An image speaks a thousand words, but can everyone listen? On image transcreation for cultural relevance

Anonymous ACL submission

Abstract

Given the rise of multimedia content, human translators increasingly focus on culturally adapting not only words but also other modalities such as images to convey the same meaning. While several applications stand to benefit from this, machine translation systems remain confined to dealing with language in speech and text. In this work, we introduce a new task of translating *images* to make them culturally relevant. First, we build three pipelines comprising state-of-the-art generative models to do the task. Next, we build a two-part evaluation dataset – (i) *concept*: comprising 600 images that are cross-culturally coherent, focusing on a single concept per image; and (ii) *application*: comprising 100 images curated from real-world applications. We conduct a multi-faceted human evaluation of translated images to assess for cultural relevance and meaning preservation. We find that as of today, image-editing models fail at this task, but can be improved by leveraging LLMs and retrievers in the loop. Best pipelines can only translate 5% of images for some countries in the easier *concept* dataset and no translation is successful for some countries in the *application* dataset, highlighting the challenging nature of the task. Our code and data is released here.[1]

1 Introduction

> *We shall try... to make not word-for-word but sense-for-sense translations.*
>
> \- Jerome (384)

Since the time ancient texts were first translated, philosophers and linguists have highlighted the need for cultural adaptation in the process (Jerome, 384; Khaldun, 1377; Dryden, 1694; Jakobson, 1959; Nida, 1964) – achieving the same "effect" on the target audience is essential (Nida, 1964). Further, with increased consumption and distribution of multimedia content, scholars in translation studies (Chaume, 2018; Ramière, 2010; Sierra, 2008) challenge the notion of simply translating words, highlighting that visuals, music, and other elements contribute equally to meaning. While each modality carries its own information, interaction between modalities creates deeper, emergent meanings. Partial translation disturbs this multimodal interaction and causes cognitive dissonance to the receptor (Esser et al., 2016). Traditionally, translation has been associated with language in speech and text. To broaden its scope to all modalities, and emphasize on the translator's creative role in the process, the term *transcreation* is seeing widespread adoption today.

Transcreation is prevalent in several fields and its precise implementation is often tied to the end-application, as shown in Figure 1. For example, in *audio-visual media* (AV), the goal is to evoke similar emotions across diverse audiences. In line with this goal, the Japanese cartoon `Doraemon` made many changes like replacing omelet-rice with pancakes, chopsticks with forks and spoons or yen notes with dollar notes, when adapting content for the US.[2] Sometimes, the translation is context-dependant, as in the US movie `Inside Out`, where bell peppers is used as a substitute for broccoli in Japan, as a vegetable that children don't like. In *education*, the goal is to create content that includes objects a child sees in their daily surroundings, known to aid learning (Hammond et al., 2020). Many worksheets already do this, where the same concepts of addition and counting are taught using different currency notes or celebration-themed worksheets, in different regions. Finally, in *advertisements and marketing*, we see global brands localize advertisements to sell the same product, a strategy proven to boost sales (Ho, 2016). `Coca-cola` is a famous example, an embodiment

