# When Cultures Meet: Multicultural Text-to-Image Generation

**Parth Bhalerao  Oana Ignat  Mounika Yalamarty  Brian Trinh**
Santa Clara University - Santa Clara, USA
*{pbhalerao, oignat}@scu.edu*

## Abstract

Text-to-image generation models have achieved strong performance in culturally homogeneous settings, yet their ability to generate *multicultural scenes*—where people and landmarks originate from different cultures—remains largely unexplored. We introduce *multicultural text-to-image generation* as a new task and present the first benchmark designed to study this setting. Our dataset contains 9,000 images spanning five countries, three age groups, two genders, 25 historical landmarks, and five languages. Using this benchmark, we analyze the behavior of state-of-the-art text-to-image models across multiple dimensions, including alignment, image quality, aesthetics, knowledge, and fairness. As one strategy for composing cultural and demographic information, we explore MosAIG, a Multi-Agent framework that enhances multicultural Image Generation by leveraging LLMs with distinct cultural personas. Our analysis shows that richer prompt composition can improve image quality and cultural grounding compared to simple prompts, while revealing substantial disparities across languages and demographic groups. We release our dataset and code at https://github.com/AIM-SCU/MosAIG.

## 1 Introduction

Societies worldwide are increasingly diverse, shaped by global travel and migration (Castles et al., 2103). This multicultural reality poses important challenges for Artificial Intelligence (AI), where robust representation of diverse populations is essential for equity and inclusivity (Hershcovich et al., 2022; Naous et al., 2024; Mihalcea et al., 2025). However, most datasets used for text-to-image generation focus on narrow demographics—predominantly Western, adult, and male—and largely depict single-culture scenarios (e.g., *a Chinese temple*, *an Indian market*) (Liu et al., 2024; Kannen et al., 2024). Such representations fail to capture common multicultural interactions, for example *a Chinese girl visiting the Golden Gate Bridge*, limiting the applicability of text-to-image models in real-world, culturally diverse settings (Hershcovich et al., 2022; Bhatia et al., 2024).

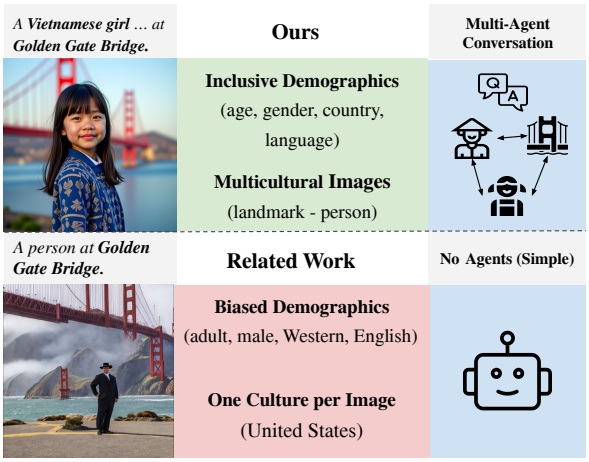



Figure 1: Most existing datasets emphasize singular cultural contexts (e.g., the Golden Gate Bridge depicted primarily with American visitors or as a standalone monument). In contrast, real-world scenes often involve people from different cultural backgrounds sharing spaces and experiences. Modeling such multicultural interactions enables richer and more realistic image generation.

Despite recent efforts to evaluate cultural awareness in vision–language models, existing benchmarks and analyses primarily consider one culture per image. To date, there has been no systematic study of *multicultural text-to-image generation*, where elements from different cultural origins—such as a person and a landmark—co-exist within the same visual scene. We introduce this setting as a new task and study how current text-to-image models handle cultural composition, demographic variation, and cross-cultural grounding.

Specifically, we examine two key dimensions: (1) the demographic attributes of the depicted person, and (2) the multicultural interaction be-



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 35808–35828

tween the person and the landmark (e.g., the Golden Gate Bridge). We consider four demographic aspects—age, gender, nationality, and language—together with cross-cultural landmarks (Figure 1). By systematically exploring these factors and their intersections, we aim to better understand the strengths and limitations of state-of-the-art text-to-image models in multicultural settings.

Our work is guided by the following research questions:

**RQ1:** How accurately do state-of-the-art text-to-image models depict people from one culture within the context of a landmark associated with a different culture?

**RQ2:** How does text-to-image generation performance vary across demographic groups and languages?

**RQ3:** What modeling strategies help improve multicultural text-to-image generation?

Contributions. **First, we release the first dataset of 9,000 images designed to study multicultural interactions**, depicting people and landmarks from different cultures across five countries, three age groups, two genders, 25 historical landmarks, and five languages. **Second, we explore `MosAIG`, a multi-agent prompting framework that decomposes cultural and demographic aspects** during caption construction as an effective strategy for addressing the task. **Finally, we analyze multicultural text-to-image generation through automated and human evaluation, revealing demographic and linguistic disparities.**

## 2 Related Work

**Cultural Evaluation in Language and Vision Models.** Recent work has made substantial progress in modeling and evaluating cultural awareness in language models through large multilingual benchmarks (Pawar et al., 2024; Romanou et al., 2025; Singh et al., 2025). In the vision–language domain, benchmarks such as CVQA (Romero et al., 2024) and GlobalRG (Bhatia et al., 2024) evaluate culturally grounded question answering, retrieval, and visual grounding. While multi-agent approaches have been explored for cross-cultural reasoning in multimodal systems (Guo et al., 2024; Han et al., 2024), prior work such as MosAIC (Bai et al., 2025) focuses on image captioning in single-culture settings, where models describe visual content post hoc. In contrast, we study *text-to-image generation in multicultural settings*, where models must jointly reason about and synthesize multiple cultural, demographic, and landmark-specific cues into a single coherent visual scene. This setting constitutes a particularly challenging test of cultural competence, as failures in grounding, representation, or composition are directly reflected in the generated images.

**Text-to-Image Generation Models and Benchmarks.** Text-to-image generation has advanced rapidly with models such as Stable Diffusion-XL (Podell et al., 2023), DALL·E 3 (Betker et al., 2023), and FLUX (Labs, 2024). Agentic approaches like GenArtist (Wang et al., 2024) focus on unified generation and editing pipelines, whereas our work emphasizes multicultural and multilingual evaluation rather than model design. Existing benchmarks, including TIFA (Hu et al., 2023), GenEval (Ghosh et al., 2024), and GenAIBench (Lin et al., 2025), primarily assess technical properties such as realism, faithfulness, and compositionality. More recent efforts, such as HEIM (Lee et al., 2024), incorporate socially situated dimensions including bias, toxicity, and aesthetics (Hartwig et al., 2024), but do not explicitly address multicultural scene composition.

**Cultural and Linguistic Gaps in Text-to-Image Generation.** Despite these advances, most text-to-image systems and evaluations remain centered on a limited set of high-resource languages, leaving many linguistic communities underserved. While models such as Taiyi-Diffusion-XL (Wu et al., 2024) and AltDiffusion (Ye et al., 2024) expand multilingual input coverage, a broader cultural gap persists (Liu et al., 2024), as existing benchmarks rarely capture cross-cultural interactions or multicultural contexts (Hershcovich et al., 2022; Mihalcea et al., 2025; Saha et al., 2025).

**Data Diversity and Cultural Competence.** Recent work has begun to assess cultural competence in text-to-image generation. For example, CUBE (Kannen et al., 2024) and TIFA (Hu et al., 2023) evaluate cultural awareness and diversity, but remains limited to single-culture depictions per image. To our knowledge, no prior work systematically studies *multicultural image generation*, where multiple cultures co-exist within the same scene. Our work addresses this gap by introducing a benchmark and analysis framework for multicultural text-to-image generation.

## 3 Multicultural Image Generation

*Culture is a multifaceted concept, meaning different things to different people at different times* (Adilazuarda et al., 2024). In this work, we adopt the definition of Nguyen et al. (2023) and focus on *visual cultural elements*, such as clothing and historical landmarks.

We introduce *multicultural image generation* as a new task that evaluates how text-to-image models represent elements from multiple cultures within a single image—specifically, a person from one cultural background depicted in the context of a landmark from another. In addition to cultural origin, we examine demographic attributes and their intersections, including age, gender, and language[1]. To address this task, we introduce MosAIG, a *novel framework* for Multi-Agent Image Generation, as illustrated in Figure 2. Our framework generates comprehensive image captions that are used to generate more accurate multicultural images using off-the-shelf image generation models. This framework is built around a multi-agent interaction model, as described below.

### 3.1 Multi-Agent Interaction Model

We introduce a multi-agent setup to emulate collaboration between demographically diverse groups. Our setup contains five agents, with specific roles: one Moderator Agent, three Social Agents, and one Summarizer Agent, as illustrated in Figure 2.

**Moderator Agent.** The Moderator Agent obtains demographic (age, gender, nationality) information about the person, the name of the landmark (e.g., Taj Mahal), and the language of the caption as input. The Moderator Agent then assigns tasks to the Social agents, instructing them to focus on the visually relevant aspects of the input information.

**Social Agents.** The Social Agents interact by asking each other relevant questions to create an image caption according to the information provided by the Moderator Agent. Each Social Agent assumes a *persona*: the first agent represents the culture of the person in the image, the second agent represents the age and gender of the person, and the last agent represents the historical landmark. Each agent generates an initial description of their persona. Then, by interacting through multiple rounds of question-answering conversations, each agent creates a more comprehensive image description.

[1]All demographics are listed in Appendix Table 1.

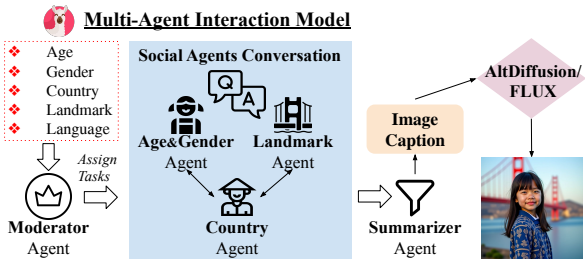



Figure 2: Overview of MosAIG, our framework for Multi-Agent Image Generation. The framework includes a multi-agent interaction model that generates an image caption from demographic information (person age, gender, country, landmark, and caption language), which is then used by an image generation model to create a multicultural image of a landmark and a person.

**Summarizer Agent.** The Summarizer Agent collects the three descriptions from the Social Agents and summarizes them into a final image caption with a maximum length of 77 tokens.

