Multi-Agent Multimodal Models for Multicultural Text to Image Generation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) demonstrate impressive performance across various multimodal tasks. However, their effectiveness in cross-cultural contexts remains limited due to the predominantly Western-centric nature of existing data and models. Meanwhile, multi-agent models have shown strong capabilities in solving complex tasks. In this paper, we evaluate the performance of LLMs in a multi-agent interaction setting for the *novel task* of multicultural image generation. Our key contributions are: (1) We introduce MosAIG, a Multi-Agent framework that enhances multicultural Image Generation by leveraging LLMs with distinct cultural personas; (2) We provide a dataset of 9,000 multicultural images spanning five countries, three age groups, two genders, 25 historical landmarks, and five languages; and (3) We demonstrate that multi-agent interactions outperform simple, no-agent models across multiple evaluation metrics, offering valuable insights for future research. Our dataset and models are available at `https://anonymous.4open.science/r/MosAIG`

1 Introduction

Societies worldwide are increasingly diverse, with people of various cultural backgrounds co-existing - an outcome amplified by global travel and migration (Castles et al., 2103). This multicultural tapestry offers both opportunities and challenges, particularly in Artificial Intelligence (AI), where robust representation of diverse groups is essential for equity and inclusivity (Hershcovich et al., 2022; Naous et al., 2023; Mihalcea et al., 2024) However, most existing datasets—especially those used for text-to-image generation—primarily focus on narrow demographics, predominantly western adult males, and frequently portray single-culture scenarios (e.g., *a Chinese temple*, *an Indian market*) (Liu et al., 2024; Kannen et al., 2024). Such limited scope fails to encompass common multicultural interactions (e.g., *a Chinese girl visiting the Golden Gate Bridge*). This limited representation affects the applicability of text-to-image generation models as they fail to accurately reflect the varied cultural and demographic landscapes of the real world (Hershcovich et al., 2022; Bhatia et al., 2024).

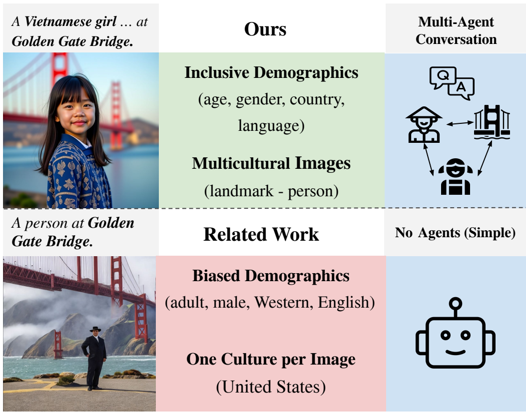


Figure 1: Most datasets used for training are dominated by singular cultural contexts (e.g., "Golden Gate Bridge" primarily depicted with American visitors or as a standalone monument). However, real-world scenarios often transcend cultural boundaries, with people from various backgrounds sharing spaces and experiences. Including images that combine multiple cultures, gender and age groups in a single scene allows models to develop a richer, more nuanced understanding of the world.

To address this gap, our work aims to enhance diversity in text-to-image generation models and datasets. We examine two critical dimensions: (1) the demographic attributes of the depicted person, and (2) the multicultural interactions between the person and the landmark (e.g., Golden Gate Bridge). To this end, we investigate four demographic aspects—age, gender, nationality, and language, while incorporating cross-cultural landmarks (Figure 1). By systematically exploring these aspects, we seek to evaluate and improve how state-of-the-art text-to-image models portray diverse populations and their interaction. Our paper aims to answer three main 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 the performance of text-to-image generation vary across different demographic groups?

RQ3: What strategies can enhance the performance of multicultural text-to-image generation?

The paper makes the following contributions. First, **we share the first dataset of 9,000 images depicting multicultural interactions, i.e., a person and a landmark from different cultures**, across five countries, three age groups, two genders, 25 historical landmarks, and five languages, that can be easily extended to other contexts. Second, we propose `MosAIG` **a novel multi-agent framework to improve multicultural text-to-image generation** across demographics and languages. Finally, we show that **our multi-agent interactions outperform simple models across multiple evaluation metrics**, and provide actionable steps for future work.

2 Related Work

Cultural Evaluation in Language and Vision Models. Research in language-based models is advancing rapidly in capturing cultural nuances through large multilingual evaluation benchmarks (Pawar et al., 2024; Romanou et al., 2024; Singh et al., 2024). In the language-vision domain, recent benchmarks like CVQA (Romero et al., 2024) and GlobalRG (Bhatia et al., 2024) focus on culturally aware question answering, retrieval, and visual grounding. Novel methods leveraging multi-agent frameworks of large multimodal models (Guo et al., 2024; Han et al., 2024) have shown further promise in enhancing cross-cultural understanding. For instance, MosAIC (Bai et al., 2024) employs a multi-agent framework for cross-cultural understanding but focuses on image captioning in single-culture contexts rather than text-to-image generation. Our work addresses this gap by examining how state-of-the-art text-to-image models handle multicultural representations within the same image.

Text-to-Image Generation Models and Benchmarks. Text-to-image generative capabilities have advanced rapidly in recent years, as evidenced by models such as Stable Diffusion-XL (Podell et al., 2023), DALLE-3 (Betker et al., 2023), or FLUX (Labs, 2024). Similar to us, GenArtist (Wang et al., 2024) uses an agentic framework. Unlike GenArtist, which focuses on unified tools for image generation and editing, our work emphasizes multi-cultural and multilingual capabilities. We design and evaluate models that handle diverse languages and cultural contexts, addressing fairness, representation, and performance across global user groups. Evaluation benchmarks like TIFA (Hu et al., 2023), GenEval (Ghosh et al., 2024), and GenAIBench (Lin et al., 2025) traditionally emphasize technical factors such as realism, text faithfulness, and compositional accuracy. More recent work, i.e., HEIM (Lee et al., 2024), extends these metrics to include socially situated aspects like toxicity, bias, and aesthetics, reflecting growing concern for the social impact of generative models (Hartwig et al., 2024).

Cultural Gap and Language Limitations in Text-to-Image Generation. Despite advancements, existing efforts predominantly focus on a narrow set of languages (e.g., English, Chinese, Japanese), leaving large user communities underserved. Recent multilingual models, such as Taiyi-Diffusion-XL (Wu et al., 2024), target Chinese text input, while AltDiffusion (Ye et al., 2024) expands language coverage to eighteen languages. However, a broader "cultural gap" persists (Liu et al., 2024), as most models and benchmarks insufficiently capture diverse cultural settings and interactions.

Data Diversity and Cultural Competence. Only recently have researchers begun to evaluate cultural competence in text-to-image models. For instance, CUBE (Kannen et al., 2024) assesses cultural awareness and diversity, yet still focuses on single-culture depictions per image. To our knowledge, no existing work systematically addresses multicultural scenarios—where multiple cultures may be represented in a single image—and rigorously evaluates the performance of state-of-the-art text-to-image systems under such conditions. Our approach aims to fill this gap by exploring how these models handle more complex, multicultural representations.

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 proposed by Nguyen et al. (2023) and focus specifically on visual cultural elements such as clothing and historical landmarks.

We propose a *novel task*, multicultural image generation, aimed at evaluating how generation models represent elements from diverse cultures within the same image, i.e., a person from one culture and a landmark from a different culture. We also analyze other demographic attributes and their intersection, such as 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.

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 used for each agent are provided in the Appendix Figure 8.

Implementation Details. The Summarizer Agent and each Social Agent are initialized as different instances of a LLaMA model[2] (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[3]. The implementation was carried out using an NVIDIA V100 GPU (32GB). More details can be found in Appendix C.

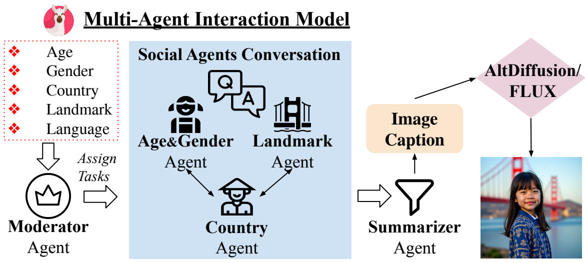


Figure 2: Overview of MosAIG, our framework for Multi-Agent Image Generation. The framework includes a multi-agent interaction model that generates an image caption from demographic information (person age, gender, country, landmark, and caption language), which is then used by an image generation model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

[1]All demographics are shown in Appendix Table 1

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

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

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[4] (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, 2019) 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[5] (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[6] (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[7].