[1] https://anonymous.4open.science/r/image-translation-6980

[2] http://tinyurl.com/doraemon-us

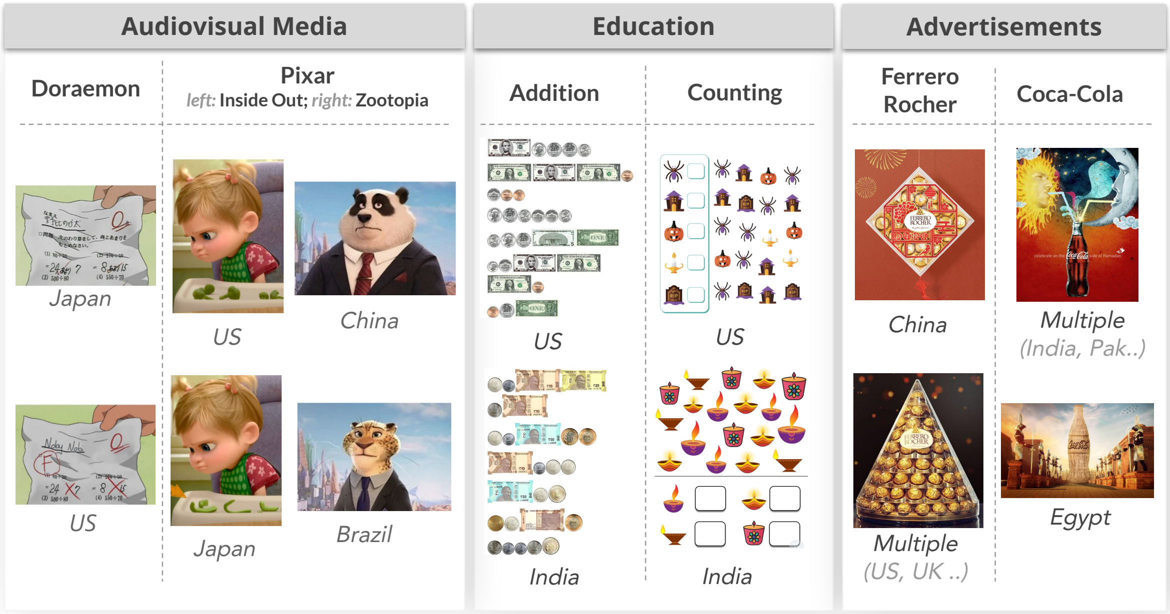


Figure 1: **Image transcreation** as done in various applications today: *a) Audiovisual (AV) media*: where several changes were made to adapt Doraemon to the US context like adding crosses and Fs in grade sheets, or in Inside Out, where broccoli is replaced with bell peppers in Japan as a vegetable that children don't like; *b) Education*: where the same concepts are taught differently in different countries, using local currencies or celebration-themed worksheets; *c) Advertisements*: where the same product is packaged and marketed differently, like in Ferrero Rocher taking the shape of a lunar festival kite in China, and that of a Christmas tree elsewhere.

of "Think Global, Act Local", that tailors its ads to resonate with local cultures and experiences and deeply connect with its audience.

Contribution 1 *(Task)*: In this paper, we take a first step towards transcreation with machine learning systems, by assessing capabilities of generative models for the task of **image transcreation** across cultural boundaries. In text-based systems alone, models struggle with translating culture-specific information, like idioms (Liu et al., 2023). Moreover, to our knowledge, automatically transcreating visual content has previously been unaddressed.

Contribution 2 *(Pipelines)*: In §2, we introduce three pipelines for this task – **a)** `e2e-instruct` *(instruction-based image-editing)*: that edits images directly following a natural language instruction; **b)** `cap-edit` *(caption → LLM edit → image edit)*: that first captions the image, makes the caption culturally relevant, and edits the original image as per the culturally-modified caption; **c)** `cap-retrieve` *(caption → LLM edit → image retrieval)*: that uses the culturally-modified caption from `cap-edit` to retrieve a natural image instead. We also experiment with GPT-4o and DALLE-3 to generate new images using culturally-modified captions (§A.4).

Contribution 3 *(Evaluation dataset)*: Given the unprecedented nature of this task, the evaluation landscape is a blank slate at present. We create an extensive and diverse evaluation dataset consisting of two parts (*concept* and *application*), as detailed in §3. *Concept* comprises 600 images across seven geographically diverse countries: Brazil, India, Japan, Nigeria, Portugal, Turkey, and United States. Five culturally salient concepts and related images are collected across a consistent set of universal categories (like food, beverages, celebrations, and so on) from each country. *Application* comprises 100 images curated from real-world applications like educational worksheets and children's literature.

Contribution 4 *(Human evaluation)*: In §4, we conduct human evaluation of images transcreated for both *concept* and *application*, across all seven countries. We find that as of today, image-editing models fail at this task, but can be improved by leveraging LLMs and retrievers in the loop. Even the best models can only successfully transcreate 5% images for Nigeria in the simpler *concept* dataset and no image transcreation is successful for some countries in the harder *application* dataset.