**Social Agents Conversation.** At the start, the three Social Agents—Country Agent, Landmark Agent, and Age-Gender Agent—receive demographic information and tasks from the Moderator Agent. The Country Agent processes nationality information and describes traditional attire, which is then evaluated by the Age-Gender Agent (e.g., "Is this attire suitable for a young female?"). Adjustments, such as modifying the color or style of a garment to suit the individual's age, are made accordingly. The Landmark Agent describes the landmark architecture, and its descriptions are refined based on feedback from the Country Agent (e.g., "How do Vietnamese visitors typically interact with this landmark?"), ensuring cultural authenticity. The Age-Gender Agent generates demographic descriptions, which are cross-checked with the Country Agent to ensure culturally appropriate accessories and mannerisms. After two rounds of conversation, the agents enhance and refine the descriptions with culturally sensitive and contextually rich details. Once the iterative improvement process is complete, the refined descriptions are passed to the Summarizer Agent, which condenses them into a final 77-token prompt capturing the cultural and contextual nuances. The prompts and implementation details are provided in Figure 9 and Appendix C.

### 3.2 Image Generation Models

We evaluate our generated image captions using two different state-of-the-art image generation models: AltDiffusion (Ye et al., 2024) and FLUX (Labs, 2024).

**AltDiffusion.** AltDiffusion[2] (Ye et al., 2024) is one of the very few multilingual open-source image generation models. The model aligns multilingual language models with diffusion models to generate high-quality images from text across multiple languages. The model builds on CLIP (Radford et al., 2021), replacing its text encoder with XLM-R (Conneau et al., 2020) and employing a two-stage training process that combines teacher learning and contrastive learning. AltDiffusion supports 18 different languages; we select five—English, German, Hindi, Spanish, and Vietnamese—based on the annotators' expertise. The model processes text inputs with a maximum length of 77 tokens.

**FLUX.** FLUX.1-dev[3] (Labs, 2024) is a state-of-the-art, widely used, open-source text-to-image model designed for English-language prompts. Due to computational constraints, we employ Flux.1 Lite[4] (Daniel Verdú, 2024), an 8B-parameter transformer model, more efficient variant distilled from FLUX.1-dev.

### 3.3 Simple vs. Multi-Agent Image Generation

Simple models generate images based on predefined captions, whereas multi-agent models utilize dynamically generated captions derived from multi-agent interactions. For instance, when provided with demographic details such as "Vietnamese" (nationality), "child" (age), "female" (gender), "Golden Gate Bridge" (landmark), and "English" (caption language), the resulting image captions differ between the two approaches. Multi-agent models generate captions that provide richer contextual information, including detailed descriptions of the landmark's architecture and surroundings, as well as a more nuanced depiction of the person's appearance, particularly focusing on clothing and facial features, as shown below[5].

**Simple caption:** *A Vietnamese girl wearing traditional attire, standing in front of the Golden Gate Bridge.*

**Multi-agent caption:** *A 12-year-old Vietnamese girl in Áo Dài, standing on the Golden Gate Bridge, with the San Francisco Bay's blue waters and the bridge's orange-red towers in the background.*

## 4 Evaluation and Results

We employ both automated metrics and human evaluation to provide a holistic and comprehensive assessment of the generated images.

### 4.1 Evaluation Metrics

We adopt a set of automated evaluation metrics that assess text-to-image generation along five complementary dimensions: **Alignment**, **Quality**, **Aesthetics**, **Knowledge**, and **Fairness**. Together, these metrics capture both technical properties of generation—such as semantic correspondence and visual fidelity—as well as socially situated aspects, including representational consistency across demographic groups (Lee et al., 2024). Given the known limitations of any single automatic metric, we combine multiple evaluators and complement them with human judgment.

**Alignment.** We measure text–image alignment using CLIPScore (Hessel et al., 2021), a widely adopted, reference-free metric that computes cosine similarity between joint text and image embeddings and enables scalable evaluation. While CLIPScore provides a useful proxy for semantic correspondence, it does not capture all aspects of visual grounding or compositional correctness. Accordingly, we interpret alignment scores cautiously and complement them with human evaluation, and we encourage future work to incorporate image-based classifiers for more direct assessment of visual attribute realization (see Limitations). CLIPScore values range from $-1$ to $+1$, with higher values indicating stronger alignment.

**Quality.** We assess image quality using the Inception Score (IS) (Salimans et al., 2016), which evaluates both visual fidelity and output diversity based on predictions from an Inception-v3 classifier. Lower scores typically reflect poor realism or limited variation, while higher scores indicate more realistic and diverse images. Although IS does not directly assess semantic correctness, it provides a complementary signal for overall visual plausibility.

**Aesthetics.** Aesthetic quality captures visual appeal beyond semantic correctness, including sharpness, color harmony, composition, and overall clarity. We use a SigLIP-based aesthetic predictor[6], which assigns scores on a 1–10 scale. This metric prioritizes perceptual attributes and may be less sensitive to semantic or cultural accuracy, making it particularly informative when interpreted alongside alignment and knowledge metrics.

**Fairness.** We evaluate fairness as the consistency

[2] https://huggingface.co/BAAI/AltDiffusion-m18

[3] https://huggingface.co/black-forest-labs/FLUX.1-dev

[4] https://huggingface.co/Freepik/flux.1-lite-8B-alpha

[5] All the captions are shown in our code repository.

[6] https://github.com/discus0434/aesthetic-predictor-v2-5

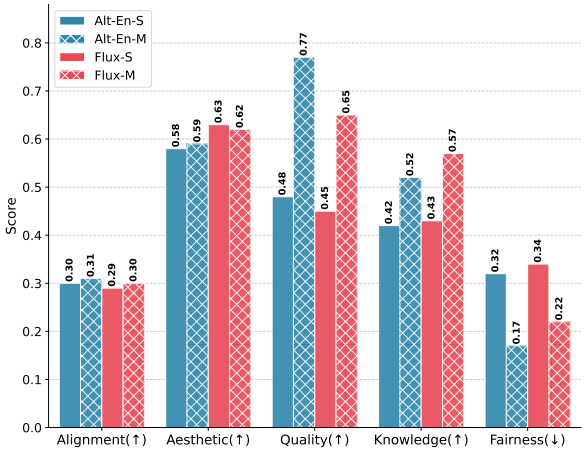



Figure 3: Multi-agent prompting improves *Quality*, *Knowledge*, and *Fairness* relative to simple prompts, while achieving comparable performance in *Alignment* and *Aesthetics*. Scores are normalized to [0–1]; higher is better except for Fairness.

of model performance across demographic substitutions. Following prior work, we modify captions by changing demographic attributes such as *gender*, *age*, or *nationality*, while keeping all other elements fixed. Given an original caption–image pair $(c, I)$, we construct a modified caption $c'$ (e.g., replacing "boy" with "girl") and generate a corresponding image $I'$. Fairness is measured as the absolute difference in alignment:

$$\Delta S = |S(c, I) - S(c', I')|.$$

Lower values of $\Delta S$ indicate more consistent behavior across demographic groups, while higher values suggest potential representational disparities. This metric captures relative changes rather than absolute bias and is best interpreted comparatively across models.

**Knowledge.** We assess world knowledge by evaluating a model's sensitivity to landmark identity. Given a caption–image pair $(c, I)$, we replace the referenced historical landmark in the caption while keeping the image fixed, yielding $(c', I)$. We compute:

$$\Delta S = S(c, I) - S(c', I).$$

A model with stronger landmark knowledge should exhibit larger alignment differences when the caption references an incorrect landmark.

### 4.2 Multi-Agent Interaction Results

Figure 3 compares multi-agent and simple prompting strategies across five evaluation dimensions. Overall, multi-agent prompting yields consistent improvements in **Quality**, **Knowledge**, and **Fairness**, while achieving comparable performance in **Alignment** and **Aesthetics**.

The largest gains are observed in **Quality**. Multi-agent models achieve substantially higher scores than simple prompts (0.77 vs. 0.48 for `Alt-En` and 0.65 vs. 0.45 for `Flux-En`), indicating more visually coherent and photorealistic outputs. We attribute these improvements to richer prompt composition that more explicitly specifies culturally grounded visual details, which appears to reduce under-specified or visually inconsistent generations. Across both prompting strategies, `Alt` consistently attains higher Quality scores than `Flux`, likely reflecting differences in background sharpness and image fidelity between the underlying generation models. In contrast, **Aesthetic** scores remain largely unchanged. This suggests that multi-agent prompting primarily affects semantic and compositional correctness rather than stylistic attributes emphasized by aesthetic predictors, which tend to prioritize surface-level visual appeal.

Multi-agent prompting also improves **Knowledge** and **Fairness**. Knowledge scores increase from 0.42 to 0.52 for `Alt-En` and from 0.43 to 0.57 for `Flux-En`, indicating stronger sensitivity to landmark-specific information when cultural context is more explicitly specified. Fairness scores—where lower values indicate smaller performance disparities—are substantially reduced (0.17 vs. 0.32 for `Alt-En` and 0.22 vs. 0.34 for `Flux-En`), suggesting more consistent behavior across demographic substitutions. These results indicate that decomposing cultural and demographic cues during prompt construction can mitigate uneven performance across social groups.

Improvements in **Alignment** are more modest and not statistically significant in aggregate. However, disaggregated analysis reveals consistent gains across several demographic dimensions, including *adults* (0.30 vs. 0.27), *females* (0.31 vs. 0.28), and multiple countries such as *Germany*, *India*, and *Vietnam* (see Appendix E). This pattern suggests that richer cultural specification can improve semantic correspondence for particular population groups, even when overall alignment scores remain similar.

### 4.3 Ablation Studies

We also perform ablation studies to assess `MosAIG's` performance across demographics.