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 automated evaluation metrics, which assess alignment, quality, aesthetics, knowledge, and fairness, ensuring a comprehensive analysis. These metrics encompass both technical factors—alignment, quality, and knowledge—as well as socially situated aspects such as fairness and aesthetics (Lee et al., 2024).

Alignment. CLIPScore (Hessel et al., 2021) measures text-to-image alignment by computing the cosine similarity between the semantic embeddings of the image and its associated text, providing an effective assessment of how well the generated image reflects the intended description. CLIPScore ranges from -1 to +1, where higher values indicate a stronger semantic alignment between the generated image and its corresponding text.

Quality. We assess the quality of generated images using the Inception Score (IS) (Salimans et al., 2016), which leverages an Inception v3 classifier to measure image fidelity and diversity. Lower scores (below 10) typically indicate poor quality or limited variation, while higher scores (10+) suggest more realistic and diverse outputs.

Aesthetic. This metric evaluates the aesthetic appeal of an image, considering factors such as visual clarity, sharpness, color vibrancy, and overall subject clarity. Aesthetic evaluation also takes into account composition, color harmony, balance, and visual complexity. To assess these aspects, we use the SigLIP-based predictor[8], which rates the aesthetics of an image on a scale from 1 to 10 (best).

Fairness. This metric evaluates the consistency of model performance when captions are modified to reference different social groups. Specifically, modifications are applied to attributes such as *gender*, *age*, and *nationality*, while keeping the rest of the caption unchanged. Given an original caption c and its corresponding image I, we construct a modified caption c' by substituting a demographic term, i.e., replacing male-gendered terms with female-gendered terms, "young" with "old" or "German"

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

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

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

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

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

with “Indian”. The corresponding modified image *I’* also reflects the demographic change.
For example, given the initial caption-image pair:
(c, I) = (*A German boy in front of Taj Mahal*, I)
modifying the gender term results in the new pair:
(c', I') =(*A German girl in front of Taj Mahal*, I')
To evaluate fairness, we compute the absolute difference in CLIPScore between the original and modified pairs:

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

where $S(c, I)$ and $S(c', I')$ denote the CLIPScores for the original and modified caption-image pairs, respectively. A fair model should exhibit minimal variation in performance across demographic groups, implying low values of ΔS. Higher values of ΔS indicate greater performance disparity, suggesting potential bias.

Knowledge. This metric evaluates the model’s knowledge of the world by analyzing its ability to recognize and distinguish historical landmarks. To assess this, we modify a given caption c by replacing one *historical landmark* with another while keeping the corresponding image I and the rest of the caption unchanged. For example, given the initial caption-image pair:
(c, I) = (*A German boy in front of Taj Mahal*, I)
modifying the landmark term results in:
(c', I) =(*A German boy in front of White House*, I)
We measure the absolute difference in CLIPScore before and after the modification:

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

A model with strong cross-cultural knowledge of historical landmarks should exhibit high performance variations when landmarks are swapped. Higher scores indicate greater knowledge, while lower scores suggest weaker landmark recognition.

4.2 Multi-Agent Interaction Results

Our multi-agent models outperform simple models in **Quality**, **Knowledge**, and **Fairness**, while scoring on par in **Alignment** and **Aesthetic**, as illustrated in Figure 3.

The most significant improvement is observed in **Quality**, where multi-agent models achieve substantially higher scores (0.77 vs. 0.48 for `Alt-En` and 0.65 vs. 0.45 for `Flux-En`). We hypothesize that this enhancement is driven by the additional contextual details provided by multi-agent interactions, leading to more visually refined outputs.

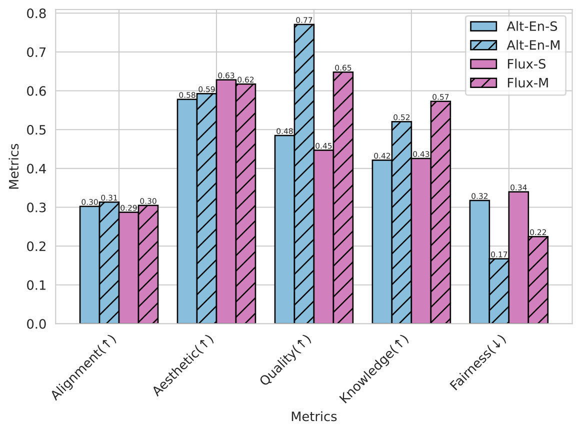


Figure 3: Our multi-agent models (Alt-En-M and Flux-M) outperform simpler models (Alt-En-S and Flux-S) in Quality, Knowledge, and Fairness, with comparable results in Alignment and Aesthetics. Scores are normalized to [0–1]; higher is better except for Fairness.

Additionally, **Quality** is consistently higher for `Alt` compared to `Flux`, likely due to the tendency of Flux-generated images to exhibit blurry backgrounds. Despite gains in Quality, **Aesthetic** scores remain similar across models. This may be because the multi-agent system primarily enhances semantic richness rather than altering the stylistic elements captured by the aesthetic metric. Furthermore, the SigLIP-based predictor may be less sensitive to semantic improvements, focusing more on surface-level visual appeal.

A notable increase is also observed in **Knowledge** (0.52 vs. 0.42 for `Alt-En` and 0.57 vs. 0.43 for `Flux-En`) and **Fairness**, where lower scores indicate better performance (0.17 vs. 0.32 for `Alt-En` and 0.22 vs. 0.34 for `Flux-En`). We attribute these improvements to the ability of the multi-agent framework to incorporate diverse perspectives, reducing social biases and encouraging a more comprehensive representation of factual and demographic information.

Although the overall improvement in **Alignment** is not statistically significant, a breakdown by demographic attributes reveals consistent gains with multi-agent models. For instance, we observe higher scores for *adults* (0.30 vs. 0.27), *females* (0.31 vs. 0.28), and countries such as *Germany* (0.30 vs. 0.27), *India* (0.31 vs. 0.28), and *Vietnam* (0.31 vs. 0.29), as detailed in Appendix E.1. These findings suggest that multi-agent systems can better capture nuanced semantic correspondences across diverse population groups.

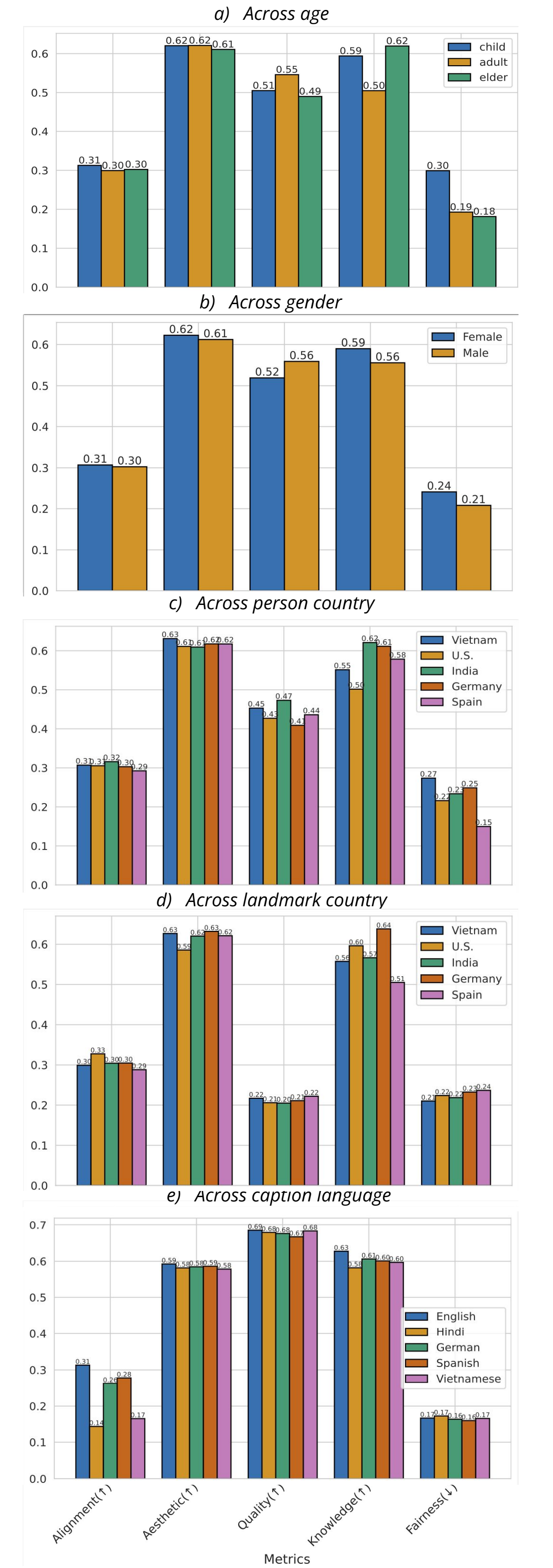


Figure 4: Ablation studies on (a) person age, (b) person gender, (c) person country, (d) landmark country, (e) caption language using the best overall model, the `Multi-agent English Flux-M` (a-d) and `Multi-agent Multilingual Alt-M` (e). Performance across all five metrics—Alignment, Aesthetic, Quality,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

4.3 Ablation Studies

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

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). Furthermore, models with English captions achieve higher Alignment than non-English (0.30 vs. 0.20) (see Figure 10).

f) Intersectionality. Examining a single demographic category, such as race or gender, may overlook 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.2.

