A Progressive Evaluation Framework for Multicultural Analysis of Story Visualization

Janak Kapuriya*, **Ali Hatami, Paul Buitelaar**
Data Science Institute
University of Galway, Ireland
{janakkumar.kapuriya, ali.hatami, paul.buitelaar}@universityofgalway.ie

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 cultural fidelity. In this study, we present a progressive evaluation framework for story visualization. We validate this framework on current text-to-image models across three languages (English, Hindi, and Chinese) on two datasets (VIST and FlintstonesSV). The proposed framework introduces three levels of cultural analysis as evaluation rubrics: 1) Basic Cultural Criteria, 2) Cultural Dimension Guidance, and 3) Cultural Examples Grounding. We evaluate story visualization by use of a novel MLLM-as-Jury approach across all three rubrics and a small-scale human evaluation only on the third rubric. We implement an MLLM-as-jury approach by aggregating scores from three different families of MLLM-as-Judge models. In our experiments, real-world stories generally receive higher cultural appropriateness scores than animated ones, with English tending to score higher than Hindi and Chinese across the evaluated models. Some examples also exhibited culturally inconsistent or stereotypical elements noted by annotators. The proposed progressive evaluation framework has therefore been shown to provide early insights into cultural misalignments in story visualization. Code for this work is made available on https://github.com/janak11111/Cultural_Eval_For_StoryViz

1 Introduction

Recent advancements in text-to-image diffusion models, such as DALLE-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 images, 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 story elements in the corresponding text. This task holds immense potential across fields such as education, animation and content creation. However, models used in story visualization are English-based, 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 methods (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, producing 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. This analysis highlights a clear gap in the generation of meaningful, culturally accurate multilingual story visualizations.

In this work, we analyze cultural fidelity in story visualization across three languages: English, Hindi, and Chinese, using two datasets: the real-world VIST dataset (Huang et al., 2016) and the animated FlintstonesSV dataset (Gupta et al., 2018). To address the cultural gap left by existing evaluation metrics such as FID (Heusel et al., 2017) and CLIPScore (Radford et al., 2021), which focus on image–image or image–text similarity and do not capture cultural aspect (Seo et al., 2025), we

*Corresponding author

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 410–427

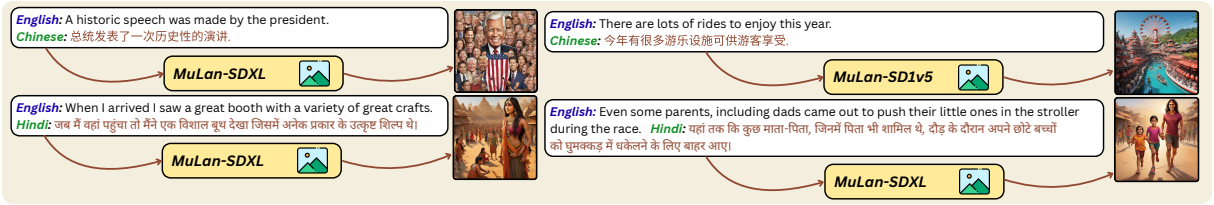

Figure 1: Cultural inconsistencies and stereotypes in generated story scenes across models and languages. **Ex–1**: The model is not interpreting the word '*president*' in Chinese as '*party leader*' in China, and generated an image of a US president. **Ex–2**: Instead of a modern craft fair, it depicts 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.

propose a Progressive Culture Evaluation Framework with three rubrics: 1) Basic Cultural Criteria, 2) Cultural Dimension Guidance, and 3) Cultural Examples Grounding. This framework enables progressive rubric criteria enrichment for culture assessment in generated stories. We evaluate models using an MLLM-as-Jury approach that aggregates ratings from three families of multimodal Judge models, and we conduct a small pilot human evaluation on rubric 3, which contains the full cultural details. These evaluations enable the identification of cultural inconsistencies and highlight cultural misalignments that occur in current story visualization models.

In this work, we analyze the following research questions:

- **RQ 1:** How does the visualization of the same story vary across different languages?
- **RQ 2:** How does story visualization adapt to different cultures?

Our contributions are as follows:

- **Culture Exploration in Story Visualization:** We investigate the role of culture in story visualization across multiple languages, examining how story narratives are interpreted and generated by models across different cultural contexts.
- **Proposed Progressive Culture Evaluation Framework:** We propose a progressive framework for cultural evaluation with three rubrics: (1) Basic Cultural Criteria, (2) Cultural Dimension Guidance, and (3) Cultural Examples Grounding, enabling increasingly detailed assessment of cultural fidelity in generated stories.
- **Multicultural Analysis on Real-world and Animated Stories:** We perform a comparative analysis between real-world stories that contain culturally rich, variable content (events, food, settings) and animated stories that have simpler, repetitive settings. This contrast allows us to evaluate the model's ability to adapt to different cultural contexts.

2 Related Work

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 only on the English language and overlook the cultural dimension. To address this gap, we conducted a multicultural analysis of story visualization in multilingual contexts using multilingual text-to-image models.

Cultural Analysis in Text-to-Image Models: Recent research has increasingly explored how cultural factors influence the behavior of text-to-image (T2I) systems. Prior studies have shown that T2I models often exhibit cultural inconsistencies when rendering culturally grounded concepts (Liu et al., 2023), prompting methods that adjust model outputs to reflect different cultural viewpoints (Ventura et al., 2024) and techniques that explicitly align generated images with specific cultural contexts (Khanuja et al., 2024). Recent work highlights that generative models may reinforce stereotypes

or misrepresent cultural elements (Bayramli et al., 2025), while others propose a structured framework as multi-agent pipelines to promote culturally appropriate imagery (Bhalerao et al., 2025). Different from these works, we investigate multicultural dimensions specifically within the task of story visualization, focusing on how text-to-image models interpret and depict narratives across different languages, aiming to uncover how cultural nuances embedded in multilingual stories influence the visual produced by T2I models.