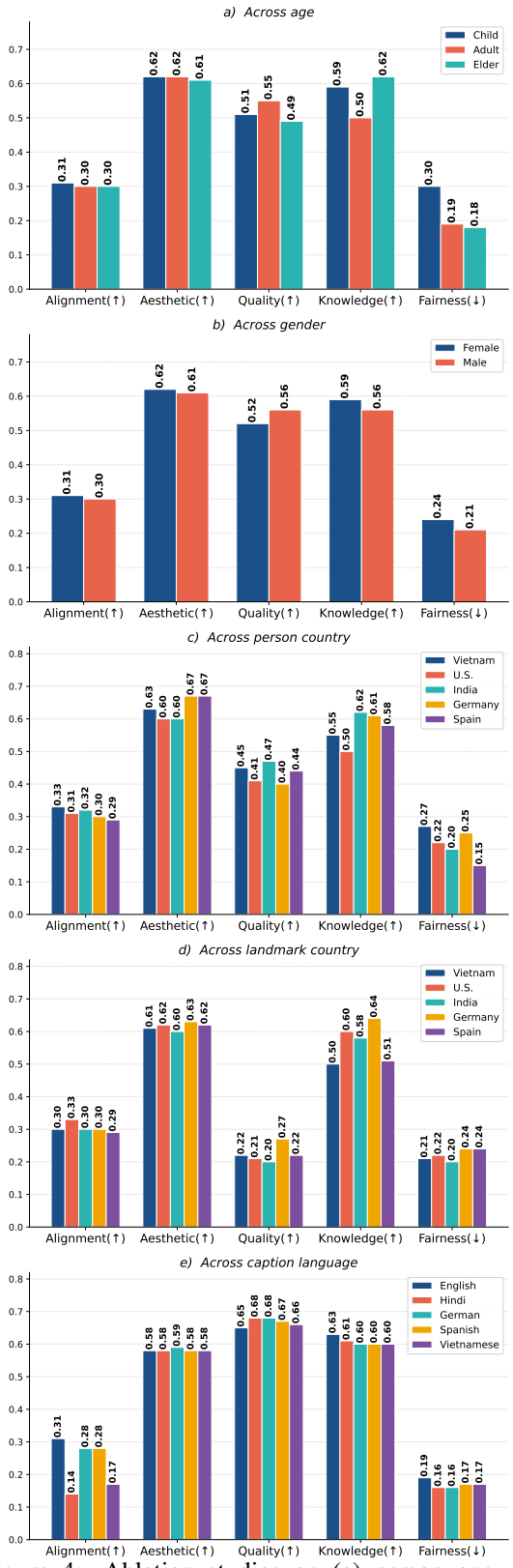



Figure 4: Ablation studies on (a) person age, (b) person gender, (c) person country, (d) landmark country, (e) caption language using the best overall model, the `Multi-agent English Flux-M` (a-d) and `Multi-agent Multilingual Alt-M` (e). Performance across all five metrics—Alignment, Aesthetic, Quality, Knowledge, and Fairness—reveals significant variation across these demographic categories.

**a) Person Age.** Figure 4 a) shows that Image Quality varies by age group, with Adults achieving the highest quality (0.55), followed by Children (0.51) and Elders (0.49). The model is also fairer when depicting Elders (0.18) and Adults (0.19) compared to Children (0.30).

**b) Person Gender.** Figure 4 b) shows that Knowledge and Image Quality varies by gender, with Males achieving higher quality (0.56) than Females (0.52). However, the model is fairer when depicting Males (0.21) than Females (0.24). The other metrics remain consistent across both groups.

**c) Person Country.** Figure 4 c) shows that model performance varies by person's country. Alignment is highest for Indian people (0.32) and lowest for Spanish people (0.29). Similarly, Image Quality is highest for Indian people (0.47) and lowest for German people (0.41). The model is also fairest when depicting Spanish (0.15) and least fair for Vietnamese (0.27).

**d) Landmark Country.** Figure 4 d) shows that model performance varies by landmark country. The most notable difference is in the Knowledge metric, with German landmarks being the most well-known (0.64), followed by U.S. (0.60), Indian (0.54), Vietnamese (0.50), and Spanish (0.51). Alignment is highest for U.S. landmarks (0.33) and lowest for Spanish landmarks (0.29).

**e) Caption Language.** Figure 4 e) shows that model performance varies by caption language, with English achieving the highest Alignment (0.31) and Knowledge (0.63), while Hindi and Vietnamese score the lowest (0.14 and 0.43, respectively). This disparity may stem from differences in training data availability, as model performance moderately correlates with dataset size (Pearson coefficient: 0.5), estimated from CommonCrawl (Wenzek et al., 2020).

As shown in Figure 5, English models (Alt-En-S, Alt-En-M) achieve substantially higher Alignment (0.30 vs. 0.20) compared to non-English models (Alt-NonEn-S, Alt-NonEn-M), while performing comparably in Aesthetics and Quality. Notably, non-English models achieve higher Knowledge (0.62 vs. 0.52) and lower Fairness scores (0.17 vs. 0.31), indicating more consistent behavior across demographic substitutions, suggesting that multilingual prompts may encode more culturally specific information despite lower overall alignment.

**f) Intersectionality.** Examining a single demographic category, such as race or gender, may over-

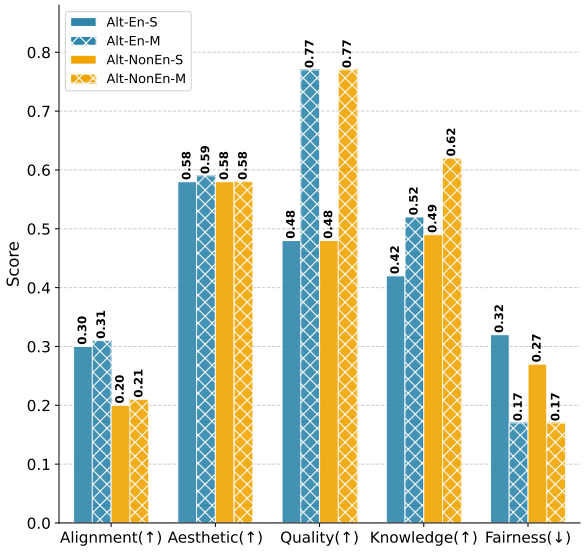



Figure 5: English vs. Multilingual Performance. Models with English captions as input (Alt-En-S, Alt-En-M) achieve higher scores than non-English (Alt-NonEn-S, Alt-NonEn-M) in Alignment (0.30 vs. 0.20), while performing comparably across Aesthetics and Quality metrics. Knowledge and Fairness performance is higher for non-English models

look nuanced inequalities (Field et al., 2021). To address this, we analyze the intersectionality of age and gender, person and landmark country, and language and person country. We measure Alignment and analyze other metrics across various demographic intersections, as detailed in Appendix E.1.

**Age and Gender.** Figure 6 (right) shows that Alignment performance varies by gender for generating adult images, with males having a lower score (0.29) compared to females (0.31). The performance for child and elder categories remains consistent across gender.

**Person and Landmark Country.** Figure 6 (left) illustrates Alignment across Person and Landmark Country. We expected higher performance when the person and landmark originate from the same country, suggesting challenges in cross-cultural representation. However, results vary by country. For instance, the highest alignment occurs when Indian or Vietnamese people visit U.S. landmarks (0.34), comparable to U.S. people at U.S. landmarks (0.33). In contrast, the lowest alignment is observed when Vietnamese people visit Spanish landmarks (0.28). All metrics are detailed in Appendix E.1.

**Language and Country.** Figure 7 shows Alignment across Person Country and Caption Language. English, Spanish, and Vietnamese captions achieve the highest performance ($\sim$ 0.3) with minimal variation across person countries. However, Hindi captions perform best for Indian people (0.17) and worst for Spanish and U.S. people (0.13). This suggests that, for certain languages, the interaction between caption language and the depicted person's culture influences Alignment in image generation.

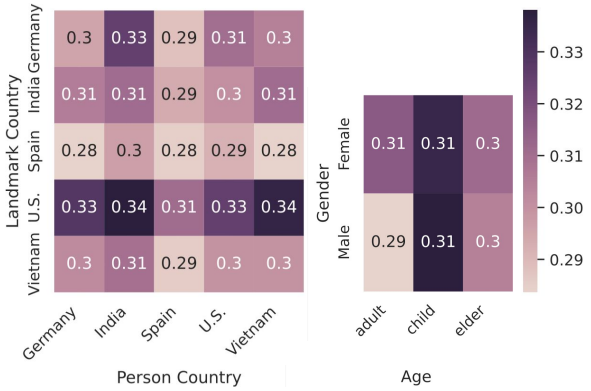



Figure 6: Alignment with best overall model, `Flux-M`, over person-landmark (left) and gender-age (right).

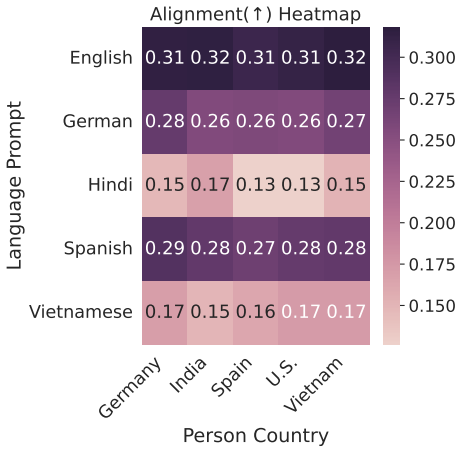



Figure 7: Alignment with best multilingual model, `Alt-M`, over image caption language and person country.

## 4.4 Human Evaluation and Error Analysis

Two annotators evaluate a subset of 300 images, covering all demographics (age, gender, country, landmark) and model settings (Alt-S, Alt-M, Flux-S, Flux-M). They assess the generated images based on three key metrics: Alignment, Quality, and Aesthetics. Following Lee et al. (2024), Quality is measured in terms of photorealism, while Aesthetics is evaluated based on subject clarity and overall visual appeal. Annotator agreement is measured using weighted Cohen's Kappa for ordinal values (Cohen, 1968), yielding scores between 0.5 and 0.6 across all three metrics, indicating moderate agreement. The complete set of human evaluation questions, along with the annotation interface, is detailed in Appendix D.

**Most Common Errors.** Across models, errors

primarily involve incorrect backgrounds and failures in human rendering. For `Flux-M`, background inaccuracies are most frequent (38/75), followed by deviations from prompt details and occasional human rendering errors (5/75), such as missing fingers or misplaced cultural markers; landmark errors are comparatively rare (2/75). In contrast, `Flux-S` exhibits substantially more landmark omissions (15/75) and increased human rendering errors (10/75), particularly for traditional attire and facial features. The `Alt` models show more severe artifacts overall, with frequent background errors (55/75), pronounced body distortions, and multiplicity errors. While `Alt-M` reduces culturally related errors (2/75), it still exhibits body distortions (15/75).

### 4.5 Qualitative Results

In Figure 8, we compare the images generated by our multi-agent framework (`Flux-M` and `Alt-M`) with those from simpler models (`Flux-S` and `Alt-S`). The second column presents images generated with Vietnamese captions using the multilingual models (`Alt-Vi-S`, `Alt-Vi-M`). Compared to the simple models, the multi-agent models perform better at generating landmarks and people. However, they still miss important details about people, such as *a person looking up*, *curly hair*, or *hair tied back with a nón lá hat*. Notably, body distortions are more pronounced in the Alt-S model. While the Flux model produces more accurate backgrounds, they tend to be blurrier compared to those in the Alt model.