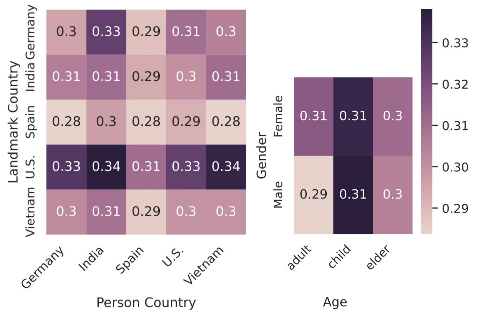


Figure 5: Alignment with best overall model, `Flux-M`, over person-landmark (left) and gender-age (right).

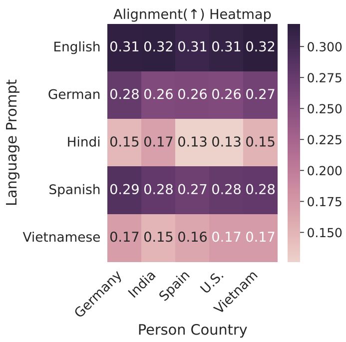


Figure 6: Alignment with best multilingual model, `Alt-M`, over image caption language and person country.

Age and Gender. Figure 5 (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 5 (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.2.

Language and Country. Figure 6 shows Alignment across Person Country and Caption Language. English, Spanish, and Vietnamese captions achieve the highest performance (~ 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.

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. The most frequent errors in the Flux-M model involve incorrect backgrounds, occurring in 38 of 75 images (38/75). Additionally, deviations from prompt descriptions are observed, along with errors in rendering human figures (5/75), such as missing fingers or incorrect cultural markers (e.g., misplacement of a bindi). Landmark-related inconsistencies are less common (2/75), and include significant omissions, such as missing faces on Mount Rushmore. In contrast, the Flux-S model exhibits a higher rate of landmark errors (15/75), such as missing the Sagrada Familia. Errors in depicting human figures also increase (10/75), particularly in rendering traditional attire and facial accuracy. The Alt models (Alt-S and Alt-M) display more pronounced inaccuracies. The most prevalent issue is incorrect backgrounds (55/75), followed by severe body distortions (e.g., three hands, elongated arms, two right feet), and multiplicity errors (e.g., two people instead of one). While the multi-agent Alt-M model reduces errors related to cultural elements (2/75), it still exhibits body distortions (15/75).

4.5 Qualitative Results

In Figure 7, 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

Figure 7: 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.

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. Additional results across demographics are in Appendix E.3.

5 Lessons Learned and Actionable Steps

Our findings provide insights into the performance of multi-agent multimodal models for multicultural image generation, highlighting key lessons and proposing actionable steps to improve accuracy and cultural representation in future models.

Prioritize Multi-Agent Models. Our analysis shows that multi-agent models generate more contextually rich and culturally nuanced images than simple models (Section 4.2). By integrating diverse perspectives through collaboration, these models enhance alignment, aesthetics, quality, and knowledge. Future research should focus on refining multi-agent frameworks to further enhance alignment, aesthetic, and representational diversity. Additionally, our framework can be extended to generate images depicting a wider range of cultural interactions—such as dancing, eating, and festivals—while featuring diverse groups. This extension would allow for a comprehensive evaluation of reasoning and action-based image generation.

Prioritize Multilingual Generation Models. Our results indicate a performance discrepancy between English and non-English prompts, with English-based generations often exhibiting higher Alignment (Figure 4 e). To ensure equitable representation across languages, future models should incorporate stronger multilingual capabilities, improving Fairness and Alignment in non-English text-to-image generation.

Develop Better Evaluation Metrics. Current evaluation metrics do not always align with qualitative assessments, particularly when surrounding elements boost scores despite incorrect Landmarks (Section 4.4). For example, an image of the Taj Mahal may score highly due to accurately depicted gardens, even if the Landmark itself is wrong. We recommend refining Alignment metrics by assigning greater weight to key elements, such as Landmarks, for more reliable assessments.

6 Conclusion

In this paper, we introduce `MosAIG`, a framework that leverages LLM agent interactions to enhance multicultural text-to-image generation. We conduct a comprehensive analysis of image generation performance across five countries, three age groups, two genders, 25 historical landmarks, and five languages, as well as their intersections. Our evaluation across five key metrics reveals significant demographic variations. Notably, our framework outperforms simple models in Quality, Knowledge, and Fairness, and shows Alignment improvements across diverse demographics. We contribute the first dataset of 9,000 images depicting multicultural interactions, specifically showcasing individuals and landmarks from different cultural backgrounds. Additionally, we open-source both our dataset and the models generated by `MosAIG`, providing a valuable resource for future research. Our dataset and models are available at: https://anonymous.4open.science/r/MosAIG

Limitations and Ethical Considerations

Limited Demographics that can be Extended. Our study focuses on a binary gender representation—male and female—while overlooking non-binary and other gender identities. Expanding future models to encompass a broader spectrum of gender identities would enhance inclusivity and fairness in image generation. Additionally, our dataset is restricted to five countries—U.S., Germany, India, Spain, and Vietnam—and five languages—English, German, Hindi, Spanish, and Vietnamese. These languages and regions are relatively well-represented in the training data, limiting our ability to evaluate model performance across less-studied linguistic and cultural groups. This highlights the need for broader validation across a more diverse set of cultures to ensure improved alignment, fairness, and reliability in cross-cultural image generation. Finally, we categorize age into three broad groups: child, adult, and elder, which may oversimplify the diversity within each age category. Further refinement of age-related categorizations could help more accurately reflect the varied experiences and characteristics of individuals across different life stages. However, our work represents the first effort to address multicultural image generation, and we deliberately focused on these demographics as a proof of concept, leveraging our personal cultural expertise. Importantly, **our open-source system is designed to be easily extended to generate images for additional countries, languages, age groups, and genders**. This flexibility ensures that our approach can be expanded to achieve broader demographic coverage in future work.

Challenges in Defining Demographic Representation. Our methodology utilizes multi-agent large language model (LLM) interactions, where each LLM simulates a unique perspective based on cultural, age, and gender attributes. While carefully designed prompts help align these models with diverse demographic contexts, identity is inherently complex and cannot be fully encapsulated through broad categorizations. Defining culture solely through national affiliation or language overlooks the vast heterogeneity of traditions, experiences, and perspectives that exist within and across borders. Relying on a limited set of demographic indicators provides only a foundational framework for understanding diversity, but it does not capture the deeper nuances that define individual and collective identities. To improve representation, future research should incorporate additional dimensions such as historical influences, societal values, traditions, and lived experiences. Expanding cultural modeling to account for attitudes, biases, and personal narratives will enable more accurate and contextually rich portrayals, ultimately enhancing both the performance and authenticity of AI-generated representations.

References

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. 2024. The power of many: Multi-agent multimodal models for cultural image captioning. *arXiv preprint arXiv:2411.11758*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 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.*

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

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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. *arXiv preprint arXiv:2402.01680*.

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, et al. 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. *arXiv preprint arXiv:2104.08718*.

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. *arXiv preprint arXiv:2407.06863*.

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, et al. 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. 2024. Why ai is weird and should not be this way: Towards ai for everyone, with everyone, by everyone. *arXiv preprint arXiv:2410.16315*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

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, et al. 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, Alfonso Amayuelas, et al. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

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, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark.

