

A Progressive Evaluation Framework for Multicultural Analysis of Story Visualization

Janak Kapuriya* Ali Hatami Paul Buitelaar

Data Science Institute, University of Galway, Galway, Ireland

{janakkumar.kapuriya, ali.hatami, paul.buitelaar}@insight-centre.org

*Corresponding Author

Abstract

Recent advancements in text-to-image generative models have improved narrative consistency in story visualization. However, current story visualization models often overlook cultural dimensions, resulting in visuals that lack authenticity and cultural fidelity. In this study, we conduct a comprehensive multicultural analysis of story visualization using current text-to-image models across multilingual settings on two datasets: FlintstonesSV and VIST. To assess cultural dimensions rigorously, we propose a Progressive Multicultural Evaluation Framework and introduce five story visualization metrics, Cultural Appropriateness, Visual Aesthetics, Cohesion, Semantic Consistency, and Object Presence, that are not addressed by existing metrics. We further automate assessment through an MLLM-as-Jury framework that approximates human judgment. Human evaluations show that models generate more coherent, visually appealing, and culturally appropriate stories for real-world datasets than for animated ones. The generated stories exhibit a stronger alignment with English-speaking cultures across all metrics except Cohesion, where Chinese performs better. In contrast, Hindi ranks lowest on all metrics except Visual Aesthetics, reflecting real-world cultural biases embedded in current models. This multicultural analysis provides a foundation for future research aimed at generating culturally appropriate and inclusive visual stories across diverse linguistic and cultural settings.

1 Introduction

Recent advancements in text-to-image diffusion models, such as DALL-E-3 (Betker et al., 2023), SDXL (Podell et al., 2023) and Imagen 3 (Baldridge et al., 2024) have achieved remarkable performance in generating photorealistic images from text prompts. These models have set new benchmarks for producing high-quality im-

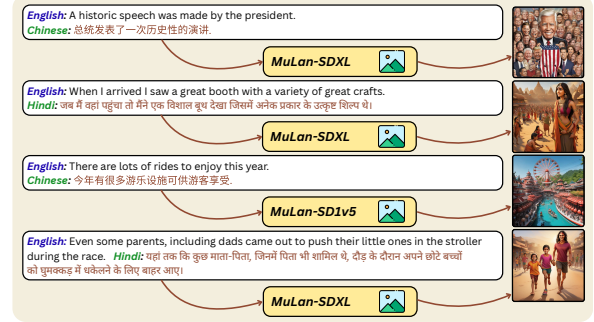


Figure 1: Cultural inconsistencies and stereotypes in generated story scenes across models and languages. **Ex-1:** Model is not interpreting the word ‘president’ in Chinese as ‘party leader’ in China, and generated an image of a US president. **Ex-2:** Instead of a modern craft fair, it depicts temples and a crowded assembly of saints, reinforcing the stereotype of an ancient *Mahakumbh* style fair. **Ex-3:** Red lantern stereotypes dominate the rides, misrepresenting Chinese culture. **Ex-4:** A Hindi cultural stereotype depicts a parent leading children in a race rather than using a stroller, with several children barefoot, reinforcing inaccurate cultural assumptions.

ages, opening up a wide array of real-world applications including controlled image editing ControlNet (Zhang et al., 2023), FLUX.1 (Labs et al., 2025) and video generation from textual prompts SORA (Brooks et al., 2024), Stable Video Diffusion (Blattmann et al., 2023) and Veo 2 (Sharma et al., 2024). Within this broader landscape, one promising and rapidly growing application is Story Visualization, which focuses on generating coherent sequences of visual scenes based on sequences of storylines. This task holds immense potential across fields such as education, animation and content creation. However, current story visualization models are predominantly advanced in English, with their capabilities in multilingual settings remaining largely unexplored.

Stories are often connected with the culture and geography they represent, serving as reflections

of the traditions, values and identities of specific communities (Bruner, 2010). Despite this, current story visualization models (Zhou et al., 2024; Mao et al., 2024; Yang et al., 2024; Zheng and Fu, 2024) generate sequences of images without adequately considering cultural dimensions, produces visuals that lack authenticity and cultural fidelity. This is critical in multilingual settings, where distinct linguistic backgrounds carry unique cultural nuances as shown in Figure 1. Addressing these aspects remains underexplored, leaving a gap in meaningful, accurate multilingual story visualization.

Building on this, we systematically conducted a multicultural analysis of story visualization across three languages English, Chinese and Hindi on the animated story dataset FlintstonesSV (Gupta et al., 2018) and the real world story dataset VIST (Huang et al., 2016), with five multilingual text-to-image generation models. To enable an in-depth evaluation of story visualization we propose five story visualization evaluation metrics such as Cohesion, Cultural Appropriateness, Object Presence, Visual Aesthetics and Semantic Consistency, which are not captured by existing evaluation methods like FID (Heusel et al., 2017) and CLIP (Radford et al., 2021). FID and CLIP metrics focus only on image-to-image or image-to-text, whereas our proposed metrics are designed for the evaluation of story visualization in regard to text-to-image sequences and cultural aspects. To more rigorously assess cultural aspects, we introduce the Progressive Multiculture Evaluation Framework, which provides a deeper evaluation of cultural fidelity in generated stories. Since human evaluation of proposed metrics at large scale is time-consuming and costly, we utilize the MLLM-as-Jury Evaluation Framework. This framework uses three Multimodal Large Language Model (MLLM) as judges. The final rating for a story sample is determined by aggregating the ratings from the three judge models.

To perform the multicultural analysis of story visualization in multilingual settings, we formulated the following research questions:

RQ 1: How does the visualization of the same story varies across different languages?

RQ 2: How does story visualization adapt to different cultures?

Our contributions are as follows:

- **Exploration of Cultural Aspects:** We conducted a comprehensive study of the cultural dimensions of story visualization in a multi-

lingual setting.

- **Proposed Story Visualization Metrics:** We propose five evaluation metrics: Cohesion, Visual Aesthetics, Cultural Appropriateness, Object Presence, Semantic Consistency tailored to capture the unique characteristics of story visualization, addressing gaps left by traditional evaluation metrics like FID and CLIP.
- **Progressive Multicultural Evaluation Framework:** This framework enables in-depth evaluation of culture by incorporating culture relevant focus-points and illustrative examples during the evaluation process.
- **MLLM-as-Jury Evaluation Framework:** We introduce the MLLM-as-Jury Evaluation Framework to enable large-scale, automated evaluation of story aspects using Expert Role-specific evaluation prompts.
- **Multicultural Analysis on Animated and Real Stories:** We analyzed story visualization on open-ended real-world story and animated story datasets.

2 Related Work

2.1 Story Visualization

Recent advancements in story visualization have improved narrative consistency through several dataset-specific training approaches. These include models such as Make-A-Story (Rahman et al., 2023), StoryGPT-V (Shen and Elhoseiny, 2023), ARLDM (Pan et al., 2024), and TemporalStory (Zheng and Fu, 2024). Other work explores long-story visualization (Yang et al., 2024) and training-free approaches like StoryDiffusion (Zhou et al., 2024) and StoryAdapter (Mao et al., 2024). However, these methods focus on English scene descriptions and overlook the cultural dimension in story visualization. To address this gap, we conducted a multicultural analysis of story visualization in multilingual contexts using multilingual text-to-image models.

2.2 Multicultural Analysis of VLMs and Text-to-Image Models

Recent research has increasingly examined cultural aspects across large language models (LLMs), vision-language models (VLMs), and text-to-image

(T2I) systems. Adilazuarda et al. (2024) highlights gaps in cultural modelling within LLMs, while Liu et al. (2023) and Ventura et al. (2024) demonstrate cultural inconsistencies in T2I generation and propose methods to adjust outputs according to cultural perspectives. Khanuja et al. (2024) propose an approach to directly align generated images with target cultural contexts. In the visual question answering domain, the CVQA and CULTUREVQA datasets show that VLMs perform poorly on culture-centric questions (Romero et al., 2024; Nayak et al., 2024), and other studies further assess cultural sensitivity in VLMs (Bhatia et al., 2024; Yadav et al., 2025; Liu et al., 2025). Bayramli et al. (2025) report that generated images can reinforce stereotypes or misrepresent cultural elements, while Bhalerao et al. (2025) proposes a multi-agent framework to produce culturally relevant imagery. In contrast to these works, our work focuses on a detailed multicultural analysis of story visualization using text-to-image models, aiming to uncover how cultural nuances specifically influence narrative visual generation across languages.

2.3 MLLM-as-Judge Models

Recent work in automatic evaluation has adopted the LLM-as-Judge paradigm, using LLMs to assess generated outputs in domain-specific tasks (Zhu et al., 2023; Kocmi and Federmann, 2023; Chiang and Lee, 2023) while (Verga et al., 2024) proposed LLM-as-Jury by aggregating judge scores from distinct families of judge models to reduce bias of single judge LLM. (Chen et al., 2024) extended the LLM-as-Judge approaches to the multimodal case, applying pairwise and batch-wise ranking across diverse datasets. However, single MLLMs-as-judge still exhibit biases in their scores. Furthermore, (Lee et al., 2024) observed that MLLMs acting as judges occasionally assign disproportionately high scores, exacerbating the bias. To overcome these issues, we extend LLM-as-Judge to LLM-as-Jury in our evaluation of multicultural story visualization. we adopt an MLLM-as-Jury framework that aggregates judgments from three different families of MLLM judges, reducing individual bias and enhancing robustness of the evaluation process.