A manual error analysis of 300 images across all demographics highlights the need for further improvements, particularly in rendering body structures and backgrounds. Models also exhibit systematic tendencies toward stereotypical representations for certain demographics, for instance defaulting to generic or exaggerated traditional attire regardless of age group, or rendering culturally ambiguous poses and accessories. Cross-cultural grounding challenges are further reflected in performance disparities across country-landmark pairs: as shown in Figure 5, Vietnamese people at Spanish landmarks consistently yield the lowest alignment scores (0.28), suggesting that certain cultural combinations remain particularly difficult for current models to represent accurately. Failure modes in multilingual captions are similarly evident in Appendix E.3, where Hindi and Vietnamese captions show increased rendering errors and cultural inaccuracies compared to English. Additional results across demographics are in Appendix E.2.

## 5 Lessons Learned and Actionable Steps

Our study surfaces insights into the challenges of *multicultural text-to-image generation* and highlights several directions for improving cultural grounding, demographic coverage, and evaluation practices in future models.

**Richer Cultural and Demographic Captions.** Multicultural generation benefits from richer prompts that clearly articulate cultural context, demographic attributes, and landmark-specific details. By integrating diverse perspectives through collaboration, multi-agent models enhance alignment, aesthetics, quality, and knowledge (Section 4.2). Future research should focus on refining multi-agent frameworks to further enhance alignment and representational diversity. Our work can also be extend to evaluate a broader range of cultural interactions—such as social activities, rituals, and everyday practices—to better assess reasoning and action-based image generation.

**Multilingual Support Remains a Bottleneck.** We observe systematic performance gaps between English and non-English prompts, with English-language generations consistently achieving higher alignment scores (Figure 4e). These disparities point to limitations in current multilingual training and evaluation practices. Improving multilingual coverage—both in training data and model architectures—is essential for achieving equitable performance across languages and cultures.

**Develop Better Evaluation Metrics.** Automated metrics do not always align with qualitative judgments in multicultural settings, particularly when visually plausible context inflates scores despite incorrect culturally salient elements (e.g., accurate surroundings but an incorrect landmark; Section 4.4). More reliable evaluation requires metrics that place greater emphasis on landmark identity, demographic attributes, and compositional correctness. Targeted automated measures, complemented by human evaluation, remain essential for accurately assessing multicultural image generation.

**Explore Stronger and More Diverse Baselines.** MosAIG establishes multi-agent collaboration as one principled strategy for multicultural prompt composition. Future work should compare against complementary approaches such as single-agent chain-of-thought (CoT) prompt expansion and


Figure 8: Comparison of generated images and captions from multi-agent (`Flux-M`, `Alt-M`) and simple models (`Flux-S`, `Alt-S`). The first two columns show where multi-agent models perform better, while the last column shows where simpler models excel. The second column depicts images generated with *Vietnamese* captions using the multilingual model `Alt` (`Alt-Vi-S`, `Alt-Vi-M`). Demographic keywords are **bolded**, and errors are marked in red.

template-based demographic augmentation, which would help disentangle the specific contribution of multi-agent decomposition from the general benefit of richer prompt specification.

**Finer-Grained Compositional Evaluation.** While CLIPScore provides a scalable alignment proxy, future work should incorporate more direct evaluation approaches such as detection-based attribute binding checks, explicit landmark classification over generated images, and LLM-as-judge VQA methods. Such measures would provide finer-grained diagnostics of cultural and demographic grounding, particularly for compositional correctness across person-landmark pairs.

## 6 Conclusion

In this paper, we introduce *multicultural text-to-image generation* as a new task for evaluating how models depict people and landmarks from different cultural backgrounds within a single image. We release MOSAIG the first benchmark for this setting, comprising 9,000 images across five countries, three age groups, two genders, 25 historical landmarks, and five languages. Our automated and human evaluations reveal substantial variation across demographics, languages, and cultural configurations, highlighting persistent gaps in multilingual and cross-cultural generation. Overall, our findings emphasize the importance of explicit cultural and demographic grounding for improving image quality, factual correctness, and representational consistency. We release our dataset and evaluation framework to support multicultural text-to-image generation: https://github.com/AIM-SCU/MosAIG.

## Limitations and Ethical Considerations

**Scope of Demographic Coverage.** Our study considers a limited set of demographic attributes, focusing on binary gender categories, three coarse age groups (child, adult, elder), and five countries and languages. These choices necessarily simplify the rich diversity of gender identities, life stages, and cultural experiences, and limit our ability to assess performance for underrepresented communities. We view this design as a proof of concept for studying multicultural image generation in a controlled setting. Importantly, our dataset and pipeline are fully open-source and designed to be easily extended to additional countries, languages, age groups, and gender identities in future work.

**Challenges in Modeling Cultural Identity.** Our approach relies on structured prompts to approximate cultural and demographic context, but identity is inherently complex and cannot be fully captured through high-level attributes such as nationality, language, age, or gender alone (Saha et al., 2025). Defining culture primarily through national affiliation risks overlooking substantial intra-cultural variation and lived experience. Future work should incorporate richer contextual dimensions—such as historical background, social practices, and personal narratives—to enable more nuanced and authentic representations.

**Limitations of Automated Alignment Metrics.** We rely on CLIPScore as a scalable, reference-free measure of text–image alignment, but this metric has several limitations. Its coarse-grained contrastive training makes it insensitive to fine-grained compositional errors, such as incorrect spa-

tial relationships or misattributed attributes, and it may assign high scores to images that contain the correct objects in incorrect configurations (Hessel et al., 2021). CLIPScore is also largely insensitive to word order and linguistic phenomena such as negation, and exhibits biases toward salient or centrally positioned objects, reducing its reliability for complex, multi-object prompts (Castro et al., 2023; Abbasi et al., 2025). We therefore encourage future work to complement CLIPScore with image-based classifiers and targeted evaluation methods that more directly assess visual grounding and demographic fidelity (Hu et al., 2023).

**Limitations of Human Evaluation Coverage.** Our human evaluation is conducted on a carefully stratified sample of 300 images covering all demographic categories, model settings, and languages, ensuring representative coverage across the experimental matrix. Expanding human evaluation to a larger sample would provide additional validation of demographic specific findings, particularly for fine-grained combinations of age, gender, country, and landmark. This is a non-trivial undertaking, as it requires annotators proficient in all five languages covered in our benchmark, and we highlight this as a valuable direction for future work to build upon.

## References


Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. 2025. Analyzing clip's performance limitations in multi-object scenarios: A controlled high-resolution study. *ArXiv*, abs/2502.19828.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Longju Bai, Angana Borah, Oana Ignat, and Rada Mihalcea. 2025. The power of many: Multi-agent multimodal models for cultural image captioning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2970–2993, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Castles, Hein de Haas, and Miller Mark J. 2103. *The Age of Migration: International Population Movements in the Modern World*.

Santiago Castro, Oana Ignat, and Rada Mihalcea. 2023. Scalable performance analysis for vision-language models. In *STARSEM*.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Javier Martín Daniel Verdú. 2024. Flux.1 lite: Distilling flux1.dev for efficient text-to-image generation.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. 2024. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.

Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Timo Ropinski, and 1 others. 2024. Evaluating text to image synthesis: Survey and taxonomy of image quality metrics. *arXiv preprint arXiv:2403.11821*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: cultural competence in text-to-image models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, and 1 others. 2024. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2025. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Bingshuai Liu, Longyue Wang, Chenyang Lyu, Yong Zhang, Jinsong Su, Shuming Shi, and Zhaopeng Tu. 2024. On the cultural gap in text-to-image generation. In *ECAI 2024*, pages 930–937. IOS Press.

Rada Mihalcea, Oana Ignat, Longju Bai, Angana Borah, Luis Chiruzzo, Zhijing Jin, Claude Kwizera, Joan Nwatu, Soujanya Poria, and Thamar Solorio. 2025. Why ai is weird and shouldn't be this way: towards ai for everyone, with everyone, by everyone. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *arXiv preprint arXiv:2411.00860*.

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.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, Marzieh Fadaee, Sara Hooker, Antoine Bosselut, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, and 40 others. 2025. INCLUDE: evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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 56 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Sougata Saha, Saurabh Kumar Pandey, and Monojit Choudhury. 2025. Meta-cultural competence: Climbing the right hill of cultural awareness. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8025–8042, Albuquerque, New Mexico. Association for Computational Linguistics.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. 2024. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37:128374–128395.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Xiaojun Wu, Dixiang Zhang, Ruyi Gan, Junyu Lu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. 2024. Taiyi-diffusion-xl: advancing bilingual text-to-image generation with large vision-language model support. *arXiv preprint arXiv:2401.14688*.

Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. 2024. Altdiffusion: A multilingual text-to-image diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6648–6656.

# A Appendix

# B Data

# C Multicultural Image Generation

## C.1 Implementation Details

The Summarizer Agent and each Social Agent are initialized as different instances of a LLaMA model[7] (Touvron et al., 2023). The Moderator Agent is a predefined function call. The agent conversation uses the CrewAI framework to establish an iterative feedback loop[8]. The implementation was carried out using an NVIDIA V100 GPU (32GB). More details can be found in Appendix C.

The multi-agent configuration processed 750 base prompts in approximately 45 minutes, while additional language variants (3,750 prompts in total) required 75 minutes using the Google Translation API. Two models—Flux and Alt-Diffusion—were used for image generation: Flux produced 750 images (768×768 pixels) in 2.5 hours with the settings: guidance scale: 4, inference steps: 30, seed: 11, averaging roughly 12 seconds per image. Alt-Diffusion was configured with the settings: guidance scale: 11, inference steps: 110, seed: 11000, and processed 3,750 images of the same resolution in 16 hours, averaging about 15 seconds per image. All processing times accounted for overhead related to model loading and image saving, ensuring consistency in image resolution (768×768 pixels) across both models.

# D Human Evaluation and Error Analysis

We rely on human annotators to assess a sample of the generated images based on three key metrics: Alignment, Quality, and Aesthetics. Following Lee et al. (2024), Quality is evaluated in terms of photorealism, while Aesthetics is assessed based on subject clarity and overall visual appeal. The complete set of human evaluation questions is outlined below. Annotators are provided with definitions (Table 2) and corresponding questions to guide their assessments. To determine whether the generated images meet their expectations, we ask annotators to rate them using a 5-point Likert scale.

