Culture In a Frame: C^{3}B as a Comic-Based Benchmark for Multimodal Culturally Awareness

Anonymous authors
Paper under double-blind review

Abstract

Cultural awareness capabilities have emerged as a critical capability for Multimodal Large Language Models (MLLMs). However, current benchmarks lack progressed difficulty in their task design and are deficient in cross-lingual tasks. Moreover, current benchmarks often use real-world images. Each real-world image typically contains one culture, making these benchmarks relatively easy for MLLMs. Based on this, we propose C^{3}B (**C**omics **C**ross-**C**ultural **B**enchmark), a novel multicultural, multitask and multilingual cultural awareness capabilities benchmark. C^{3}B comprises over 2000 images and over 18000 QA pairs, constructed on three tasks with progressed difficulties, from basic visual recognition to higher-level cultural conflict understanding, and finally to cultural content generation. We conducted evaluations on 11 open-source MLLMs, revealing a significant performance gap between MLLMs and human performance. The gap demonstrates that C^{3}B poses substantial challenges for current MLLMs, encouraging future research to advance the cultural awareness capabilities of MLLMs.

1 Introduction

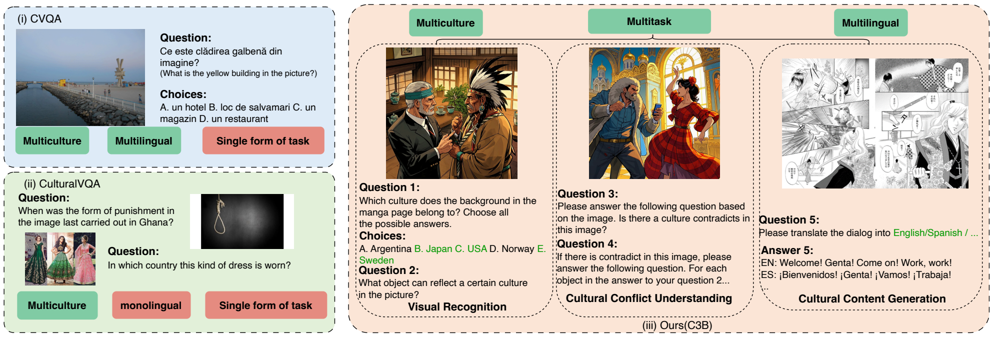

(i) CVQA
Question:
Ce este clădirea galbenă din imagine?
(What is the yellow building in the picture?)
Choices:
A. un hotel B. loc de salvamari C. un magazin D. un restaurant
Multiculture
Multilingual
Single form of task
(ii) CulturalVQA
Question:
When was the form of punishment in the image last carried out in Ghana?
Question:
In which country this kind of dress is worn?
Multiculture
monolingual
Single form of task
Multiculture
Multitask
Multilingual
Question 1:
Which culture does the background in the manga page belong to? Choose all the possible answers.
Choices:
A. Argentina B. Japan C. USA D. Norway E. Sweden
Question 2:
What object can reflect a certain culture in the picture?
Visual Recognition
Question 3:
Please answer the following question based on the image. Is there a culture contradicts in this image?
Question 4:
If there is contradict in this image, please answer the following question. For each object in the answer to your question 2...
Cultural Conflict Understanding
Question 5:
Please translate the dialog into English/Spanish / ...
Answer 5:
EN: Welcome! Genta! Come on! Work, work!
ES: ¡Bienvenidos! ¡Genta! ¡Vamos! ¡Trabaja!
...
Cultural Content Generation
(iii) Ours(C3B)

Figure 1: **Comparison between C^{3}B and previous culture awareness capability benchmarks**. In comparison with existing benchmarks for cultural awareness capabilities, C^{3}B is compatible with multicultural, multilingual, and multitask contexts, thereby facilitating a more thorough evaluation.

Multimodal Large Language Models (MLLMs) have become more and more important in many aspects of daily living, such as machine translation (Chen et al., 2025), image captioning (Anantha Ramakrishnan et al., 2025) and visual question answering (Huynh et al., 2025). Users who interact with these models often discover such a situation: current models perform well in a Western-centric culture context but perform badly in non-Western culture contexts (Singh et al., 2025; Nayak et al., 2024; AlKhamissi et al., 2024; Burda-Lassen et al., 2025; Naous et al., 2024). This imbalance shows that MLLMs need to improve cultural awareness capabilities which refers to capabilities of MLLMs to understand and process cultural contexts (Pawar et al., 2024).

Benchmarks are essential to build MLLMs with strong cultural awareness capabilities (Cohen and Howe, 1988; Reuel et al., 2024). While existing benchmarks have laid a foundational framework for evaluation, there remains room for improvement. From a multicultral perspective, existing benchmarks for evaluating cultural awareness capabilities of MLLMs mainly focus on real-world images (Arora et al., 2025; Romero et al., 2024; Xu et al., 2025). A single real-world image typically contains one culture, making these benchmarks relatively easy for MLLMs. From a multitask perspective, most of these benchmarks include one question per data sample, which makes it difficult to evaluate MLLMs across multiple dimensions. From a multilingual perspective, languages are carriers of cultural meaning (Kramsch, 2014). A concept may have no direct equivalent expression across languages. Therefore, adding multi-lingual tasks to benchmarks will introduce appropriate complexity, enabling us to evaluate MLLMs more comprehensively.

To address these issues, we propose C^{3}B (**C**omics **C**ross-**C**ultural **B**enchmark), a novel multicultural, multitask and multilingual cultural awareness capabilities benchmark, containing 2220 images and 18789 QA pairs. In contrast to real-world images, we select comics as the primary medium in our benchmark. Comics differ from real-world images: they often depict a fictional scene. Real-world images often tied to the specific, singular cultural contexts of real-life scenarios, but fictional scenes in comics are free from such constraints. This enables comics to condense numerous cultures into a single frame, creating a more complex context, raising the bar for evaluation. C^{3}B consists of 3 tasks that form a logical chain. Based on difficulty levels, these tasks progress from basic visual recognition to higher-level cultural conflict understanding, and finally to cultural content generation. This multitask arrangement enables a more comprehensive evaluation of MLLMs. In cultural generation task, C^{3}B incorporates 5 languages (Japanese, Russian, Thai, English, and Spanish), reducing the limitations of previous monolingual benchmarks. The differences between C^{3}B and previous cultural awareness capability benchmarks are presented in Figure 1.

In this study, we conduct a comprehensive evaluation of 11 open-source MLLMs on C^{3}B. The results confirm the value of C^{3}B, as they reveal a significant performance gap between MLLMs and human performance. Furthermore, the results provide critical insights into the current state of cultural awareness capabilities of MLLMs. Specifically, MLLMs should enhance their understanding of lesser-known cultures and their abilities to process cultural conflicts. Our contributions can be summarized as follows:

1. We propose C^{3}B, a novel comic-centric, multicultural, multitask and multilingual cultural awareness capabilities benchmark.
2. C^{3}B incorporates 3 tasks with escalating difficulty, evaluating cultural awareness capabilities of MLLMs comprehensively through progressively challenging tasks.
3. We benchmark 11 MLLMs with C^{3}B, which presents an initial set of evaluations on this benchmark, establishing a baseline for future research on MLLMs with strong cultural awareness capabilities.

2 Related Work

2.1 Multimodal Large Language Models

Recent years have witnessed rapid advancements in Multimodal Large Language Models (MLLMs). The multimodal capabilities of MLLMs enable large language models to address a broader range of tasks (Caffagni et al., 2024). Models like BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023b; 2024; 2023a) and Qwen-VL (Bai et al., 2023) perform well in tasks such as visual question answering (Dong et al., 2024) and multimodal machine translation (Chen et al., 2025). Moreover, InternLM-XC2.5 (Zhang et al., 2024) has good long-contextual input and output capabilities, which enables many advanced features, such as high resolution image understanding. Llama3.2 series models (Grattafiori et al., 2024) are optimized for many multimodal tasks, becoming baseline for many methods.

2.2 Multimodal Benchmark on Cultural Awareness Capabilities

Recent multimodal benchmarks on cultural awareness capabilities mainly focus on real-world images to evaluate MLLMs. Among various recent benchmarks (Romero et al., 2024; Burda-Lassen et al., 2025; Arora et al., 2025; Schneider et al., 2025; Yang et al., 2025), CVQA (Romero et al.,

2024) adopts a multiple-choice format to evaluate the cultural awareness capabilities of MLLMs using real-world images. Notably, CVQA covers 30 languages, enabling evaluation from a multilingual perspective. CVQA relies on a single form of task, restricting its ability to comprehensively evaluate the cultural awareness capabilities of MLLMs across diverse interaction scenarios. CulturalVQA (Nayak et al., 2024) also evaluates MLLMs with real-world images, covering 11 countries in 2378 images. More recently, GIMMICK (Schneider et al., 2025) incorporates six tasks for evaluating the cultural awareness capabilities of MLLMs. The images included still exhibit relatively low cultural density. Recently, GLOBALRG (Bhatia et al., 2024) has utilized images sourced from 50 countries and integrated visual grounding tasks to benchmark the cultural awareness capabilities of MLLMs. However, this framework does not account for multilingual settings. Moreover, ALM-bench (Vayani et al., 2025) encompasses over 100 languages and 19 different cultural domains, yet the images contained ALM-bench are still real-world images. AlKhamissi et al. (2025) explores the development of more sophisticated cultural benchmarks, providing a complementary lens for multimodal cultural reasoning. In parallel, Karamolegkou et al. (2024) introduces a culture-centric evaluation subset wherein a subset of images is annotated with multiple cultural associations. Overall, C^{3}B integrates the advantages of these benchmarks, forming a multitask, multicultural, and multilingual benchmark.

