Cultural Image Translation for Human Figures

Anjishnu Mukherjee* Chahat Raj*
Department of Computer Science
George Mason University, Fairfax, VA, USA
{amukher6, craj}@gmu.edu

Abstract

We present an approach for adapting the cultural representation of human figures in generated images to obtain images more representative of the target culture while maintaining structural similarity with the original generated image. While diffusion models with text guidance can be used for this purpose, the images from such approaches often misappropriate gender, age, or other culturally relevant artifacts in the image. We first create a small-scale, high-quality dataset for this task using diffusion models and then propose a simple method that combines computer vision techniques to get strong performance on this task as measured with both quantitative metrics and also a qualitative study. We further analyze different cases where this method makes errors objectively to understand the shortcomings of this approach. [1]

1. Introduction

Culture is loosely defined in literature as a group of people who believe in some common ideologies and practices. However, operationalizing this and other definitions of culture in computer science is hard because the boundaries between what constitutes a group are so narrow. Recent advances have seen growing interest in understanding how culture influences language models and human behavior, including language, art, and decision-making [1–3, 10]. One application of this research is in transcreation, i.e., content adaptation while preserving context to ensure a better match to the target audience. [7] and [12] describe this problem broadly and propose different solutions to transcreate images with cultural context.

The challenge, however, with the current approaches is that they approach the translation of an image as a problem of the whole and not as a composition of its parts. In this work, we consider this challenge in more depth by focusing on the human figures present in these images. We have observed in results from existing approaches that human figures tend to be a particularly difficult case leading to many errors where the faces are completely deformed or unrecognizable. We want to fix this problem using simple solutions that can be combined to support the reverse diffusion process for better creating these images (Figure 1).

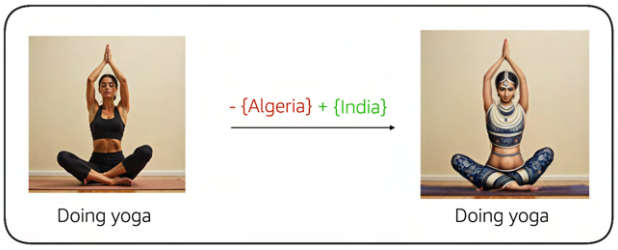

Figure 1. We propose an approach to translate human figures depicting a particular day-to-day activity from one cultural context to another.

Our main contributions are as follows:

- **Dataset:** We introduce a small-scale dataset of 440 images generated using Flux.1-dev, a 12B rectified flow transformer, focused on humans from 20 different cultural backgrounds performing day-to-day activities. We also validate a random sample of this data for quality and cultural relevance.
- **Framework:** We propose a modular approach, CULTUREADAPTHUMAN(Figure 5), to adapt human-like figures in images to different cultural contexts.
- **Benchmark:** We use our approach to edit all possible pairs of images (8,361 edits) and measure performance using three well-defined metrics from existing literature, along with human studies, to validate our analysis.
- **Error analysis:** We perform a thorough analysis of all the cases where the edits from our approach do not meet quantitative thresholds compared to the average to understand possible future areas of improvement.

[1] Dataset and code are available: https://github.com/iamshnoo/cultural-humans

2. Data

We create a dataset of images representing human figures performing daily activities. This involves four steps - choosing countries and activities, developing a prompt that combines the different components into a cultural persona that can be portrayed as a human figure, and prompting a rectified flow transformer model.

2.1. Countries

We chose four continents (Asia, Africa, Europe, and the Americas) ranked in order of population density to ensure that our data is representative of the distribution of people from all parts of the world (Figure 2). From each continent, we chose five countries to get 20 overall by ranking the most visited places in each continent to denote a sense of cross-cultural interaction happening in those places. These countries would benefit most from tourist guides with culturally adapted content for visitors from different countries, for example.

The final list of countries in our dataset includes Africa (Morocco, Egypt, South Africa, Tunisia, Algeria), Asia (Thailand, China, Japan, Malaysia, India), Europe (France, Spain, Italy, Turkey, United Kingdom), and the Americas (United States, Mexico, Canada, Brazil, Argentina).

2.2. Activities

We chose 21 activities, including different day-to-day activities and hobbies that are universal to different extents in which people from different countries are involved.

The full list of activities includes cleaning, cooking, cycling, dancing, doing yoga, drawing, eating, exercising, hiking, ironing, jogging, kickboxing, knitting, meditating, painting, playing soccer, playing tennis, reading, sculpting, sleeping, and swimming.

2.3. Prompt

After multiple rounds of iteration, we developed two prompts consistently, resulting in front-facing images of human figures involved in our mentioned activities.

We consider two settings for data generation:

- **Setting 1:** This involves images where the person from a particular country is involved in a particular activity from our list of activities (Prompt 3).
- **Setting 2:** This involves images where we only have a headshot of a person from a particular country, not involved in any activity (Prompt 4).

3. Approach

Our proposed approach for cultural image translation involves a series of systematic steps, as illustrated in Figure 5. The pipeline leverages a combination of object detection, segmentation, and fine-tuned generative models to perform accurate cultural adaptations while preserving structural integrity. Next, we describe each step in detail.