**Alignment.** We ask the annotators to rate how well the image matches the description.

[7] https://huggingface.co/meta-llama/Llama-3.1-8B

[8] https://www.crewai.com/open-source

| Age | Gender | Country | Landmark |
|---|---|---|---|
| Child/ Adult/ Elder | Female/Male | Germany | Cologne Cathedral<br>Reichstag Building<br>Neuschwanstein Castle<br>Brandenburg Gate<br>Holocaust Memorial |
| | | India | Taj Mahal<br>Lotus Temple<br>Gateway of India<br>India Gate<br>Charminar |
| | | Spain | Sagrada Familia<br>Alhambra<br>Guggenheim Museum<br>Roman Theater of Cartagena<br>Royal Palace of Madrid |
| | | U.S. | White House<br>Statue of Liberty<br>Mount Rushmore<br>Golden Gate Bridge<br>Lincoln Memorial |
| | | Vietnam | Meridian Gate of Hu´<br>Independence Palace<br>One Pillar Pagoda<br>Ho Chi Minh Mausoleum<br>Thien Mu Pagoda |

Table 1: Demographics Overview: 3 Age groups, 2 Genders, 5 Countries, and 25 Landmarks

**How well does the image match the description?**

1. Does not match at all
2. Has significant discrepancies
3. Has several minor discrepancies
4. Has a few minor discrepancies
5. Matches exactly

**Quality.** We ask the annotators to rate how photorealistic the generated images are.

**Determine if the following image is AI-generated or real.**

1. AI-generated photo.
2. Probably an AI-generated photo, but photorealistic.
3. Neutral.
4. Probably a real photo, but with irregular textures and shapes.
5. Real photo.

**Aesthetics.** To evaluate the overall aesthetics, we ask annotators to provide a holistic assessment of the image's visual appeal by rating its aesthetic quality.

**How aesthetically pleasing is the image?**

1. I find the image ugly.
2. The image has a lot of flaws, but it's not completely unappealing.
3. I find the image neither ugly nor aesthetically pleasing.
4. The image is aesthetically pleasing and is nice to look at.
5. The image is aesthetically stunning. I can look at it all day.

| Conv. Round | Agent Role | Prompt |
|---|---|---|
| Round 1 | Country Agent | SYSTEM: You are a {nationality} person from {country} who knows the culture of this country well. USER: Provide a visual description of culturally appropriate traditional clothing, accessories, and colors, for the {nationality} person. Focus on specific materials, key cultural patterns, and symbolic colors. Your response must be under 25 words. \nASSISTANT: |
| | Landmark Agent | SYSTEM: You are a person who has visited {place} many times and know this landmark well. USER: Provide a visual description of its architectural features, colors, and environmental details. Your response must be under 25 words. \nASSISTANT: |
| | Age-Gender Agent | SYSTEM: You are a {age_gender_combined} and can describe traits of this person well. USER: Provide a visual description of attire, accessories, and physical details. Focus on skin, body, hair texture, and accessories. Your response must be under 25 words. \nASSISTANT: |
| Round 2 | Country Agent | SYSTEM: You are a {nationality} person from {country}. USER: Enhance the persona description by addressing: 'How would a person's clothing harmonize with the colors of {place}?'. Ensure cultural significance is highlighted. \nASSISTANT: |
| | Landmark Agent | SYSTEM: You are a person who knows {place} well. USER: Enhance the place description by addressing: 'What visual elements of {place} would complement the persona's attire?'. Limit to under 25 words. \nASSISTANT: |
| | Age-Gender Agent | SYSTEM: You are a {age_gender_combined}. USER: Enhance the age-gender description by addressing: 'What attire adjustments could reflect age-appropriate traits for a {nationality} {age_gender_combined}?'. Ensure specific details on attire and physical traits. \nASSISTANT: |
| Round 3 | Summarizer Agent | SYSTEM: You excel at crafting concise visual prompts. USER: Give a final prompt in a single line under 48 words and under 77 tokens strictly. Ensure the words {nationality} and {age_gender_combined} of the person and other descriptions with the {place} background are mentioned explicitly in the final prompt. \nASSISTANT: |

Figure 9: Our Multi-agent Framework Prompts

| Aspect | Definition |
|---|---|
| Alignment | Is the image semantically correct given the text (text-image alignment)? |
| Quality | Do the generated images look like real photographs? |
| Aesthetic | Is the image aesthetically pleasing? |
| Fairness | Does the model exhibit performance disparities across social groups (e.g., gender, dialect) |
| Knowledge | Does the model have knowledge about the world or domains? |

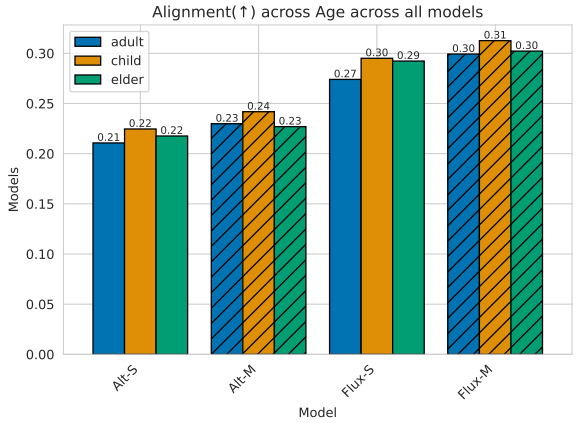
Table 2: Evaluation Aspects of Text-to-Image Models

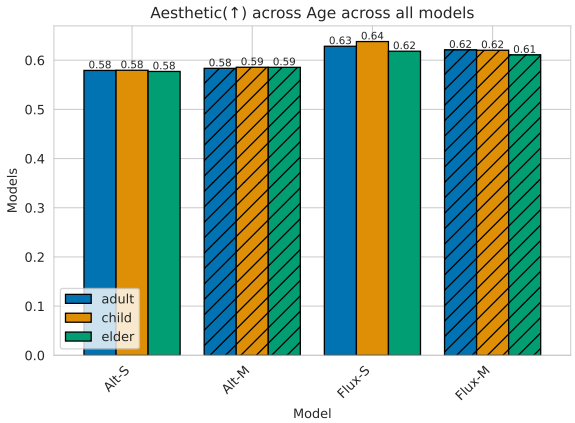
# E Results

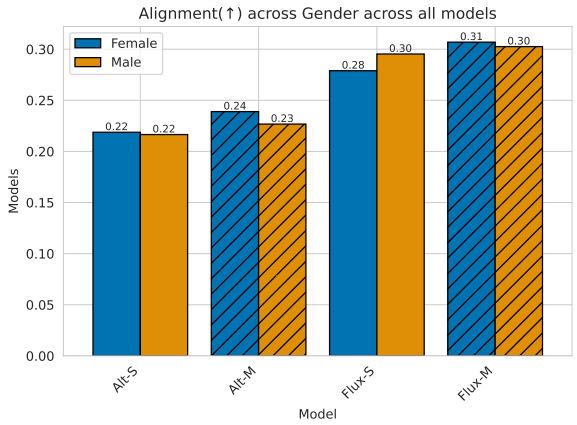
## E.1 Intersectionality

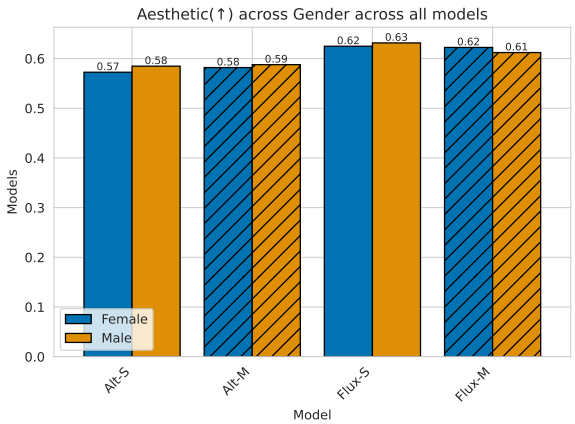
## E.2 Qualitative Results

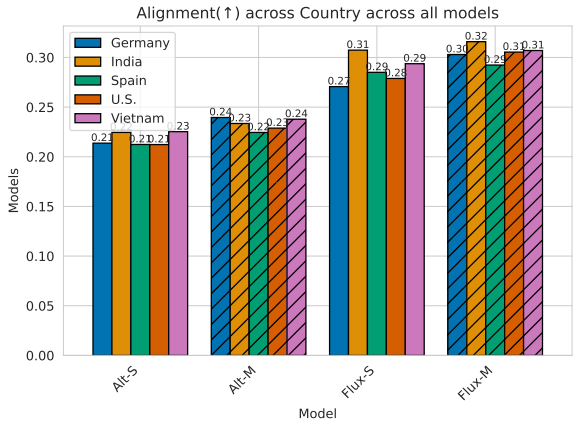

Figure 10: Human Annotation Interface for manually evaluating the images across all models.