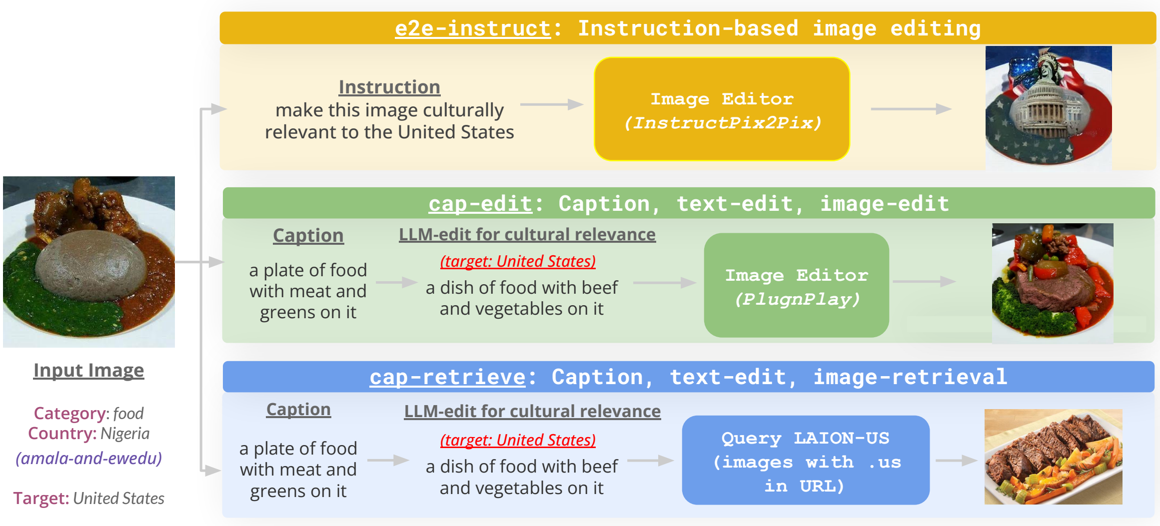


Figure 2: *Pipelines to transcreate images:* `e2e-instruct` takes as input the original image and a natural language instruction; `cap-edit` first captions the image, uses a LLM to edit the caption for cultural relevance, and edits the original image using the LLM-edit as instruction; and `cap-retrieve` uses this LLM-edit to retrieve a natural image from a country-specific image dataset. Given the unprecedented nature of this task, we create pipelines using pre-existing SOTA models, and benchmark them on our newly created test set.

2 Pipelines for Image Transcreation

We introduce three pipelines for image transcreation comprising of state-of-the-art generative models. The code to run all pipelines with exact prompts used can be found in Table D.2. An overview of each pipeline is in Figure 2.

2.1 `e2e-instruct`: Instruction-based editing

First, we use out-of-the-box instruction-based image editing models to translate the image in one pass. Specifically, we use InstructPix2Pix (Brooks et al., 2023), a model that allows users to define edits using natural language, as opposed to other models requiring text labels, captions, segmentation masks, example output images and so on.[3]

We feed in the original image and instruct the model to *make the image culturally relevant to* `COUNTRY`, following a similar prompt format as that used to train the model. This pipeline is simple and flexible, but relies heavily on the image models' ability to perform culturally relevant edits, which it is currently incapable of doing, as discussed in §4.

2.2 `cap-edit`: Caption, text-edit, image-edit

Our second approach is a modular pipeline that offloads some of the requirement of cultural understanding from image editing models to large language models (LLMs). LLMs have been trained on trillions of tokens of text (Touvron et al., 2023; Achiam et al., 2023), and exhibit at least a certain degree of cultural awareness (Arora et al., 2022). Concretely, we adopt a method that first performs image captioning, edits the caption for cultural relevance using an LLM, and then edits the image using an instruction-based image editing model. In experiments, we use InstructBLIP-FlanT5-XXL[4] (Li et al., 2023) as the image captioner, GPT-3.5[5] for caption transformation, and PlugnPlay as the image editing model (Tumanyan et al., 2023).

2.3 `cap-retrieve`: Caption, edit, retrieve

In `cap-edit`, the final output is sometimes not reflective of how the concept naturally appears in the target country, due to image-editing models being trained to strictly preserve spatial layout (§A.2). Hence, here we rely on retrieval from a country-specific image database instead. Concretely, we first caption the image and edit the caption for cultural relevance, similar to `cap-edit`. Next, we use the LLM-edited caption to query country-specific subsets of LAION (Schuhmann et al., 2022). These subsets are created by parsing image URLs and categorizing them based on the country-code top-level domain they contain. For example, URLs featuring ".in" are assigned to the India subset, those with ".jp" are grouped into the Japan subset, etc.