2.3 Multimodal Benchmarks on Comics

Multimodal benchmarks centered on comics primarily focus on basic visual tasks. For Western-style comics, the eBDtheque dataset (Guérin et al., 2013) was the first publicly released comic dataset, featuring spatial and semantic annotations for 100 pages of Western comics. The COMICS dataset (Iyyer et al., 2017) includes over 1.2 million comic panels, offering resources for future research. For Japanese-style comics, Manga109 (Matsui et al., 2016) comprises 21,142 comics pages, with a primary focus on multimedia applications. More recently, CoMix (Vivoli et al., 2024) integrates comics into a new dataset but remains focused on visual tasks such as speaker identification and character naming.

Given the current landscape of multimodal benchmarks for cultural awareness capabilities and comics, we propose C^{3}B. The differences between C^{3}B and previous works are presented in Table 1. From the table, we observe that benchmarks for evaluating cultural awareness capabilities have not simultaneously integrated multicultural images, multitask settings, and multilingual tasks. Additionally, none of them incorporate progressive difficulty tiers. Regarding comic-centered datasets, most are not designed for cultural evaluation purposes.

Table 1: **The difference between C^{3}B and previous works.** We analyze the works along three dimensions: whether the dataset is **Multicultural** for every image, **Multitask** for every data sample, **Multilingual**, whether its primary task is **Cultural Awareness Capabilities** and whether the tasks within have **Progressed Difficulty**.

Benchmarks	Multicultural	Multitask	Multilingual	Cultural Awareness Capabilities	Progressed Difficulty
CVQA (Romero et al., 2024)	✗	✗	✓	✓	✗
MOSAIC-1.5k (Burda-Lassen et al., 2025)	✗	✗	✗	✓	✗
CoMix (Vivoli et al., 2024)	✓	✗	✗	✗	✗
eBDtheque (Guérin et al., 2013)	✓	✗	✗	✗	✗
Manga109 (Matsui et al., 2016)	✓	✗	✗	✗	✗
CulturalVQA (Arora et al., 2025)	✗	✗	✗	✓	✗
GIMMICK (Schneider et al., 2025)	✓	✓	✗	✓	✗
C^{3}B	✓	✓	✓	✓	✓

3 C^{3}B: Comics Cross-Cultural Understanding Benchmark

C^{3}B is a novel comic-centric benchmark designed to comprehensively evaluate MLLMs' cultural awareness capabilities. C^{3}B features a multicultural diversity, a multitask setting and a multilingual coverage. To comprehensively evaluate MLLMs, C^{3}B incorporates three tasks with increasing difficulty (detailed in Section 3.1). For data construction (detailed in Section 3.2), we design a multi-agent method to create culturally rich comics and their annotations, ensuring both efficiency and quality of images. In Section 3.3, some necessary statistics of C^{3}B is presented. An overview of C^{3}B is presented in Figure 2. Some data samples of C^{3}B is presented in Appendix B.

3.1 Tasks

We design 3 tasks with escalating difficulty to evaluate different dimensions of cultural awareness capabilities of MLLMs. The first task, Culture-aware Object Extraction (Extraction@Culture), evaluates the visual recognition and basic cultural understanding capabilities of MLLMs. The second task, Cultural-conflict Object Detection (Conflict@Culture), focuses on evaluating their abilities to understand cultural conflicts. The third task, Culturally-aligned Content Generation (Generation@Culture), measures their multilingual generation capabilities when provided multimodal cultural contexts. The question template for each task is presented in Appendix A.

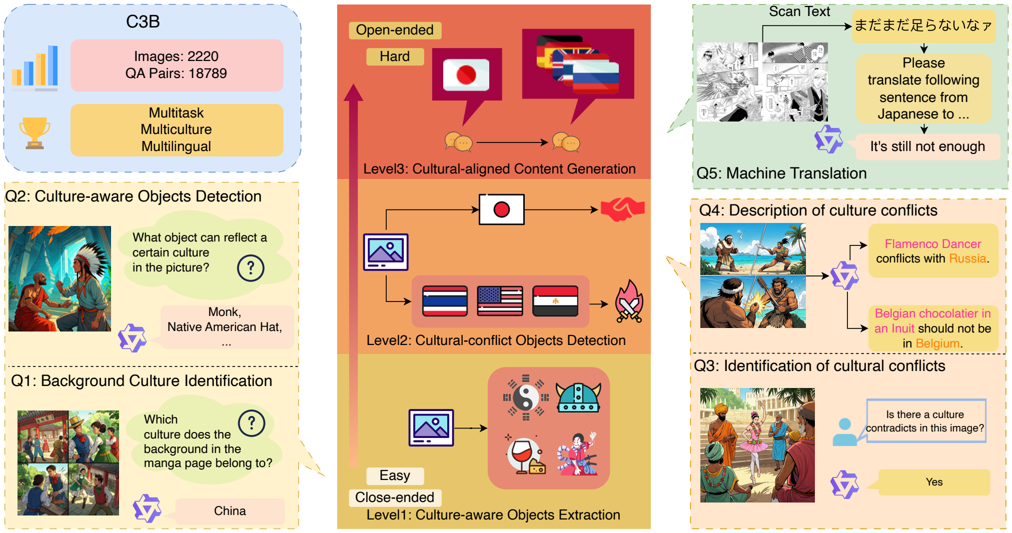

C3B
Images: 2220
QA Pairs: 18789
Multitask
Multiculture
Multilingual
Q2: Culture-aware Objects Detection
What object can reflect a certain culture in the picture?
?
Monk,
Native American Hat,
...
Q1: Background Culture Identification
Which culture does the background in the manga page belong to?
?
China
Open-ended
Hard
Level3: Cultural-aligned Content Generation
Level2: Cultural-conflict Objects Detection
Easy
Close-ended
Level1: Culture-aware Objects Extraction
Scan Text
まだまだ足らないなァ
Please translate following sentence from Japanese to ...
It's still not enough
Q5: Machine Translation
Q4: Description of culture conflicts
Flamenco Dancer conflicts with Russia.
Belgian chocolatier in an Inuit should not be in Belgium.
Q3: Identification of cultural conflicts
Is there a culture contradicts in this image?
Yes

Figure 2: **Overview of C^{3}B**. C^{3}B evaluates MLLMs across three dimensions: Object Identification (foundational vision capability based on culture), Conflict Identification (cultural conflict understanding), and Culturally-aligned Content Generation (comprehensive cultural generation).

Culture-aware Objects Extraction (Extraction@Culture) This task requires the MLLM not only to identify objects within the image but also to understand whether or not the object is related to a specific cultural context. In this task, we set two questions:

1. **Question 1 (Q1): Background Culture Identification** We task MLLMs with identifying the background culture of comic pages. Multiple valid answers may exist for a single question.
2. **Question 2 (Q2): Culture-aware Objects Detection** In this question, MLLMs are required to identify all culturally representative objects in the image. We set up multiple options, with each option including several culturally representative objects. The MLLM needs to choose an option that contains exactly all the objects present in the image.

Cultural-conflict Objects Detection (Conflict@Culture) This task is designed to evaluate MLLMs' capabilities in identifying culturally conflicts within comics. In C^{3}B, if two distinct cultures are depicted within a single image, the instance is classified as a cultural conflict. In this task, we also set two questions:

1. **Question 3 (Q3): Identification of cultural conflicts** In this question, MLLMs must determine whether there is cultural conflict in the presented image.
2. **Question 4 (Q4): Description of culture conflicts** When cultural conflict is detected, the MLLM is subsequently required to specify which objects of the answer to Q2 contradict the culture to the answer of Q1.

Culturally-aligned Content Generation (Generaion@Culture) This task is designed to evaluate the cultural generation capabilities of MLLMs. As for the specific task form, we select machine translation. We have set up translation tasks for the language pairs of Japanese-English (JA-

EN), Japanese-Russian (JA-RU), Japanese-German (JA-DE), Japanese-Thai (JA-TH), and Japanese-Spanish (JA-ES). These language pairs roughly cover languages from various continents, enabling a comprehensive evaluation of the overall translation capabilities of MLLMs.

3.2 Construction of C^{3}B

The construction process of C^{3}B comprises two parts: image collection and annotation. An overview of the entire process is presented in Figure 3.

Image Collection For task Extraction@Culture and Conflict@Culture, we design a comic generation pipeline with the help of doubao APIs[1]. The pipeline consists of two key stages: (1) prompt generation to specify cultural conflict scenarios, followed by (2) image creation based on generated prompts. This process is illustrated in Figure 3a. Subsequent to prompt generation, manual verification is performed on the generated prompts to assess the presence of harmful content.