3 Methodology

3.1 Story Translation

To support multilingual story visualization, each story narrative $s_i^{(\text{English})}$, originally written in En-

glish, is translated to Hindi and Chinese as shown in stage 1 of Figure 2, using the NLLB-200 (Team et al., 2022) translation model. This results in a multilingual set of narratives $s_i^{(\ell)}$ for each language $\ell \in \mathcal{L} = \{\text{English, Hindi, Chinese}\}$. These translated narratives serve as inputs for the subsequent multilingual T2I generation. A detailed analysis of the translation quality, including reference-free COMET-QE (Rei et al., 2020) scores for both datasets, is provided in Appendix A.1.

3.2 Story Visualization using Multilingual T2I Models

As shown in stage 2 of Figure 2, Let M denote a multilingual T2I model. We generate the image corresponding to each narrative $s_i^{(\ell)}$ in language ℓ independently¹ using the function $I_i^{(\ell)} = \mathcal{F}(M, s_i^{(\ell)}, T, \gamma)$ for all $\ell \in \mathcal{L}$, where T is the number of denoising steps and γ is the guidance scale.

3.3 Proposed Story Visualization Evaluation Metrics

To evaluate the quality of the generated text-to-image sequences, which are not captured by existing metrics FID and CLIP, we propose two types of evaluation metrics: (1) **Sequential Scene Evaluation** and (2) **Individual Scene Evaluation**.

- **Sequential Scene Evaluation:**

- **Cultural Appropriateness:** Measures how accurately the generated visuals represent the cultural styles, symbols and contexts described in the scene captions over the full story.
- **Cohesion:** Measures the coherence of visual elements across all scenes in a story, ensuring smooth transitions and logical relationships between scenes.
- **Visual Aesthetics:** Evaluates the overall appeal of the visuals in terms of clarity, color, and sharpness over the full story.

- **Individual Scene Evaluation:**

- **Object Presence:** Evaluates the inclusion of all essential objects mentioned in the captions for each scene.

¹We generated each scene independently due to unavailability of multilingual story visualization models and our primary objective is to analyze culture rather than cross-scene consistency.

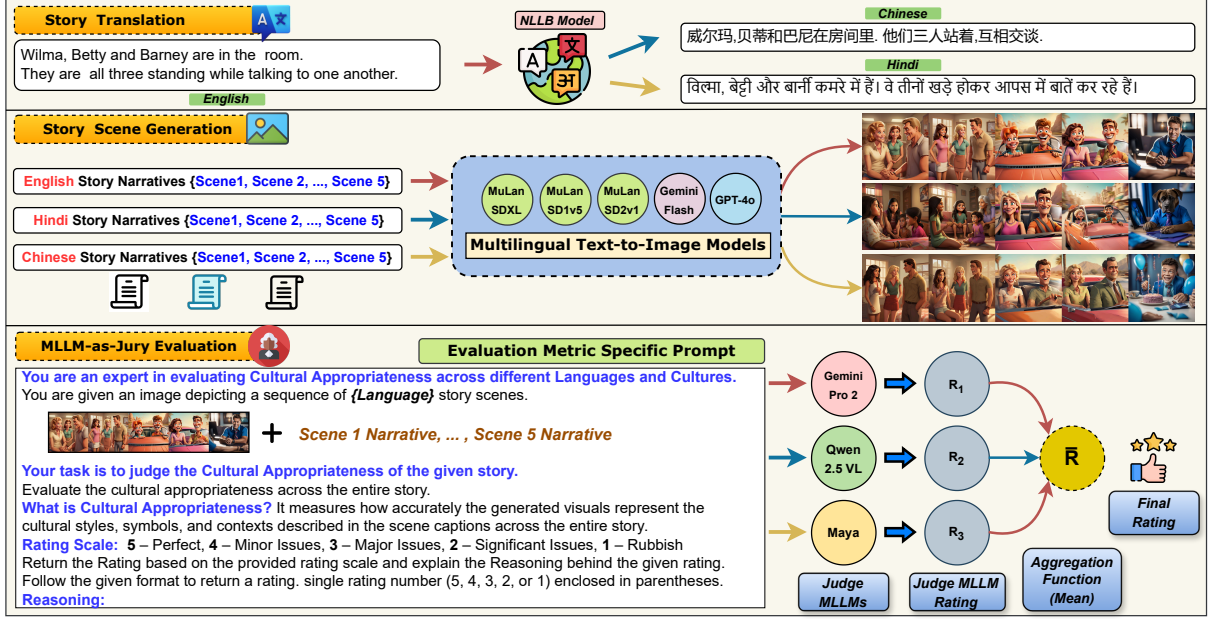


Figure 2: Three-stage framework for multicultural analysis of story visualization using text-to-image models.

- **Semantic Consistency:** Assesses how well the visual content aligns with the semantic meaning and details provided in the text descriptions for each scene.

Sequential evaluation metrics assess the entire story by considering all image-narrative pairs and assigning a single score (1–5) for overall quality. In contrast, individual scene metrics evaluate each pair independently, yielding five separate scores per story. We then take an average of five scores as the final score. **Rating Scheme:** 5 – Perfect, 4 – Minor Issues, 3 – Major Issues, 2 – Significant Issues, 1 – Rubbish.

3.4 Progressive Multiculture Evaluation Framework

To perform a detailed evaluation of cultural aspects in generated story images, this framework incrementally enriches the cultural evaluation criteria across three levels:

- **Version 1 (V1):** Culture Definition ($\mathcal{C}$) - evaluates culture appropriateness based on the just definition.
- **Version 2 (V2):** Culture Definition ($\mathcal{C}$) + Focus Points ($\mathcal{F}$) - introduces specific cultural dimensions to pay attention to during the evaluation.
Focus Points ($\mathcal{F}$): Background Objects, Facial Features, Infrastructures, Apparel

- **Version 3 (V3):** Culture Definition ($\mathcal{C}$) + Focus Points ($\mathcal{F}$) + Illustrative Examples ($\mathcal{E}$) - provides concrete examples for each focus point to guide consistent and precise assessment.

Illustrative Examples ($\mathcal{E}$)

1. Background Objects: Assess whether the depicted objects represent the target culture setting described in the scenes, focusing on nearby objects, furniture, decorations and other contextual details.

2. Character’s Facial Features: Evaluate whether the facial structures align with the diverse traits commonly found in the target culture. Avoid assumptions about stereotypical features.

3. Infrastructures: Consider whether the settings, such as architectural elements, are appropriate for the target culture.

4. Apparel: Assess whether the clothing aligns with traditional or contemporary styles representative of the Appropriate culture.

Formally, let $S = \{s_1, s_2, \dots, s_n\}$ denote the set of n generated story samples. Each story s_i is evaluated using a progressively detailed evaluation mapping $\mathbb{E}_v(s_i)$, where $v \in \{1, 2, 3\}$ represents the evaluation version:

$$\mathbb{E}_1(s_i) = g(s_i | \mathcal{C})$$

$$\mathbb{E}_2(s_i) = g(s_i | \mathcal{C}, \mathcal{F})$$

$$\mathbb{E}_3(s_i) = g(s_i | \mathcal{C}, \mathcal{F}, \mathcal{E})$$

The goal of this progressive structure is to evaluate whether richer prompts lead to more accurate and rigorous culture assessments of the generated story content.

3.5 MLLM-as-Jury Evaluation Framework

To enable large-scale, automatic assessment of story visualization metrics, we propose a *MLLM-as-Jury* evaluation framework that leverages Expert Role-specific prompting² shown in stage 3 of Figure 2. Each story is represented as a sequence of narrative and scene image pairs, $S = \{(n_1, i_1), (n_2, i_2), \dots, (n_T, i_T)\}$, where n_t is the narrative and i_t is the generated image from multilingual T2I models for the t^{th} scene. An Expert Role-specific prompt $\mathcal{P}$ fed to three different MLLM judge models:

$$\mathcal{M}_1 = J_{\text{Gemini}}, \quad \mathcal{M}_2 = J_{\text{Qwen}}, \quad \mathcal{M}_3 = J_{\text{Maya}}.$$

Each judge model produces a rating for the particular evaluation metric, denoted $r_1 = \mathcal{M}_1(P)$, $r_2 = \mathcal{M}_2(P)$ and $r_3 = \mathcal{M}_3(P)$. The final rating is computed by a mean aggregation function over the three judges' ratings given by $R_{\text{pred}} = \frac{1}{3}(r_1 + r_2 + r_3)$.

We used the mean as an aggregation function rather than the median because the median removes extreme scores, which can unintentionally overweight a single model's behavior. We focus here on integrating different judge models' complementary perspectives to mitigate overall bias in assessment.

The three models were selected for their complementary multimodal and multilingual strengths. **Gemini-Pro 2.0** is strong in multimodal multilingual reasoning, while **Qwen2.5-VL** (Bai et al., 2025) excels at image understanding and multilingual instruction-following, and **Maya** (Alam et al., 2024) performs well in multilingual and culturally sensitive tasks.

3.6 Human Evaluation Process

To assess the quality and cultural relevance of generated story images, we conducted a human evaluation with native English, Chinese, and Hindi speakers. Three fluent native speakers per language (a total of nine) volunteered from the research community. Their linguistic and cultural backgrounds ensured reliable evaluation. Five male and four female evaluators participated, offering diverse

perspectives. Each received 25 randomly selected story samples, consisting of five sequences of (image, caption) pairs spanning different models and datasets. Evaluation followed the Sequential and Individual Scene criteria. In addition to ratings, evaluators provided qualitative feedback, supporting a nuanced analysis of visual and cultural effectiveness in multilingual story generation. Further details on agreement scores and evaluation guidelines are provided in Appendix A.2.