[3] https://www.timothybrooks.com/instruct-pix2pix

[4] https://huggingface.co/Salesforce/instructblip-flan-t5-xxl

[5] https://platform.openai.com/docs/models/gpt-3-5

Figure 3: *Concept* dataset: We select seven geographically diverse countries and universal categories that are cross-culturally comprehensive. Annotators native to selected countries give us 5 concepts and associated images that are culturally salient for the speaking population of their country.

3 Evaluation Dataset

We design a two-part dataset where the first (*concept*) is meant to serve as a research prototype, while the second (*application*) is grounded in real-world applications like those in Figure 1.

3.1 *Concept* dataset

We collect images for a set of universal categories, across seven countries (Figure 3). We follow the annotation protocol of MaRVL (Liu et al., 2021) for which people local to a region drive the entire annotation process, ensuring the collected data accurately captures their lived experiences. Concretely, our collection process is as follows:

Country Selection: We select seven geographically diverse countries: Brazil, India, Japan, Nigeria, Portugal, Turkey, and United States. But do geographic borders dictate cultural ones? Cultures constantly change and are hybrid at any point in time (Hall, 2015). However, audiovisual adaptation is most often equated with national boundaries (Moran, 2009; Keinonen, 2016), given the significant influence of history, policy, and state regulations on media consumption within countries (Steemers and D'Arma, 2012). Further, from a practical perspective, ML systems need data, whose source can be geographically tagged and segregated. While the ultimate goal is to adapt to individual experiences that shape cultural contexts, focusing on the national level serves as a practical starting point.

Category Selection: Ideally, datasets for different cultures should reflect most salient concepts as they naturally occur in that culture, while retaining some thematic coherence for comparability Liu et al. (2021). Hence, we opt for a list of universal concepts that are cross-culturally comprehensive, as laid out in the Intercontinental Dictionary Series (Key and Bernard Comrie, 2015).

Concept Selection: We hire five people who are intimately familiar with the culture of each of the countries above, and ask them to list five culturally salient concepts, such that they are **a)** commonly seen or representative in the speaking population of the language; and **b)** ideally, are physical and concrete (details in §B). Aggregating all responses, we retain top-5 most frequent concepts in each category, for each country.

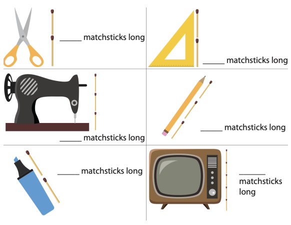
Post-Filtering: The selected concepts and images are additionally verified by 3 native speakers, and those without a majority voting (< 2) are filtered out. We obtain 85 images per country, which become roughly 580 images overall, post-filtering.

3.2 *Application* dataset

The second part of the dataset is curated from real-world applications (*education* and *literature*), a choice guided by availability of data resources.

Education: Research suggests that incorporating objects in a child's surrounding and grounding content in their culture aids learning (Council et al., 2015). Looking at math worksheets for grades 1-3, we find this to be true. We source worksheets from K5 Learning,[6] a US-based learning platform. The transcreation process is tied to the task here, and may not be as straightforward as replacing currency notes in Figure 1. For example, in the left below, the model must find differently-colored elements while retaining the count of each colored object during transcreation, or on the right, where its necessary to find objects that can be measured using the chosen replacement for a matchstick.

(a) *Task (counting):* Count the number of blue, violet, red and yellow christmas balls



(b) *Task (measurement):* Use matchsticks to measure objects

Literature: We curate images from Bloom Library,[7] a digital library of stories for children released for research purposes by Leong et al. (2022).[8] Dealing with a sequence of images is out-of-scope of our current work, hence we collect the first image in each story along with its text that is later used to guide the transcreation. We manually select roughly 60 images out of 400 from the *eng* subset, making sure the selected images are of high quality and de-duplicated (Figure 5).

3.3 Why the two-part dataset?

Even though our eventual goal is to transcreate images for real-world applications, real-world scenes are complex, comprising of multiple interacting objects, and have application-specific constraints, making the task harder. For example, in Figure 4b, one is constrained to find objects of a specific length that can be measured using a matchstick.

[6]https://www.k5learning.com/free-worksheets-for-kids We obtain permission to use and distribute the worksheets for non-commercial research purposes from the publisher.
[7]https://bloomlibrary.org/
[8]https://huggingface.co/datasets/sil-ai/bloom-vist

Figure 5: *Story text:* My mom bought rice.