For the Generation@Culture task, we source images from Manga109 (Matsui et al., 2016). We manually selected 1197 comics images that are closely related to culture and contain more culturally representative objects as well as cultural conflicts.

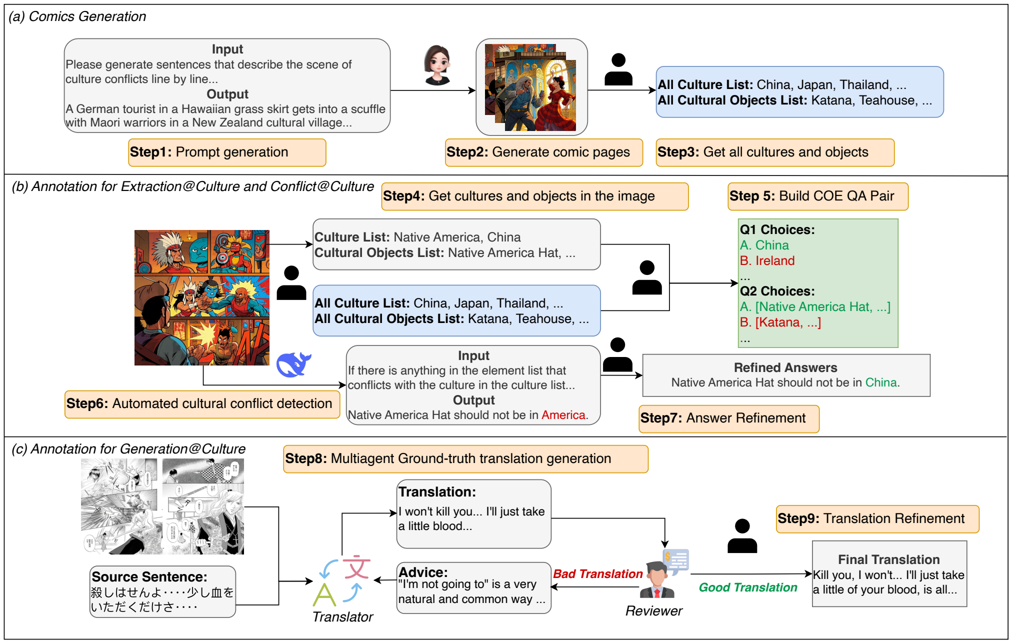

(a) Comics Generation
Input
Please generate sentences that describe the scene of culture conflicts line by line...
Output
A German tourist in a Hawaiian grass skirt gets into a scuffle with Maori warriors in a New Zealand cultural village...
Step1: Prompt generation
Step2: Generate comic pages
All Culture List: China, Japan, Thailand, ...
All Cultural Objects List: Katana, Teahouse, ...
Step3: Get all cultures and objects
(b) Annotation for Extraction@Culture and Conflict@Culture
Step4: Get cultures and objects in the image
Step 5: Build COE QA Pair
Culture List: Native America, China
Cultural Objects List: Native America Hat, ...
All Culture List: China, Japan, Thailand, ...
All Cultural Objects List: Katana, Teahouse, ...
Q1 Choices:
A. China
B. Ireland
...
Q2 Choices:
A. [Native America Hat, ...]
B. [Katana, ...]
...
Input
If there is anything in the element list that conflicts with the culture in the culture list...
Output
Native America Hat should not be in America.
Refined Answers
Native America Hat should not be in China.
Step6: Automated cultural conflict detection
Step7: Answer Refinement
(c) Annotation for Generation@Culture
Step8: Multiagent Ground-truth translation generation
Translation:
I won't kill you... I'll just take a little blood...
Step9: Translation Refinement
Source Sentence:
殺しはせんよ‥‥少し血をいただくだけさ‥‥
Advice:
"I'm not going to" is a very natural and common way ...
Bad Translation
Good Translation
Final Translation
Kill you, I won't... I'll just take a little of your blood, is all...
Translator
Reviewer

Figure 3: **The construction process of C^{3}B**. The process contains 3 steps: Comics Generation, Annotation for Extraction@Culture and Conflict@Culture and Annotation for Generation@Culture.

Annotation The annotation for the tasks of Extraction@Culture and Conflict@Culture refers to the process of QA pairs creation (illustrated in Figure 3b). For the task Extraction@Culture, we first manually compiled two lists: one containing all distinct cultures present in the comic pages, and the other listing all culturally representative objects. Next, we manually created QA pairs for each comic page. For Q1, we identified the cultures presented in a given image, then randomly selected additional cultures from the precompiled culture list to form a total of 5 options. For Q2, we first manually created a list of all culturally representative objects in the image, then generated 5 options by randomly modifying this list by either adding one object, deleting one object, or deleting two objects.

For the task Conflict@Culture, we implemented an annotation pipeline consisting of two stages with the help of Deepseek-V3 (DeepSeek-AI, 2024) (illustrated in Figure 3c). The prompt setting is presented in Appendix C.1. The details of the pipeline is:

[1] https://www.volcengine.com/

1. **Automated Cultural Conflict Detection:** Given the background cultures and culturally representative objects in the comics page, Deepseek will analyze each object to identify if the object conflicts with one of the background cultures. After this, it generates either a formatted conflict description or "No".
2. **Manual Verification and Correction:** We manually inspect all generated results, focusing primarily on verifying whether the model misjudged the existence of conflicts. Subsequent to this, we check if the generated formatted conflict descriptions contained formatting inconsistencies or culture-related inaccuracies. The examples of these two kinds of errors are presented in Appendix D.

For the task Generation@Culture, annotation refers to the process of creating ground-truth translations. We design a multi-agent process to generate ground-truth translations, which involves two specialized agents: (1) a Translator responsible for generating translations, and (2) a Reviewer responsible for checking the translation based on the input sentence and giving suggestions. The Translator first provides a rough translation based on the extracted dialog from Manga109. If the Reviewer regards the translation as good translation, we will conduct manual verification to finalize the translation. Otherwise, the Reviewer will examine three specific types of potential errors: contextual inconsistencies, basic translation errors, and culture-related inaccuracies. The suggestions generated by the Reviewer are subsequently integrated into the prompt. The Translator then uses it to produce a new translation. In the annotation process, we employed DeepSeek-V3 as the base model, with the prompt used presented in C.2. More details on annotation are presented in E.

3.3 Data Statistics

To ensure a comprehensive evaluation, C^{3}B includes a total of 2,220 images and more than 18000 QA pairs. An overall statistics of C^{3}B is presented in Table 2.

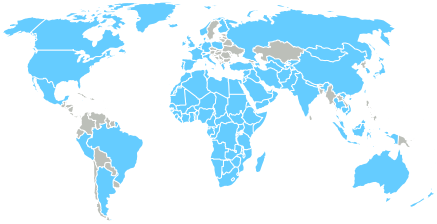


Figure 4: **The cultures C^{3}B covers are presented in a world map.** Regions shaded in blue indicate that the culture is included in C^{3}B.

Table 2: **Statistics of C^{3}B.**

Statistic	Number
Images	2220
- Extraction@Culture& Conflict@Culture	1023
- Generation@Culture	1197
- No. Japanese/Western Style Images	1697/523
QA Pairs	18789
- Extraction@Culture	1023
- Conflict@Culture	1023
- Generation@Culture	16743
- No. All Culture Covered	77
- Min/Max Possible Answers of Extraction@Culture	1/5

In task Extraction@Culture, there are 1023 images. Each image corresponds to one QA pair. In Q1, the number of correct numbers ranges from 1 to 5 among 5 candidate options. This variability increases the difficulty for MLLMs to process the task.

In task Conflict@Culture, the images used are the same as task Extraction@Culture, covering both Japanese comics and American comic styles. This design choice is intentional because the Extraction@Culture and Conflict@Culture tasks evaluate the visual signal processing capabilities of MLLMs. In these two tasks, introducing comics of different styles would make the process more complicated and challenging.

In task Generation@Culture, each QA pair consists of a source sentence and its ground-truth translation in 5 languages. All 16743 QA pairs cover 77 distinct cultures (visualized in Figure 4), enabling us to comprehensively benchmark the understanding different cultures of MLLMs.

4 C^{3}B Evaluation on Existing LLMs

Models To evaluate performance of MLLMs on our C^{3}B benchmark, we choose 11 open-source MLLMs. Specifically, we select SPHINX (Lin et al., 2023), Monkey (Li et al., 2024), MiniGPT-v2 (Chen et al., 2023), mPLUG-Ow13 (Ye et al., 2024), LLaVA family models (Liu et al., 2023b; 2024; 2023a), InternLM-XC2.5 (Zhang et al., 2024), Llama3.2 (Grattafiori et al., 2024), Qwen2.5-VL (Bai et al., 2025) and InternVL2 (Chen et al., 2024).

Metrics For task Extraction@Culture and Conflict@Culture, we use accuracy (ACC) as the evaluation metric. For Q4, because it is constructed based on models' answers to Q1 and Q2, we have designed a composite ACC metric CACC:

$$\text{CACC}(Q_4) = a \times \text{ACC}(Q_1) + b \times \text{ACC}(Q_2) + c \times \text{ACC}(Q_4) \tag{1}$$

where a, b, c represent preset weighting parameters, ACC(Q_i) denotes the accuracy score for Question i. These weighting parameters quantify the respective contributions to the final evaluation. In our experiments, the values of a, b, c are 0.3, 0.3 and 0.4. Q1 and Q2 contribute approximately equally to Q4, while Q4 should dominate. To simultaneously consider the contributions of the three tasks, we set the hyperparameters in this way.