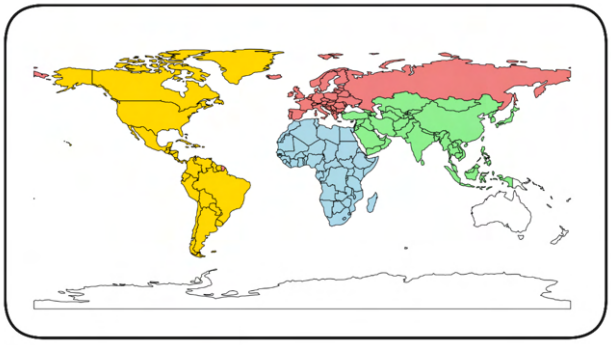

Figure 2. We chose the four most populated continents and then sampled the five most popular countries from each.

Prompt for Setting 1

```
A person from {country} engaged in
{activity}, with their face visible.
```

Figure 3. Our prompt to generate images where the person is from a specific country and involved in a specific activity.

Prompt for Setting 2

```
A person from {country}, wearing culturally
relevant clothing, facing front, with their
face clearly visible.
```

Figure 4. Our prompt to generate images for the control setting, where the person is just from a specific country but not involved in any activity.

Face Detection We begin by detecting all faces in the input image using a fine-tuned YOLOv8 model [5]. This step accurately localizes facial regions, ensuring a robust foundation for subsequent editing processes. The detected face serves as the focal point for targeted adaptations.

Age and Gender Detection Once the faces are localized, age and gender classification is performed to extract demographic attributes. We use image classification models trained on the FairFace data [6] for this purpose. This information provides context for selecting culturally appropriate adaptations that align with the subject's characteristics.

Face Masking The detected facial region is masked to isolate it from the surrounding objects. Face masking is a crucial preprocessing step that allows precise facial edits

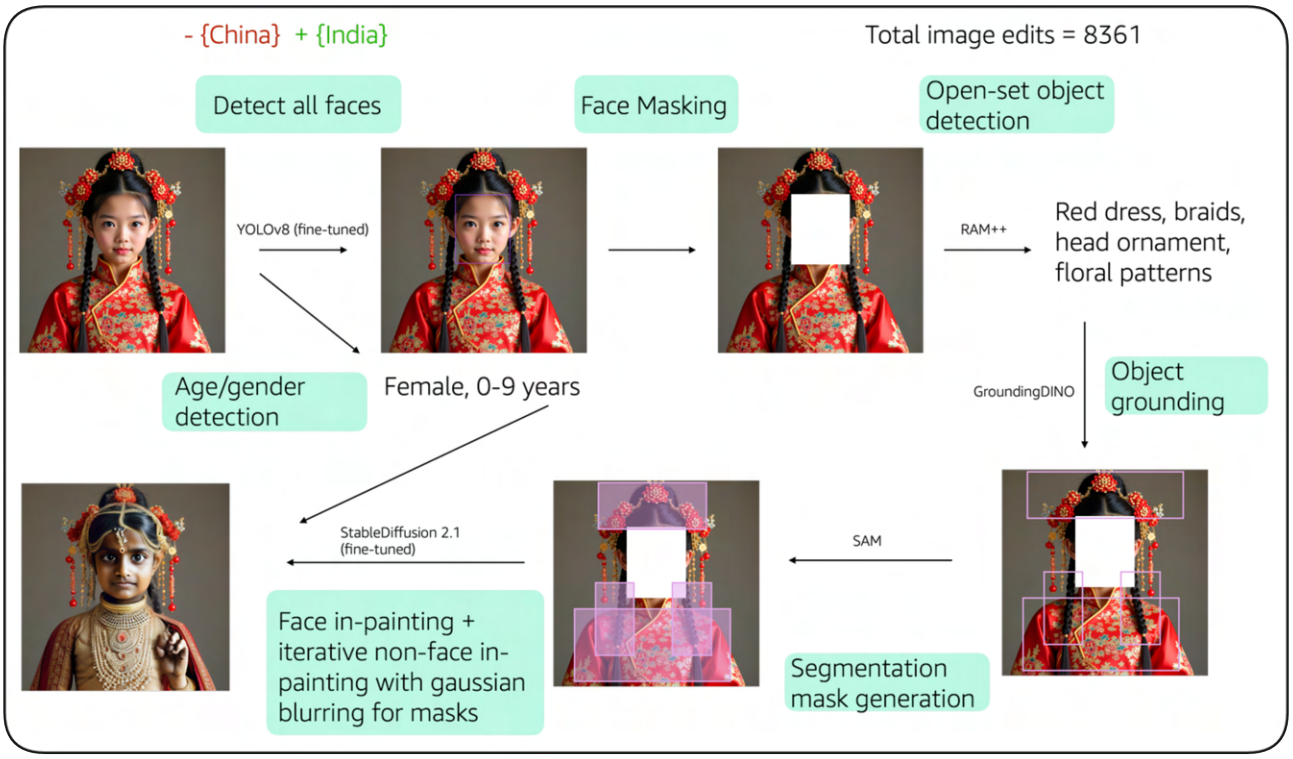

Figure 5. Proposed Approach for Cultural Image Translation.

while ensuring non-face regions remain unaffected during the cultural translation process.