4 Experiments

4.1 Choices of Language and Datasets

To analyze the cultural aspects of story visualization, we selected three languages: English (original language of the datasets used), Hindi and Chinese. These choices were guided by several factors, like representation of distinct geographical and cultural regions, the availability of high-quality translation models, compatibility of MLLM used in the MLLM-as-Jury framework and accessibility of annotators for human evaluation.

We analyze story visualization across cultural contexts with two datasets: VIST (real-world) and FlintstonesSV (animated). We performed experiments on randomly sampled 500 stories from each dataset, where each story has 5 (image, caption) pairs. VIST (Huang et al., 2016) features photo sequences from Flickr albums of everyday events, consisting of 50,000 stories and FlintstonesSV (Gupta et al., 2018), based on the classic animated series "*The Flintstones*", includes over 24,000 image-caption pairs focused on seven main characters in varied scenes.

4.2 Multilingual Text-to-Image Models

To generate story images from scene narratives in different languages, we use the multilingual T2I diffusion model MuLan³ (Xing et al., 2024). We use three MuLan variants: **1)** SD1v5, **2)** SD2v1 and **3)** SDXL, representing progressively diverse open-source stable diffusion versions along with two closed-source models, **4)** GPT-4o and **5)** Gemini-Flash 2.5, known for their strong image generation capability across languages.

5 Results

MLLM-as-Jury results on Test set: As shown in Table 1, SDXL consistently outperforms SD2v1

²MLLM-as-Jury evaluation prompts are given in the Appendix.

³Model Configuration for all experiments is provided in the Appendix

Models	Cultural Appr. (V3) ↑			Visual Aesthetics ↑			Cohesion ↑			Semantic Consistency ↑			Object Presence ↑		
	English	Hindi	Chinese	English	Hindi	Chinese	English	Hindi	Chinese	English	Hindi	Chinese	English	Hindi	Chinese
FlintstonesSV															
MuLan SD2v1	3.84	3.18	3.09	3.79	3.73	3.76	3.31	3.11	3.33	4.02	3.46	3.89	4.46	4.05	4.47
MuLan SD1v5	3.86	3.21	3.16	3.86	3.76	3.84	3.35	3.19	3.35	4.09	3.58	3.83	4.46	4.15	4.44
MuLan SDXL	3.87	3.40	3.18	4.05	3.97	4.01	3.35	3.19	3.43	4.15	3.72	4.05	4.50	4.20	4.52
VIST															
MuLan SD2v1	3.94	3.45	3.49	3.96	3.84	3.94	3.70	3.39	3.46	4.29	3.83	3.76	4.57	4.46	4.45
MuLan SD1v5	4.03	3.82	3.90	4.02	3.89	4.06	3.81	3.51	3.59	4.37	4.02	3.96	4.59	4.45	4.49
MuLan SDXL	4.12	3.89	3.95	4.16	4.03	4.13	3.90	3.58	3.69	4.46	4.18	4.16	4.65	4.56	4.59

Table 1: Results of MLLM-as-Jury across three different models on 500 test samples on two datasets across multiple evaluation metrics and languages with score range (1-5).

and SD1v5 across all metrics, languages, and datasets. Scores are generally higher on VIST than on FlintstonesSV, for all three languages, indicating that models handle real-world content better than animated scenes. English receives the highest scores overall, while Chinese and Hindi lag behind, with Hindi performing the lowest in Cohesion and Cultural Appropriateness.

Table 2 shows that, when averaged across all metrics, English again achieves the highest scores across datasets and models. Gemini leads in human evaluations, followed by GPT-4o, while MuLan models perform poorly in Hindi. VIST continues to outperform FlintstonesSV for English and Hindi, whereas Chinese results are less consistent. In the MLLM-as-Jury evaluation, MuLan SDXL outperforms the other two open-source models. While English shows higher results than Hindi and Chinese, the VIST dataset yields higher scores than FlintstonesSV across all three languages. The MLLM-as-Jury ablation study results are provided in Appendix Section A.3.

Models	FlintstonesSV			VIST		
	English	Hindi	Chinese	English	Hindi	Chinese
Human Evaluation ↑						
MuLan SD2v1	2.88	2.35	2.72	3.28	2.58	2.82
MuLan SD1v5	2.85	2.50	2.73	3.12	2.76	2.51
MuLan SDXL	3.01	2.22	3.03	3.16	2.70	2.95
GPT-4o	3.74	2.89	3.35	4.00	3.99	3.88
Gemini Pro 2.5	4.33	3.53	3.64	4.19	4.18	4.22
MLLM-as-Jury Evaluation ↑						
MuLan SD2v1	3.97	3.52	3.72	4.17	3.85	3.87
MuLan SD1v5	4.01	3.58	3.73	4.23	3.92	4.02
MuLan SDXL	4.08	3.69	3.84	4.32	4.06	4.12

Table 2: Average scores across metrics from Human Evaluation (top) and MLLM-as-Jury Evaluation (bottom), spanning models, languages and datasets with score range (1-5).

Human Evaluation Results: Figure 3 illustrates that English scores are high in FlintstonesSV, especially for GPT-4o and Gemini. Open-source models perform poorly, except SDXL in Cultural Appropriateness. Hindi scores drop, with only Gemini

scoring well in Object Presence and Semantic Consistency. Chinese results vary: Gemini scores low in Cohesion, where GPT-4o leads. In VIST, similar trends hold: English scores highest, Hindi improves slightly and Chinese results are better. GPT-4o and Gemini lead, while SD1v5 and SDXL show moderate performance. Closed-source models perform best across five metrics, while VIST results outperform FlintstonesSV.

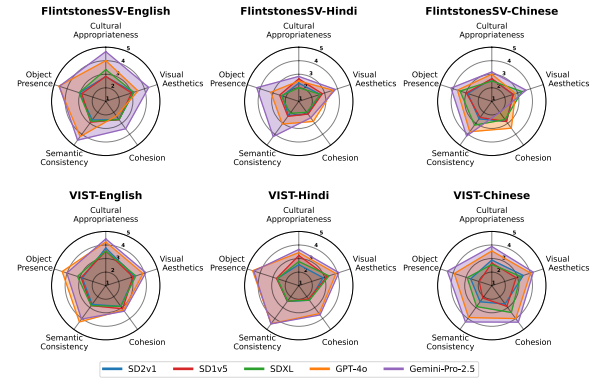


Figure 3: Human evaluation results across datasets, languages and evaluation metrics.

CLIP/FID Metrics Results: As shown in Table 5 (see Appendix), CLIP and FID scores are reported for story images generated by three open-source multilingual text-to-image models. FlintstonesSV outputs achieve higher CLIP scores than VIST, indicating better alignment with scene descriptions due to its consistent generation of animated visual structure, while VIST’s open-ended real-world scenes result in more diverse generations. In contrast, FlintstonesSV shows higher FID distances from ground-truth animated scenes, likely due to stylistic and texture variations, whereas VIST images are closer in distribution to real-world references. Overall, English performs best across both metrics, followed by Hindi and Chinese for CLIP, and Chinese and Hindi for FID. The conflicting trends highlight that CLIP and FID capture differ-

ent aspects of quality, making it difficult to rely solely on either metric for final evaluation.

Progressive Multiculture Evaluation Results: Figure 4 shows cultural appropriateness scores across three levels. For English, V1 scores highest over all models while decreasing its performance with V2 and V3, showing stricter judgments in evaluation with added focus points and illustrative examples. For Chinese and Hindi, scores remain stable between V1 to V3, indicating that the added structure has minimal impact because of the underlying Judge models’ limited capability to evaluate cultural aspects in these languages by enriching prompts. Model SDXL outperforms SD2v1 and SD1v5 on both datasets, while overall scores high on VIST compared to FlintstonesSV, showing generated story adapts culture better in real-world dataset.

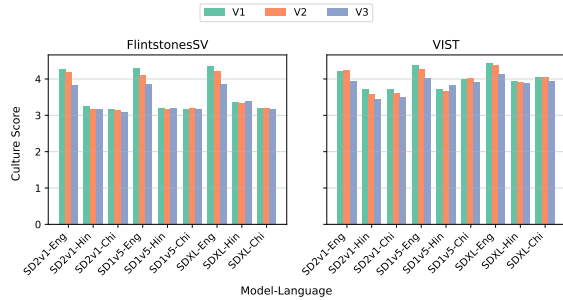


Figure 4: Progressive Culture evaluation scores by MLLM-as-Jury across three levels on FlintstonesSV and VIST datasets.

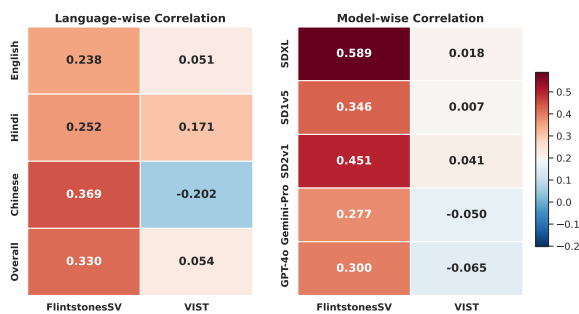


Figure 5: Correlation between Human and MLLM-as-Jury evaluation on FlintstonesSV and VIST datasets

Correlation between Human and MLLM-as-Jury Evaluation: Figure 5 shows Spearman Rank Correlation between Human and MLLM-as-Jury evaluation. In FlintstonesSV, the MLLM-as-Jury evaluation aligns moderately with human evaluation across all languages and models. This is because FlintstonesSV features cartoon-based, fixed

characters and repetitive story settings (e.g., cave, car, room, backyard). In contrast, VIST exhibits a lower correlation across languages and models, as it contains open-ended stories where scenes vary widely, and each story introduces new people, places, activities, ceremonies, and relationships. As shown in Figure 14 (see Appendix), the rating distribution of MLLM-as-Jury on VIST tends to produce higher scores compared to FlintstonesSV, where ratings are more dispersed and better aligned with human evaluations. The lower variance in MLLM-as-Jury ratings on VIST leads to a reduced Spearman correlation with human assessments. Overall, this analysis indicates that while MLLM-as-Jury can approximate human judgment effectively in simple, structured tasks, it tends to overestimate scores in complex or open-ended scenarios, making it a dataset-dependent automatic evaluation method.