For Generation@Culture, we adopt the conventional metric BLEU (Papineni et al., 2002), supplemented by COMET (Rei et al., 2022) and BLEURT (Sellam et al., 2020) to align with current standards in LLM-based translation research.

Evaluation Settings All evaluations were conducted on a Ubuntu server equipped with an H20-NVLink GPU featuring 96GB of memory. We adhered strictly to the official inference example codes provided by the respective model developers.

5 Experiment Results

5.1 Main Results for task Extraction@Culture and Conflict@Culture

Table 3: **Main Results of task Extraction@Culture and Conflict@Culture in C^{3}B**. For task Conflict@Culture, two types of accuracy metrics are used to evaluate model performance. Specifically, Q4 is constructed based on the answer of MLLM to Q1 and Q2. The calculation process of **CACC** of Q4 considers the answers to Q1 and Q2. In contrast, **ACC** refers to the accuracy of Q4 calculated without referencing answers to Q1 and Q2. The highest performance is marked in **bold**.

Methods	Extraction@Culture		Conflict@Culture	Conflict@Culture (Q4)	
	Q1 (ACC)	Q2 (ACC)	Q3 (ACC)	ACC	CACC
SPHINX	29.9	5.28	**69.2**	0.16	10.6
Monkey	28.4	5.08	33.0	0.19	10.1
MiniGPT-v2	18.0	6.84	66.3	0.51	7.7
mPLUG-Owl3	24.9	15.2	30.8	2.03	12.8
LLaVA1.5-7B	32.5	2.93	56.3	0.00	10.6
LLaVA-NeXT	16.5	39.8	0.88	0.00	16.9
LLaVA-OV	39.0	32.4	53.1	**3.87**	23.0
InternLM-XC2.5	46.0	50.9	68.5	1.94	29.8
Llama3.2	46.0	**59.0**	44.9	2.76	32.6
Qwen2.5-VL	**53.7**	55.9	63.1	3.20	**34.2**
InternVL2	46.0	50.9	68.5	0.01	29.1

Table 3 presents the results for task Extraction@Culture and Conflict@Culture. Qwen2.5-VL demonstrates optimal performance, with Q1 outperforming the second-ranked models (InternLM-XC2.5, Llama3.2 and InternVL2) by 7.7 points and Q4 achieving a 1.6-point lead over the second-ranked model (Llama3.2). We also find that all models achieve extremely low **ACC** in Q4 due to the influence from Q1 and Q2.

As for task Extraction@Culture, in Q1, LLaVA-NeXT performs the worst, because it tends to describe the image rather than answer the question directly. We name this phenomenon as "Turn-a-deaf-ear" (shown in Appendix F.1). In Q2, Llama3.2 performs the best, exceeding average performance by 72.7%. In this task, LLaVA1.5-7B performs the worst, because it keeps answering "A". We name this failure pattern as "Take-a-shot-in-the-dark" and an example is shown in Appendix F.2.

As for task Conflict@Culture, in Q3, SPHINX performs the best, exceeding the second-ranked model (InternLM-XC2.5 and InternVL2) by 0.7 points. In Q4, we find that models from the LLaVA series yield considerably lower results. Specifically, both LLaVA1.5-7B and LLaVA-NeXT achieves 0.00 ACC. Through case study (detailed in Appendix F.3), we find that, LLaVA-NeXT persistently outputs "Nothing", indicating a lack of cultural conflict comprehension capability, while LLaVA1.5-7B cannot follow instructions properly, and we name this phenomenon as "stubbornness".

5.2 Main Results for task Generation@Culture

Table 4 shows the results for task Generation@Culture. The results show that Qwen-2.5-VL demonstrates the most robust multilingual capabilities, as it outperforms other models across all cultural generation tasks. In contrast, MiniGPT-v2 exhibits notably poor performance. Specifically, it achieves a BLEU score of 0 in most tasks. This underperformance can be attributed to the weak instruction-following ability, as presented in Appendix F.4. This error pattern might be caused by the poor understanding capability of comic pages. Additionally, LLaVA-NeXT tends to repeat the source sentence when handling the tasks except JA-EN. Among all tasks, the performance of all models in JA-TH is the poorest while they all perform the best in JA-EN tasks. These results indicate that the multilingual capabilities when given cultural contexts needs to be enhanced, and greater support for low-resource languages should also be strengthened.

Table 4: **Main Result of task Generation@Culture**. The result is presented in the format BLEU/COMET/BLEURT. The highest performance is marked in **bold**.

Methods	JA-EN	JA-DE	JA-RU	JA-ES	JA-TH
SPHINX	3.71/63.2/42.3	1.12/54.3/33.1	0.27/45.0/19.3	1.62/56.9/27.5	0.02/38.7/16.1
Monkey	3.88/62.0/43.3	1.62/55.5/33.5	0.73/48.2/17.7	2.81/56.3/29.3	0.63/50.2/18.0
MiniGPT-v2	0.03/44.5/30.9	0.00/32.1/15.8	0.00/30.0/13.9	0.00/36.5/21.6	0.00/32.2/14.5
mPLUG-Owl3	6.21/58.4/39.8	4.97/56.0/36.8	3.97/57.6/33.5	4.75/55.7/34.6	2.02/51.9/25.2
LLaVA1.5-7B	5.94/62.0/41.0	2.74/54.2/29.7	1.39/53.8/23.1	4.23/62.2/38.4	1.13/44.9/10.2
LLaVA-NeXT	4.88/61.2/41.2	0.00/46.4/10.5	0.14/41.8/11.1	0.00/48.7/6.87	0.28/42.1/12.5
LLaVA-OV	6.00/58.3/35.2	3.37/49.6/24.7	2.78/49.3/23.0	3.24/53.4/26.5	3.77/42.8/13.8
InternLM-XC2.5	4.99/66.6/49.9	1.08/59.2/42.0	1.90/63.4/40.6	2.33/65.3/44.7	0.53/53.7/27.3
Llama3.2	5.05/54.6/35.2	4.27/51.3/31.8	1.70/47.0/20.1	5.73/55.0/31.1	0.99/46.9/17.4
Qwen2.5-VL	**13.2/70.9/53.8**	**12.0/67.7/52.1**	**8.74/69.8/48.5**	**14.5/72.0/54.3**	**9.72/67.8/42.2**
InternVL2	7.20/66.6/47.6	3.52/58.5/40.4	3.33/63.8/40.3	5.32/66.8/46.9	0.74/45.3/24.5

5.3 Impact for Q4 from Q1 and Q2

In C^{3}B, answering Q4 requires models to correctly answer Q1 and Q2. To evaluate the influence of Q1 and Q2 on Q4, we first calculate the correlation coefficients (Formula 2) between Q1 and Q4 and between Q2 and Q4, which yielded values of 0.56 and 0.51, respectively. The result demonstrates that Q1 and Q2 exhibit a moderate correlations with Q4.

$$R(Q_i, Q_4) = \frac{\mathrm{Cov}(\mathrm{ACC}(Q_i), \mathrm{ACC}(Q_4))}{\sqrt{\mathrm{Var}(\mathrm{ACC}(Q_i))\mathrm{Var}(\mathrm{ACC}(Q_4))}} \tag{2}$$

where R represents the correlation coefficient between two questions, Cov refers to the covariance, Var refers to the variance, and $\mathrm{ACC}(Q_i)$ is the accuracy value of Question i.

Moreover, we conducted an ablation study on Q1 and Q2. We added the answers to Q1/Q2 into the prompt template of Q4 to observe the impact on model performance. **CACC** represents the average performance of all 11 MLLMs, and the results are presented in Table 5. The results indicate that incorporating the Q1 answer into the Q4 prompt will enhance performance (+0.003), but adding the Q2 answer will not change the performance. The greatest performance boost is observed when both Q1 and Q2 answers are included (+0.533). Although Q1 and Q2 show moderate correlations with Q4 (0.56 and 0.51), the ablation results reveal only marginal improvement when adding Q1/Q2 answers. This suggests that the positive correlations mainly reflect an intrinsic consistency across related questions, while the direct prompt incorporation of Q1/Q2 answers is not always effectively leveraged by MLLMs.

Table 5: **Performance of Q4 when answers to Q1 and/or Q2 are omitted.** Q1/Q2 Answer means that only Q1/Q2 answer is provided. Q1&Q2 Answer means both Q1 and Q2 answers are provided.