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 a single judge LLM. LLM-as-Judge has been extended to the multimodal case and compared scorewise, pairwise and batch-wise ranking across diverse datasets (Chen et al., 2024). 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 MLLM-as-Judge to MLLM-as-Jury to analyze multicultural story visualization. We adopt an MLLM-as-Jury framework that aggregates judgements by reducing individual bias and enhancing the robustness of the evaluation.

3 Methodology

Our approach consists of three steps: story translation, story visualization using multilingual text-to-image models, and evaluation of the generated stories using the proposed progressive evaluation framework with MLLM-as-Jury approach. Each step is described in detail in the following subsections.

3.1 Story Translation

To support multilingual story visualization, each story narrative $s_i^{(1)}$, originally written in English, is translated to Hindi and Chinese as shown in Figure 2, using the `NLLB-200` (Team et al., 2022) translation model, chosen for its empirically demonstrated robustness and high translation quality across typologically diverse language pairs, as well as its open-access availability enabling reproducible research. This results in a multilingual set of narratives $s_i^{(\ell)}$ for each language $\ell \in \mathcal{L} = \{\text{English}, \text{Hindi}, \text{Chinese}\}$. We treat machine translation as a controlled preprocessing step and fix the translation model across all conditions. These translated narratives serve as inputs for the subsequent multilingual T2I generation. We reported reference-free translation quality scores computed using COMET-QE[1] on both the VIST and FlintstonesSV datasets. COMET-QE allows us to estimate translation quality without requiring gold-standard reference translations. As shown in Table 1, VIST consistently achieves higher COMET-QE scores than FlintstonesSV for both translation directions, with the average scores indicating slightly stronger performance for English–to-Hindi compared to English–to-Chinese.

Dataset	Eng → Hi (↑)	Eng → Ch (↑)
FlintstonesSV	0.4537	0.3902
VIST	0.8224	0.7432
Average	**0.6380**	**0.5667**

Table 1: Reference-free COMET-QE translation quality scores.

3.2 Story Visualization using Multilingual T2I Models

We generate the image corresponding to each story narrative $s_i^{(\ell)}$ in language ℓ independently without condition on previous images/story narratives using the multilingual text-to-image models denoted by function $I_i^{(\ell)} = \{(M, s_i^{(\ell)}, T, \gamma)$ for all $\ell \in \mathcal{L}$, where T is the number of denoising steps and γ is the guidance scale. We generated each scene independently due to the unavailability of multilingual story visualization models, and our primary objective is to analyze culture rather than cross-scene consistency.

3.3 MLLM-as-Jury Evaluation Framework

We introduced an MLLM-as-Jury to evaluate generated stories. MLLM-as-Jury leverages Expert Role-specific prompting shown in stage 3 of Figure 2. Each story is represented as a sequence of story narrative and image pairs, $S = \{(n_1, i_1), (n_2, i_2), \ldots, (n_T, i_T)\}$, 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 P fed to three different

[1] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

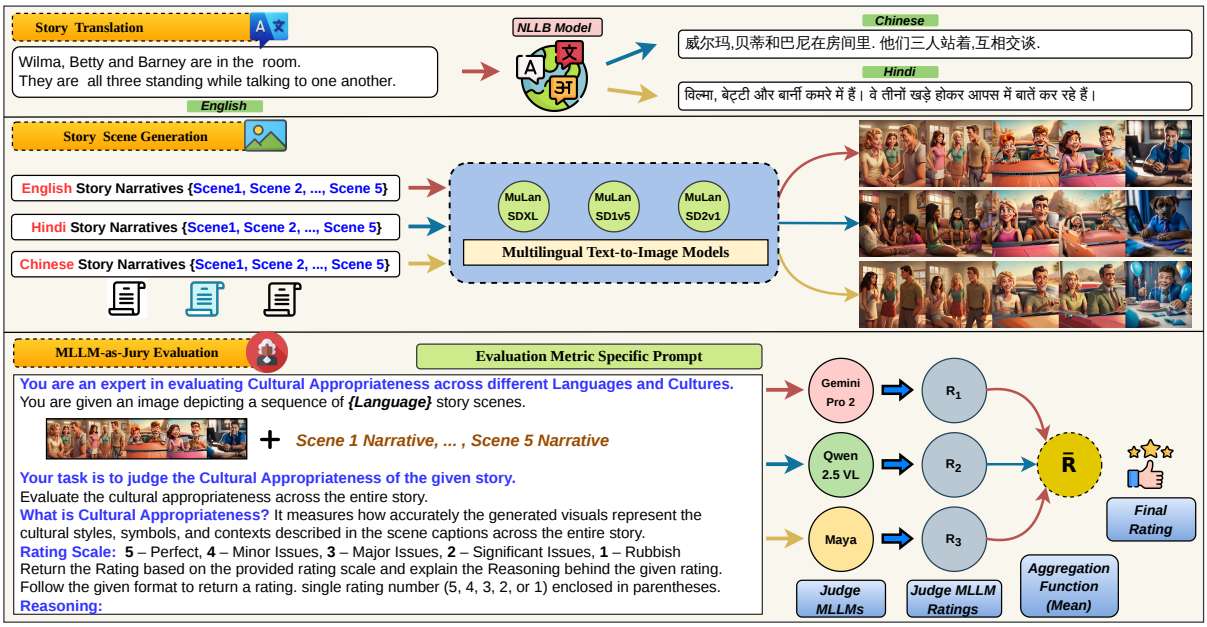

Figure 2: Three-stage framework for multicultural analysis of story visualization

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, 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 $R_{\text{pred}} = \frac{1}{3}(r_1 + r_2 + r_3)$. Example is given in the appendix Figure 11. In the evaluation prompt, scene narratives are in the native languages, with the rest of the prompt instructions in English.

