F (247) = 873
* **4樓:** 4A (65) + 4B (133) + 4C (134) + 4D (22) + 4E (12) + 4F (87) = 453
* **5樓:** 5A (196) + 5B (155) + 5C (82) + 5D (406) + 5E (62) + 5F (80) = 981

3. **比較總用電量:** 比較各樓層的總用電量，找出數值最高的樓層。

* 1樓: 0
* 2樓: 757
* 3樓: 873
* 4樓: 453
* **5樓: 981**

4. **結論:** 5樓的總用電量最高 (981)，因此 **5樓用電最多**。

Gemini 2.0 flash ensemble judge score: 9.5

Figure 33: A comparison between Gemini-2.0-Pro and Gemini-2.0-Flash-Thinking. While Gemini-2.0-Pro correctly identified the room with the highest individual usage (room 5D, 406 units), it initially misunderstood the question by focusing on a single room rather than total usage per floor. However with Gemini 2.0 Flash Thinking a multi modal reasoning model, it accurately calculated each floor's total electricity consumption and correctly concluded that the 5th floor had the highest usage at 981 units.