Method	CACC
Base Prompt	19.231
+ Q1 Answer	19.234
+ Q2 Answer	19.231
+ Q1&Q2 Answer	19.764

5.4 Generated Comic Pages Analysis

To evaluate C^{3}B's cultural diversity, we compute three measuresCulture Density Per Image (CDPI), Cultural Breadth Intensity (CBI), and Coverage-Adjusted Density (CAD)and compare them against cultural QA datasets. CDPI is defined as the quantification of the average number of distinct cultures depicted within an image. CBI is designed to measure both cultural density and the total number of cultures depicted in an image. CAD incorporates log-scaling to alleviate excessive rewards arising from large cultural lists. The CDPI, CBI and CAD of a dataset are presented in Equation 3, 4 and 5. In the equations D represents a complete image dataset, which is $\{I_1, I_2, \cdots, I_n\}$, $\text{CultureInImage}(I_i)$ denotes the number of cultures in the image I_i, N_{cultures} refers to the number of cultures occurring in the dataset and $|D|$ as the cardinality of D, corresponding to the total number of images.

$$\text{CDPI}(D) = \frac{1}{|D|} \sum_{i=1}^{n} \text{CultureInImage}(I_i) \tag{3}$$

$$\text{CBI}(D) = \text{CDPI}(D) \times N_{\text{cultures}} \tag{4}$$

$$\text{CAD}(D) = \text{CDPI}(D) \times \log_2(N_{\text{cultures}} + 1) \tag{5}$$

The results are presented in Table 6, indicating that C^{3}B shows significantly higher cultural diversity compared to similar datasets. This highlights its improved ability to evaluate cultural awareness capabilities of MLLMs.

Table 6: **Statistics of different culture-related datasets**.

Dataset	CDPI	CBI	CAD
CVQA	1.00	30	4.91
CulturalVQA	1.00	11	3.58
MOSAIC-1.5k	1.00	43	5.46
GIMMICK	1.00	144	7.18
C^{3}B	**2.28**	**175.56**	**14.29**

5.5 Scores of Different Culture

To examine how well MLLMs can recognize diverse cultures, we conducted a culture-specific evaluation of each model's correctness to mono-culture QA task (Q1). The results, visualized in Figure 5, indicate that the performance of MLLMs differs substantially across cultural groups. In detail, representative cultures (e.g., Cambodia and Japan) are reliably identified by the models, while lesser-known cultures (e.g., Finnish and Somalia) exhibit notably higher error rates. The results show that the cultural awareness capabilities of MLLMs for lesser-known cultures should be enhanced.

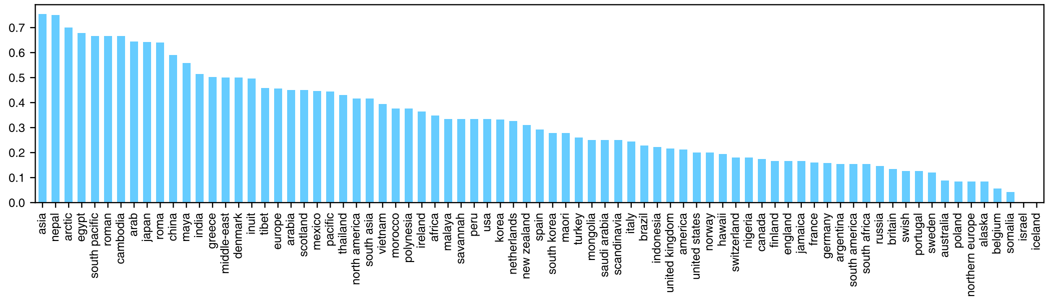


Figure 5: **Model Accuracy on Q1 by Culture.**

5.6 Human-Model Analysis

To validate the effectiveness of C^{3}B, we conducted a human evaluation (more detail in Appendix G). First, We categorized task difficulties based on the number of distinct cultures in each image: 1 (easy), 2/3 (medium), and 4/5 (hard). This tiered classification allows evaluation across different complexity levels. Next, we randomly selected 100 questions per tier. For each tier, we calculated two metrics: the average accuracy performance of human and MLLMs.

Table 7: **Result of Human-Model Analysis**. Q4 is constructed based on the answer of MLLM to Q1 and Q2. The calculation process of ACC of Q4 considers the answers to Q1 and Q2. In contrast, **ACC** refers to the accuracy of Q4 calculated without referencing models' answers to Q1 and Q2. **IRA** stands for Inter-rater Agreement Score. **PTA** stands for Per-tier Time-to-answer.

Method	Q1	Q2	Q3	Q4		IRA	PTA(Seconds)
	ACC	ACC	ACC	ACC	CACC		
		Easy Questions					
Human Evaluation	92.0	77.0	100.0	60.0	74.7	95%	27.01
Model Performance	48.0	28.8	50.5	1.90	23.8	-	-
		Medium Questions					
Human Evaluation	78.0	60.0	100.0	45.0	59.4	91%	52.92
Model Performance	32.6	27.1	54.0	1.69	18.6	-	-
		Hard Questions					
Human Evaluation	69.0	51.0	100.0	35.0	50.0	93%	60.61
Model Performance	25.7	25.5	41.7	1.27	15.9	-	-

The results are presented in Table 7. Three graduate student volunteers participated in the experiment, with each completing all the questions. We then calculated their average accuracy to quantify their performance.

We find that the human performance is significantly better than that of MLLMs, particularly in Q3, where the human performance achieves all 100% accuracy. Constrained by the performance in Q1 and Q2, the human ACC result is relatively low, yet it still exceeds the MLLMs' performance. The results show that C^{3}B is challenging for MLLMs and the cultural awareness capabilities of MLLMs need to be improved.

6 Conclusion

We propose C^{3}B, a novel comic-based, multicultural, multitask, and multilingual cultural awareness capabilities benchmark. Applying comics as the core medium, we enable comprehensive evaluation of cultural awareness capabilities of MLLMs, with the dataset encompassing a large number of images, QA pairs, and cultural contexts. C^{3}B contains three tasks with progressed difficulties, from basic visual recognition to higher-level cultural conflict understanding, and finally to cultural content generation. We benchmarked 11 open-source MLLMs on C^{3}B, revealing a significant performance gap between MLLMs and human performance. Specifically, current MLLMs lack proficiency in understanding less well-known cultures and processing cultural conflicts, highlighting areas for improvement. We anticipate that C^{3}B will serve as a critical tool to support and advance research on MLLMs' cultural awareness capabilities.

Ethics Statement

Our C^{3}B benchmark incorporates images created by Doubao. Through a rigorous human annotation and verification process (Appendix E), we have made every effort to mitigate the majority of cultural biases and stereotypes. Prior to the creation of images, we manually checked all prompts and removed harmful content, thereby ensuring the high quality of the created images.

Regarding the definition of cultural conflict, we have explicitly clarified that the cultural conflict in our task refers to co-occurrence conflict rather than aggressive conflict. Our work is not intended to cause any harm to individuals or groups.

References

B. AlKhamissi, M. ElNokrashy, M. Alkhamissi, and M. Diab. Investigating cultural alignment of large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.671. URL https://aclanthology.org/2024.acl-long.671/.

M. AlKhamissi, Y. Xiao, B. AlKhamissi, and M. Diab. Hire your anthropologist! rethinking culture benchmarks through an anthropological lens, 2025. URL https://arxiv.org/abs/2510.05931.

A. Anantha Ramakrishnan, A. A. Ramakrishnan, and D. Lee. RONA: Pragmatically diverse image captioning with coherence relations. In V. Padmakumar, K. Gero, T. Wambsganss, S. Sterman, T.-H. Huang, D. Zhou, and J. Chung, editors, *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 74–86, Albuquerque, New Mexico, US, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-239-8. doi: 10.18653/v1/2025.in2writing-1.8. URL https://aclanthology.org/2025.in2writing-1.8/.

S. Arora, M. Karpinska, H.-T. Chen, I. Bhattacharjee, M. Iyyer, and E. Choi. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.578. URL https://aclanthology.org/2025.acl-long.578/.

J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

M. Bhatia, S. Ravi, A. Chinchure, E. Hwang, and V. Shwartz. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.385. URL https://aclanthology.org/2024.emnlp-main.385/.

O. Burda-Lassen, A. Chadha, S. Goswami, and V. Jain. How culturally aware are vision-language models?, 2025. URL https://arxiv.org/abs/2405.17475.

D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, L. Baraldi, M. Cornia, and R. Cucchiara. The revolution of multimodal large language models: A survey. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.807. URL https://aclanthology.org/2024.findings-acl.807/.

A. Chen, Y. Song, K. Chen, X. Bai, M. Yang, L. Nie, J. Liu, T. Zhao, and M. Zhang. Make imagination clearer! stable diffusion-based visual imagination for multimodal machine translation. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26567–26583, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1289. URL https://aclanthology.org/2025.acl-long.1289/.

J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023. URL https://arxiv.org/abs/2310.09478.

Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

P. R. Cohen and A. E. Howe. How evaluation guides ai research: The message still counts more than the medium. *AI Magazine*, 9(4):35, Dec. 1988. doi: 10.1609/aimag.v9i4.952. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/952.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

R. Dong, C. Han, Y. Peng, Z. Qi, Z. Ge, J. Yang, L. Zhao, J. Sun, H. Zhou, H. Wei, X. Kong, X. Zhang, K. Ma, and L. Yi. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=y01KGvd9Bw.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan,

N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, and Z. Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

C. Guérin, C. Rigaud, A. Mercier, F. Ammar-Boudjelal, K. Bertet, A. Bouju, J.-C. Burie, G. Louis, J.-M. Ogier, and A. Revel. ebdtheque: a representative database of comics. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1145–1149, 2013.