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 MLLM-as-Judge 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.

To evaluate culture in generated stories, we propose a cultural evaluation framework. This framework assesses the complete story by considering all image–narrative pairs and assigning a single rating score (1–5) that reflects the overall quality of the generated sequence. The **Rating scheme** used is : **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 culture in generated stories, this framework incrementally enriches the cultural evaluation criteria across three levels:

- **Basic Cultural Criteria** ($\mathcal{C}$): Evaluates `Cultural Appropriateness` based on the basic cultural criteria derived directly from the story context. `Cultural Appropriateness` measures how accurately the generated visuals represent the cultural styles, symbols, and contexts described in the scene captions throughout the entire story.
- **Cultural Dimension Guidance**: Basic Cultural Criteria ($\mathcal{C}$) + Focus Points ($\mathcal{F}$) - adds a specific cultural dimension to pay attention to with basic cultural criteria during the evaluation.
  Focus Points ($\mathcal{F}$)**:** Background Objects, Facial Features, Infrastructures, Apparel
- **Cultural Examples Grounding**: Basic Cultural Criteria ($\mathcal{C}$) + Cultural Dimension Guidance ($\mathcal{F}$) + Illustrative Examples ($\mathcal{E}$) - provides concrete examples for each focus point to guide consistent and precise assessment.

Illustrative Examples ($\mathcal{E}$): Below are the illustrative examples for each of the four cultural focus points.

- **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.
- **Facial Features**: Evaluate whether the facial structures align with the diverse traits commonly found in the target culture. Avoid assumptions about stereotypical features.
- **Infrastructures**: Consider whether the settings, such as architectural elements, are appropriate for the target culture
- **Apparel**: Assess whether clothing aligns with traditional or contemporary styles representative of the appropriate culture.

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

$$\mathbb{E}_1(s_i) = g(s_i \mid \mathcal{C})$$
$$\mathbb{E}_2(s_i) = g(s_i \mid \mathcal{C}, \mathcal{F})$$
$$\mathbb{E}_3(s_i) = g(s_i \mid \mathcal{C}, \mathcal{F}, \mathcal{E})$$

The goal of this progressive evaluation framework is to evaluate whether richer prompts lead to more accurate and rigorous culture assessments of the generated story content. See Section A.2 in the Appendix for detailed information on the prompts used in our experiments.

3.5 Human Evaluation

To assess the cultural relevance of generated visual stories, we conducted a preliminary human evaluation with native speakers in English, Chinese, and Hindi. 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.

We randomly selected 15 story samples from different models and datasets. Each sample consisted of five (image, story narrative) pairs. Of these, 6 samples were drawn from the FlintstonesSV dataset and 9 from the VIST dataset. For each of the three text-to-image models, we selected 2 samples from FlintstonesSV and 3 samples from VIST, ensuring balanced coverage across models and datasets. For each language, three annotators were provided with the same set of 15 samples and independently evaluated the cultural appropriateness of each story using the rating scheme described in Section 3.3. The main question given to annotators was to rate the story based on how culturally appropriate the generated story was. Examples of human annotations from one annotator for each of the three languages are shown in Appendix Figures 12, 13, and 14.

The scores from the three annotators were aggregated using the mean, and the resulting values are reported in Section 5.2. Given the focus on a multilingual, cross-cultural setting, this limited-scale design is intended as an initial feasibility study rather than a large-scale human evaluation. In addition to ratings, evaluators provided qualitative feedback, supporting a nuanced analysis of visual and cultural effectiveness in multilingual story generation.

Table 2: Inter-rater reliability results across languages for Cultural Appropriateness.

Language	κ	Within-1	ICC(2, 3)	95% CI
English	0.05	0.69	0.16	[-0.47, 0.47]
Hindi	0.17	0.78	0.39	[0.05, 0.60]
Chinese	0.07	0.60	0.18	[-0.15, 0.38]

Given the subjectivity of cultural judgments, we assess inter-rater reliability using quadratic-weighted Cohen's κ, Within-1 agreement (W1), and Intra-class correlation ICC(2,3) with 95% confidence intervals (Table 2). As expected for nuanced cultural evaluation, κ values remain low, whereas W1 and ICC offer a more stable picture of annotator consistency. English and Hindi exhibit higher W1 scores, indicating strong near-agreement, while Chinese shows comparatively lower W1 and ICC, reflecting greater variability in how annotators interpret cultural appropriateness. Overall, W1 and ICC suggest reasonably consistent scoring patterns across raters, whereas the lower κ values mainly capture the inherent difficulty and subjectivity of fine-grained cultural assessment.