With *concept*, we build a prototype which has the following features: **a)** *diverse*: images are collected across 7 geographically spread-out countries; **b)** *single concept or object per image*: making it easier to analyse model errors when one image represents a concept in isolation; **c)** *loose constraints on output*: the goal is simply to increase cultural relevance while staying within bounds of the universal category.

Below, we discuss how all models face difficulties even with *concept*, further strengthening the need for it in evaluation.

4 Human Evaluation and Quantitative Metrics

Evaluation of image-editing models typically relies on quantitative metrics and qualitative analysis of a few select samples.[9] While image-editing focuses on image quality and how closely the edit follows the instruction, image-transcreation comes with additional requirements such as cultural relevance, meaning preservation, and so on. Hence, we design an extensive questionnaire and conduct human evaluation to assess the quality of *all* generated images, across both parts of the dataset (Table 1). Evaluators are shown the source image and the three pipeline outputs in a single instance, (Figure 11). This ensures that scores capture relative differences across pipelines. Further, the order of pipeline outputs is randomized so as to not bias the ratings.

[9]Some skip a quantitative evaluation altogether as in Hertz et al. (2022).

ID	Question	Property	Applications	Performance
		Concept Dataset		
C0	Is there any visual change in the generated image compared to the original image?	visual-change	None (*helps filter non-edits*)	e2e-instruct cap-edit cap-retrieve
C1	Is the generated image from the same semantic category as the original image?	semantic-equivalence	AV (Zootopia); Education	e2e-instruct cap-edit cap-retrieve
C2	Does the generated image maintain spatial layout of the original image?	spatial-layout	AV (Doraemon, Inside Out)	e2e-instruct cap-edit cap-retrieve
C3	Does the image seem like it came from your country/ is representative of your culture?	culture-concept	AV, Education, Ads	e2e-instruct cap-edit cap-retrieve
C4	Does the generated image reflect naturally occurring scenes/objects?	naturalness	Ads (Ferrero Rocher)	e2e-instruct cap-edit cap-retrieve
C5	Is this image offensive to you, or is likely offensive to someone from your culture?	offensiveness	All	e2e-instruct cap-edit cap-retrieve
-	For edited images, is the change meaningful (C1) and culturally relevant (C3)?	meaningful-edit	All	e2e-instruct cap-edit cap-retrieve
		Application Dataset		
E/S0	Is there any visual change in the generated image compared to the original image?	visual-change	None (*helps filter non-edits*)	e2e-instruct cap-edit cap-retrieve
E1	Can the generated image be used to teach the concept of the worksheet?	education-task	Education	e2e-instruct cap-edit cap-retrieve
S1	Would the generated image match the text of the story in a children's storybook?	story-text	AV, Literature	e2e-instruct cap-edit cap-retrieve
E/S2	Does the image seem like it came from your country/is representative of your culture?	culture-application	All	e2e-instruct cap-edit cap-retrieve
-	For edited images, is the change meaningful (E/S1) and culturally relevant (E/S2)?	meaningful-edit	All	e2e-instruct cap-edit cap-retrieve

Table 1: Questions asked for evaluation, the applications a model with this property would benefit (examples from Figure 1), and the pipeline ranking for the property tested (first second third).

4.1 *Questions and Findings*: Concept

End Goal: To transcreate the image such that the final image: **a)** belongs to the same universal category as the original (like food, animals etc.), and **b)** has higher cultural relevance than the original image, for a given target country.

However, note that we ask many more questions on layout preservation, offensiveness etc, since different applications may have different constraints on the output, as shown in Table 1. A summary of responses are below, while detailed analyses of responses can be found in §D:

C0: Is there any visual change in the generated image, when compared with the source image? `cap-retrieve` maximally edits images, with roughly 90% scoring 5 (Figure 6); `e2e-instruct` makes no edit sometimes, with 40-60% images scoring 1; and `cap-edit` lies mid-way.

C1: If an edit is made, is it meaningful? For images with **C0** $>$ 2, (indicating some visual changes), we observe that `cap-edit`'s changes maximally retain the universal category, for ex., a food item from country A is changed to another food item from country B; whereas `e2e-instruct` often makes meaningless edits like pasting flag colors of the target country on the image (§A.1). `cap-retrieve` is highly variable; for some countries (India, US), it is better than `cap-edit` and for some (Nigeria), it is very noisy.