N. D. Huynh, M. R. Bouadjenek, S. Aryal, I. Razzak, and H. Hacid. Visual question answering: from early developments to recent advances – a survey, 2025. URL https://arxiv.org/abs/2501.03939.

M. Iyyer, V. Manjunatha, A. Guha, Y. Vyas, J. Boyd-Graber, H. Daumé, and L. Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6478–6487, 2017. doi: 10.1109/CVPR.2017.686.

A. Karamolegkou, P. Rust, R. Cui, Y. Cao, A. Søgaard, and D. Hershcovich. Vision-language models under cultural and inclusive considerations. In N. Soni, L. Flek, A. Sharma, D. Yang, S. Hooker, and H. A. Schwartz, editors, *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, pages 53–66, TBD, Aug. 2024. ACL. doi: 10.18653/v1/2024.hucllm-1.5. URL https://aclanthology.org/2024.hucllm-1.5/.

C. Kramsch. Language and culture. *AILA Review*, 27(1):30–55, 2014. ISSN 1461-0213. doi: https://doi.org/10.1075/aila.27.02kra. URL https://www.jbe-platform.com/content/journals/10.1075/aila.27.02kra.

J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Z. Li, B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, and X. Bai. Monkey: Image resolution and text label are important things for large multi-modal models, 2024. URL https://arxiv.org/abs/2311.06607.

Z. Lin, C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. Chen, J. Han, S. Huang, Y. Zhang, X. He, H. Li, and Y. Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models, 2023. URL https://arxiv.org/abs/2311.07575.

H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2023a.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning, 2023b.

H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):2181121838, Nov. 2016. ISSN 1573-7721. doi: 10.1007/s11042-016-4020-z. URL http://dx.doi.org/10.1007/s11042-016-4020-z.

T. Naous, M. J. Ryan, A. Ritter, and W. Xu. Having beer after prayer? measuring cultural bias in large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.862. URL https://aclanthology.org/2024.acl-long.862/.

S. Nayak, K. Jain, R. Awal, S. Reddy, S. V. Steenkiste, L. A. Hendricks, K. Stanczak, and A. Agrawal. Benchmarking vision language models for cultural understanding. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.329. URL https://aclanthology.org/2024.emnlp-main.329/.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

S. Pawar, J. Park, J. Jin, A. Arora, J. Myung, S. Yadav, F. G. Haznitrama, I. Song, A. Oh, and I. Augenstein. Survey of cultural awareness in language models: Text and beyond, 2024. URL https://arxiv.org/abs/2411.00860.

R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52/.

A. Reuel, A. Hardy, C. Smith, M. Lamparth, M. Hardy, and M. J. Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 21763–21813. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/26889e8359e7ef8a7f5d77457364ca55-Paper-Datasets_and_Benchmarks_Track.pdf.

D. Romero, C. Lyu, H. A. Wibowo, T. Lynn, I. Hamed, A. N. Kishore, A. Mandal, A. Dragonetti, A. Abzaliev, A. L. Tonja, B. F. Balcha, C. Whitehouse, C. Salamea, D. J. Velasco, D. I. Adelani, D. L. Meur, E. Villa-Cueva, F. Koto, F. Farooqui, F. Belcavello, G. Batnasan, G. Vallejo, G. Caulfield, G. Ivetta, H. Song, H. B. Ademtew, H. Maina, H. Lovenia, I. A. Azime, J. C. B. Cruz, J. Gala, J. Geng, J.-G. Ortiz-Barajas, J. Baek,

J. Dunstan, L. A. Alemany, K. R. Y. Nagasinghe, L. Benotti, L. F. D'Haro, M. Viridiano, M. Estecha-Garitagoitia, M. C. B. Cabrera, M. Rodríguez-Cantelar, M. Jouitteau, M. Mihaylov, M. F. M. Imam, M. F. Adilazuarda, M. Gochoo, M.-E. Otgonbold, N. Etori, O. Niyomugisha, P. M. Silva, P. Chitale, R. Dabre, R. Chevi, R. Zhang, R. Diandaru, S. Cahyawijaya, S. Góngora, S. Jeong, S. Purkayastha, T. Kuribayashi, T. Clifford, T. Jayakumar, T. T. Torrent, T. Ehsan, V. Araujo, Y. Kementchedjhieva, Z. Burzo, Z. W. Lim, Z. X. Yong, O. Ignat, J. Nwatu, R. Mihalcea, T. Solorio, and A. F. Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024. URL https://arxiv.org/abs/2406.05967.

F. Schneider, C. Holtermann, C. Biemann, and A. Lauscher. GIMMICK: Globally inclusive multimodal multitask cultural knowledge benchmarking. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9605–9668, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.500. URL https://aclanthology.org/2025.findings-acl.500/.

T. Sellam, D. Das, and A. Parikh. BLEURT: Learning robust metrics for text generation. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

S. Singh, A. Romanou, C. Fourrier, D. I. Adelani, J. G. Ngui, D. Vila-Suero, P. Limkonchotiwat, K. Marchisio, W. Q. Leong, Y. Susanto, R. Ng, S. Longpre, S. Ruder, W.-Y. Ko, A. Bosselut, A. Oh, A. Martins, L. Choshen, D. Ippolito, E. Ferrante, M. Fadaee, B. Ermis, and S. Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919/.

A. Vayani, D. Dissanayake, H. Watawana, N. Ahsan, N. Sasikumar, O. Thawakar, H. B. Ademtew, Y. Hmaiti, A. Kumar, K. Kuckreja, M. Maslych, W. A. Ghallabi, M. Mihaylov, C. Qin, A. M. Shaker, M. Zhang, M. K. Ihsani, A. Esplana, M. Gokani, S. Mirkin, H. Singh, A. Srivastava, E. Hamerlik, F. A. Izzati, F. A. Maani, S. Cavada, J. Chim, R. Gupta, S. Manjunath, K. Zhumakhanova, F. H. Rabevohitra, A. Amirudin, M. Ridzuan, D. Kareem, K. More, K. Li, P. Shakya, M. Saad, A. Ghasemaghaei, A. Djanibekov, D. Azizov, B. Jankovic, N. Bhatia, A. Cabrera, J. Obando-Ceron, O. Otieno, F. Farestam, M. Rabbani, S. Baliah, S. Sanjeev, A. Shtanchaev, M. Fatima, T. Nguyen, A. Kareem, T. Aremu, N. Xavier, A. Bhatkal, H. Toyin, A. Chadha, H. Cholakkal, R. M. Anwer, M. Felsberg, J. Laaksonen, T. Solorio, M. Choudhury, I. Laptev, M. Shah, S. Khan, and F. Khan. All languages matter: Evaluating lmms on culturally diverse 100 languages, 2025. URL https://arxiv.org/abs/2411.16508.

E. Vivoli, M. Bertini, and D. Karatzas. Comix: A comprehensive benchmark for multi-task comic understanding, 2024. URL https://arxiv.org/abs/2407.03550.

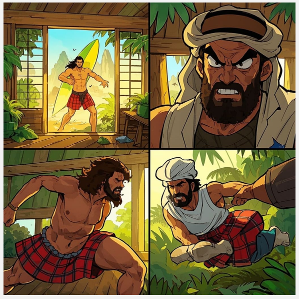
P. Xu, Y. Wang, S. Zhang, X. Zhou, X. Li, Y. Yuan, F. Li, S. Zhou, X. Wang, Y. Zhang, and H. Zhao. Tcc-bench: Benchmarking the traditional chinese culture understanding capabilities of mllms, 2025. URL https://arxiv.org/abs/2505.11275.

S. Yang, D. Zhang, J. Ren, Z. Xu, X. Zhang, Y. Song, H. Lin, and F. Xia. Cultural bias matters: A cross-cultural benchmark dataset and sentiment-enriched model for understanding multimodal metaphors. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26301–26317, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1275. URL https://aclanthology.org/2025.acl-long.1275/.

J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. URL https://arxiv.org/abs/2408.04840.

P. Zhang, X. Dong, Y. Zang, Y. Cao, R. Qian, L. Chen, Q. Guo, H. Duan, B. Wang, L. Ouyang, S. Zhang, W. Zhang, Y. Li, Y. Gao, P. Sun, X. Zhang, W. Li, J. Li, W. Wang, H. Yan, C. He, X. Zhang, K. Chen, J. Dai, Y. Qiao, D. Lin, and J. Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output, 2024. URL https://arxiv.org/abs/2407.03320.

Input Image:



Q1: Which culture does the background in the comics page belong to? Choose all the possible answers.

Choices for Q1:

A. Cambodia
B. Greece
C. Brazil
D. Australia
E. China

Answer to Q1: C D

Q2: What object can reect a certain culture in the picture? You should choose the most appropriate answer.

Choices for Q2:

A. ['An Australian surfer', ' kilt', ' Arabian sheikhs in a Brazilian', ' rainforest hut', 'indian sari']

B. ['An Australian surfer', ' kilt', ' Arabian sheikhs in a Brazilian', ' rainforest hut', 'batik robe has a clash']

C. ['An Australian surfer', ' kilt', ' Arabian sheikhs in a Brazilian', ' rainforest hut', 'a group of inuit elders']

D. ['An Australian surfer', ' kilt', ' Arabian sheikhs in a Brazilian', ' rainforest hut']

E. ['An Australian surfer', ' kilt', ' Arabian sheikhs in a Brazilian']

Answer to Q2: D

Q3: lease answer the following question based on the image. Is there a culture contradicts in this image?