4 Experimental Setup

Choices of Language: To analyze the cultural aspect in story visualization, we selected three lan-

Models	Basic Cultural Criteria			Cultural Dimension Guidance			Cultural Examples Grounding		
	English	Hindi	Chinese	English	Hindi	Chinese	English	Hindi	Chinese
FlintstonesSV									
MuLan-SD2v1	$4.27_{\pm 0.03}$	$3.24_{\pm 0.04}$	$3.17_{\pm 0.04}$	$4.18_{\pm 0.03}$	$3.16_{\pm 0.03}$	$3.13_{\pm 0.03}$	$3.84_{\pm 0.03}$	$3.18_{\pm 0.03}$	$3.09_{\pm 0.02}$
MuLan-SD1v5	$4.30_{\pm 0.02}$	$3.20_{\pm 0.04}$	$3.17_{\pm 0.04}$	$4.10_{\pm 0.03}$	$3.16_{\pm 0.04}$	$3.19_{\pm 0.03}$	$3.86_{\pm 0.02}$	$3.21_{\pm 0.03}$	$3.16_{\pm 0.02}$
MuLan-SDXL	$\mathbf{4.34}^{*}_{\pm 0.02}$	$\mathbf{3.35}^{*}_{\pm 0.04}$	$\mathbf{3.20}_{\pm 0.04}$	$\mathbf{4.21}^{*}_{\pm 0.03}$	$\mathbf{3.34}^{*}_{\pm 0.04}$	$\mathbf{3.20}_{\pm 0.03}$	$\mathbf{3.87}_{\pm 0.02}$	$\mathbf{3.40}^{*}_{\pm 0.03}$	$\mathbf{3.18}_{\pm 0.02}$
VIST									
MuLan-SD2v1	$4.31_{\pm 0.03}$	$3.71_{\pm 0.04}$	$3.72_{\pm 0.04}$	$4.23_{\pm 0.03}$	$3.59_{\pm 0.04}$	$3.61_{\pm 0.04}$	$3.94_{\pm 0.04}$	$3.45_{\pm 0.03}$	$3.49_{\pm 0.04}$
MuLan-SD1v5	$4.37_{\pm 0.03}$	$3.72_{\pm 0.04}$	$3.99_{\pm 0.03}$	$4.27_{\pm 0.03}$	$3.66_{\pm 0.04}$	$4.03_{\pm 0.03}$	$4.03_{\pm 0.03}$	$3.82_{\pm 0.42}$	$3.90_{\pm 0.04}$
MuLan-SDXL	$\mathbf{4.44}^{*}_{\pm 0.03}$	$\mathbf{3.93}^{*}_{\pm 0.03}$	$\mathbf{4.04}^{*}_{\pm 0.03}$	$\mathbf{4.38}^{*}_{\pm 0.02}$	$\mathbf{3.91}^{*}_{\pm 0.04}$	$\mathbf{4.06}_{\pm 0.03}$	$\mathbf{4.12}^{*}_{\pm 0.03}$	$\mathbf{3.89}^{*}_{\pm 0.04}$	$\mathbf{3.95}^{*}_{\pm 0.04}$

Table 3: MLLM-as-Jury Cultural Appropriateness results with rounded confidence intervals shown as subscripts. Marker * indicates Wilcoxon significance at $p < 0.05$, where the model is statistically better than all other models.

guages: 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.

Choices of Datasets: We analyze story visualization across cultural contexts with two datasets: VIST (real-world) and FlintstonesSV (animated). VIST (Huang et al., 2016) features photo sequences from Flickr albums of everyday events, consisting of 50,000 stories, while 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. We performed experiments on randomly sampled 500 story samples from both datasets, where each story sample has 5 (image, caption) pairs.

Multilingual Text-to-Image Models: To generate story images from story narratives in different languages, we use the multilingual T2I diffusion model MuLan (Xing et al., 2024). We use three MuLan variants: 1) `MuLan-SD1v5`, 2) `MuLan-SD2v1` and 3) `MuLan-SDXL`, representing progressively diverse open-source stable diffusion versions. See Section A.3 in the Appendix for inference configurations for MLLM-as-Judge models (Gemini Pro 2.0, QwenVL 2.5, Maya) and text-to-image MuLan-based models (SDXL, SD2v1, SD1v5).

5 Results

5.1 MLLM-as-Jury Results

As shown in Table 3, MuLan-SDXL consistently achieves the highest cultural appropriateness scores across all three evaluation levels (Basic Cultural Criteria, Cultural Dimension Guidance, and Cultural Examples Grounding), datasets, and languages. SD1v5 generally performs second-best, while SD2v1 performs lowest. In several settings, SDXL's improvements are marked with (*), indicating statistically significant gains over both baselines ($p < 0.05$). Across datasets, VIST scores are higher than FlintstonesSV for every language and model, suggesting that MLLM-as-Jury models find real-world stories easier to evaluate for cultural appropriateness than animated stories, likely due to more familiar cultural, visual, and narrative cues. This improvement is especially notable for Hindi and Chinese, which show larger score increases on VIST compared to FlintstonesSV.

Across languages, English achieves the highest cultural appropriateness scores across all evaluation levels, followed by Hindi and Chinese. Hindi and Chinese benefit more on VIST, where their scores are closer to English than those on FlintstonesSV. However, Hindi remains the lowest under the Basic Cultural Criteria on FlintstonesSV, while Chinese is lowest at the Cultural Example Grounding stage on VIST.

For English, scores decrease progressively when moving from Basic Criteria to Dimension Guidance to Examples Grounding. This indicates that introducing more structured criteria and cultural examples makes the evaluation stricter, causing the judge model to penalize mismatches more heavily. In contrast, Hindi and Chinese remain relatively stable across all three levels on both datasets, showing only minor fluctuations. This stability suggests that the added structure has minimal impact for these languages, likely because the MLLM-as-Jury models possess limited cultural knowledge for underrepresented languages, resulting in similar judgments regardless of added guidance and examples. Finally, confidence intervals are consistently nar-

Figure 3: Qualitative visualization of the same story generated in different languages by MuLan-SDXL.

row (0.02–0.04), indicating stable jury predictions across samples. Overall, the results show clear model ranking, consistent dataset effects, and stable behavior across languages, while also highlighting where cultural understanding remains challenging for the jury models. We also provide the ablation study of MLLM-as-Judge versus MLLM-as-Jury across language and dataset in the Appendix section A.1.