Open-Set Object Detection We employ the Recognize-Anything-Plus (RAM++) model [14] to detect culturally relevant objects and features within the image. This step identifies elements such as clothing (e.g., red dress), accessories (e.g., braids, head ornaments), and patterns (e.g., floral designs). These objects are critical for conveying cultural context during the adaptation process.

Object Grounding Using GroundingDINO [11], the detected objects are grounded within the spatial layout of the image using bounding boxes. This step establishes clear boundaries for cultural features, enabling accurate localization and preparation for segmentation and inpainting.

Segmentation Mask Generation The SAM (Segment Anything Model [8]) generates fine-grained segmentation masks for the grounded objects, providing pixel-level delineation of cultural attributes, such as attire and accessories, and enabling precise editing of specific image regions.

Face and Non-Face Inpainting Finally, we apply a fine-tuned Stable Diffusion 2.1 model [13] to perform face inpainting and iterative non-face inpainting. Gaussian blurring is used on the generated masks to ensure smooth transitions between edited and unedited regions. This step replaces masked-out regions with culturally appropriate adaptations while preserving the overall structural and visual coherence of the image.

An important aspect of this step is that we in-paint masks, one at a time, iteratively. This ensures we don't have to fill in a large region composed of overlapping masks directly in one step, a problem faced by previous work [12].

Summary The proposed approach integrates face detection, object detection, segmentation, and generative inpainting to achieve seamless cultural adaptations. Each pipeline component builds upon the previous step, ensuring the adaptations are both structurally faithful and culturally relevant, as depicted in Figure 5.

4. Evaluation

Let image I_1 correspond to country C_1 (the source image), and I_2 represent its adaptation for country C_2 (the target image). The CLIPScore [4] for an image-country pair is denoted as $S(I, C)$, which measures the alignment between an image I and the cultural representation C. Using this score, we define two deltas that quantify the changes introduced during adaptation:

$$\Delta_1 = S(I_2, C_1) - S(I_1, C_1), \quad (1)$$
$$\Delta_2 = S(I_2, C_2) - S(I_1, C_2). \quad (2)$$

These deltas form the basis for evaluating how well the adaptation moves away from the source culture and aligns with the target culture.

4.1. Primary Metric: M_1

The primary metric, M_1, evaluates the percentage of cases where the following condition is satisfied:

$$\Delta_1 < 0 \text{ and } \Delta_2 > 0. \quad (3)$$

This condition signifies successful adaptation, where the target image I_2 moves closer to the target country's cultural features C_2 while diverging from the source country's representation C_1. A negative Δ_1 indicates reduced alignment with the source country, while a positive Δ_2 confirms improved alignment with the target country.

The M_1 metric directly tracks the core objective of the adaptation process. High M_1 values indicate that the adaptations consistently move toward the intended cultural context without retaining unnecessary source-specific features. In our evaluation, countries such as Japan, India, and China achieved high M_1 scores, while lower scores for Spain and South Africa highlight cases where adaptations struggled to move away from the source cultural representation.

4.2. Secondary Metric: M_2

The secondary metric, M_2, measures the net gain in cultural relevance by comparing the difference between Δ_2 and Δ_1:

$$M_2 = \Delta_2 - \Delta_1. \quad (4)$$

A positive M_2 indicates that the target image I_2 has moved closer to the target country's cultural features than it has moved away from the source. Intuitively, this metric captures the relative success of embedding target cultural features while accounting for any lingering influence from the source culture.

While M_2 provides valuable insights, it is less effective when Δ_1 is positive, suggesting the adaptation retains strong alignment with the source culture. A high M_2 value can be misleading in such cases. Nevertheless, M_2 remains a useful complementary metric for understanding cultural adaptation performance.

4.3. Structural Similarity (SSIM)

To evaluate structural fidelity between the source image I_1 and the adapted image I_2, we compute the similarity of their DINO-ViT embeddings. Instead of relying on pixel-wise comparisons, we use DINO-ViT embeddings to capture high-level visual features. The structural similarity is defined as:

$$\text{SSIM}(I_1, I_2) = \text{cosine\_similarity}(\phi(I_1), \phi(I_2)), \quad (5)$$

where $\phi(I)$ represents the DINO-ViT embedding of image I, and cosine similarity is calculated as:

$$\text{cosine\_similarity}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}. \quad (6)$$

Here, $\mathbf{a}$ and $\mathbf{b}$ are the embedding vectors of I_1 and I_2, respectively. SSIM values close to 1 indicate strong structural similarity, while lower values suggest structural deviations between the source and target images.

By relying on DINO-ViT embeddings, SSIM evaluates whether the source image's high-level structure, pose, and orientation are preserved in the adapted image. In our experiments, SSIM scores were consistently high, reflecting robust structural fidelity. However, slight drops were observed in dynamic activities such as hiking and kickboxing, where complex poses proved challenging to maintain.

4.4. Composite Metric

To achieve a holistic evaluation of the adaptation process, we combine SSIM and M_2 into a unified metric that balances structural fidelity with cultural alignment. The combined score is defined as:

$$\text{Unified Score} = w_1 \cdot \text{SSIM} + w_2 \cdot M_2, \quad (7)$$

where w_1 and w_2 are the weights assigned to SSIM and M_2, respectively. Different combinations of these weights lead to slightly different countries ordering towards the lower end of the performance spectrum. Still, the countries at the top are usually consistent in performance with most combinations of weights (Figure 7).