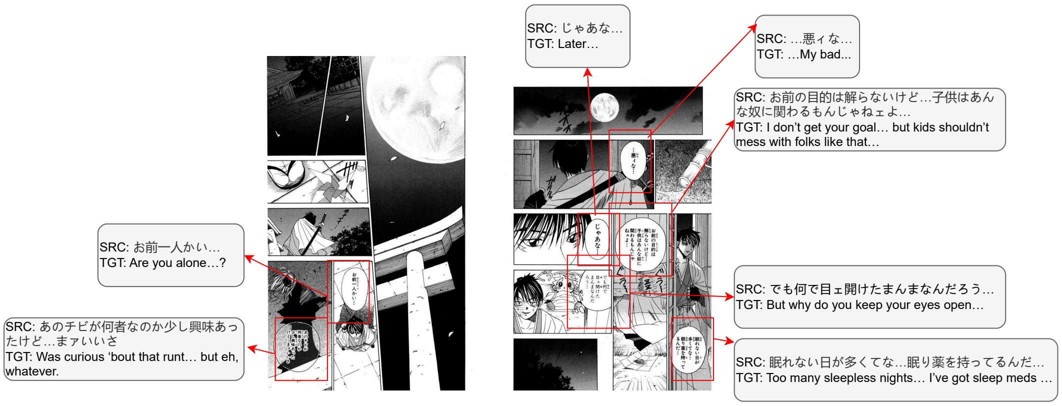
Answer to Q3: Yes.

Q4: If there is contradict in this image, please answer the following question. For each object in ['An Australian surfer', ' kilt', ' Arabian sheikhs in a Brazilian', ' rainforest hut'], if that object and a certain culture in ["Brazil", "Australia"] conflict with each other, please output it. The format is "Something should not be in Some Culture". For example, "Katana should not be in America", "Chinese calligrapher should not be in Africa". If there isn't contradict, output nothing.

Answer to Q4:

1. An Australian surfer should not be in Brazil.
2. kilt should not be in Australia.
3. kilt should not be in Brazil.
4. Arabian sheikhs in a Brazilian should not be in Australia.
5. rainforest hut should not be in Australia.

Figure 6: **A data sample of Extraction@Culture and Conflict@Culture**.



SRC: じゃあな…
TGT: Later…

SRC: …悪ィな…
TGT: …My bad...

SRC: お前の目的は解らないけど…子供はあんな奴に関わるもんじゃねェよ…
TGT: I don’t get your goal… but kids shouldn’t mess with folks like that…

SRC: でも何で目ェ開けたまんまなんだろう…
TGT: But why do you keep your eyes open…

SRC: 眠れない日が多くてな…眠り薬を持ってるんだ…
TGT: Too many sleepless nights… I’ve got sleep meds …

SRC: お前一人かい…
TGT: Are you alone…?

SRC: あのチビが何者なのか少し興味あったけど…まァいいさ
TGT: Was curious ‘bout that runt… but eh, whatever.

Figure 7: **A data sample of Generation@Culture**. During each inference, a single source sentence and the comic page are input to the MLLM.

A QUESTION SETTING FOR EACH TASK

A.1 EXTRACTION@CULTURE

Question 1: Which culture does the background in the comics page belong to? Choose all the possible answers.
Question 2: What object can reflect a certain culture in the picture? You should choose the most appropriate answer.

A.2 CONFLICT@CULTURE

Question 3: Please answer the following question based on the image. Is there a culture contradicts in this image?
Question 4: If there is contradict in this image, please answer the following question. For each object in the A_2, if that object and a certain culture in A_1 conflict with each other, please output it. The format is "Something should not be in Some Culture". For example, "Katana should not be in America", "Chinese calligrapher should not be in Africa". If there isn’t contradict, output nothing.

where A_1 refers to the list of cultures of the MLLM’s answer to Q1, and A_2 refers to the list of cultural representative objects of the MLLM’s answer to Q2.

A.3 GENERATION@CULTURE

Translate following sentences from Japanese into English. Output the result line by line. The sentences are:

B SOME DATA SAMPLE FROM C^{3}B

B.1 EXTRACTION@CULTURE AND CONFLICT@CULTURE

An example of Extraction@Culture and Conflict@Culture is presented in Figure 6.

B.2 GENERATION@CULTURE

An example of Generation@Culture is presented in Figure 7.

C PROMPT SETTING FOR DATA COLLECTION AND ANNOTATION

C.1 PROMPT SETTING FOR QA PAIR CREATION OF CONFLICT@CULTURE TASK

> Only output the answer. You need to check if there is anything in the element list that conflicts with the culture in the culture list. If there is, you should output the result as "Something should not be in Somewhere.". Otherwise, just output "No".
> For an element, if it conflicts with any one of the culture in the list, you should count it as aconflict. For example, the culture list is [Japan, China] and the element list is [katana, Chinese tea]. Then you should output "katana should not be in China" and "Chinese tea should not bein Japan".
> As for acceptable results. For example, some acceptable results are "Katana should not be inAmerica", "Chinese calligrapher should not be in Africa" and "No".
> Now output the result for: The element list is l_e. The culture list is l_p.

Figure 8: **The prompt setting for QA pair creation in Conflict@Culture Task**.

In the prompt setting of Conflict@Culture (presented in Figure 8), l_e and l_p denote the list of culturally representative objects and culture of image annotated in task Extraction@Culture respectively.

C.2 PROMPT SETTING FOR ANNOTATION IN GENERATION@CULTURE

The model we apply in annotation in Generation@Culture for creating C^{3}B is DeepSeek-R1 (DeepSeek-AI, 2025). The prompt we use for Translator is presented in Figure 9. The prompt we use for Reviewer is presented in Figure 10. Upon receiving the review feedback, the prompt provided to the Translator is presented in Figure 11.

> Output the answer line by line. Please only output the answer. Translate following sentence from Japanese to English: s.

Figure 9: **The prompt used for the Translator**, where s denotes the source sentence that is to be translated.

> Output the answer line by line. Only output the translation, not suggestions. Please only output the modified translation. You need to pay attention to whether there are any inconsistencies in context, whether there are culturally related errors and whether there are any translation errors. The source is:s and the translation is:t.

Figure 10: **The prompt used for the Reviewer**, where t denotes the translation provided by the Translator.

> Output the answer line by line. Please only output the answer. Translate following sentence from Japanese to English: s. Here is some advice: a

Figure 11: **The prompt used for the Translator once the review is provided**, where a denotes the review generated by the Reviewer.

> **Input**: Only output the answer. You need to check if there is anything in the element list that conflicts with the culture in the culture list. If there is, you should output the result as "Something should not be in Somewhere.". Otherwise, just output "No".
> For an element, if it conflicts with any one of the culture in the list, you should count it as a conflict. For example, the culture list is [Japan, China] and the element list is [katana, Chinese tea]. Then you should output "katana should not be in China" and "Chinese tea should not bein Japan".
> As for acceptable results. For example, some acceptable results are "Katana should not be in America", "Chinese calligrapher should not be in Africa" and "No".
> Now output the result for: The element list is [China, Japan, Brazil]. The culture list is [Chinese calligrapher, Japanese Katana, Brazilian Dancer].
> **Output with Error**: In China, Brazilian dancers are rare...
> **Corrected Output**: Brazilian dancers should not be in China.

Figure 12: **An example of formatting inconsistency.** The format provided is shaded in blue, and the output with error doens't follow it.

D Examples of Errors In Manual Verification in Conflict@Culture

D.1 Formatting inconsistency

This type of error occurs when the format of the conflict descriptions generated by the models fails to conform to the format provided. An example is provided in Figure 12.

D.2 Culture-related Errors

This type of error occurs when the conflict descriptions generated by the models contains cultural errors. An example is provided in Figure 13.

E Annotation Details

All annotations were performed by three graduate students and two undergraduate students, with distinct division of labor across annotation tasks. All of the annotators come from Asian countries. Among the five annotators, three are graduate students engaged in culture-related research. The quality of annotation is ensured by credible online resources (e.g. Wikipedia).

For the annotation of cultural lists and cultural object lists in Image Collection phase, the two undergraduate students were tasked with this work, where each was assigned 512 images (accounting for 50% of the total dataset) for annotation. Subsequent to this initial annotation phase, the three graduate students conducted a comprehensive review of the annotated content and rectified any annotation errors identified based on the prompts which are used to generate images. The criteria for this phase are defined as follows:

1. **Fidelity**: An object is counted if it evokes a perceived association with a specific cultural context.

Input: Only output the answer. You need to check if there is anything in the element list that conflicts with the culture in the culture list. If there is, you should output the result as "Something should not be in Somewhere.". Otherwise, just output "No".
For an element, if it conflicts with any one of the culture in the list, you should count it as a conflict. For example, the culture list is [Japan, China] and the element list is [katana, Chinese tea]. Then you should output "katana should not be in China" and "Chinese tea should not bein Japan".
As for acceptable results. For example, some acceptable results are "Katana should not be in America", "Chinese calligrapher should not be in Africa" and "No".
Now output the result for: The element list is [China, Japan, Brazil]. The culture list is [Chinese calligrapher, Japanese Katana, Brazilian Dancer].
Output with Error: Japanese Katana should not be in Japan.
Corrected Output: Japanese Katana should not be in China. Japanese Katana should not be in Brazil.