6 Discussion

In this section, we address the two research questions focusing visualisation of the same story in different languages and cultural adaptation.



Figure 6: Visualization of the same story in different languages generated by MuLan-SDXL model.

RQ1: How does the visualization of the same story vary across different languages? As we can see from Figure 6, the generated image sequences across different languages reveal various cultural elements including hairstyles, facial features and apparel. In scene 1, for a family get-together, in Chinese, the food appears as traditional Chinese cuisine, whereas in the English setting, a cake-cutting and in Hindi, it is a casual get-together. These examples show that the same model generates culturally different visuals across different languages. Additionally, the concept of “success” is represented differently across languages: in Hindi, it is shown as an individual achievement, in Chinese as a joint success with a partner, while in English, a collective family success.

Figure 13 (see Appendix) compares visualiza-

tions of the same story generated by different models in the same language. For the Chinese story on the left (a visit to a craft fair from VIST), open models such as SD1v5, SDXL, and SD2v1 retain strong cultural cues such as traditional clothing, facial features, food items, shop signs, and calligraphy. GPT-4o preserves some Chinese elements only in Scenes 3 and 5, and Gemini-Pro does so mainly in Scenes 2 and 5. The remaining scenes resemble Western settings. For the Hindi story on the right (Wilma and Fred talking indoors from FlintstoneSV), only SDXL introduces partial Indian cues, though even there the characters and environments remain largely Westernized. Across the other models, traditional Indian attire, features, and household settings are mostly absent, showing a clear bias toward Western visual culture.

Table 1 and Figure 3 show that models generate more culturally appropriate visual stories for real-world narratives in VIST than for animated stories in FlintstonesSV. VIST’s open-ended structure gives models flexibility to adapt cultural cues across languages, whereas FlintstonesSV is anchored to Western character identities (Fred, Wilma, Betty, Barney). As also seen in Figure 13 for Hindi outputs, these fixed references restrict the models’ ability to produce culturally aligned visuals, making animated datasets inherently less adaptable than real-world ones.

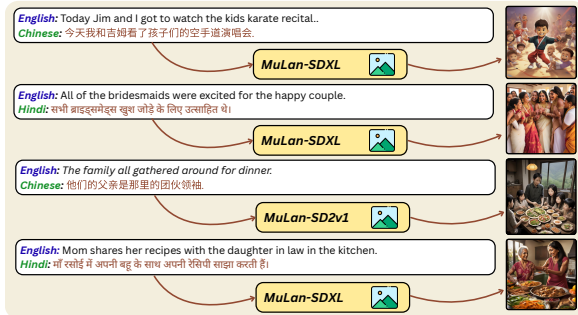


Figure 7: Model outputs demonstrating cultural adaptation

RQ2: How does story visualization adapt to different cultures? From our multicultural analysis, we found that current models generate Western cultural elements most accurately, while their adaptation of Chinese and Hindi cultures is moderate and low, respectively. For example, in Figure 7, **Ex1:** the story scene in Chinese about karate adapts to Chinese karate clothes. **Ex2:** The story scene in Hindi about a wedding, where the scene adapts to an Indian wedding with culturally relevant ‘saari’

dressess for the bride and bridesmaids. **Ex3:** The story scene in Chinese about a family dinner, where the scene adapts to Chinese food with a traditional light bulb. **Ex4:** The story scene in Hindi about a mother sharing a recipe with her daughter-in-law in the kitchen. The scene adapts to Indian food, steel utensils, and Indian clothes. These examples show a minimal level of cultural adaptation by default in multilingual T2I models. This analysis further underscores the need to develop more culturally grounded story visualization models.

7 Culture Error Analysis

We identified several errors during the evaluation of cultural aspects in the generated stories. Figure 11 (see Appendix) shows a Hindi story depicting a family vacation with friends. In Scene 1, the absence of footwear, while Scene 4 exhibits Western attire, shows a misrepresentation of Indian culture. Similarly, Figure 12 (see Appendix) presents an example from a Chinese story about families enjoying a game together. In Scene 2, the flag closely resembles the American flag. In Scene 4, the male character’s exaggerated height over the woman and the girl reflects a stereotypical “dominant male” portrayal, which is inconsistent with typical Chinese family depictions. Scene 5 further displays characters with narrowed eyes and nearly identical facial expressions, signalling culturally inappropriate and stereotyped visuals.

8 Conclusion

We studied the multicultural analysis of story visualization on real-world and animated datasets. Our study reveals that current multilingual T2I models exhibit a strong bias towards Western culture. Human and MLLM-as-Jury evaluation show that models generate high culturally appropriate stories in English, moderate in Chinese and low for Hindi. More specifically, on a real-world story dataset, models generate better culturally relevant stories compared to an animated dataset. Our Progressive Cultural Evaluation framework shows that providing cultural focus points and illustrative examples helps to evaluate detailed cultural aspects in English. However, for Chinese and Hindi, not much benefit is observed, reflecting the underlying Judge models’ limited ability to accurately identify cultural aspects in different languages. MLLM-as-Jury evaluation results approximate human judgment with higher correlation on the FlintstonesSV

dataset than on the open-ended story dataset VIST. These findings highlight the need to develop culturally grounded story visualization approaches that faithfully represent diverse cultures and languages.

Limitations

This work has several limitations. **1)** For multilingual story visualization, we generate each scene independently using multilingual T2I models, as our primary focus is on the cultural aspects of each scene rather than the narrative coherence across scenes. **2)** The scope of multicultural analysis in this study is limited to three languages: English, Chinese and Hindi. We plan to expand this analysis in future work by incorporating a more diverse set of languages and cultures. **3)** While this work focuses on analyzing cultural aspects, our future efforts will address these discrepancies by developing culturally grounded methods that adapt generated stories to the target culture.

Acknowledgments

This research work is funded by Research Ireland under Grant Number SFI/12/RC/2289_P2 (Insight), co-funded by the European Regional Development Fund.

References

- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and modeling "culture" in llms: A survey](#). *Preprint*, arXiv:2403.15412.
- Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, and 1 others. 2024. Maya: An instruction finetuned multilingual multimodal model. *arXiv preprint arXiv:2412.07112*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, and 1 others. 2024. Imagen 3. *arXiv preprint arXiv:2408.07009*.
- Zahra Bayramli, Ayhan Suleymanzade, Na Min An, Huzama Ahmad, Eunsu Kim, Junyeong Park, James Thorne, and Alice Oh. 2025. [Diffusion models through a global lens: Are they culturally inclusive?](#) *Preprint*, arXiv:2502.08914.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8.
- Parth Bhalerao, Mounika Yalamarty, Brian Trinh, and Oana Ignat. 2025. [Multi-agent multimodal models for multicultural text to image generation](#). *Preprint*, arXiv:2502.15972.
- Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. [From local concepts to universals: Evaluating the multicultural understanding of vision-language models](#). *Preprint*, arXiv:2407.00263.
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, and 1 others. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.
- Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. [Video generation models as world simulators](#).
- Jerome Bruner. 2010. Narrative, culture, and mind. *Telling stories: Language, narrative, and social life*, 46:49.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, pages 6562–6595.
- Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.
- Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. 2018. [Imagine this! scripts to compositions to videos](#). *Preprint*, arXiv:1804.03608.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet

- Kohli, Dhruv Batra, and 1 others. 2016. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*.
- Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. [An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. [Flux.1 kontext: Flow matching for in-context image generation and editing in latent space](#). *Preprint*, arXiv:2506.15742.
- Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11286–11315.
- Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2023. [On the cultural gap in text-to-image generation](#). *Preprint*, arXiv:2307.02971.
- Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. [CultureVLM: Characterizing and improving cultural understanding of vision-language models for over 100 countries](#). *Preprint*, arXiv:2501.01282.
- Jiawei Mao, Xiaoke Huang, Yunfei Xie, Yuanqi Chang, Mude Hui, Bingjie Xu, and Yuyin Zhou. 2024. Story-adapter: A training-free iterative framework for long story visualization. *CoRR*.
- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Steenkiste, Lisa Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790.
- Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. 2024. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2920–2930.
- Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2493–2502.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. [Cvqa: Culturally-diverse multilingual visual question answering benchmark](#). *Preprint*, arXiv:2406.05967.
- Abhishek Sharma, Adams Yu, Ali Razavi, Andeep Toor, Andrew Pierson, Ankush Gupta, Austin Waters, Aäron van den Oord, Daniel Tanis, Dumitru Erhan, Eric Lau, Eleni Shaw, Gabe Barth-Maron, Greg Shaw, Han Zhang, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, and 38 others. 2024. [Veo](#).
- Xiaoqian Shen and Mohamed Elhoseiny. 2023. Large language models as consistent story visualizers. *arXiv preprint arXiv:2312.02252*.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2024. [Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models](#). *Preprint*, arXiv:2310.01929.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Sen Xing, Muyan Zhong, Zeqiang Lai, Liangchen Li, Jiawen Liu, Yaohui Wang, Jifeng Dai, and Wenhai Wang. 2024. Mulan: Adapting multilingual diffusion models for hundreds of languages with negligible cost. *arXiv preprint arXiv:2412.01271*.