The choice of M_2 over M_1 is driven by its simplicity and comparability to SSIM. Unlike M_1, which evaluates a binary condition, M_2 provides a continuous measure of net improvement in cultural alignment. Furthermore, the scores for M_2 are on a similar scale to SSIM, eliminating the need for complex weight balancing to account for differences in magnitude. This makes M_2 a more practical and interpretable choice for inclusion in the unified score.

The unified score provides a balanced assessment of the adaptation performance, ensuring that structural integrity and cultural relevance are considered. High unified scores across most country pairs validate our approach's effectiveness in preserving the source image's structure while embedding culturally appropriate features. However, occasional trade-offs are observed: structural fidelity (SSIM) may come at the cost of lower cultural adaptation (low M_2) and vice versa. Addressing these trade-offs remains a focus for future improvements.

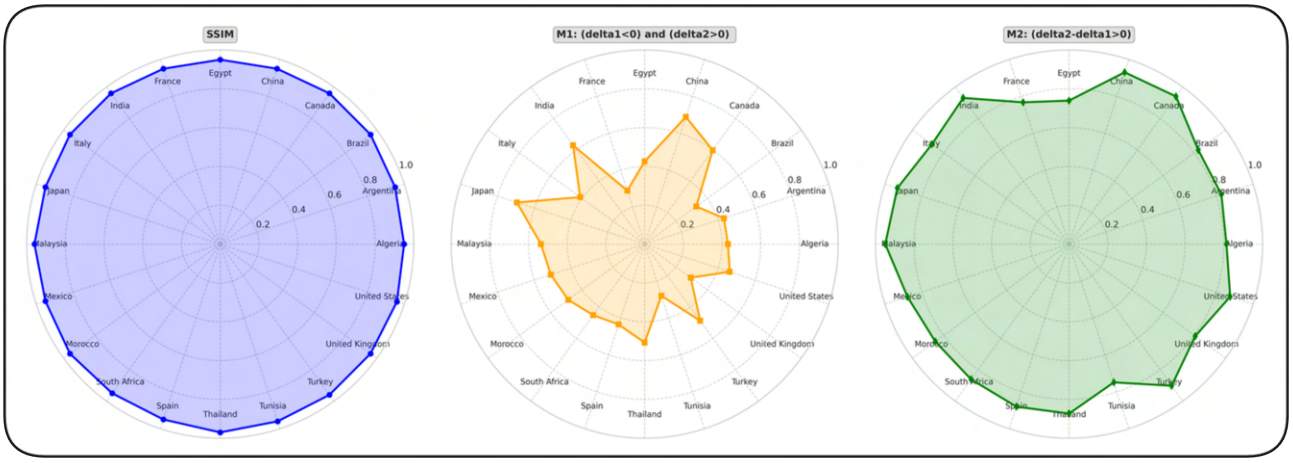

Figure 6. Radar charts showing performance across countries for SSIM (structural fidelity), M1 (cultural divergence and alignment), and M2 (net cultural alignment improvement). SSIM remains consistent, M1 varies, and M2 shows strong overall performance.

5. Results

Our results demonstrate the effectiveness of the proposed approach in balancing structural similarity and cultural adaptation, as evaluated using three key metrics: Structural Similarity Index (SSIM), CLIPScore Delta M1, and CLIPScore Delta M2.

5.1. SSIM, M1, M2

Structural Similarity The SSIM scores demonstrate a strong level of structural similarity between the adapted and source images, highlighting the preservation of essential features such as pose, body alignment, and overall composition (Figure 6). The SSIM radar plot demonstrates consistently high scores across all country pairs, with values predominantly close to 1.0. This indicates that the structural integrity of the adapted images is well-preserved, maintaining features such as pose, body orientation, and composition. Most countries exhibit near-perfect SSIM scores, signifying minimal distortion in structure during adaptation.

Primary Metric M1 measures the alignment of the adapted image with the source country's context, ensuring that cultural distortions are minimized. The M1 plot (Figure 6) shows countries such as China, India, and Japan with higher M1 values, indicating that the adaptations remain aligned with the source cultural context. However, countries like Brazil, France, the United Kingdom, and Tunisia display noticeably lower M1 scores, suggesting that the adaptations for these regions occasionally introduce deviations or distortions that compromise source cultural similarity.

Secondary Metric The M2 plot (Figure 6) evaluates the cultural relevance of the target images, with scores reflecting alignment with the target country's cultural context. With an average M2 score of 0.8, the results highlight significant improvements in cultural relevance over the source images. Countries such as China, Canada, India, and the United States achieve near-optimal M2 scores, indicating successful embedding of culturally appropriate features in the adaptations. Conversely, lower scores for regions like Tunisia, France, and Egypt suggest difficulties in accurately incorporating target-specific cultural details, possibly due to oversimplified or stereotypical representations.

5.2. Composite Score

The composite scores for different country pairs under varying trade-offs between SSIM and M2 are presented in Figure 7. Each of the three bar plots corresponds to different weightings of SSIM and M2, revealing important patterns in performance across regions.