Figure 13: **An example of culture-related errors.** In our example, Japanese katana should not be in the country except Japan.

2. **Granularity**: For an object associated with a culture shared across multiple countries, a regional-level description shall be adopted.
3. **Security**: An image shall be excluded if it contains aggressive depictions, which refer to racist content, sexual explicitness, and other analogous content that may cause harm to individuals.

Following the annotation rectification phase, we calculated that 237 out of the total 1,023 images underwent revisions. The inter-annotator agreement score was determined to be 76.8%.

For the annotation of question-answer (QA) pairs corresponding to the three tasks, all five annotators participated in the verification phase. Initially, each annotator independently revised the annotation results; subsequent to this individual revision step, a joint review of each question was conducted collectively by the entire annotation team. Finally, the questions corresponding to 755 images were manually revised, resulting in a correction rate of 73.8%. In this phase, the criteria are defined as follows:

1. **Culture Fidelity**: All annotators were instructed to retrieve relevant information from credible online sources (e.g. Wikipedia) to verify the accuracy of the specific forms of cultural objects depicted in the image. In cases where inaccuracies were identified, the annotations were revised based on the outcomes of group discussions.

The guidelines for the annotation of cultural lists and cultural object lists are presented as follows:

1. **Fidelity & Granularity**: If an element in the image shows a country's cultural traits, please label it with the country name. If it reflects a regional cultural feature, please use the region or continent namelike "Middle East," "Asia," "South Pacific," and so on.
2. **Security**: If there's any stereotype or offensive content in the image or prompt, mark it for removal.
3. **Example**: When labeling elements in the image, try to keep the labels as close as possible to the prompt used to generate the image. Here's an example: "A Brazilian footballer in a Scottish kilt has a shouting match with Chinese ping-pong players in a Beijing sports complex." In such case, if the elements mentioned in the prompt appear in the image and can be matched to the entities in the prompt (e.g., the footballer in the image corresponds to a "Brazilian footballer," etc.), label them exactly as "Brazilian Footballer," "Scottish Kilt," and so on.
4. **Credibility**: For the cultural elements in the image, look them up on Wikipedia if you are not familiar with the culture. You can label them only if they match the relevant information there; otherwise, skip the labeling.

F Case Study

During evaluation, we observed several unexpected behavioral patterns that led to suboptimal question-answering performance in MLLMs.

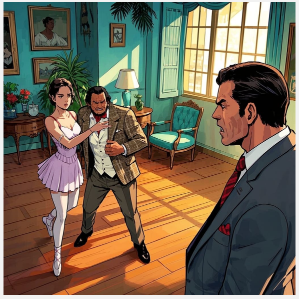
F.1 Error Cases in Q1

"Turn-a-deaf-ear": This behavioral pattern is particularly occurred in LLaVA-NeXT, where the model frequently defaulted to describing the image rather than direct question-answering when presented with clear instructions. We name this behavoir as "Turn-a-deaf-ear". An example is provided in Figure 14.

We assume that this behavioral pattern likely stems from the fine-tuning dataset's compositional bias, which predominantly trains MLLMs for image description tasks rather than image-grounded question answering.

We need to mention that, in Q3, LLaVA-NeXT fails to predict the result properly due to the same reason as in Q1.

Input Image:



Input: Which culture does the background in the comics page belong to? Choose all the possible answers. Please only generate answer.
A. Australia
B. France
C. Middle-east
D. Indonesia
E. South america
Ground-truth Answer: A,B,C
Output: The background in the comics page appears to be inspired by a setting that could be from a variety.

Figure 14: **An example of "Turn-a-deaf-ear" is presented.** The MLLM is asked to output the correct choices but it outputs the detail of the comics page instead.

F.2 Error Case in Q2

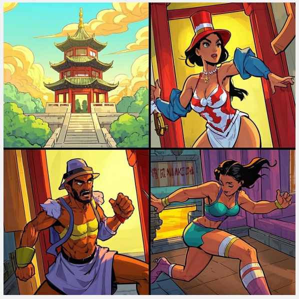
"Take-a-shot-in-the-dark": This behavioral pattern was particularly observed in LLaVA1.5-7B when the model attempted to answer Q2. For this question, the model exhibited a tendency to output "A" as the response.

We calculated the frequency that LLaVA1.5-7B outputs the choice "A" and the result is 78.4, which says that the model cannot understand the cultures properly and tends to output an answer.

F.3 Error Case in Q4

Keep answering "Nothing": This behavioral pattern is particularly occurred in LLaVA-NeXT, where the model always answering "Nothing", which shows a lack of cultural conflict comprehension capability. An example is presented in Figure 15.

Input Image:



Input: If there is contradict in this image, please answer the following question. For each object in ['Brazilian samba dancer', ' Swiss dirndl engages', ' dispute with Turkish Janissaries', ' Chinese pagoda'], if that object and a certain culture in the [Turkey, Switzeland, Brazil, China] conflict with each other, please output it. The format is "Something should not be in Some Culture". For example, "Katana should not be in America", "Chinese calligrapher should not be in Africa". If there isn't contradict, output nothing.
Ground-truth Answer: "Swiss dirndl engages should not be in Brazil, · · · "
Output: Nothing.

Figure 15: **An example of keeping answering nothing**. MLLM keeps answering "nothing" instead of the description of cultural conflicts.

"Stubbornness": This behavioral pattern is particularly occurred in LLaVA-1.5-7B, where the model frequently defaulted to follow the instruction without thinking. We name this behavior as "Stubbornness". An example is provided in Figure 16.

F.4 Error Case In Q5

Poor instruction following capabilities of MiniGPT: For Question 5, our instruction to MLLMs was to directly output the translated sentence without including any redundant explanatory content. Nevertheless, MiniGPT tended to elaborate on how the translation was performed and also repeated the identical sentence a second time. An example is presented in Figure 17.

G Details on Human-Model Analysis

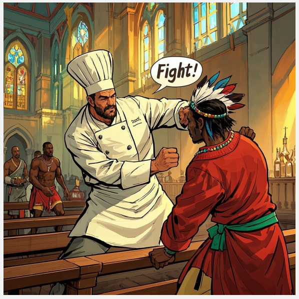
Regarding the human-model analysis, three undergraduate students volunteered to participate. Each question was answered separately, with a one-hour time limit per question, and was presented in the same format used for the MLLMs.

H Effect of OCR Integration on MT Task Performance

We conducted additional experiments to evaluate the impact of OCR on the Generation@Culture task. Our experimental settings are as follows: one group directly adopted the pre-extracted text from Manga109, while the other group utilized OCR[2] technology for text extraction. We use BLEU as the evaluation metric The results are presented in Table 8.

[2] https://github.com/PaddlePaddle/PaddleOCR

Input Image:



Input: If there is contradict in this image, please answer the following question. For each object in ['French chef in a Native', ' American headdress is wrestling', ' African tribal chiefs', ' Russian Orthodox church'], if that object and a certain culture in the [France, America, Russia] conflict with each other, please output it. The format is "Something should not be in Some Culture". For example, "Katana should not be in America", "Chinese calligrapher should not be in Africa". If there isn't contradict, output nothing.
Ground-truth Answer: "French chef in a Native should not be in United States.", · · · .
Output:Something should not be in Some Culture.

Figure 16: **An example of "Stubbornness" is presented.** The MLLM fails to generate the required substitutions, leaving both the placeholder terms 'Something' and 'Some Culture' unmodified in its output.

Sure, here are the translations of the sentences you provided:
1. kuru na!: This is an interjection used to express excitement or surprise, similar to "Wow!" or "Whoa!" in English.
2. kuru na!: This is another way of expressing excitement or surprise, similar to "Wow!" or "Whoa!"

Figure 17: **An example of poor instruction following capability of MiniGPT**. Our instruction clearly required MiniGPT to directly output the target translations without additional content. Nevertheless, the model responded with redundant descriptive text (e.g., explaining the sentence type and emotional connotation) and duplicated translations for the identical Japanese input. Both issues indicate its failure to adhere to the constraint of instruction.

From the results, it is evident that OCR significantly degrades model performance. Since we did not utilize OCR in our experiments, the observed lower model performance is not attributable to OCR-related factors.

Table 8: **Main Result of effect of OCR on MT tasks**. The highest performance is marked in **bold**.

Methods	**JA-DE**		**JA-RU**		**JA-ES**		**JA-TH**	
	OCR	**No OCR**	**OCR**	**No OCR**	**OCR**	**No OCR**	**OCR**	**No OCR**
Qwen2.5-VL	0.92	**12.0**	0.97	**8.74**	1.32	**14.5**	1.22	**9.72**
LLaVA1.5-7B	0.64	**2.74**	0.24	**1.39**	0.62	**4.23**	0.32	**1.13**
Llama3.2	0.33	**4.27**	0.22	**1.70**	0.45	**5.73**	2.73	**0.99**