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 I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 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 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.

How well does the image match the description?

1. Does not match at all
2. Has significant discrepancies
3. Has several minor discrepancies

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

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

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 8: 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

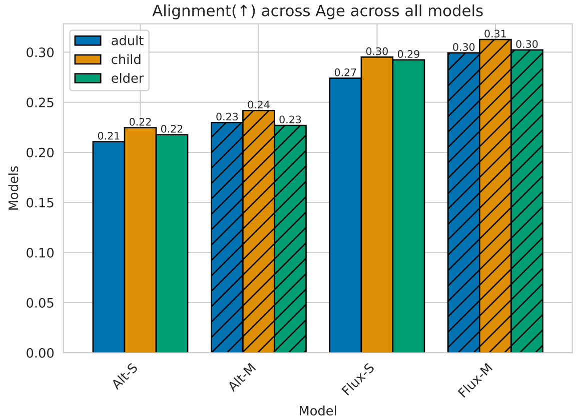
E Results

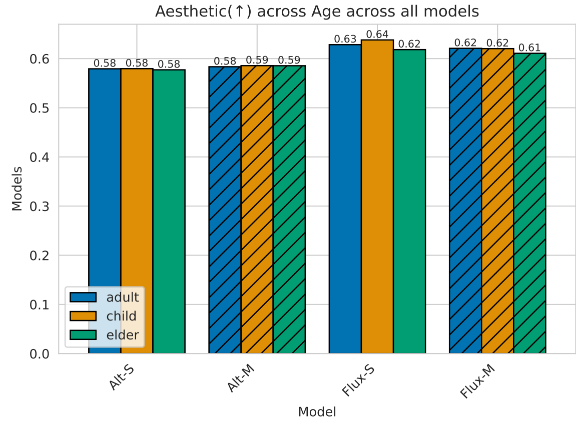
E.1 Across Metrics and Demographics, across All Models

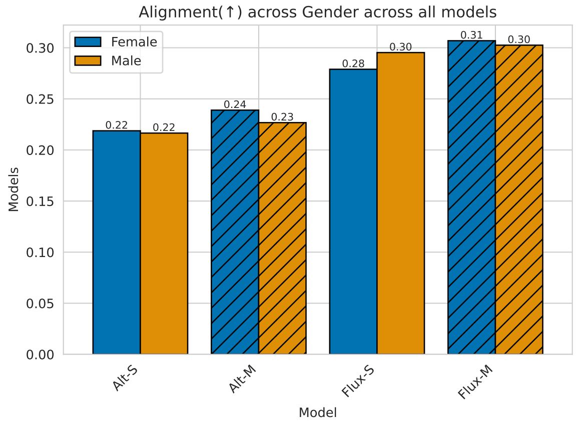
E.2 Intersectionality

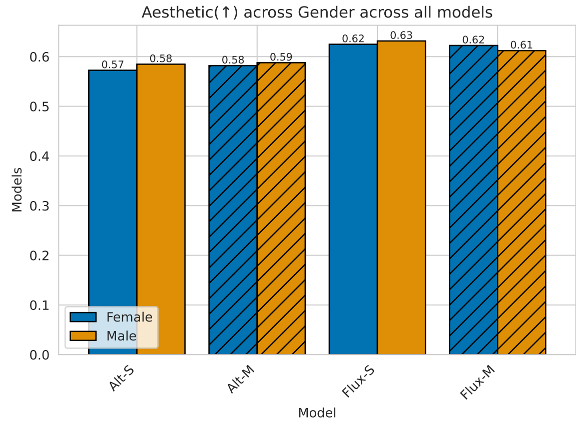
E.3 Qualitative Results

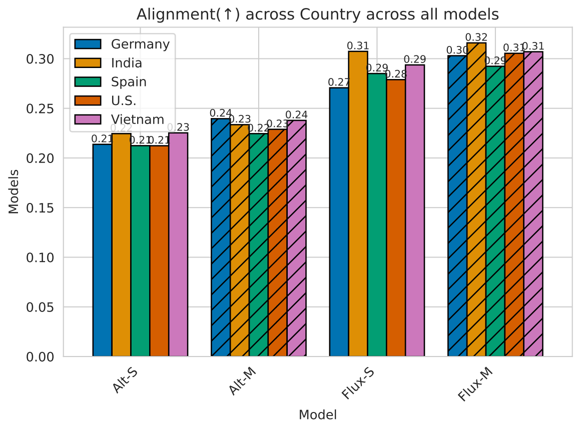
Figure 9: Human Annotation Interface for manually evaluating the images across all models.

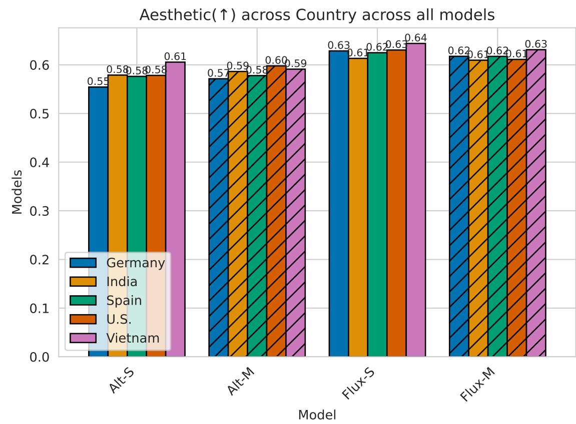

Alignment(↑) across Age across all models
adult
child
elder
Models
Model
Alt-S
Alt-M
Flux-S
Flux-M

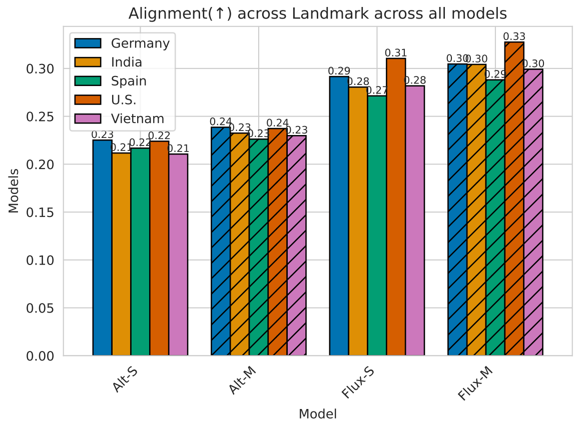

Aesthetic(↑) across Age across all models
adult
child
elder
Models
Model
Alt-S
Alt-M
Flux-S
Flux-M

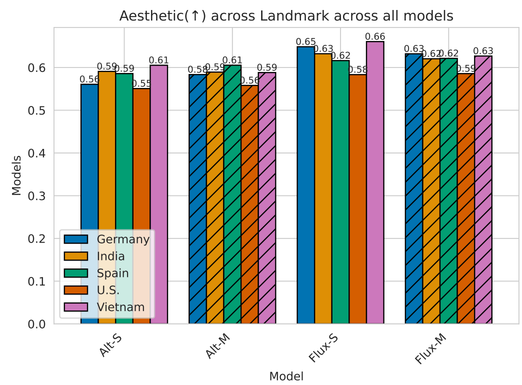

Alignment(↑) across Gender across all models
Female
Male
Models
Model
Alt-S
Alt-M
Flux-S
Flux-M

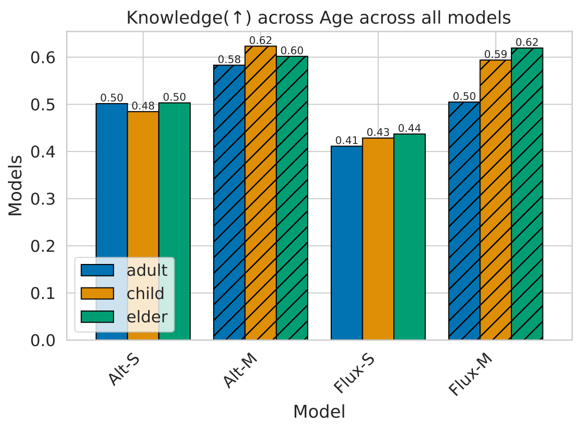

Aesthetic(↑) across Gender across all models
Female
Male
Models
Model
Alt-S
Alt-M
Flux-S
Flux-M