High SSIM, Low M2 In the leftmost plot (Figure 7), where SSIM is prioritized (SSIM: 0.7, M2: 0.3), the composite scores emphasize structural similarity over cultural alignment. Countries like Malaysia, Canada, and India achieve the highest scores, indicating strong preservation of structural integrity. However, lower-ranked countries, such as the United Kingdom, Algeria, France, Egypt, and Tunisia, show lower success. This suggests that while the model maintains pose and spatial structure, it struggles to incorporate culturally relevant elements for these regions.

Balanced SSIM and M2 In the middle plot (Figure 7), where weights are more balanced (SSIM: 0.4, M2: 0.6), the composite scores show a slight shift. Countries previously ranked lower, such as Argentina and Algeria, achieve moderate improvements, suggesting the model better embeds target cultural features when greater emphasis is placed on

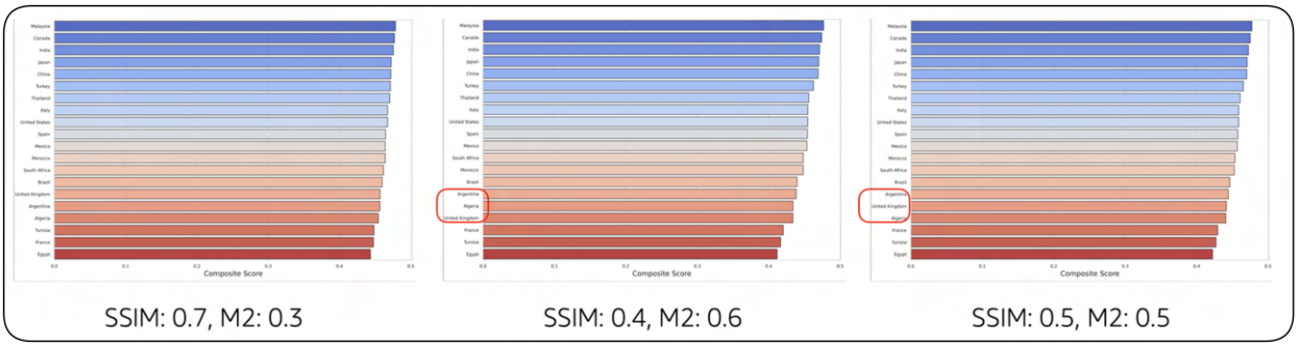

Figure 7. Composite scores across countries under varying trade-offs between SSIM (structural fidelity) and M2 (cultural alignment). Results highlight performance variability, with consistent top performers (e.g., Malaysia, Canada, India) and underperformers (e.g., Argentina, United Kingdom).

M2, as compared with UK. However, Egypt and Tunisia remain at the bottom, indicating challenges in capturing cultural representations for these regions.

Equal SSIM and M2 The rightmost plot (Figure 7) assigns equal weights to SSIM and M2 (SSIM: 0.5, M2: 0.5). The performance for high-ranking countries such as Malaysia, Canada, and India remains strong, while middle-tier regions like Mexico, Morocco, and Brazil show improved alignment. Interestingly, countries such as the United Kingdom and Algeria still rank near the bottom, suggesting that structural fidelity and cultural relevance remain difficult to optimize simultaneously for these specific pairs.

5.3. Activity-based Performance

Figure 8 presents the performance of the proposed approach across various activities in our dataset, evaluated using SSIM, M_1, and M_2. Each activity type is represented along the x-axis, while the y-axis shows each metric's normalized performance scores (ranging from 0 to 1).

Structural Similarity SSIM scores are consistently high (above 0.8) across most activities, reflecting robust structural preservation. Activities involving static poses, such as cleaning, drawing, knitting, and meditating, achieve near-perfect SSIM scores. This suggests that the model maintains structure when the poses are less dynamic. For activities with complex or dynamic poses, such as kickboxing, playing soccer, and sleeping, SSIM scores are slightly lower. This indicates structural fidelity is more challenging when the activity involves movement or unconventional body alignments.

Primary Metric The M_1 scores exhibit more significant variability across activities, with values ranging from 0.5 to 0.9. Activities such as cooking, dancing, and hiking achieve higher M_1 scores, indicating effective cultural embedding

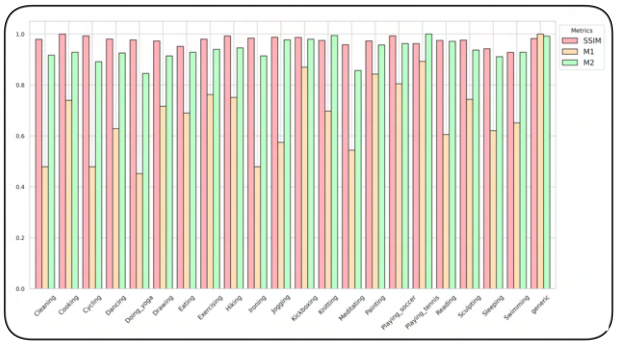

Figure 8. Performance across activities for SSIM (structural fidelity), M1 (cultural divergence and alignment), and M2 (net cultural alignment improvement). SSIM remains consistently high across static activities (e.g., cleaning, knitting, painting), while M1 and M2 exhibit variability, particularly in dynamic or ambiguous activities (e.g., ironing, hiking, playing tennis).

with minimal retention of source features. Conversely, ironing, sleeping, and playing tennis display lower M_1 scores. These lower values suggest difficulties in accurately moving away from the source cultural representation, potentially due to oversimplified adaptations or stereotypical artifacts.