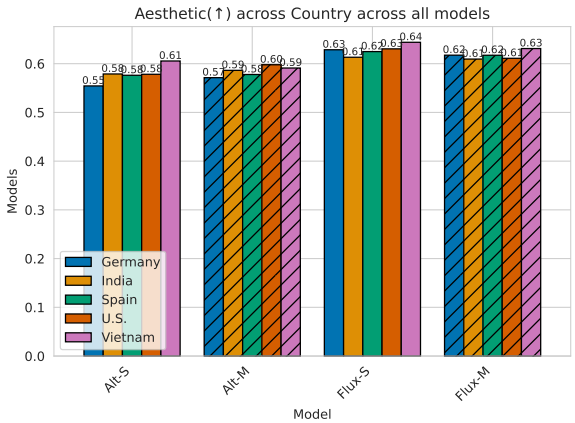

Alignment(↑) across Age across all models
adult
child
elder
Models
Alt-S
Alt-M
Flux-S
Flux-M
Model

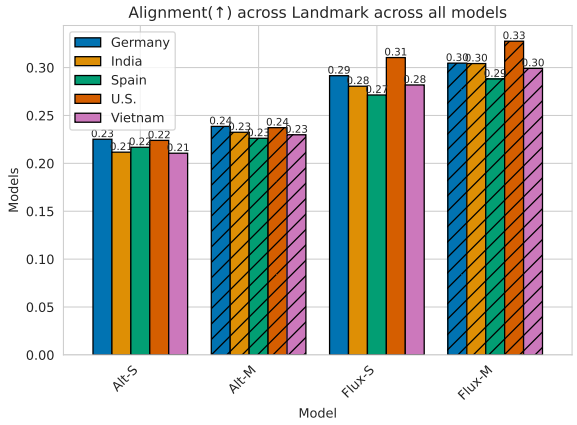

Aesthetic(↑) across Age across all models
adult
child
elder
Models
Alt-S
Alt-M
Flux-S
Flux-M
Model

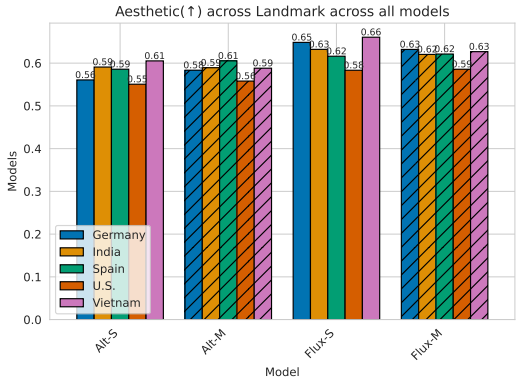

Alignment(↑) across Gender across all models
Female
Male
Models
Alt-S
Alt-M
Flux-S
Flux-M
Model

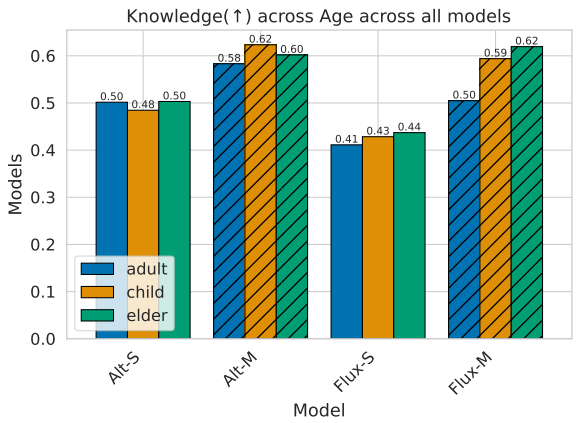

Aesthetic(↑) across Gender across all models
Female
Male
Models
Alt-S
Alt-M
Flux-S
Flux-M
Model

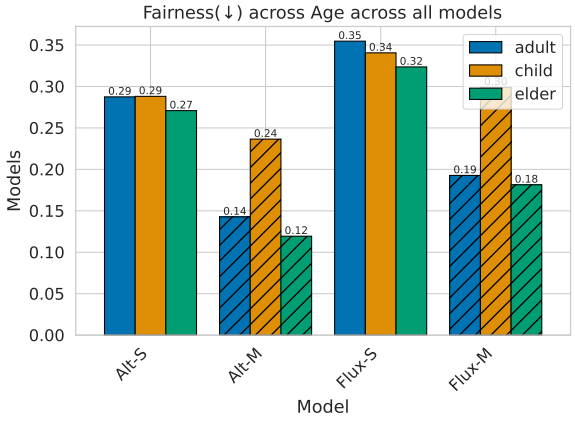

Alignment(↑) across Country across all models
Germany
India
Spain
U.S.
Vietnam
Models
Alt-S
Alt-M
Flux-S
Flux-M
Model

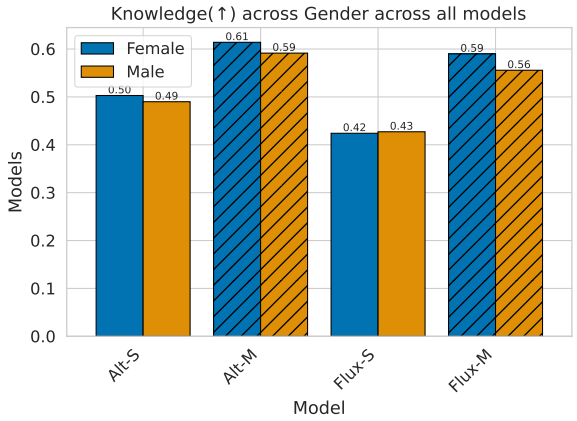

Aesthetic(↑) across Country across all models
Germany
India
Spain
U.S.
Vietnam
Models
Alt-S
Alt-M
Flux-S
Flux-M
Model

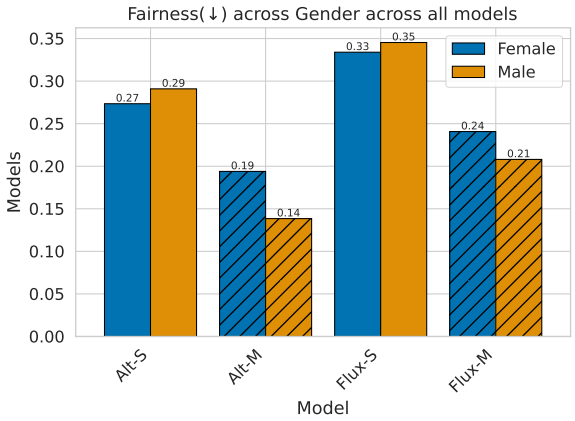

Alignment(↑) across Landmark across all models
Germany
India
Spain
U.S.
Vietnam
Models
Alt-S
Alt-M
Flux-S
Flux-M
Model

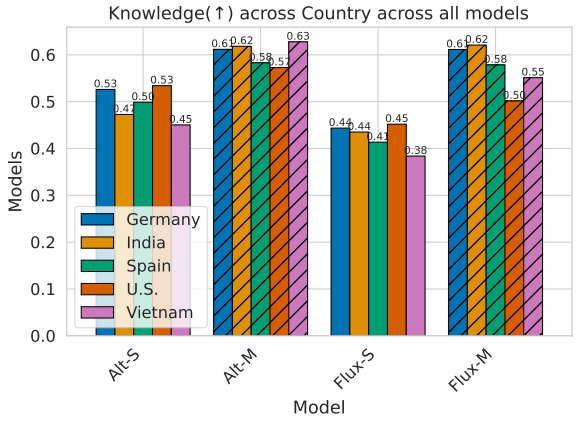

Aesthetic(↑) across Landmark across all models
Germany
India
Spain
U.S.
Vietnam
Models
Alt-S
Alt-M
Flux-S
Flux-M
Model

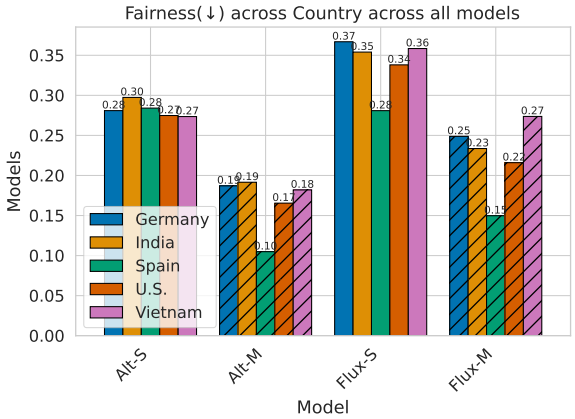

Knowledge(↑) across Age across all models
Models
Model
adult
child
elder
Alt-S
Alt-M
Flux-S
Flux-M

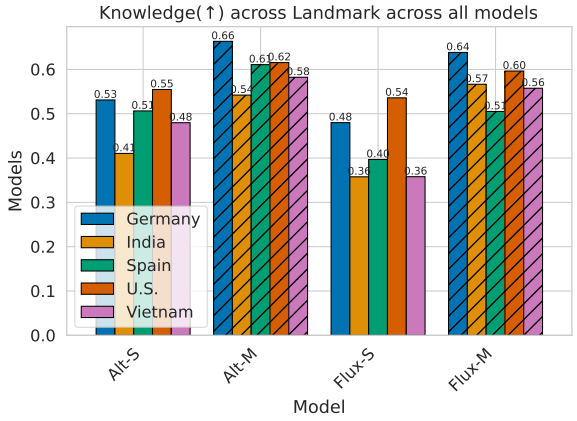

Fairness(↓) across Age across all models
Models
Model
adult
child
elder
Alt-S
Alt-M
Flux-S
Flux-M

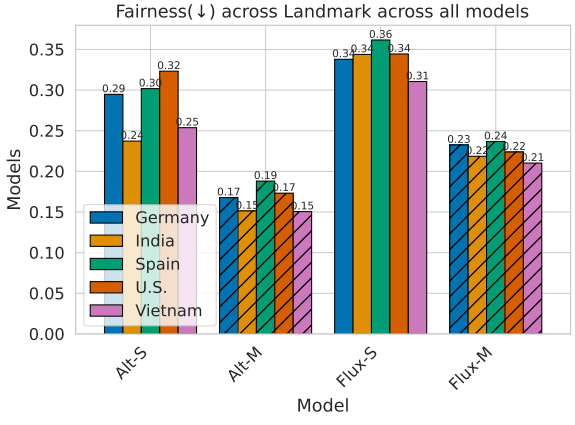

Knowledge(↑) across Gender across all models
Models
Model
Female
Male
Alt-S
Alt-M
Flux-S
Flux-M

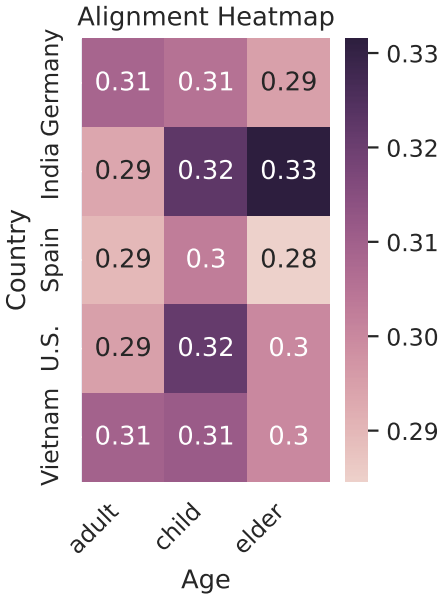

Fairness(↓) across Gender across all models
Models
Model
Female
Male
Alt-S
Alt-M
Flux-S
Flux-M