Srishti Yadav, Zhi Zhang, Daniel Hershcovich, and Ekaterina Shutova. 2025. *Beyond words: Exploring cultural value sensitivity in multimodal models*. pages 7592–7608.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847.

Sixiao Zheng and Yanwei Fu. 2024. Temporalstory: Enhancing consistency in story visualization using spatial-temporal attention. *arXiv e-prints*, pages arXiv–2407.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. 2024. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Appendix

This appendix details the translation details, human evaluation process with examples, culture error analysis, MLLM-as-Jury ablation study, Human and MLLM-as-Jury correlation details, experimental setup details, and prompts utilized within the MLLM-as-Jury evaluation framework.

A.1 Reference Free Translation

This section presents reference-free translation quality scores computed using COMET-QE⁴, a quality estimation model for machine translation. We evaluate English–Chinese and English–Hindi translations generated by the NLLB-200 model

on both the VIST and FlintstonesSV datasets. COMET-QE allows us to estimate translation quality without requiring gold-standard reference translations. The scores reported here provide a reliable, model-based approximation of fluency and adequacy in the translated outputs, supporting systematic comparison across languages and datasets. As shown in Table 3, VIST consistently achieves higher COMET-QE scores than FlintstonesSV for both translation directions, with the average scores indicating slightly stronger performance for English–Chinese compared to English–Hindi.

Dataset	Eng -> Hindi (↑)	Eng -> Chinese (↑)
FlintstonesSV	0.4537	0.3902
VIST	0.8224	0.7432
Average	0.5667	0.6380

Table 3: Reference-free COMET-QE translation quality scores for the NLLB-200 model on the VIST and FlintstonesSV datasets for English–Hindi and English–Chinese translation.

A.2 Human Evaluation

We conducted a human evaluation on random samples. 25 test samples spanned over 5 models and 2 datasets with three evaluators per language: English, Hindi and Chinese. The evaluation focused on five key metrics: Cultural Appropriateness, Visual Aesthetics, Cohesion, Semantic Consistency and Object Presence. Given the subjective nature of these metrics, inter-rater agreement was measured by weighted Cohen’s κ scores shown in Table 4 were generally low to moderate across languages, which is typical for nuanced human judgment tasks.

Sequential metrics such as Cultural Appropriateness, Visual Aesthetics and Cohesion were rated with a single score over the entire story, which contributed to lower agreement among annotators due to their subjective nature. In contrast, individual scene metrics like Semantic Consistency and Object Presence were evaluated separately for each scene, leading to higher inter-annotator agreement as this more granular approach reduces ambiguity and improves consistency. Figures 8, 9, and 10 show examples of human evaluation results for the same story across three sequential and two individual scene evaluation metrics in English, Hindi, and Chinese, respectively.

⁴<https://github.com/Unbabel/COMET>

Metric	English (κ)	Hindi (κ)	Chinese (κ)
Cultural Appropriateness	0.239	0.262	0.164
Visual Aesthetics	0.143	0.263	0.118
Cohesion	0.257	0.231	0.280
Semantic Consistency	0.551	0.610	0.422
Object Presence	0.525	0.605	0.340

Table 4: Weighted Cohen’s κ scores for each evaluation metric across English, Hindi and Chinese.

Our goal was to capture a broad spectrum of expert opinions rather than enforce consensus, so we do not emphasize inter-rater reliability but instead present aggregated expert ratings across all metrics (see Figure 3 in the main paper). All evaluators were volunteers with relevant academic backgrounds and received a detailed briefing on rating guidelines to promote a consistent understanding of the evaluation metrics across languages.

Models	CLIP $\uparrow$			FID $\downarrow$		
	English	Hindi	Chinese	English	Hindi	Chinese
FlintstonesSV						
MuLan SD2v1	29.18	25.73	27.54	216.23	230.63	236.97
MuLan SD1v5	28.71	26.32	27.18	251.07	252.19	259.31
MuLan SDXL	27.94	25.50	27.31	239.52	242.31	250.70
Average	28.61	25.85	27.34	235.61	242.31	250.70
VIST						
MuLan SD2v1	25.62	23.95	24.22	86.21	94.11	103.95
MuLan SD1v5	25.77	24.05	24.03	87.42	97.41	113.27
MuLan SDXL	25.82	23.88	24.51	89.23	93.61	96.82
Average	25.74	23.96	24.25	87.62	95.04	104.68

Table 5: Results of CLIP and FID score on story generated on VIST and FlintstonesSV datasets

English Story Narratives

- 1) When I arrived I saw a great booth with a variety of great crafts.
- 2) I was so excited to be heading to the crafts fair.
- 3) I stopped at chatted at my friend Beth's booth for a bit.
- 4) There were even booths set up for all of the kids.
- 5) I found some awesome crafts at the fair, I'm really happy that I went.



Metric	Evaluation Type	Score
Cultural Appropriateness	Sequential	4
Visual Aesthetics	Sequential	3
Cohesion	Sequential	3
Semantic Consistency	Individual Scene	44434
Object Presence	Individual Scene	43433

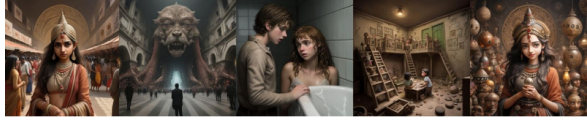
Figure 8: An example of human evaluation results for an English story across five evaluation metrics.

A.3 MLLM-as-Jury Ablation Study

This section compares the single-judge models with the jury method to examine how individual judges

Translated Hindi Story Narratives

- 1) मैं बहुत उत्साहित था शिल्प मेले में जा रहा है करने के लिए.
- 2) जब मैं वहां पहुंचा तो मैंने एक विशाल बुथ देखा जिसमें अनेक प्रकार के उत्कृष्ट शिल्प थे।
- 3) मैं अपने दोस्त बेथ के बुथ पर एक छोटे से बात करने के लिए बंद कर दिया.
- 4) यहाँ तक कि सभी बच्चों के लिए भी बुथ बनाए गए थे।
- 5) मुझे मेले में कुछ बेहतरीन शिल्प मिले, मैं बहुत खुश हूँ कि मैं गया।



Metric	Evaluation Type	Score
Cultural Appropriateness	Sequential	3
Visual Aesthetics	Sequential	3
Cohesion	Sequential	2
Semantic Consistency	Individual Scene	42114
Object Presence	Individual Scene	42223

Figure 9: An example of human evaluation results for a Hindi story across five evaluation metrics.

Translated Chinese Story Narratives

- 1) 我很高兴能参加手工艺博览会.
- 2) 我看到一个展位.
- 3) 我停下来聊聊我的朋友贝丝的摊位一会儿.
- 4) 孩子们甚至为所有人设置了摊位.
- 5) 我在展会上发现了一些很棒的手工艺品.



Metric	Evaluation Type	Score
Cultural Appropriateness	Sequential	3
Visual Aesthetics	Sequential	3
Cohesion	Sequential	2
Semantic Consistency	Individual Scene	33443
Object Presence	Individual Scene	34233

Figure 10: An example of human evaluation results for a Chinese story across five evaluation metrics.

contribute across languages, metrics, and datasets. For Semantic Consistency in Figure 17 and Object Presence in Figure 18 show that single-judge scores and jury scores are largely aligned across all languages. For Cohesion, Figure 19 shows that Gemini consistently assigns lower scores than Maya and Qwen across all languages, particularly on FlintstonesSV compared to VIST. The jury method mitigates this bias by averaging the judges’ outputs. For Visual Aesthetics, Gemini again scores moderately lower than Qwen and Maya, though the gap is smaller than in the Cohesion metric, as seen in Figure 20.

A different pattern emerges for Cultural Appropriateness (Figures 21, 22, and 23). Gemini assigns the lowest scores on Hindi and Chinese, indicating weak cultural fidelity in these languages in generated stories, and FlintstonesSV receives consistently lower scores than VIST, showing real-world story adapts to culture more than animated ones. As prompt complexity increases from V1 to V2 and

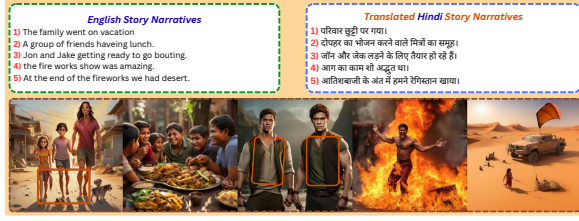


Figure 11: Hindi Error Analysis on the Visualization of the same story in different languages generated by MuLan-SDXL model depicted with orange boxes.



Figure 12: Chinese Error Analysis on the Visualization of the same story in different languages generated by MuLan-SDXL model depicted with orange boxes.

from V2 to V3, the scores of Gemini and Qwen decrease, indicating that the added cultural focus points and illustrative examples enforce a stricter evaluation of cultural aspects. In contrast, Maya’s scores remain largely stable. For English, Gemini’s scores are higher than for Hindi and Chinese. Overall, Maya assigns substantially higher scores than the other two judges at V2 and V3, especially on VIST. The jury method reduces this imbalance by averaging the three judges’ scores, though some outlier influence persists on VIST, as reflected in the rating distribution shown in Figure 14. Overall, across all metrics, the MLLM-as-Jury approach provides a more stable and reliable automatic evaluation for rubric-based story visualization by reducing individual model bias and leveraging complementary judging capabilities. Figure 15 shows a schematic representation of the MLLM-as-Jury evaluation on Cultural Appropriateness V3 in Chinese using the three MLLMs.