Secondary Metric The M_2 scores are generally high across all activities, with values consistently ranging between 0.7 and 1.0. This demonstrates the model's strength in embedding target cultural attributes. Activities like drawing, eating, painting, and sculpting achieve the highest M_2 scores, suggesting that these activities provide clearer cultural cues (e.g., attire, setting) for the model to incorporate. Lower M_2 scores are observed for activities like sleeping, ironing, and playing soccer, where the cultural relevance may be harder to embed due to limited semantic variation or ambiguity in visual features.

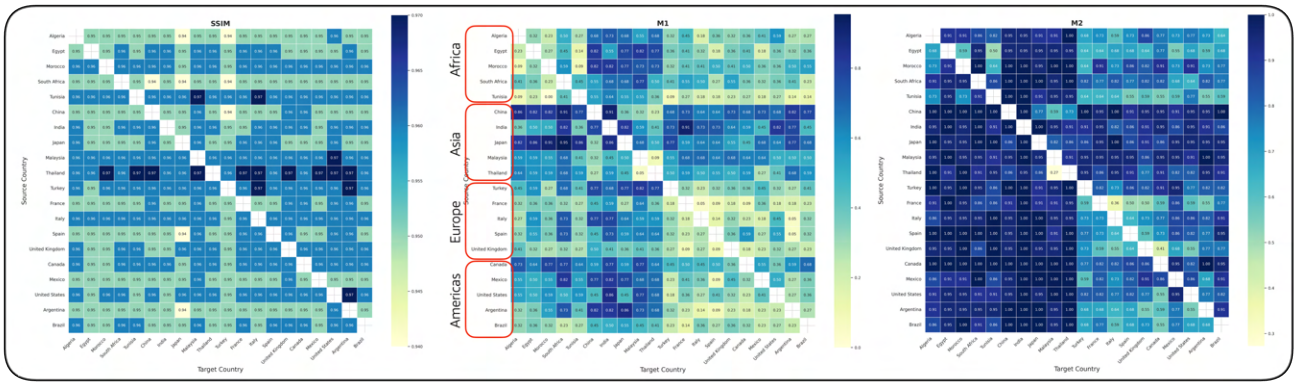

Figure 9. Heatmaps showing performance across source-target country pairs for SSIM (structural fidelity), M1 (cultural divergence and alignment), and M2 (net cultural alignment improvement). SSIM (left) remains consistently high across pairs. M1 (middle) shows variability, with stronger performance in Asia. M2 (right) demonstrates overall high scores, though challenges remain in Africa and Europe.

5.4. Country Pair Interaction

Figure 9 presents the performance comparison heatmaps of source-target country pairs across the three metrics.

Structural Similarity The leftmost heatmap (Figure 9) shows the SSIM values across source-target country pairs. SSIM scores are consistently high across most country pairs, ranging between 0.94 and 0.97. This shows that the model effectively preserves the structural integrity of the source images during the cultural adaptation process. Regions like Asia (e.g., China, Japan, and India) and the Americas (e.g., Canada and the United States) exhibit particularly high SSIM scores, indicating minimal structural distortion in adaptations for these target countries. Conversely, slightly lower SSIM scores are observed for countries such as Argentina, Algeria, and South Africa. This suggests that the model struggles to maintain pose and structural coherence when adapting images to these regions, possibly due to the complexity or uniqueness of cultural features.

Primary Metric The middle heatmap (Figure 9) displays M_1 scores, which evaluate whether the adapted image diverges from the source country's features and moves closer to the target cultural features. The heatmap reveals significant variability in M_1 scores, with values ranging from 0.2 to 0.8 across different country pairs. Asia consistently achieves higher M_1 scores, with regions such as China, India, Japan, and Malaysia performing particularly well. This shows the model's strong ability to embed target cultural features while reducing alignment with the source culture for these countries. In contrast, countries from Africa (e.g., Algeria, Egypt, and Tunisia) and parts of Europe (e.g., UK, France) display lower M_1 scores, suggesting challenges in achieving cultural divergence from the source images, which may stem from the presence of stereotypical representations or insufficient adaptation of cultural features.

	Source	Target	Similarity
Yes	43	41	46
No	4	8	3
Maybe	3	1	1

Table 1. Human Evaluation of (A) Cultural Relevance of source image to source country and (B) of target image to target country and (C) Structural similarity between source and target.

Secondary Metric The rightmost heatmap (Figure 9) shows the M_2 scores, which measure the net gain in cultural relevance by comparing the alignment of the adapted image with the target country's cultural features. The M_2 scores are generally high across most country pairs, particularly for target countries in Asia (e.g., China, India, Japan) and the Americas (e.g., Canada, Mexico, and the United States). These results reflect the model's success in embedding culturally relevant features for these regions. A notable observation is the diagonal dominance, where source-target pairs involving the same country achieve perfect M_2 scores (value = 1.0).