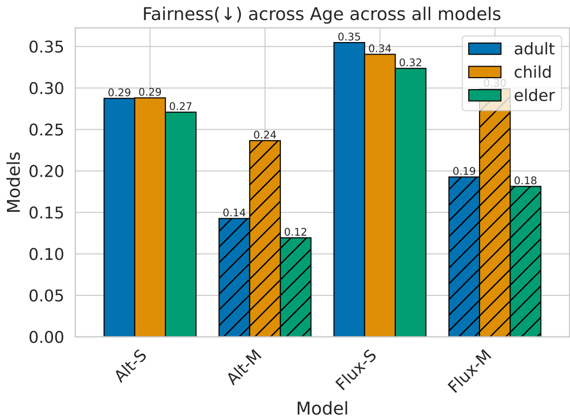

Alignment(↑) across Country across all models
Germany
India
Spain
U.S.
Vietnam
Models
Model
Alt-S
Alt-M
Flux-S
Flux-M

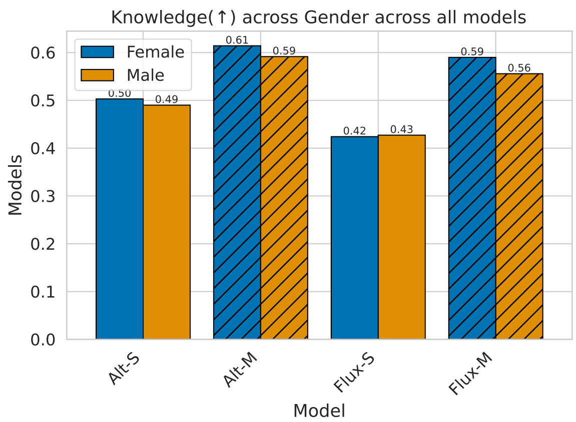

Aesthetic(↑) across Country across all models
Germany
India
Spain
U.S.
Vietnam
Models
Model
Alt-S
Alt-M
Flux-S
Flux-M

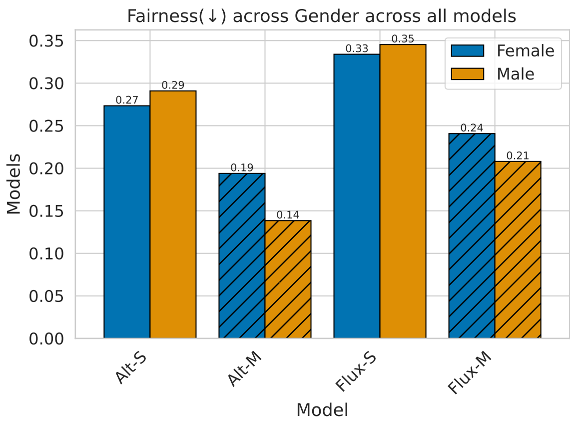

Alignment(↑) across Landmark across all models
Germany
India
Spain
U.S.
Vietnam
Models
Model
Alt-S
Alt-M
Flux-S
Flux-M

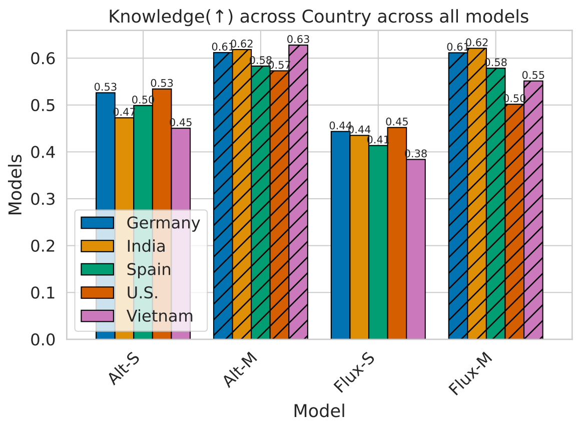

Aesthetic(↑) across Landmark across all models
Germany
India
Spain
U.S.
Vietnam
Models
Model
Alt-S
Alt-M
Flux-S
Flux-M

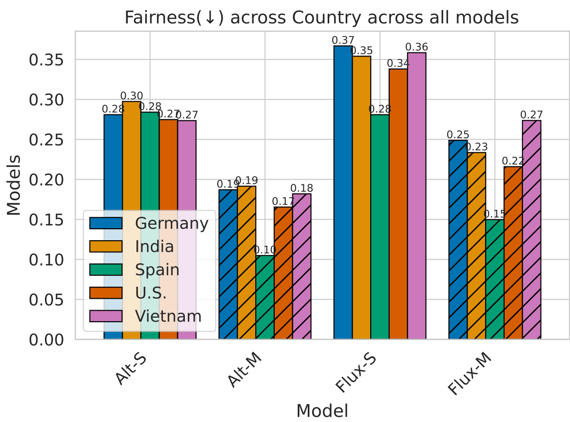

Knowledge(↑) across Age across all models
Models
Model
adult
child
elder
Alt-S
Alt-M
Flux-S
Flux-M

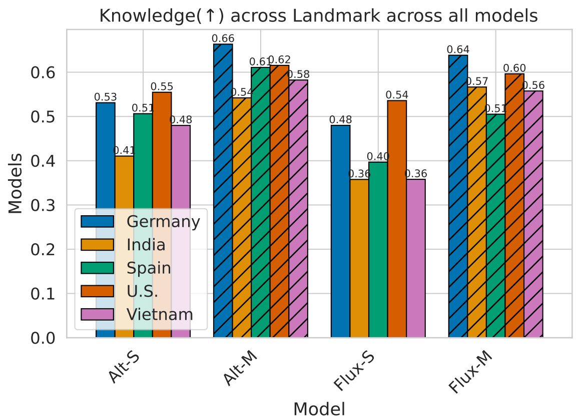

Fairness(↓) across Age across all models
Models
Model
adult
child
elder
Alt-S
Alt-M
Flux-S
Flux-M

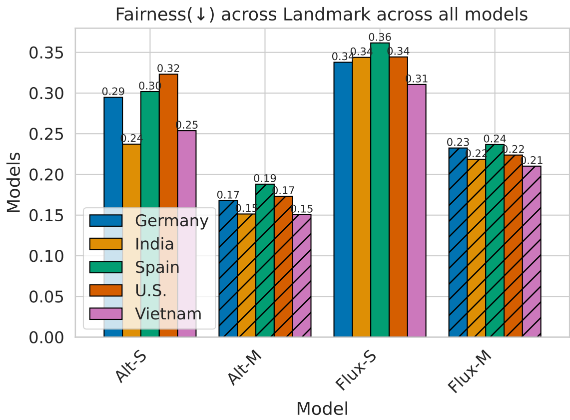

Knowledge(↑) across Gender across all models
Models
Model
Female
Male
Alt-S
Alt-M
Flux-S
Flux-M