A.4 Prompts used in MLLM-as-Jury Evaluation

This section presents the prompts used in our MLLM-as-Jury evaluation framework. Figure 24 shows the prompt for assessing semantic consistency, which ensures logical alignment between visual and textual content. Figure 25 contains the prompt for evaluating object presence, verifying whether the described entities appear in the image. Visual aesthetics are evaluated using the prompt in

Figure 26, which focuses on image quality and composition. Cohesion is assessed through the prompt in Figure 27, which measures narrative flow and continuity.

Our Progressive Multicultural Evaluation framework enables culturally grounded assessment through three prompt versions. Version 1 (V1) (Figure 28) introduces a culture definition that establishes the criteria for cultural representation. Version 2 (V2) (Figure 29) extends this with focused cultural aspects. Version 3 (V3) (Figure 30) further refines the evaluation by adding illustrative examples, providing concrete references that help MLLMs produce more accurate and consistent cultural judgments.

A.5 MLLM-as-Jury Models Configuration

In this section, we describe the configuration of our MLLM-as-Jury evaluation framework, designed to enable large-scale, automated assessment of various story aspects. As introduced in the main paper, this framework leverages three Multimodal Large Language Models (MLLMs): Gemini Pro 2.0, Qwen2.5-VL and Maya. Table 6 presents the configuration details and parameters used for each model. Using diverse models in the evaluation process makes the framework more robust, as it combines different strengths and perspectives in assessing story quality.

Parameter	Gemini Pro 2.0	QwenVL 2.5	Maya
Access	Gemini API	GPU Inference	GPU Inference
Max Tokens	512	512	512
Sampling	False	False	False
Temperature	0	0	0
Prompt	Custom	Custom	Custom
Seed	42	42	42
Hardware	RTX A6000 48GB GPU	RTX A6000 48GB GPU	RTX A6000 48GB GPU

Table 6: Inference configuration for Gemini Pro 2.0, QwenVL 2.5 and Maya models.

A.6 Human vs MLLM-as-Jury Correlation

Figure 16 shows scatter plots comparing human ratings with MLLM-as-Jury ratings for the FlintstonesSV and VIST datasets across three languages. For FlintstonesSV, the spread of ratings lies moderately along the diagonal, indicating a moderate correlation with human evaluations, as also shown



Figure 13: Visualization of the same story generated by different models in the same language. **Left:** Chinese **Right:** Hindi

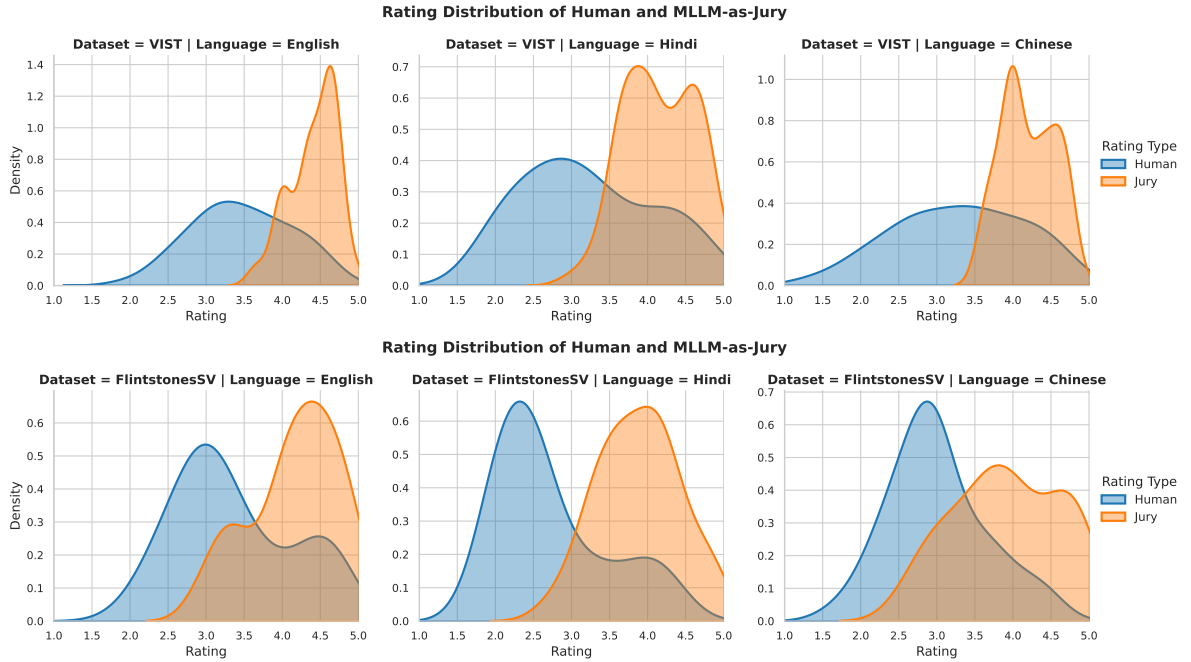


Figure 14: Rating distribution of Human and MLLM-as-Jury method across dataset and languages.

Parameter	SDXL	SD2v1	SD1v5
Multilingual Adapter	MuLan	MuLan	MuLan
Guidance Scale	7	7	7
Inference Steps	50	50	50
Framework Used	Diffusers	Diffusers	Diffusers
Seed	12345	12345	12345
Hardware	RTX A6000 48GB GPU	RTX A6000 48GB GPU	RTX A6000 48GB GPU

Table 7: Inference configuration for Text-to-Image models used: MuLan-SDXL, MuLan-SD2v1 and MuLan-SD1v5

in Figure 5. In contrast, for VIST, the rating distribution is more horizontally spread on the higher side of the MLLM-as-Jury scores, with no clear diagonal trend, resulting in very low correlation.

This suggests that MLLM-as-Jury is less critical for open-ended, contextually rich, and culturally nuanced scenes, often assigning higher scores. However, for simpler tasks with fixed settings, such as FlintstonesSV, it evaluates more critically and aligns more closely with human judgments. Overall, the results show that MLLM-as-Jury performance is dataset-dependent and influenced by the choice of underlying judge models.

A.7 Multilingual Text-to-Image Models Configuration

This section provides the configuration details for the multilingual text-to-image (T2I) models used to generate story from scene-level narratives in various languages. All models are based on the MuLan framework, which uses language adapters to extend English-centric diffusion models for multilingual

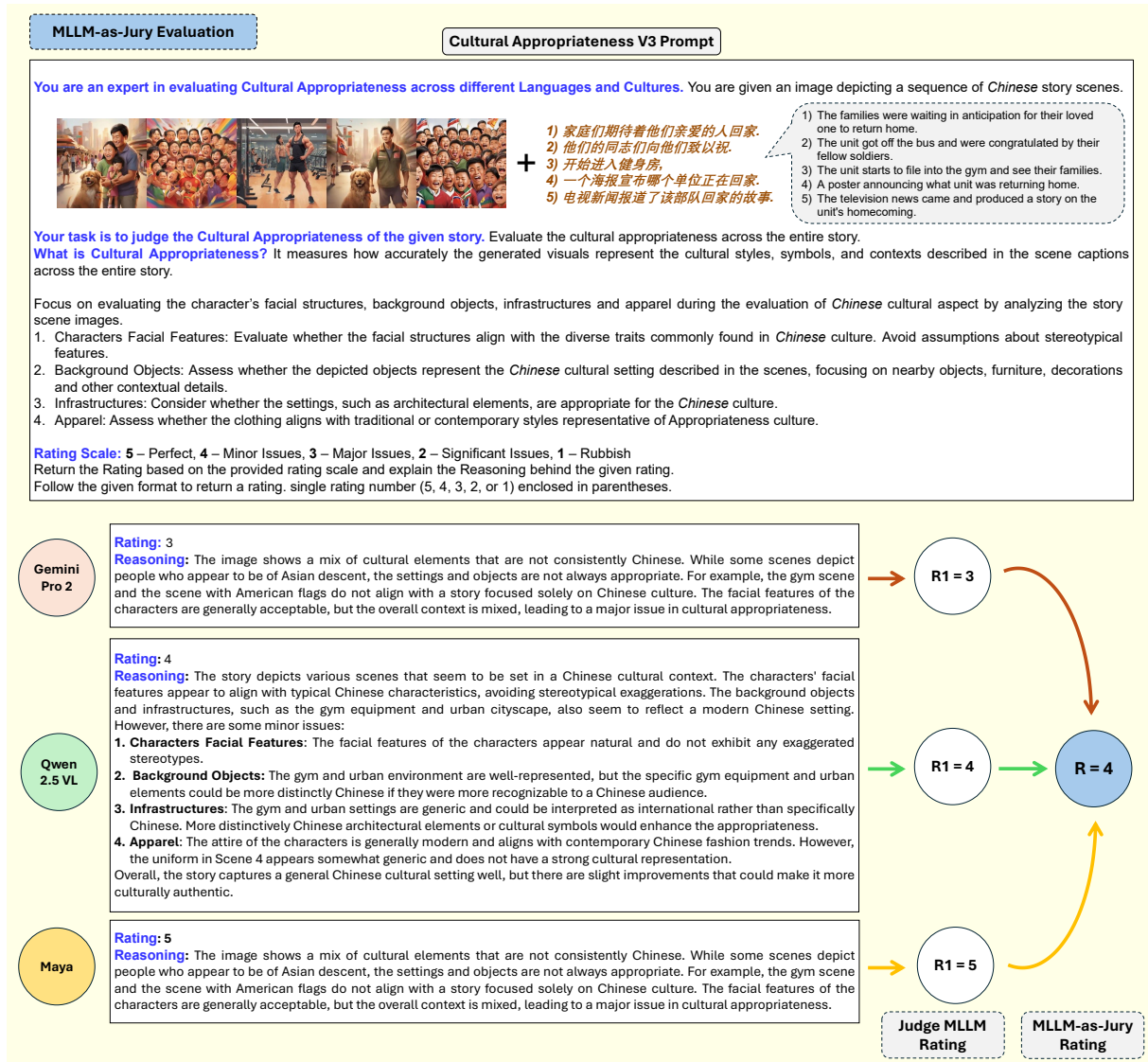


Figure 15: Schematic representation of the MLLM-as-Jury evaluation on Cultural Appropriateness V3 in Chinese using three MLLMs: Gemini Pro 2, Qwen 2.5 VL, and Maya.