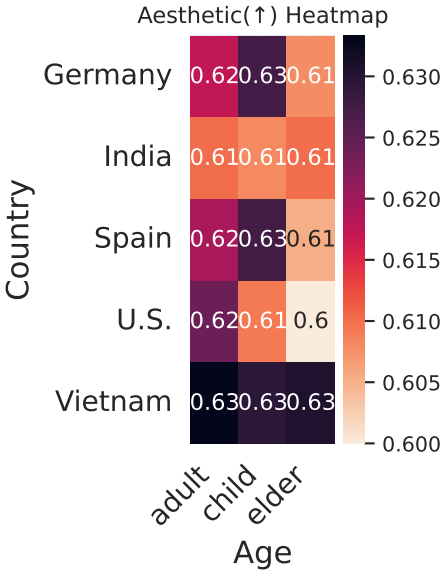

Knowledge(↑) across Country across all models
Models
Model
Germany
India
Spain
U.S.
Vietnam
Alt-S
Alt-M
Flux-S
Flux-M

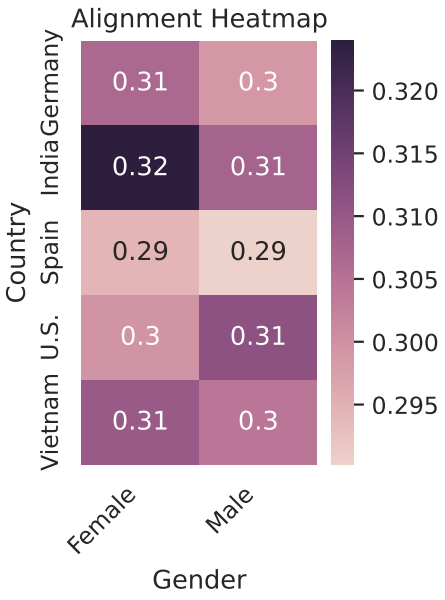

Fairness(↓) across Country across all models
Models
Model
Germany
India
Spain
U.S.
Vietnam
Alt-S
Alt-M
Flux-S
Flux-M

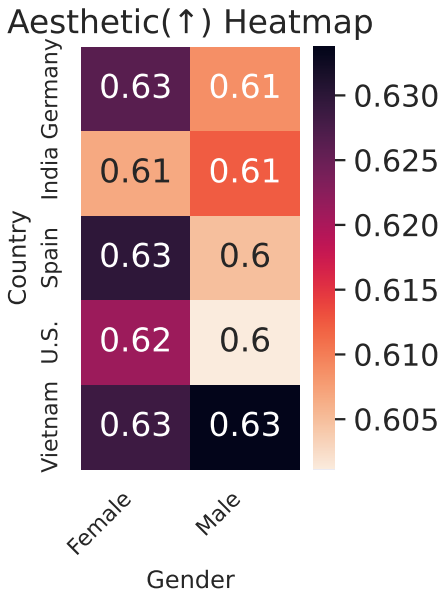

Knowledge(↑) across Landmark across all models
Models
Model
Germany
India
Spain
U.S.
Vietnam
Alt-S
Alt-M
Flux-S
Flux-M


Fairness(↓) across Landmark across all models
Models
Model
Germany
India
Spain
U.S.
Vietnam
Alt-S
Alt-M
Flux-S
Flux-M

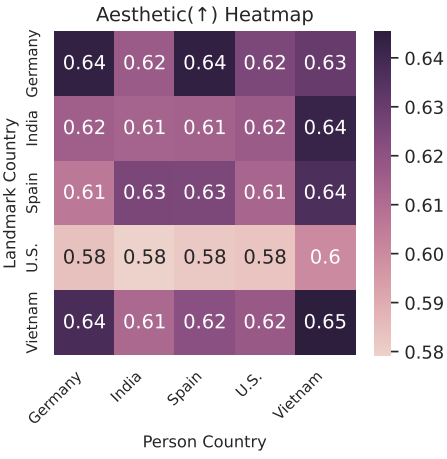

Alignment Heatmap
Country
Germany
India
Spain
U.S.
Vietnam
0.31 0.31 0.29
0.29 0.32 0.33
0.29 0.3 0.28
0.29 0.32 0.3
0.31 0.31 0.3
0.33
0.32
0.31
0.30
0.29
adult
child
elder
Age

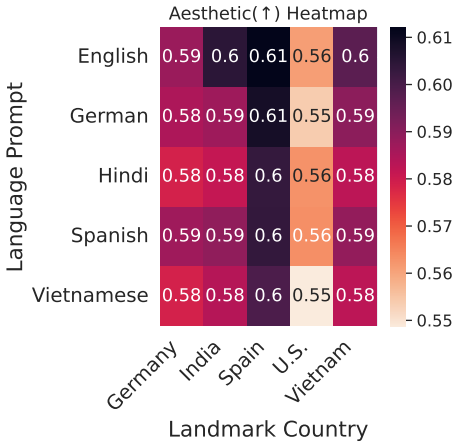

Aesthetic(↑) Heatmap
Country
Germany
India
Spain
U.S.
Vietnam
0.62 0.63 0.61
0.61 0.61 0.61
0.62 0.63 0.61
0.62 0.61 0.6
0.63 0.63 0.63
0.630
0.625
0.620
0.615
0.610
0.605
0.600
adult
child
elder
Age

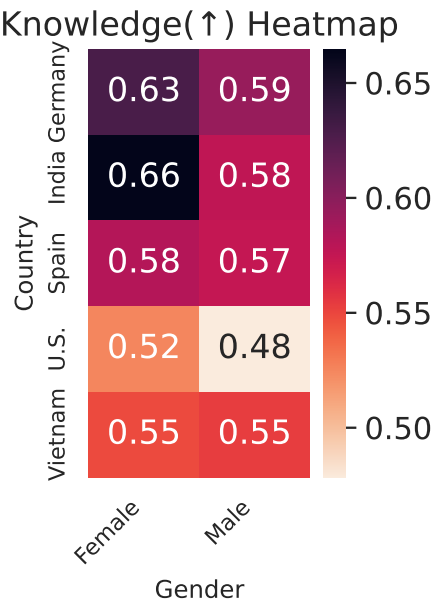

Alignment Heatmap
Country
Germany
India
Spain
U.S.
Vietnam
0.31 0.3
0.32 0.31
0.29 0.29
0.3 0.31
0.31 0.3
0.320
0.315
0.310
0.305
0.300
0.295
Female
Male
Gender

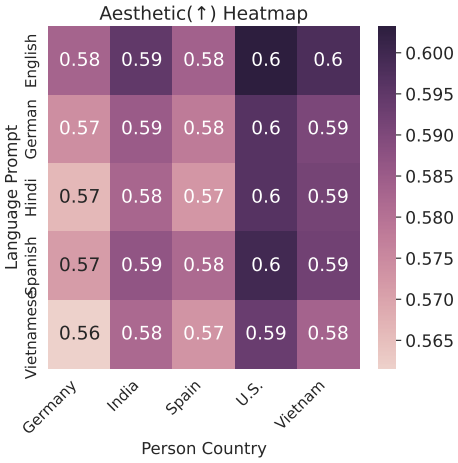

Aesthetic(↑) Heatmap
Country
Germany
India
Spain
U.S.
Vietnam
0.63 0.61
0.61 0.61
0.63 0.6
0.62 0.6
0.63 0.63
0.630
0.625
0.620
0.615
0.610
0.605
Female
Male
Gender

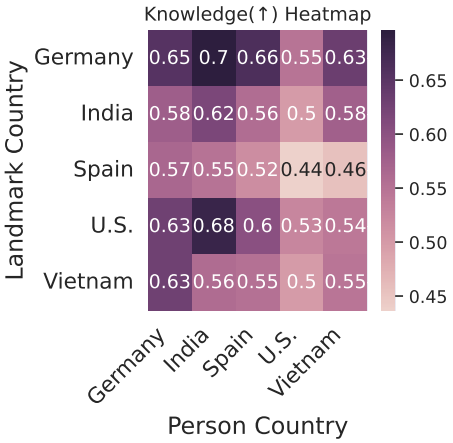

Alignment Heatmap
Language Prompt
English
German
Hindi
Spanish
Vietnamese
0.32 0.3 0.31 0.32 0.32
0.27 0.26 0.26 0.27 0.26
0.14 0.17 0.14 0.14 0.14
0.29 0.28 0.27 0.29 0.27
0.17 0.16 0.16 0.17 0.17
0.300
0.275
0.250
0.225
0.200
0.175
0.150
Germany
India
Spain
U.S.
Vietnam
Landmark Country

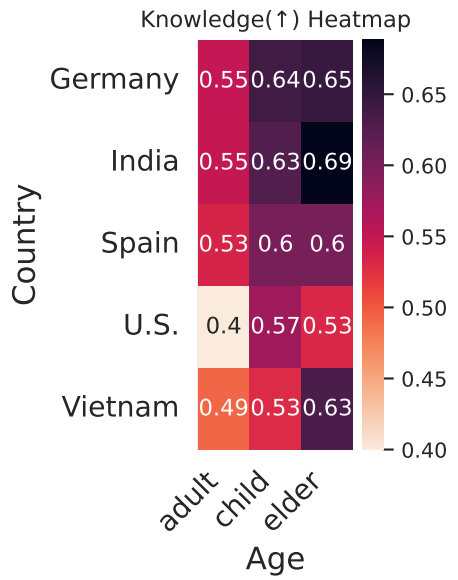

Aesthetic(↑) Heatmap
Landmark Country
Germany
India
Spain
U.S.
Vietnam
0.64 0.62 0.64 0.62 0.63
0.62 0.61 0.61 0.62 0.64
0.61 0.63 0.63 0.61 0.64
0.58 0.58 0.58 0.58 0.6
0.64 0.61 0.62 0.62 0.65
0.64
0.63
0.62
0.61
0.60
0.59
0.58
Germany
India
Spain
U.S.
Vietnam
Person Country

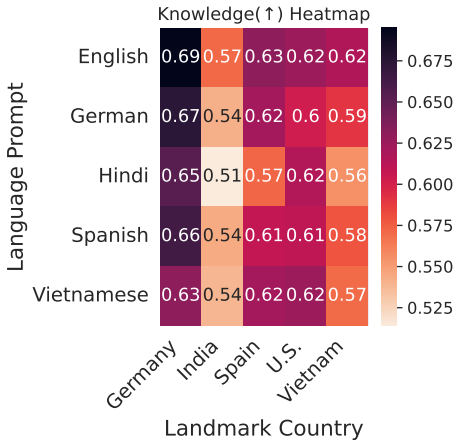

Aesthetic(↑) Heatmap
Language Prompt
English 0.59 0.6 0.61 0.56 0.6
German 0.58 0.59 0.61 0.55 0.59
Hindi 0.58 0.58 0.6 0.56 0.58
Spanish 0.59 0.59 0.6 0.56 0.59
Vietnamese 0.58 0.58 0.6 0.55 0.58
0.61 0.60 0.59 0.58 0.57 0.56 0.55
Germany India Spain U.S. Vietnam
Landmark Country