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

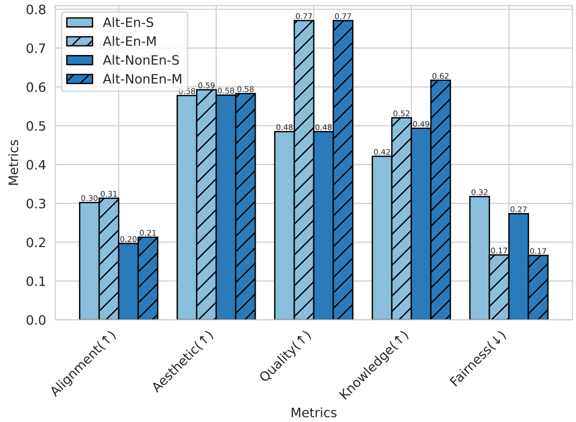
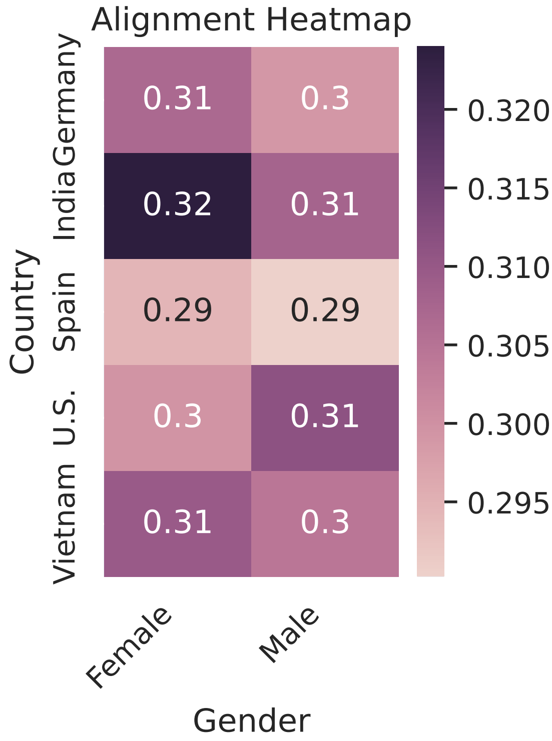
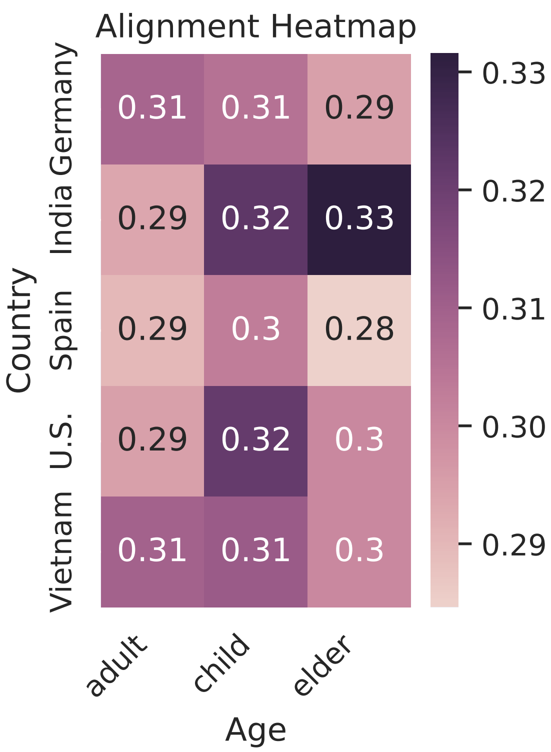
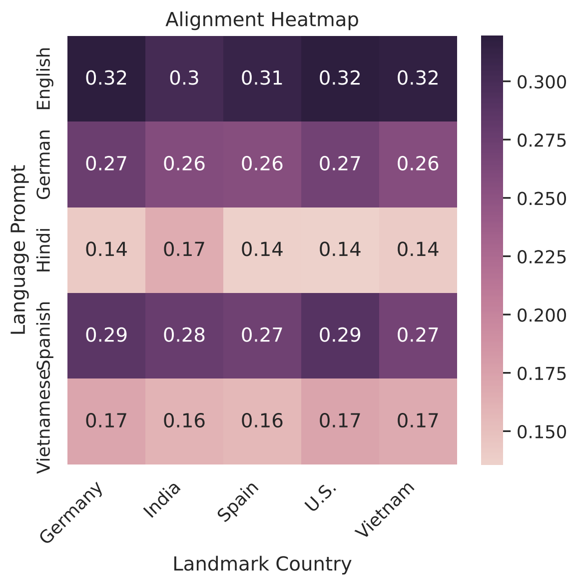
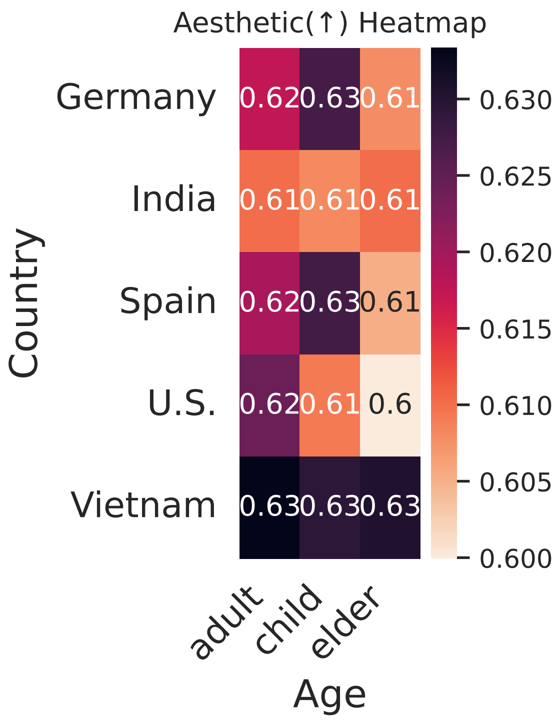


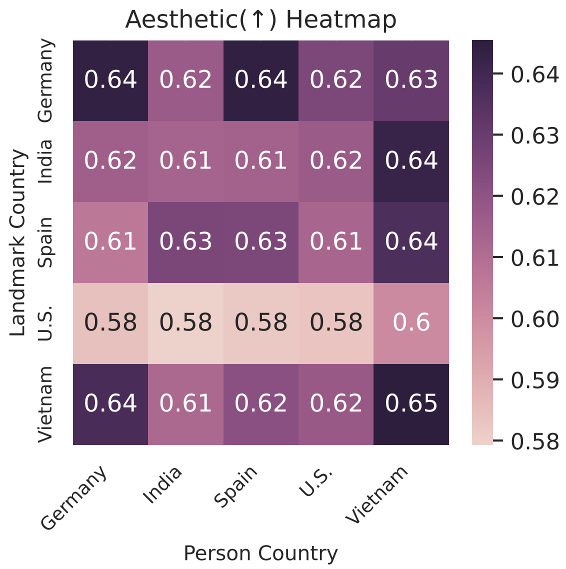
Figure 10: English vs. Multilingual Performance. Models with English captions as input (Alt-En-S, Alt-En-M) achieve higher scores than non-English (Alt-NonEn-S, Alt-NonEn-M) in Alignment (0.30 vs. 0.20), while performing comparably across Aesthetics and Quality metrics. Knowledge and Fairness performance is higher for non-English models