prompts. Three variants of MuLan-enabled Stable Diffusion were used in our experiments: SD1v5, SD2v1 and SDXL. These models differ in architecture and image resolution, enabling a range of visual generation capabilities. Table 7 outlines the inference configurations for each model, including resolution, scheduler type and hardware settings. The remaining two models, GPT-4o and Gemini-Flash-2.0, were used with their default settings through the chat interface.

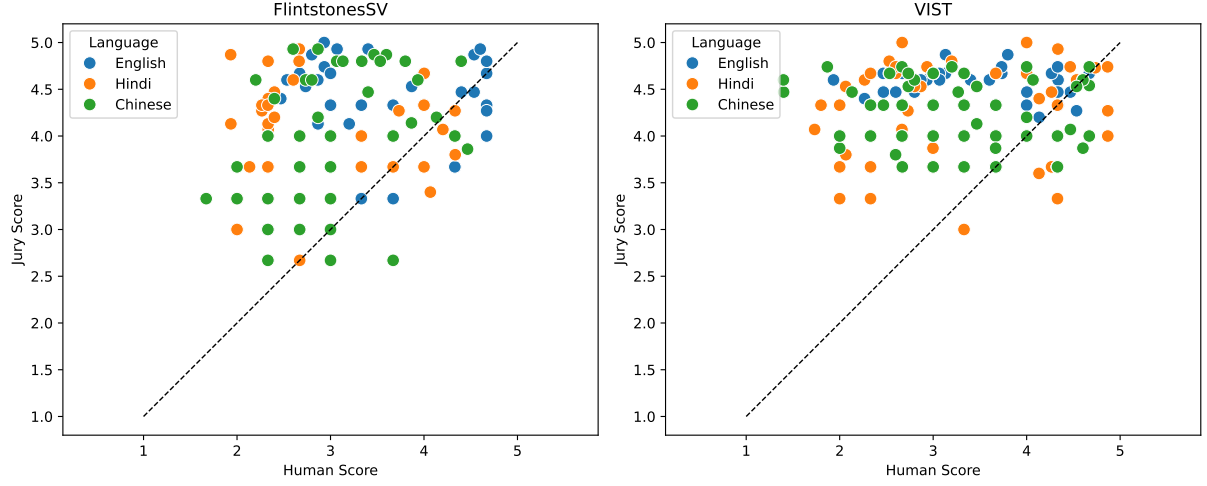


Figure 16: Scatter plot of ratings of Human and MLLM-as-Jury evaluation on FlintstonesSV and VIST datasets.

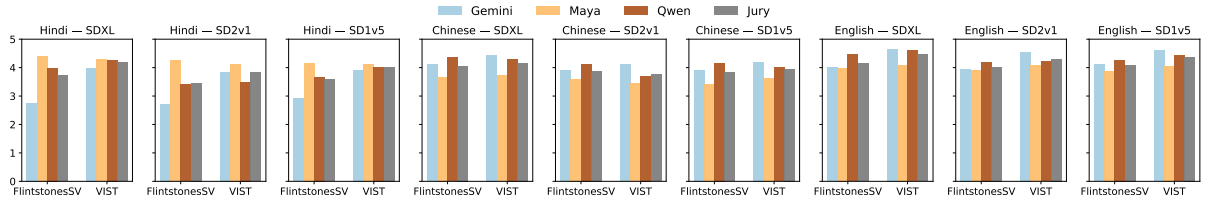


Figure 17: Semantic Consistency Jury vs Judge Results

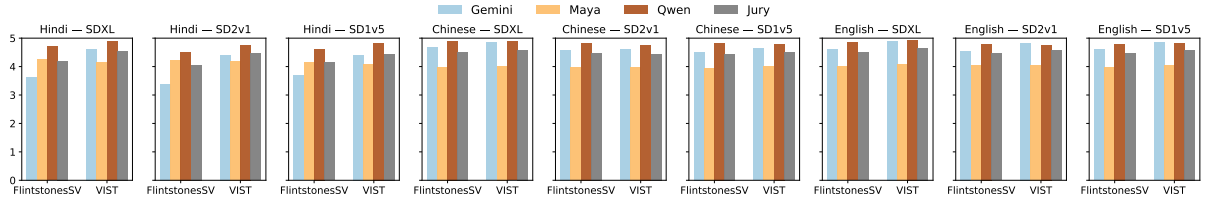


Figure 18: Object Presence Jury vs Judge Results

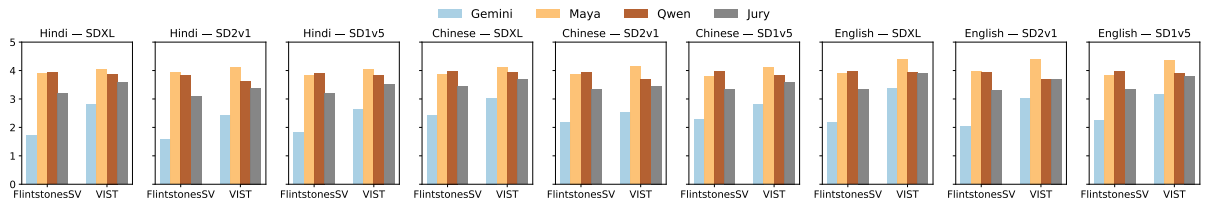


Figure 19: Cohesion Jury vs Judge Results

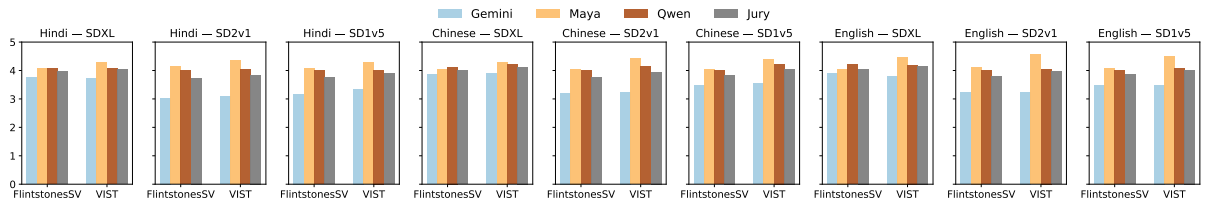


Figure 20: Visual Aesthetics Jury vs Judge Results

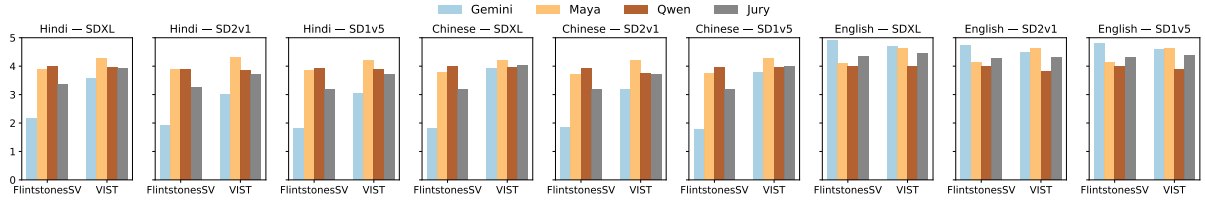


Figure 21: Cultural Appropriateness V1 - Jury vs Judge Results

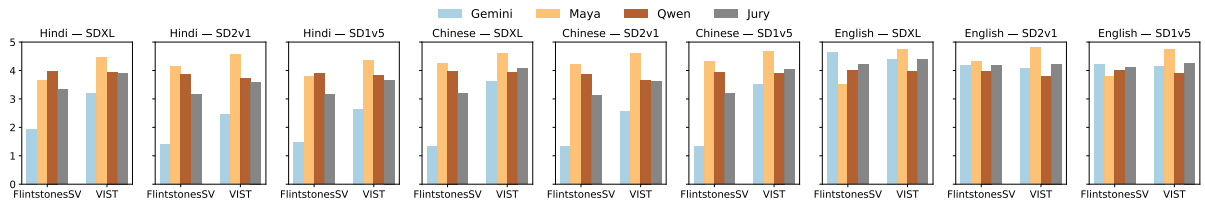


Figure 22: Cultural Appropriateness V2 - Jury vs Judge Results

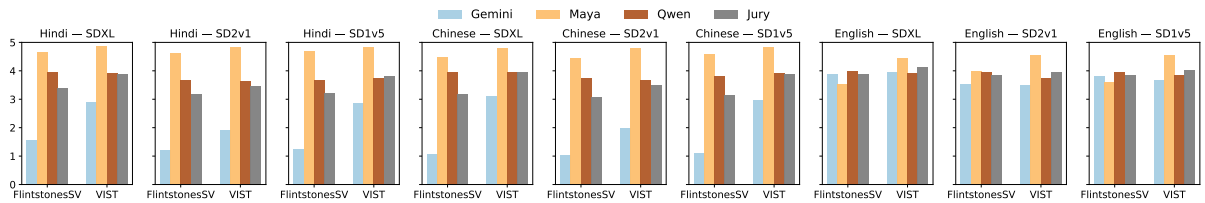


Figure 23: Cultural Appropriateness V3 - Jury vs Judge Results

Semantic Consistency Prompt

You are an expert in evaluating **Semantic Consistency in Visual Stories**.