C3: Are the edited images more culturally relevant than the original image? Here, we compare the change in the final image's cultural relevance score with the original image (Figure 6). `cap-retrieve` has the highest % of images with a positive change, followed by `cap-edit` after a relatively large gap, while `e2e-instruct` performs worst. This shows that offloading the cultural translation to LLMs generally helps, and natural images are highly preferred over edited images when assessing for culture.

C1+C3: What proportion of images are successfully transcreated? We define **C0** > 2 & **C1** > 2 & $\mathbf{C3}_{edited} > \mathbf{C3}_{original}$ as the criteria for a successful transcreation. Best pipelines can only transcreate 5% images for some countries (Nigeria); while the accuracy is 30% for some others (Japan), indicating that this task is far from solved.

4.2 *Questions and Findings*: Application

End Goal (*Education*): To transcreate such that the final image: **a)** can be used to teach the same concept as the original image (like counting); **b)** has higher cultural relevance than the original image, for a given target country.

Figure 6: *Human ratings for the concept dataset*: Our primary goal is to test whether the *edited image belongs to the same universal category* as the original image (**C1**) and whether it *increases cultural relevance* (**C3**). We plot the count of images that can do both above (**C1+C3**), and observe that the best pipeline's performance ranges between 5% (Nigeria) to 30% (Japan).

End Goal (*Stories*): To transcreate such that the final image: **a)** matches the text of the story; **b)** has higher cultural relevance than the original image, for a given target country.

Observations: Overall, responses to individual questions are similar to as observed for the `concept` dataset. The task here is much harder than simply transcreating within a universal category like in `concept` because of which no image is successfully transcreated by any pipeline for some countries (Portugal). In Figure 7 we see a sample output where `e2e-instruct` makes the cherries a red that resembles the Japan flag, and `cap-edit` is a successful transcreation because even though there is a semantic drift from cherries to flowers, the worksheet can be used to teach counting. Detailed results are in §D.1.

4.3 Quantitative Metrics

For image-editing, these typically capture how closely the edited image matches – **(i)** the original image; and **(ii)** the edit instruction. Following suit, we calculate two metrics:

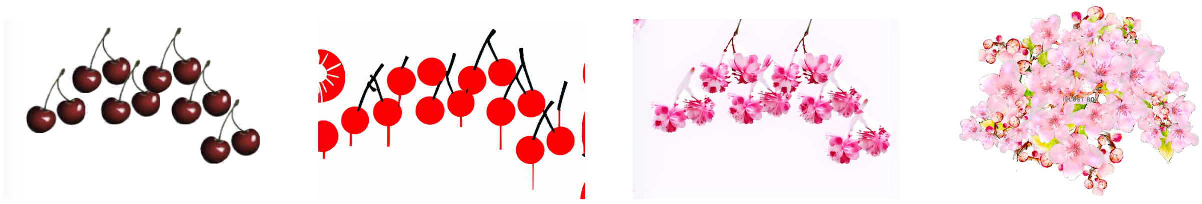
a) *image-similarity*: we embed the original image and each of the generated images using DiNO-ViT (Caron et al., 2021) and measure cosine similarity
b) *country-relevance*: we embed the text – `This image is culturally relevant to {COUNTRY}`, and the edited images using CLIP (Radford et al., 2021) and calculate their cosine similarity.

We present results for both metrics in Figures 20 and 21. A discussion on correlation of these metrics with human evaluation is in §C.

We find that overall for *image-similarity*, `e2e-instruct` scores highest, closely followed by `cap-edit`, while `cap-retrieve` lags behind, consistent with human ratings. However, note that our goal here is to have the right trade-off between image-similarity and the naturalness of the edited image (which cannot be captured by this metric). Figure 8 shows an example of the final image having a high similarity with the original, but nonetheless looks unnatural.

For the *country-relevance score*, we observe that it has a high recall but low precision. These scores are positively correlated with human ratings for **C3:** `cultural-relevance`, but this metric also scores images containing stereotypical artifacts (such as the ones discussed in §A.1) high on cultural relevance.