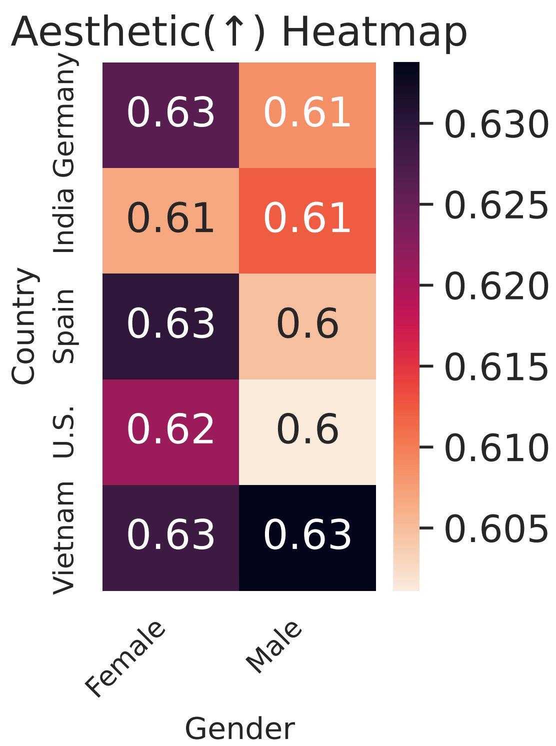


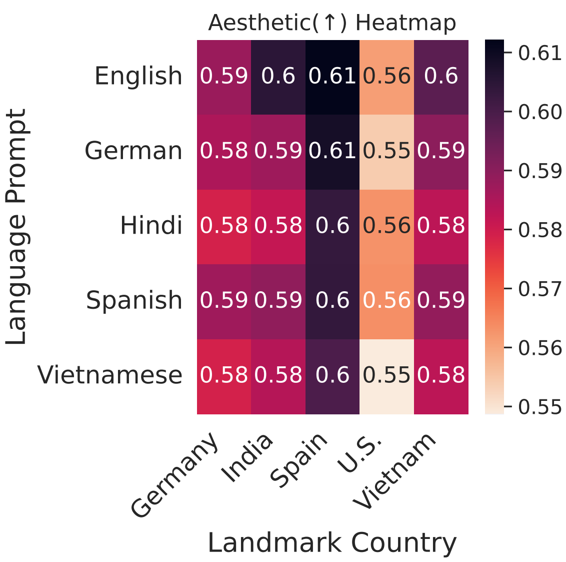


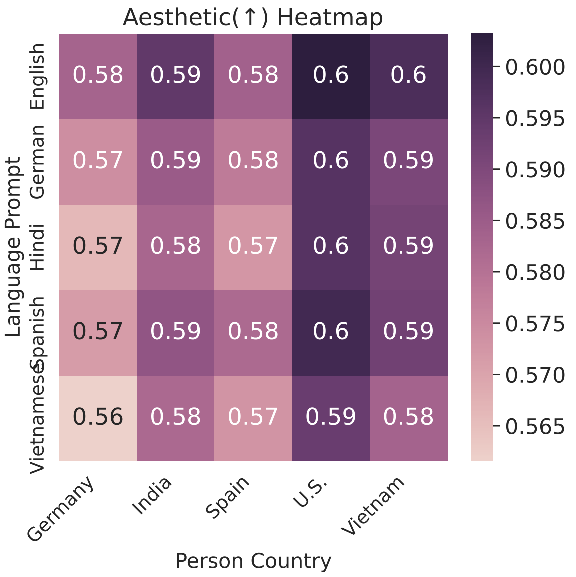


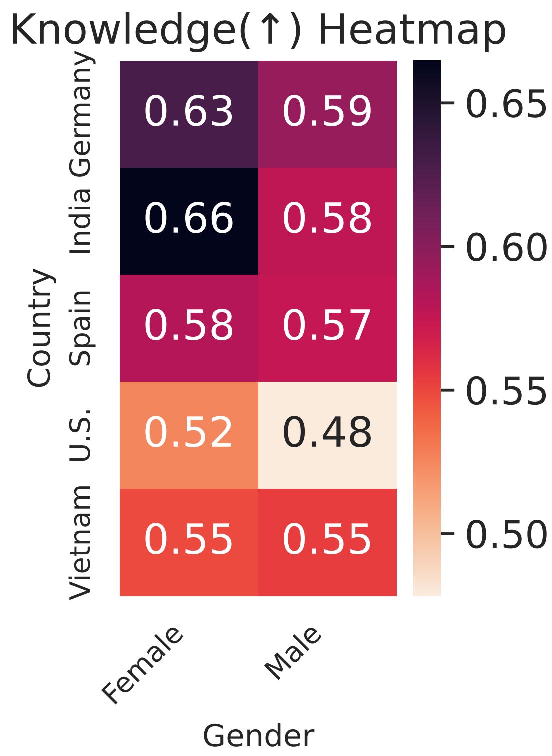

Aesthetic(↑) Heatmap
Country
Germany
India
Spain
U.S.
Vietnam
0.62 0.63 0.61
0.61 0.61 0.61
0.62 0.63 0.61
0.62 0.61 0.6
0.63 0.63 0.63
adult
child
elder
Age
0.630
0.625
0.620
0.615
0.610
0.605
0.600

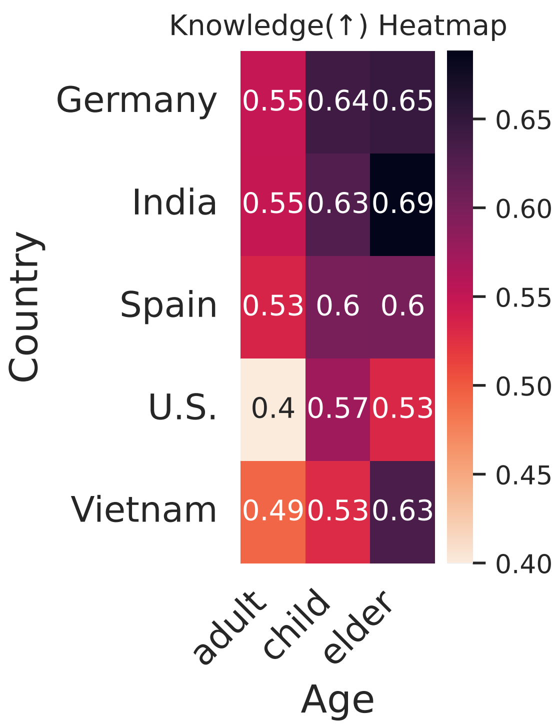

Aesthetic(↑) Heatmap
Landmark Country
Germany
India
Spain
U.S.
Vietnam
0.64 0.62 0.64 0.62 0.63
0.62 0.61 0.61 0.62 0.64
0.61 0.63 0.63 0.61 0.64
0.58 0.58 0.58 0.58 0.6
0.64 0.61 0.62 0.62 0.65
Germany
India
Spain
U.S.
Vietnam
Person Country
0.64
0.63
0.62
0.61
0.60
0.59
0.58

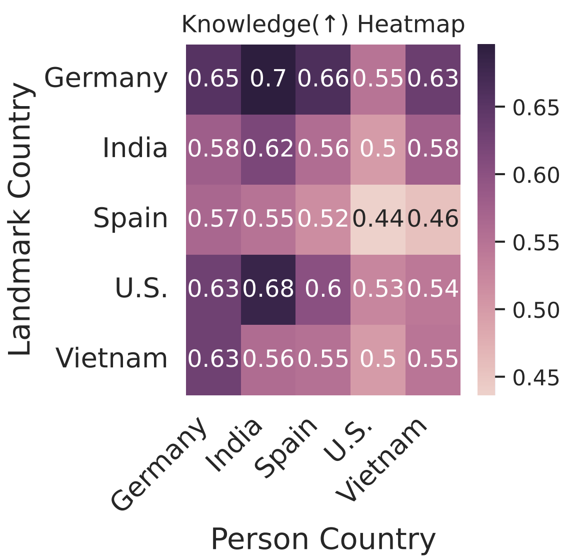

Aesthetic(↑) Heatmap
Country
Germany
India
Spain
U.S.
Vietnam
0.63 0.61
0.61 0.61
0.63 0.6
0.62 0.6
0.63 0.63
Female
Male
Gender
0.630
0.625
0.620
0.615
0.610
0.605

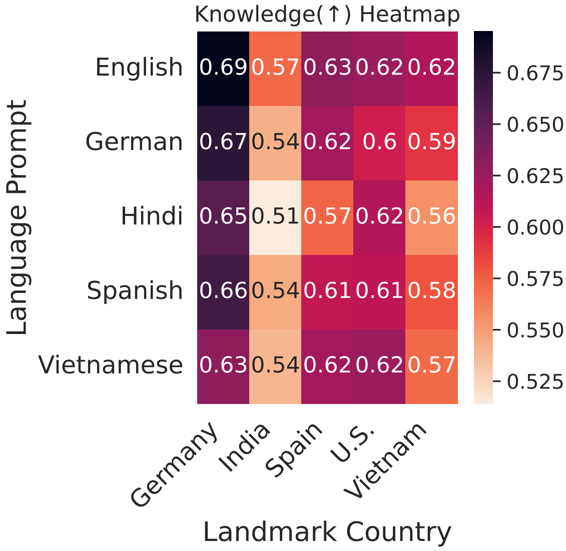

Aesthetic(↑) Heatmap
Language Prompt
English
German
Hindi
Spanish
Vietnamese
0.59 0.6 0.61 0.56 0.6
0.58 0.59 0.61 0.55 0.59
0.58 0.58 0.6 0.56 0.58
0.59 0.59 0.6 0.56 0.59
0.58 0.58 0.6 0.55 0.58
Germany
India
Spain
U.S.
Vietnam
Landmark Country
0.61
0.60
0.59
0.58
0.57
0.56
0.55

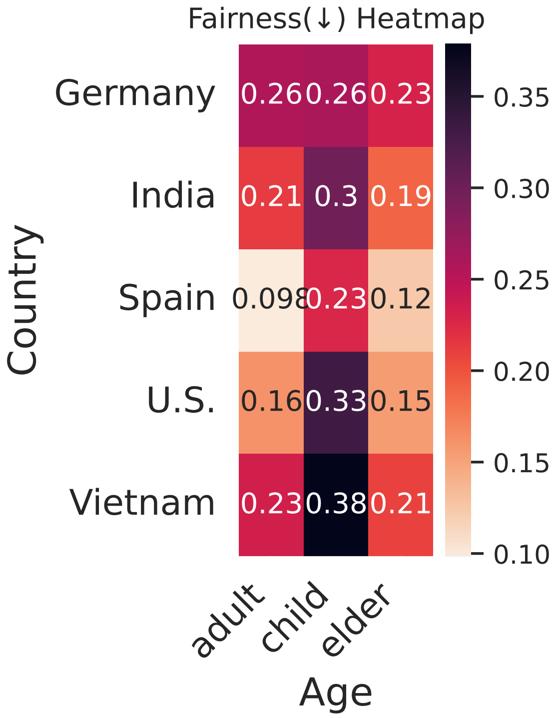

Aesthetic(↑) Heatmap
Language Prompt
English
German
Hindi
Spanish
Vietnamese
0.58 0.59 0.58 0.6 0.6
0.57 0.59 0.58 0.6 0.59
0.57 0.58 0.57 0.6 0.59
0.57 0.59 0.58 0.6 0.59
0.56 0.58 0.57 0.59 0.58
0.600
0.595
0.590
0.585
0.580
0.575
0.570
0.565
Germany
India
Spain
U.S.
Vietnam
Person Country