5.2 Human Evaluation Results

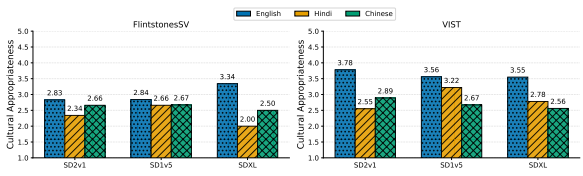

Figure 4: Human evaluation scores across datasets, languages, and models.

Our human evaluation pilot study results, shown in Figure 4, indicate that English stories receive the highest cultural appropriateness scores across all three models on both datasets. Hindi and Chinese follow with lower and more variable scores, with Hindi slightly higher than Chinese on the VIST dataset, while Chinese performs slightly better than Hindi on FlintstonesSV. SDXL achieves the strongest performance for English overall. However, differences across models for Hindi and Chinese remain modest within this small sample. Scores on VIST are consistently higher than those on FlintstonesSV, suggesting that real-world story narratives provide clearer and more interpretable cultural cues than cartoon-style scenes. Overall, as this is a small-scale pilot study, these findings should be considered exploratory rather than conclusive, although they highlight observable differences in cultural alignment across languages, models, and datasets.

6 Discussion

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

RQ1: How does the visualization of the same story vary across different languages? As we can see from Figure 3, 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. Additionally, the concept of "success" is represented differently across languages in scene 5: 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. These examples show that the same model generates culturally different visuals across different languages.

RQ2: How does story visualization adapt to different cultures? We can see in Figure 5, it shows the culture adaptation examples. **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 '*saari*' dresses for the bride and bridesmaids. **Ex3:** The story scene in Chinese about a family dinner, where the scene adapts to Chinese food. **Ex4:** The story scene in Hindi about a mother sharing a recipe with her daughter-in-law

Figure 5: Cultural adaptation in multilingual story visualization using the MuLan-SDXL model on VIST stories.

in the kitchen. The scene adapts to Indian food, steel utensils, and Indian clothes. These examples show a default culture adaptation in multilingual T2I models.

7 Error Analysis

This section analyzes errors in multilingual text-to-image story visualization, focusing first on translation errors and then on cultural errors.

7.1 Translation Errors

We manually analyzed a subset of multilingual story narratives used for text-to-image story visualization and identified several translation inconsistencies that may affect multimodal alignment. Figure 6 shows examples of errors in English-to-Hindi and English-to-Chinese translation. In the Hindi example, the phrase "*pirate*" is translated as "समुद्री डाकू" (literally "*sea bandit*"), which is semantically correct but represents a literal decomposition rather than a direct lexical equivalent. While this does not introduce a major error, it reflects a shift in expression that may influence visual grounding. In contrast, the Chinese translations exhibit clearer semantic errors. For instance, the word "*romaine*" (referring to a type of lettuce) is incorrectly translated as "罗马", which refers to the "*Rome*" and "*romance culture*", resulting in a complete loss of the intended meaning. These inconsistencies can lead to mismatches between text and generated visuals, thereby impacting the performance of text-to-image story visualization.

Figure 6: Examples of translation errors propagated in MuLan-SDXL model. **(Left)**: English-to-Hindi **Right**: English-to-Chinese. The red-colored words highlight lexical translation errors.

7.2 Culture Errors

We identified several errors during the evaluation of cultural aspects in the generated stories. Figure 7 illustrates examples from both Hindi and Chinese story generations. The top image shows a Hindi story depicting a family vacation with friends. In Scene 1, the absence of footwear, while Scene 3 exhibits culturally inconsistent clothing, demonstrates misrepresentations of Indian culture. The bottom image presents 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 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, signaling culturally inappropriate and stereotyped visuals.

Figure 7: Examples of cultural inconsistencies in generated story scenes: **Top:** Hindi, **Bottom:** Chinese.

8 Conclusion

In this work, we introduced a progressive evaluation framework for assessing cultural alignment in multilingual text-to-image story visualization. Across experiments on real-world (VIST) and animated (FlintstonesSV) datasets in English, Hindi,

and Chinese, we observed a consistent trend in which English narratives tend to receive higher cultural appropriateness scores than Hindi and Chinese. This pattern appears in both human evaluations and the MLLM-as-Jury framework, indicating preliminary consistency. Additionally, real-world stories generally achieve higher cultural alignment than animated ones, suggesting that dataset characteristics influence cultural fidelity. A small number of outputs showed culturally inconsistent patterns or stereotypical elements noted by annotators. Overall, our findings demonstrate that cultural alignment in story visualization is sensitive to language, dataset type, and model design. The proposed framework provides a structured way to detect such misalignment, offering a foundation for developing more culture aware generative models and fine-grained evaluation systems.

Limitations

This study has several limitations. **1)** The scope of our multicultural analysis is restricted to three languages, due to resource and annotation constraints. **2)** Our work concentrates on evaluating cultural fidelity using progressively detailed rubrics, rather than proposing techniques to mitigate cultural inconsistencies or stereotypes. Developing mitigation methods is our future work. **3)** Our human evaluation is a small pilot study, and with only a limited sample, the analysis should be interpreted as exploratory.