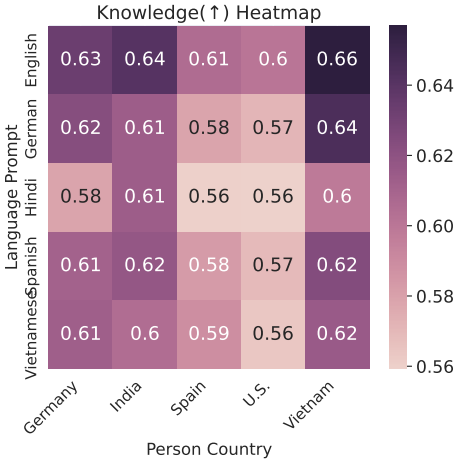

Knowledge(↑) Heatmap
Country
India 0.63 0.59
Germany 0.66 0.58
Spain 0.58 0.57
U.S. 0.52 0.48
Vietnam 0.55 0.55
0.65 0.60 0.55 0.50
Female Male
Gender

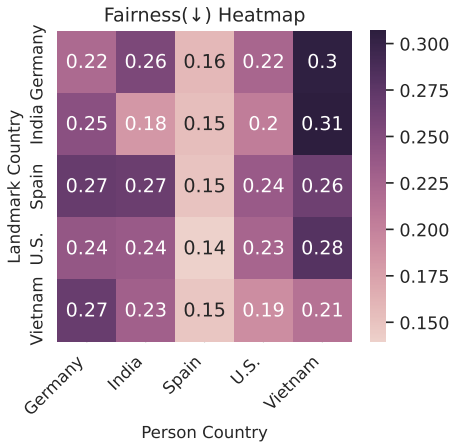

Aesthetic(↑) Heatmap
Language Prompt
English 0.58 0.59 0.58 0.6 0.6
German 0.57 0.59 0.58 0.6 0.59
Hindi 0.57 0.58 0.57 0.6 0.59
Spanish 0.57 0.59 0.58 0.6 0.59
Vietnamese 0.56 0.58 0.57 0.59 0.58
0.600 0.595 0.590 0.585 0.580 0.575 0.570 0.565
Germany India Spain U.S. Vietnam
Person Country

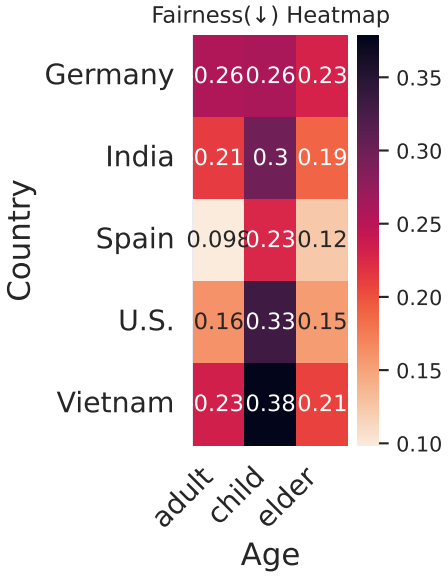

Knowledge(↑) Heatmap
Landmark Country
Germany 0.65 0.7 0.66 0.55 0.63
India 0.58 0.62 0.56 0.5 0.58
Spain 0.57 0.55 0.52 0.44 0.46
U.S. 0.63 0.68 0.6 0.53 0.54
Vietnam 0.63 0.56 0.55 0.5 0.55
0.65 0.60 0.55 0.50 0.45
Germany India Spain U.S. Vietnam
Person Country

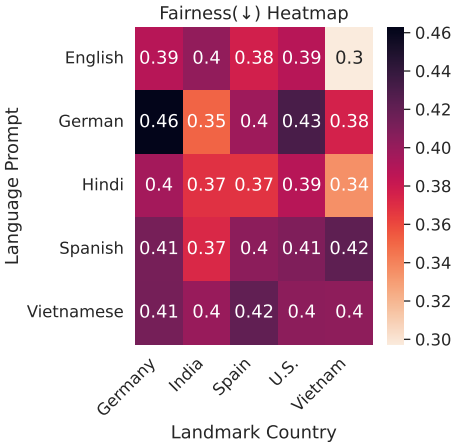

Knowledge(↑) Heatmap
Country
Germany 0.55 0.64 0.65
India 0.55 0.63 0.69
Spain 0.53 0.6 0.6
U.S. 0.4 0.57 0.53
Vietnam 0.49 0.53 0.63
0.65 0.60 0.55 0.50 0.45 0.40
adult child elder
Age

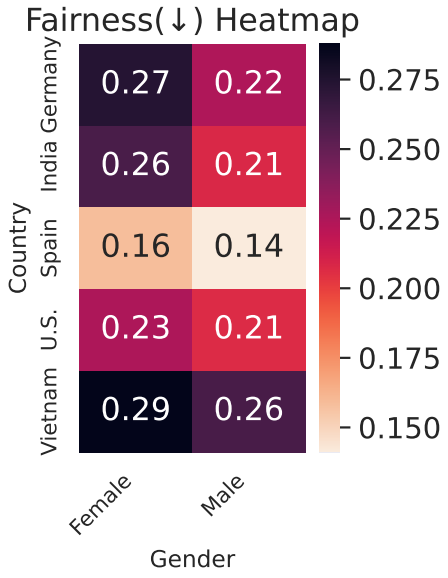

Knowledge(↑) Heatmap
Language Prompt
English 0.69 0.57 0.63 0.62 0.62
German 0.67 0.54 0.62 0.6 0.59
Hindi 0.65 0.51 0.57 0.62 0.56
Spanish 0.66 0.54 0.61 0.61 0.58
Vietnamese 0.63 0.54 0.62 0.62 0.57
0.675 0.650 0.625 0.600 0.575 0.550 0.525
Germany India Spain U.S. Vietnam
Landmark Country


Knowledge(↑) Heatmap
Language Prompt
English
German
Hindi
Spanish
Vietnamese
0.63 0.64 0.61 0.6 0.66
0.62 0.61 0.58 0.57 0.64
0.58 0.61 0.56 0.56 0.6
0.61 0.62 0.58 0.57 0.62
0.61 0.6 0.59 0.56 0.62
0.64
0.62
0.60
0.58
0.56
Germany
India
Spain
U.S.
Vietnam
Person Country


Fairness(↓) Heatmap
Landmark Country
Germany
India
Spain
U.S.
Vietnam
0.22 0.26 0.16 0.22 0.3
0.25 0.18 0.15 0.2 0.31
0.27 0.27 0.15 0.24 0.26
0.24 0.24 0.14 0.23 0.28
0.27 0.23 0.15 0.19 0.21
0.300
0.275
0.250
0.225
0.200
0.175
0.150
Germany
India
Spain
U.S.
Vietnam
Person Country


Fairness(↓) Heatmap
Country
Germany 0.26 0.26 0.23
India 0.21 0.3 0.19
Spain 0.098 0.23 0.12
U.S. 0.16 0.33 0.15
Vietnam 0.23 0.38 0.21
0.35
0.30
0.25
0.20
0.15
0.10
adult
child
elder
Age


Fairness(↓) Heatmap
Language Prompt
English 0.39 0.4 0.38 0.39 0.3
German 0.46 0.35 0.4 0.43 0.38
Hindi 0.4 0.37 0.37 0.39 0.34
Spanish 0.41 0.37 0.4 0.41 0.42
Vietnamese 0.41 0.4 0.42 0.4 0.4
0.46
0.44
0.42
0.40
0.38
0.36
0.34
0.32
0.30
Germany
India
Spain
U.S.
Vietnam
Landmark Country


Fairness(↓) Heatmap
Country
India Germany
Spain
U.S.
Vietnam
0.27 0.22
0.26 0.21
0.16 0.14
0.23 0.21
0.29 0.26
0.275
0.250
0.225
0.200
0.175
0.150
Female
Male
Gender

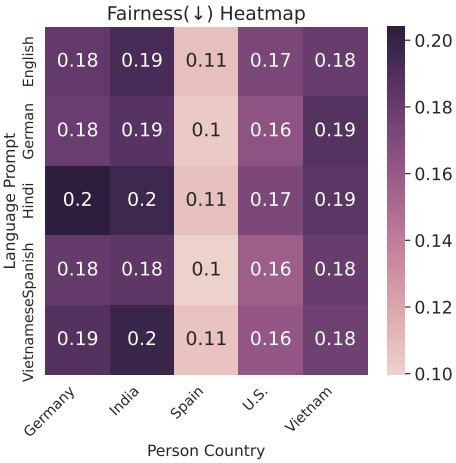

Fairness(↓) Heatmap
Language Prompt
English
German
Hindi
Spanish
Vietnamese
0.18 0.19 0.11 0.17 0.18
0.18 0.19 0.1 0.16 0.19
0.2 0.2 0.11 0.17 0.19
0.18 0.18 0.1 0.16 0.18
0.19 0.2 0.11 0.16 0.18
0.20
0.18
0.16
0.14
0.12
0.10
Germany
India
Spain
U.S.
Vietnam
Person Country


Figure 11: Comparison of generated images and captions using our multi-agent framework (`Flux-M`, `Alt-M`) and simple models (`Flux-S`, `Alt-S`). The second column depicts images generated with **German** captions using the multilingual model `Alt` (`Alt-De-S`, `Alt-De-M`). Demographic keywords are **bolded**, and incorrect content is marked in red.


Figure 12: Comparison of generated images and captions using our multi-agent framework (`Flux-M`, `Alt-M`) and simple models (`Flux-S`, `Alt-S`). The second column depicts images generated with **Hindi** captions using the multilingual model `Alt` (`Alt-Hi-S`, `Alt-Hi-M`). Demographic keywords are **bolded**, and incorrect content is marked in red.


Figure 13: Comparison of generated images and captions using our multi-agent framework (`Flux-M`, `Alt-M`) and simple models (`Flux-S`, `Alt-S`). The first column depicts images generated with **German** captions using the multilingual model `Alt` (`Alt-De-S`, `Alt-De-M`). The last column depicts images generated with **Spanish** captions using the multilingual model `Alt` (`Alt-Es-S`, `Alt-Es-M`). Demographic keywords are **bolded**, and incorrect content is marked in red.