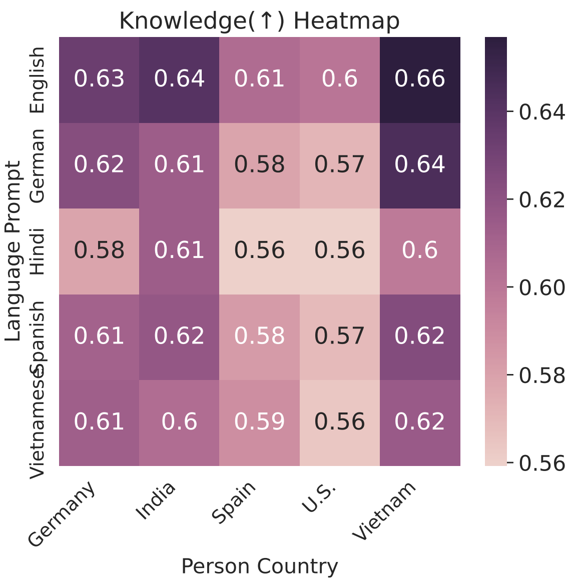

Knowledge(↑) Heatmap
Country
Germany
India
Spain
U.S.
Vietnam
0.63 0.59
0.66 0.58
0.58 0.57
0.52 0.48
0.55 0.55
0.65
0.60
0.55
0.50
Female
Male
Gender

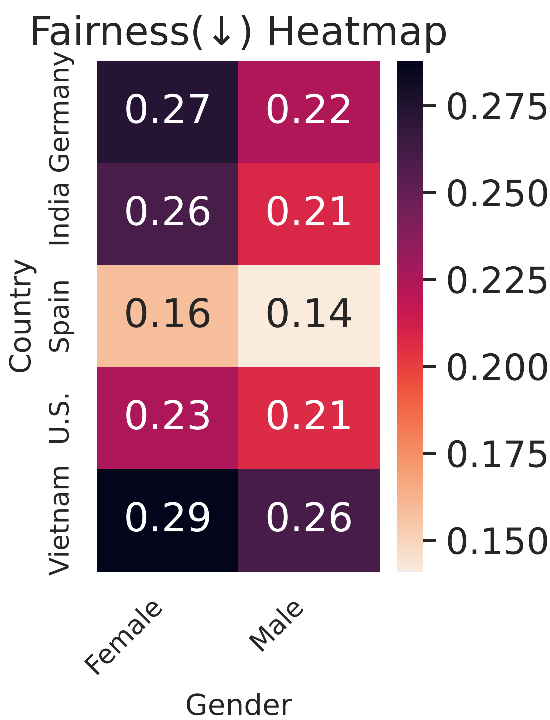

Knowledge(↑) Heatmap
Country
Germany
India
Spain
U.S.
Vietnam
0.55 0.64 0.65
0.55 0.63 0.69
0.53 0.6 0.6
0.4 0.57 0.53
0.49 0.53 0.63
0.65
0.60
0.55
0.50
0.45
0.40
adult
child
elder
Age

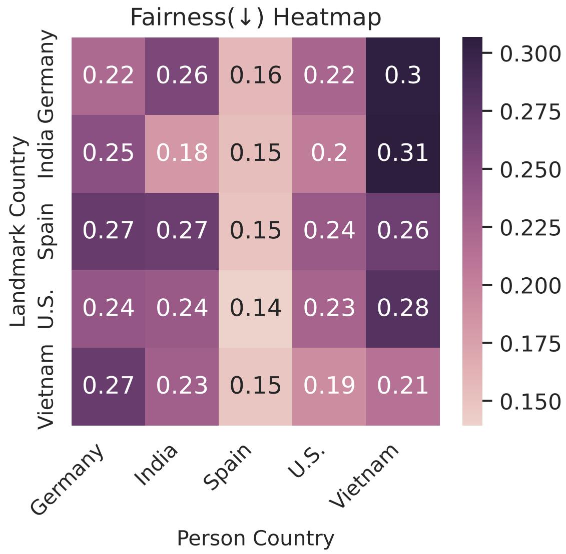

Knowledge(↑) Heatmap
Landmark Country
Germany
India
Spain
U.S.
Vietnam
0.65 0.7 0.66 0.55 0.63
0.58 0.62 0.56 0.5 0.58
0.57 0.55 0.52 0.44 0.46
0.63 0.68 0.6 0.53 0.54
0.63 0.56 0.55 0.5 0.55
0.65
0.60
0.55
0.50
0.45
Germany
India
Spain
U.S.
Vietnam
Person Country

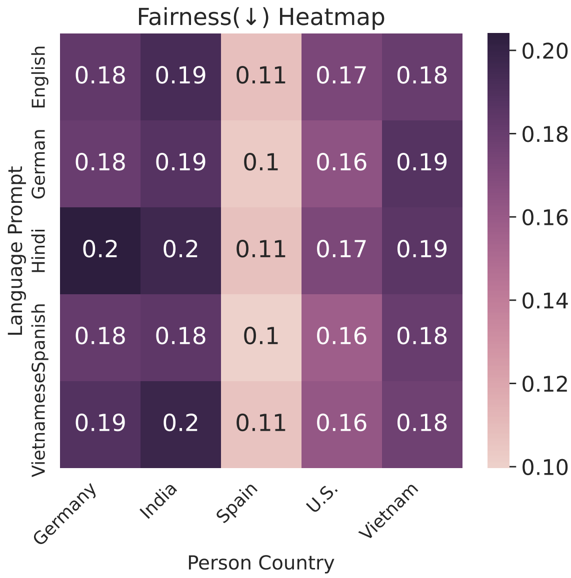

Knowledge(↑) Heatmap
Language Prompt
English 0.69 0.57 0.63 0.62 0.62
German 0.67 0.54 0.62 0.6 0.59
Hindi 0.65 0.51 0.57 0.62 0.56
Spanish 0.66 0.54 0.61 0.61 0.58
Vietnamese 0.63 0.54 0.62 0.62 0.57
Germany India Spain U.S. Vietnam
Landmark Country
0.675 0.650 0.625 0.600 0.575 0.550 0.525


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
adult child elder
Age
0.35 0.30 0.25 0.20 0.15 0.10


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


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


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
Germany
India
Spain
U.S.
Vietnam
Person Country
0.300
0.275
0.250
0.225
0.200
0.175
0.150


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
Germany
India
Spain
U.S.
Vietnam
Person Country
0.20
0.18
0.16
0.14
0.12
0.10

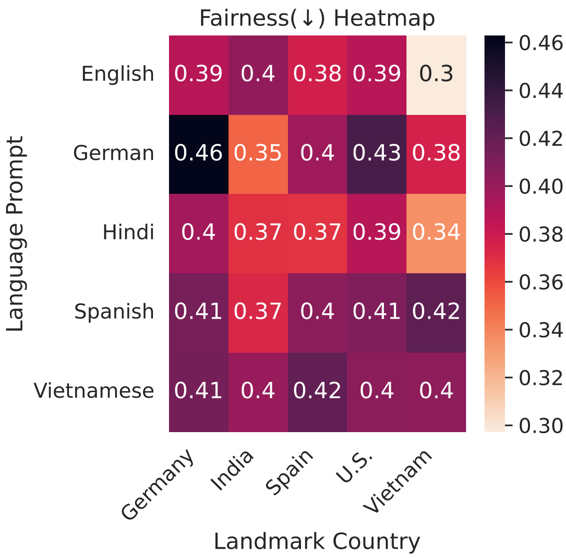

Fairness(↓) Heatmap
Language Prompt
English
German
Hindi
Spanish
Vietnamese
0.39 0.4 0.38 0.39 0.3
0.46 0.35 0.4 0.43 0.38
0.4 0.37 0.37 0.39 0.34
0.41 0.37 0.4 0.41 0.42
0.41 0.4 0.42 0.4 0.4
Germany
India
Spain
U.S.
Vietnam
Landmark Country
0.46
0.44
0.42
0.40
0.38
0.36
0.34
0.32
0.30

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.