9 Ethics statement

This work uses only publicly available datasets and open-source text-to-image models. Our cultural evaluation focuses only on assessing the behavior of generative models and not on evaluating or judging any real cultural or social groups. Any discussed cultural inconsistencies or stereotypes reflect model outputs rather than human communities.

Acknowledgment

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

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT for grammar correction and spelling checks.

References

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, Sibo 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.

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.

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*.

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.

Huichan Seo, Sieun Choi, Minki Hong, Yi Zhou, Junseo Kim, Lukman Ismaila, Naome Etori, Mehul Agarwal, Zhixuan Liu, Jihie Kim, and 1 others. 2025. Exposing blindspots: Cultural bias evaluation in generative image models. *arXiv preprint arXiv:2510.20042*.

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 Heffernan, 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*.

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 MLLM-as-Jury ablation study, experimental setup config details, and prompts utilized within the MLLM-as-Jury evaluation framework.

A.1 MLLM-as-Jury Ablation Study

This section compares the single-judge models with the jury method to examine how individual judges contribute across languages, metrics, and datasets. A different pattern emerges for cultural appropriateness with different rubrics as shown in Figure 8, Figure 9, and Figure 10. Across all rubrics, Maya consistently produces the highest scores, especially for English, while Qwen generally assigns the lowest scores for Hindi and Chinese, contrary to Gemini being the lowest. FlintstonesSV receives lower cultural scores than VIST for all judges, but the magnitude of this drop varies: it is largest for Gemini and Qwen and smallest for Maya, indicating that animated stories are more culturally challenging for some judges than others.

As rubric complexity increases from Basic Cultural Criteria (V1) to Cultural Dimension Guidance (V2) and then to Cultural Examples Grounding (V3), Gemini and Qwen exhibit noticeable score decreases, particularly for Hindi and Chinese, reflecting stricter evaluation when cultural dimensions and examples are explicitly required. In contrast, Maya remains relatively stable, leading to increasing disagreement among the three judges under more detailed criteria. This divergence is most pronounced in Cultural Examples Grounding, where Maya assigns scores above 4.5 for Hindi and Chinese, while Gemini drops to around or below 2 on FlintstonesSV, revealing strong model-specific sensitivities to culturally grounded prompts. The jury approach moderates these extremes by averaging the outputs of the three judges, consistently producing scores that fall between the highest (Maya) and lowest (Gemini). This reduces variance and mitigates judge-specific bias, particularly for Hindi and Chinese. For English, however, score variance across different rubrics remains relatively stable. Overall, the MLLM-as-Jury method provides a more stable and reliable approach to automatic cultural evaluation by smoothing model disagreements while still capturing relative trends across languages, datasets, and rubric levels.

A.2 Prompts used in MLLM-as-Jury Evaluation

This section presents the prompts used in our MLLM-as-Jury evaluation framework. Our Progressive Multicultural Evaluation framework enables culturally grounded assessment through three prompt versions. **Basic Cultural Criteria** (Figure 15), **Cultural Dimension Guidance** (Figure 16) and **Cultural Examples Grounding** (Figure 17). Adding three levels of progressive cultural context to the rubric enables a more detailed evaluation of culture in story visualization.

A.3 Text-to-Image and MLLM-as-Judge Models Configuration

This section describes the configurations of the models used in our multilingual text-to-image story generation and MLLM-based evaluation framework. For story generation, we employ multilingual text-to-image (T2I) models built on the MuLan framework, which extends English-centric diffusion models to handle multilingual prompts using language adapters. Three variants of MuLan-enabled Stable Diffusion were used: SD1v5, SD2v1, and SDXL. These models vary in architecture and image resolution, enabling a spectrum of visual generation capabilities. Table 4 outlines

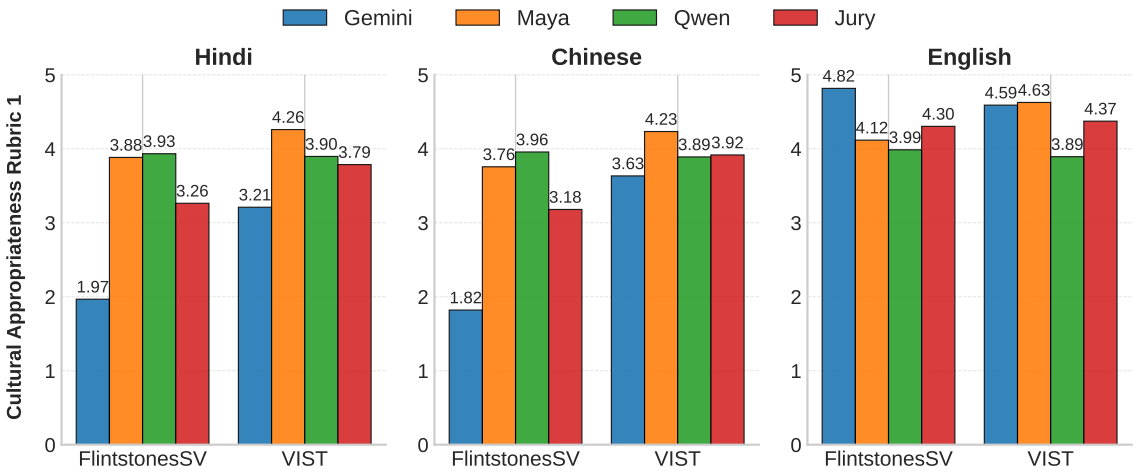

Figure 8: Jury vs Judge ablation on Cultural Appropriateness (Basic Cultural Criteria)