5.5. Human Evaluation

To further validate the performance of our approach, we conducted human evaluations on a small representative sample of 50 edited images. The evaluation focuses on three key aspects: Structural Similarity, Cultural Relevance of the Target, and Cultural Relevance of the Source. The results of the human evaluations are presented in Table 1.

Most of the images (43) were deemed culturally relevant, highlighting the quality of the original dataset in representing source cultural features. A small number of images (4) lacked clear cultural relevance. This could arise due to

the source images' ambiguous or generic cultural markers. Three images, marked as "Maybe" reflect cases where human evaluators were uncertain about the cultural relevance.

Most images (41) were considered culturally relevant, indicating a high success rate for the proposed approach embedding target cultural features. Eight images, marked as "No" suggest that these adaptations failed to incorporate culturally appropriate elements. Such failures may arise from the introduction of stereotypical features or semantic misinterpretations. One image, marked as "Maybe" reflects uncertainty in cultural alignment, possibly due to subtle or unclear cultural features.

Most images (46) were also deemed structurally similar, demonstrating the robustness of the model in preserving pose, spatial composition, and visual coherence during adaptation. Three images indicated structural inconsistencies such as distortions or pose mismatches that compromised fidelity. One image, marked as "Maybe" reflected minor deviations where structural similarity was uncertain but not completely lost.

6. Error Cases

We analyze pairs of source and target images after applying our editing pipeline to determine error cases where the results can be further improved qualitatively. We discuss the different kinds of errors next.

6.1. Source Image Generation Errors

Source image generation encounters issues when the generated facial features and cultural artifacts fail to represent the majority population of the depicted culture (Figure 10). For instance, a South African image may result in facial structures and skin colors that do not represent the country's demographic. Similarly, the generated images may deviate entirely from the intended context, such as producing a sculpture instead of a person sculpting.

To address demographic and cultural inaccuracies in source image generation, we can use CLIP-based semantic supervision to enforce cultural relevance and activity alignment during generation by comparing textual prompts and generated images.

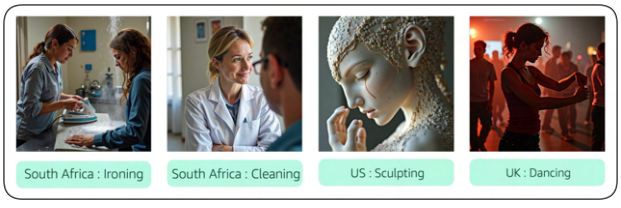

Figure 10. Image generation errors (source image).

6.2. Target Image Adaptation Errors

Stereotypical Attire When translating cultural elements into target images, the model often produces stereotypical or cartoonish representations (Figure 11). For example, in cases where an Indian athlete is portrayed, the attire might reflect non-athletic traditional clothing that is culturally inappropriate for the activity. Similarly, depicting a crowned figure for an Argentinian-to-UK adaptation introduces irrelevant stereotypes by portraying the person wearing a crown.

To mitigate stereotypical or cartoonish representation in target images, we may use SAM to segment out culturally inappropriate attire and replace it with realistic, context-appropriate clothing using text-to-image generation models. We could also incorporate pose-aware generation tools (e.g., ControlNet with Stable Diffusion) conditioned on detected activities to ensure attire aligns with the activity.

Figure 11. Image generation errors: Stereotypical Attire.

Pose Inconsistency Pose inconsistency arises when the adapted image fails to align the body structure or orientation of the target individual with the source image (Figure 12). Examples include individuals facing backwards, making facial adaptation and cultural verification difficult. This is shown in scenarios like hiking, where the back-facing figure obstructs judgment on cultural relevance.

This can be fixed using OpenPose for pose estimation and integrating pose maps as conditioning inputs for the generative model to ensure pose consistency. We can also utilize diffusion-based pose refinement methods to iteratively correct body orientation and maintain activity-relevant structural integrity.

Figure 12. Image generation errors: Pose Inconsistency.

Unnatural Images These frequently occur when the adaptation introduces unrealistic or non-human features (Figure 13). For example, some adaptations may result in textures resembling sculptures or cartoonish characters, as seen in swimming or painting scenarios.

We can use GAN-based post-processing (e.g., StyleGAN2 refinement) to correct unrealistic artifacts and enhance realism.

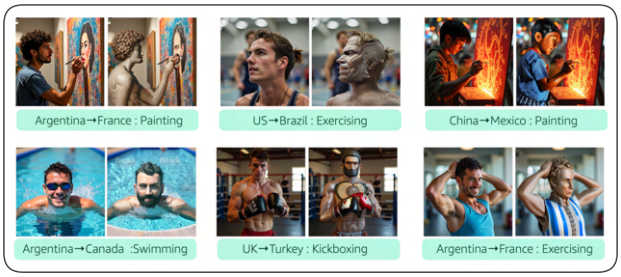

Figure 13. Image generation errors: Unnatural Images.

Distorted Images Distorted images are termed to those where facial features appear exaggerated or deformed (Figure 14). Adaptations, such as a person cleaning with a distorted face or gender misappropriation in sports activities, illustrate these errors.