You are given an image depicting a sequence of **{language}** story scenes.

{language} Story Scene Descriptions:

Scene 1: {scene_descriptions[0]}

Scene 2: {scene_descriptions[1]}

Scene 3: {scene_descriptions[2]}

Scene 4: {scene_descriptions[3]}

Scene 5: {scene_descriptions[4]}

Your task is to carefully check if **scene description is semantically consistent with scene image or not**.

Rating Scale:

5 - Perfect

4 - Minor Issues

3 - Major Issues

2 - Significant Issues

1 - Rubbish

Return the Rating based on the provided rating scale and explain the Reasoning behind the given rating.

If any scene description is not semantically meaningful to scene image, lower the rating accordingly. Do not give a perfect score unless its fully semantically consistent.

Strictly first return the rating in the format **(X, X, X, X, X)**, where each X is a single rating number **(5, 4, 3, 2, or 1)** corresponding to each story scene followed by reasoning.

Reasoning:

Figure 24: Prompt used for evaluating Semantic Consistency in MLLM-as-Jury Evaluation

Object Presence Prompt

You are an expert in evaluating **Object Presence in Visual Stories**.

You are given an image depicting a sequence of **{language}** story scenes.

{language} Story Scene Descriptions:

Scene 1: {scene_descriptions[0]}

Scene 2: {scene_descriptions[1]}

Scene 3: {scene_descriptions[2]}

Scene 4: {scene_descriptions[3]}

Scene 5: {scene_descriptions[4]}

Your task is to carefully check if every **object mentioned in the scene description is clearly present and visible** in the corresponding scene image.

Rating Scale:

5 - Perfect

4 - Minor Issues

3 - Major Issues

2 - Significant Issues

1 - Rubbish

Return the Rating based on the provided rating scale and explain the Reasoning behind the given rating.

If any object is missing or unclear, lower the rating accordingly. Do not give a perfect score unless all objects are fully and clearly present.

Strictly first return the rating in the format **(X, X, X, X, X)**, where each X is a single rating number **(5, 4, 3, 2, or 1)** corresponding to each story scene followed by reasoning.

Reasoning:

Figure 25: Prompt used for evaluating Object Presence in MLLM-as-Jury Evaluation

Visual Aesthetics Prompt

You are an expert in evaluating **Visual Aesthetics in Visual Stories**..

You are given an image depicting a sequence of **{language}** story scenes.

{language} Story Scene Descriptions:

Scene 1: {scene_descriptions[0]}

Scene 2: {scene_descriptions[1]}

Scene 3: {scene_descriptions[2]}

Scene 4: {scene_descriptions[3]}

Scene 5: {scene_descriptions[4]}

Your task is to judge the **Visual Aesthetics** of the given story. Evaluate the Visual Aesthetics across the entire story.

What is Visual Aesthetics? It measures the coherence and logical flow between scenes, ensuring smooth transitions and meaningful relationships in the visual narrative.

Rating Scale:

5 - Perfect

4 - Minor Issues

3 - Major Issues

2 - Significant Issues

1 - Rubbish

Return the Rating based on the provided rating scale and explain the Reasoning behind the given rating.

Follow the below given format to return Rating.

single rating number (5, 4, 3, 2, or 1) enclosed in parentheses.

Reasoning:

Figure 26: Prompt used for evaluating Visual Aesthetics in MLLM-as-Jury Evaluation

Cohesion Prompt

You are an expert in evaluating **Cohesion in Visual Stories**.

You are given an image depicting a sequence of **{language}** story scenes.

{language} Story Scene Descriptions:

Scene 1: {scene_descriptions[0]}

Scene 2: {scene_descriptions[1]}

Scene 3: {scene_descriptions[2]}

Scene 4: {scene_descriptions[3]}

Scene 5: {scene_descriptions[4]}

Your task is to evaluate the **Cohesion** of the given story across all scenes.

What is Cohesion? It measures the coherence and logical flow between scenes, ensuring smooth transitions and meaningful relationships in the visual narrative.

Rating Scale:

5 - Perfect

4 - Minor Issues

3 - Major Issues

2 - Significant Issues

1 - Rubbish

Return the Rating based on the provided rating scale and explain the Reasoning behind the given rating.

Follow the below given format to return Rating.

single rating number (5, 4, 3, 2, or 1) enclosed in parentheses.

Reasoning:

Figure 27: Prompt used for evaluating Cohesion in MLLM-as-Jury Evaluation

Cultural Appropriateness V1 Prompt

You are an expert in evaluating **Cultural Appropriateness across different Languages and Cultures**.

You are given an image depicting a sequence of **{language}** story scenes.

{language} Story Scene Descriptions:

Scene 1: {scene_descriptions[0]}

Scene 2: {scene_descriptions[1]}

Scene 3: {scene_descriptions[2]}

Scene 4: {scene_descriptions[3]}

Scene 5: {scene_descriptions[4]}

Your task is to judge the **Cultural Appropriateness** of the given story. Evaluate the Cultural Appropriateness across the entire story.

What is Cultural Appropriateness? It measures how accurately the generated visuals represent the cultural styles, symbols and contexts described in the scene captions across the entire story.

Rating Scale:

5 – Perfect

4 – Minor Issues

3 – Major Issues

2 – Significant Issues

1 – Rubbish

Return the Rating based on the provided rating scale and explain the Reasoning behind the given rating.

Follow the below given format to return Rating.

single rating number (5, 4, 3, 2, or 1) enclosed in parentheses.

Reasoning:

Figure 28: Prompt used for evaluating Cultural Appropriateness level V1 in MLLM-as-Jury Evaluation

Cultural Appropriateness V2 Prompt

You are an expert in evaluating **Cultural Appropriateness** across different **Languages and Cultures**.

You are given an image depicting a sequence of **{language}** story scenes.

{language} Story Scene Descriptions:

Scene 1: {scene_descriptions[0]}

Scene 2: {scene_descriptions[1]}

Scene 3: {scene_descriptions[2]}

Scene 4: {scene_descriptions[3]}

Scene 5: {scene_descriptions[4]}

Your task is to judge the **Cultural Appropriateness** of the given story. Evaluate the Cultural Appropriateness across the entire story.

What is Cultural Appropriateness? It measures how accurately the generated visuals represent the cultural styles, symbols and contexts described in the scene captions across the entire story.

Focus on evaluating the character's facial structures, background objects, infrastructures and apparel during the evaluation of **{target}** cultural Aspect by analyzing the story scene images.

Rating Scale:

5 - Perfect

4 - Minor Issues

3 - Major Issues

2 - Significant Issues

1 - Rubbish

Return the Rating based on the provided rating scale and explain the Reasoning behind the given rating.

Follow the below given format to return Rating.

single rating number (5, 4, 3, 2, or 1) enclosed in parentheses.

Reasoning:

Figure 29: Prompt used for evaluating Cultural Appropriateness level V2 in MLLM-as-Jury Evaluation

Cultural Appropriateness V3 Prompt

You are an expert in evaluating **Cultural Appropriateness** across different **Languages and Cultures**.

You are given an image depicting a sequence of **{language}** story scenes.

{language} Story Scene Descriptions:

Scene 1: {scene_descriptions[0]}

Scene 2: {scene_descriptions[1]}

Scene 3: {scene_descriptions[2]}

Scene 4: {scene_descriptions[3]}

Scene 5: {scene_descriptions[4]}

Your task is to judge the **Cultural Appropriateness** of the given story. Evaluate the Cultural Appropriateness across the entire story.

What is Cultural Appropriateness? It measures how accurately the generated visuals represent the cultural styles, symbols and contexts described in the scene captions across the entire story.

Focus on evaluating the character's facial structures, background objects, infrastructures and apparel during the evaluation of **{target}** cultural aspect by analyzing the story scene images.

1. **Characters' Facial Features:** Evaluate whether the facial structures align with the diverse traits commonly found in **{target}** culture. Avoid assumptions about stereotypical features.
2. **Background Objects:** Assess whether the depicted objects represent the **{target}** cultural setting described in the scenes, focusing on nearby objects, furniture, decorations and other contextual details.
3. **Infrastructures:** Consider whether the settings, such as architectural elements, are appropriate for the **{target}** culture.
4. **Apparel:** Assess whether the clothing aligns with traditional or contemporary styles representative of **Appropriate** culture.

Rating Scale:

5 - Perfect

4 - Minor Issues

3 - Major Issues

2 - Significant Issues

1 - Rubbish

Return the Rating based on the provided rating scale and explain the Reasoning behind the given rating.

Follow the below given format to return Rating.

single rating number (5, 4, 3, 2, or 1) enclosed in parentheses.

Reasoning:

Figure 30: Prompt used for evaluating Cultural Appropriateness level V3 in MLLM-as-Jury Evaluation