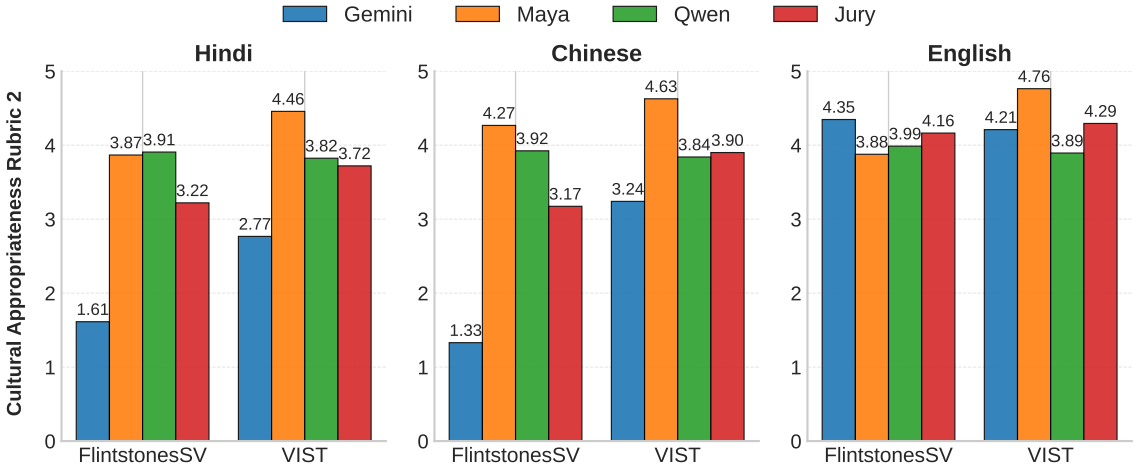

Figure 9: Jury vs Judge ablation on Cultural Appropriateness (Cultural Dimension Guidance)

the inference settings for each model, including resolution, scheduler type, and hardware specifications.

For evaluation, we use an MLLM-as-Jury framework to perform large-scale, automated assessment of generated stories. This framework leverages three Multimodal Large Language Models (MLLMs): `Gemini Pro 2.0`, `Qwen2.5-VL`, and `Maya`. Table 4 also summarizes the inference configurations for these models, including access method, maximum tokens, sampling settings, temperature, prompt type, random seed, and hardware setup. Using diverse generation and evaluation models ensures both robust visual outputs and comprehensive, multi-perspective quality assessment.

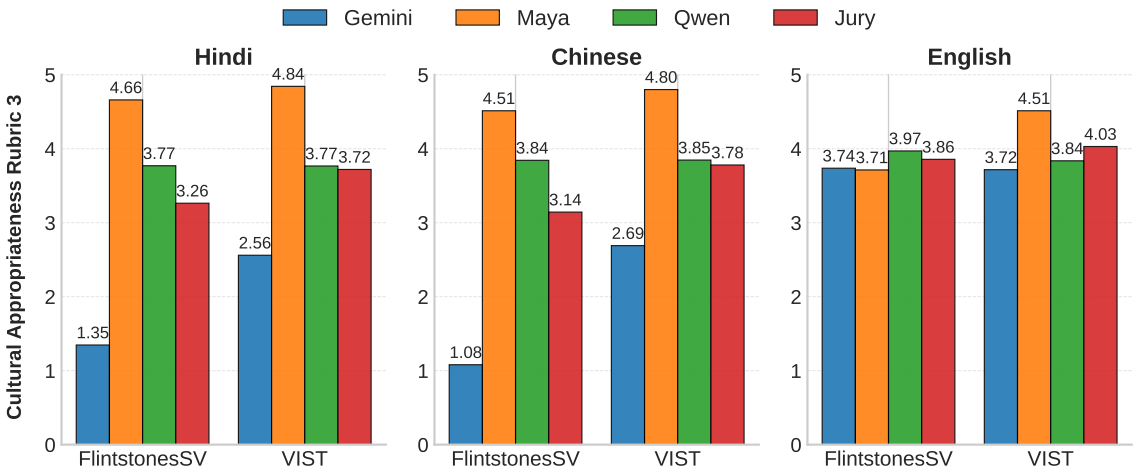

Figure 10: Jury vs Judge ablation on Cultural Appropriateness (Cultural Examples Grounding).

MLLM-as-Judge Models (Gemini Pro 2.0, QwenVL 2.5, Maya)		**Text-to-Image Models (SDXL, SD2v1, SD1v5)**	
Parameter	**Value**	**Parameter**	**Value**
Access	Gemini API / GPU Inference	Multilingual Adapter	MuLan
Max Tokens	512	Guidance Scale	7
Sampling	False	Inference Steps	50
Temperature	0	Framework Used	Diffusers
Prompt	Custom	Seed	12345
Seed	42	Hardware	RTX A6000 48GB GPU
Hardware	RTX A6000 48GB GPU		

Table 4: Inference configurations for MLLM-as-Judge models (Gemini Pro 2.0, QwenVL 2.5, Maya) and text-to-image MuLan-based models (SDXL, SD2v1, SD1v5).