We can use facial landmark models (e.g., MediaPipe or OpenFace) to detect and enforce correct facial proportions during generation to reduce facial distortions and structural inconsistencies. Another post-processing approach is to segment distorted facial regions using SAM and repair them using fine-tuned face inpainting models.

Figure 14. Image generation errors: Distorted Images.

Activity Mismatch Activity mismatch occurs when the generated images fail to represent the target image performing the same activity as in the source image (Figure 15). Examples include portraying an individual wearing knitted clothing instead of engaging in knitting or misrepresenting eating activities.

To fix this issue, action recognition models can validate and align generated images with the intended activity description. Another way is to compare generated images with textual descriptions of the activity using CLIP similarity to identify and flag mismatches for iterative correction.

Figure 15. Image generation errors: Activity Mismatch.

7. Related Work

The growing interest in culturally aware models has inspired various aspects of our research. [9] propose a data augmentation approach using semantic graphs to enhance cultural components in captions. However, their method often results in inconsistencies at object boundaries when cultural artifacts are copied and pasted into images. Similarly, [7] formalize the task of image transcreation, but their pipelines can produce images that differ significantly from the source or not at all. [12] proposes a CULTUREADAPT pipeline that maintains image coherence by generating semantic masks with bounding boxes and using diffusion-based inpainting. Still, their work inpaints all masked artifacts together does not use precise semantic masks, depends on a closed source model for object tag detection, and has poor performance for specifically human figures. Our method takes into account the different shortcomings of all related work and proposes a simple yet effective solution that achieves good results quantitatively and qualitatively.

8. Resources

All our code is available on GitHub: `github.com/iamshnoo/cultural-humans`. The dataset and image edits can be accessed here: `link to data`.

We use commonly available libraries like numpy, pandas, seaborn, matplotlib, tqdm, and pytorch to set up the basic framework for our code. We access model weights for the FairFace based classifiers and GroundingDINO and the StableDiffusion inpainting models from the transformers and diffusers libraries. For the RAM++ model, we refer to its source repository (`https://github.com/xinyu1205/recognize-anything`), and similarly, for the SAM model, we use the segment anything library. For the YOLO models for face detection, we used the Ultralytics library. We have included a detailed README and our repository, with steps to set up and run the codebase.

9. What We Have Learned

We did pair coding for most parts of the project, with each person taking turns to lead some portion of the work.

In this section, we individually share our learnings from this work.

Anjishnu I gained a deeper understanding of designing modular image-editing pipelines using computer vision techniques like segmentation, object detection, and inpainting. Implementing evaluation metrics, especially integrating SSIM and CLIP-based metrics, taught me how to balance structural fidelity and semantic relevance.

Chahat I learned how to generate high-quality datasets using text-to-image diffusion models, iterating on prompts to achieve consistent and culturally relevant outputs. Analyzing results across multiple metrics and conducting error analysis helped me understand the challenges in balancing quantitative evaluation with qualitative insights for real-world image adaptation tasks.

10. Conclusion

In this work, we present CULTUREADAPTHUMAN, a modular approach for cultural translation of human figures in images that effectively balances structural fidelity and cultural alignment. By leveraging a combination of object detection, segmentation, and generative inpainting, our method preserves the structural integrity of human figures while accurately embedding culturally relevant attributes. We validate our approach on a high-quality dataset of 440 images representing human figures from 20 countries performing 21 activities, achieving strong performance across both quantitative metrics and human evaluations. Our detailed error analysis highlights the model's ability to handle diverse cultural adaptations while identifying challenges such as pose inconsistencies, stereotypical attire, and activity mismatches. Future work will focus on addressing these limitations using more advanced computer vision techniques, improving cultural grounding, and scaling the method to larger datasets for broader generalization.

References

[1] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in llms: A survey. *ArXiv preprint*, abs/2403.15412, 2024. 1

[2] Xiao Ge, Chunchen Xu, Daigo Misaki, Hazel Rose Markus, and Jeanne L Tsai. How culture shapes what people want from ai. *ArXiv preprint*, abs/2403.05104, 2024. 1

[3] 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. Challenges and strategies in cross-cultural NLP. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. 1

[4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 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, 2021. Association for Computational Linguistics. 3

[5] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Tao, L. Kwon, et al. YOLOv8: Ultralytics Implementation of YOLO for Object Detection. *ArXiv preprint*, abs/2301.10821, 2023. [Online]. Available: https://arxiv.org/abs/2301.10821. 2

[6] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 2

[7] Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. *ArXiv preprint*, abs/2404.01247, 2024. 1, 9

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[9] Zhi Li and Yin Zhang. Cultural concept adaptation on multimodal reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 262–276, Singapore, 2023. Association for Computational Linguistics. 9

[10] Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *ArXiv preprint*, abs/2406.03930, 2024. 1

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv preprint*, abs/2303.05499, 2023. 3

[12] Anjishnu Mukherjee, Ziwei Zhu, and Antonios Anastasopoulos. Crossroads of continents: Automated artifact extraction for cultural adaptation with large multimodal models. *ArXiv preprint*, abs/2407.02067, 2024. 1, 3, 9

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *ArXiv preprint*, abs/2112.10752, 2021. 3

[14] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 3