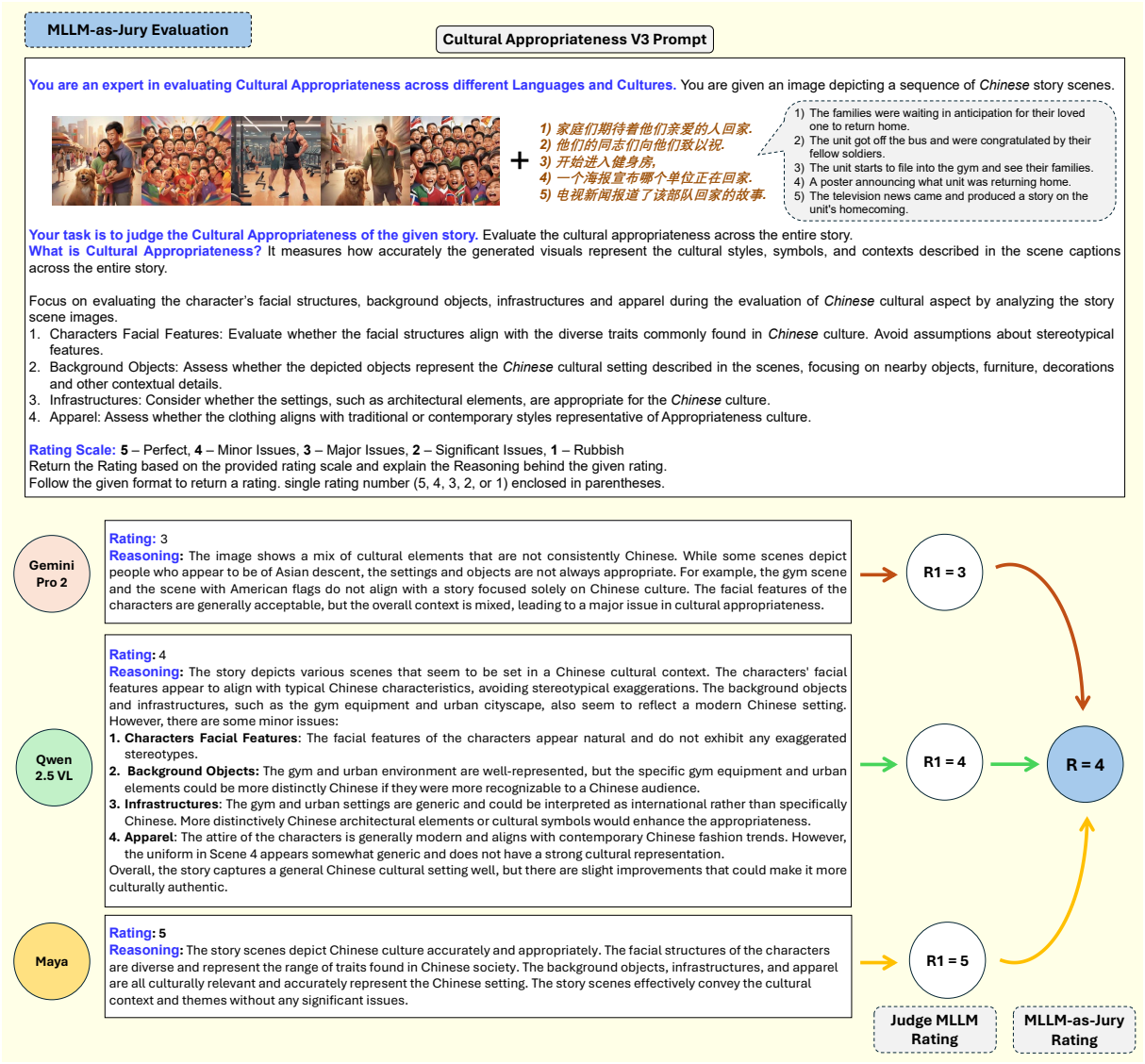

Figure 11: Schematic representation of the MLLM-as-Jury evaluation on Cultural Appropriateness V3 in Chinese using three MLLMs: Gemini Pro 2, Qwen 2.5 VL, and Maya.

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.

Figure 12: An example of human evaluation for an English story using the Cultural Appropriateness metric, with a score of 4 out of 5.

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,

Figure 13: An example of human evaluation for an Chinese story using the Cultural Appropriateness metric, with a score of 4 out of 5.

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 बूथ देखा जिसमें अनेक प्रकार के उत्कृष्ट शिल्प थे।
3) मैं अपने दोस्त बेथ के बूथ पर एक छोटे से बात करने के लिए बंद कर दिया.
4) यहाँ तक कि सभी बच्चों के लिए भी बूथ बनाए गए थे।
5) मुझे मेले में कुछ बेहतरीन शिल्प मिले, मैं बहुत खुश हूँ कि मैं गया।

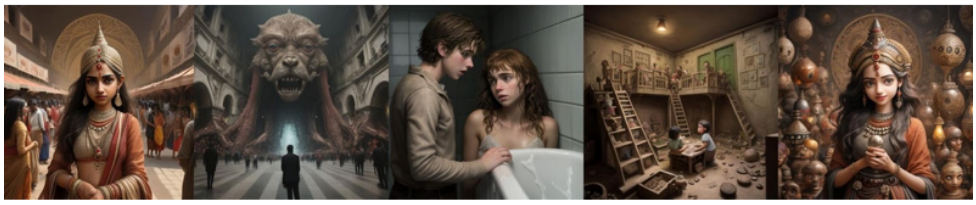

Figure 14: An example of human evaluation for an Hindi story using the Cultural Appropriateness metric, with a score of 3 out of 5.

Basic Cultural Criteria 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 15: Prompt used for evaluating Cultural Appropriateness Rubric 1 in MLLM-as-Jury Evaluation

Cultural Dimension Guidance 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 16: Prompt used for evaluating Cultural Appropriateness Rubric 2 in MLLM-as-Jury Evaluation

Cultural Examples Grounding 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 17: Prompt used for evaluating Cultural Appropriateness Rubric 3 in MLLM-as-Jury Evaluation