Our findings above indicate that quantitative metrics cannot sufficiently capture the quality of transcreation of an image, and developing a BLEU-equivalent, but for images, would be necessary to



(a) original (b) `e2e-instruct` (c) `cap-edit` (d) `cap-retrieve`

Figure 7: *Application:* Education; *Target:* Japan — *Task*: count the number of cherries. `cap-edit` is a successful transcreation despite the semantic drift from a fruit to a flower.

make measurable progress on this task.

5 Related Work

Cultural diversity in image generation: Several recent works investigate cultural awareness of text-to-image (T2I) systems typically highlighting biases towards certain cultures. Hutchinson et al. (2022) highlight how under-specified prompts show gender and western cultural biases, Jha et al. (2024) analyse regional stereotypical markers in generated images, Naik and Nushi (2023) discuss occupational biases of neutral prompts and personality trait associations with limited groups of people, Cho et al. (2023) reveal skin-tone biases and Bird et al. (2023) discuss associated risks of these biases for society. Some other works focus on ways to probe for and evaluate cultural relevance of generated images. Ventura et al. (2023) derive prompt templates to unlock the cultural knowledge in T2I systems, and Hall et al. (2023) evaluate the realism and diversity of T2I systems when prompted to generate objects from across the world. While all of these works are targeted towards assessing and mitigating cultural biases in pre-trained models, our work is targeted towards an *application* (i.e. transcreating visual content) that would benefit by such efforts that improve the cultural understanding and diversity of image generation models.

Image-editing models have evolved over the years from being capable of single editing tasks like style transfer (Gatys et al., 2015, 2016) to handling multiple such tasks in one model (Isola et al., 2017; Choi et al., 2018; Huang et al., 2018; Ojha et al., 2021). Today, their capabilities range from performing targeted editing that preserves spatial layout, local in-painting, to edits that can follow natural language instructions (Brooks et al., 2023). We choose InstructPix2Pix (Brooks et al., 2023) to experiment with, given its flexibility to prompt with natural language instructions, as opposed to other models requiring text labels, captions, segmentation masks, example output images and so on. It has also consistently been one of the most downloaded image-editing models on HuggingFace.[10] As discussed in Section 4 however, these models are only capable of making color, shape and style changes, and lack a deeper understanding of natural language. No image-editing works have tackled the semantically complex task of cultural transcreation. We hope that our work paves the way to building image-editing models that truly understand natural language, which can benefit multiple applications, including ours.

6 Conclusion

In this paper, we introduce a new task of **image transcreation** with machine learning systems, where we culturally adapt visual content to suit a target audience. Translation has traditionally been limited to language, but with increased consumption of multimedia content, translating *all* modes in a coherent way is essential. We build three pipelines comprising state-of-the-art generation models, and show that end-to-end image editing models are incapable of understanding cultural contexts, but using LLMs and retrievers in the loop helps boost performance. We create a challenging two-part evaluation dataset: (i) *concept* which is simple, cross-culturally coherent, and diverse; and (ii) *application* which is curated from education and stories. We conduct an extensive human evaluation and show that even the best models can only translate 5% images for select countries (like Nigeria) in the easier *concept* dataset and no image transcreation is successful for some countries (like Portugal) in the harder *application* dataset. Our code and data is released to facilitate future work in this new, exciting line of research.

[10] https://huggingface.co/models?pipeline_tag=image-to-image&sort=downloads

7 Limitations

Categorizing culture based on country: In §3, we acknowledge that cultures do not follow geographic boundaries. It varies at an individual level and is shaped by one's own life experiences. However, the content of several multimedia resources is often influenced by state regulations and policies decided at the national level. Further, a nation has long history which ties people together and influences their languages, customs and way of life. Finally, from a practical standpoint, data for machine learning systems can be segregated based on physical boundaries by geo-tagging it. All these factors convinced us that approaching this problem from a nation-level would be a good starting point. Eventually, we'd like to build something that can learn from individual user interaction, and adapt to varied and ever-evolving cultures.

Limited coverage of languages and countries under study: In this work, we consider seven geographically diverse countries given time and budget constraints involved in data collection and human evaluation. Our choices were also motivated by availability of annotators on the crowd-sourcing platform we use, Upwork. Further, in `cap-edit` and `cap-retrieve`, we only explore captioning in English. This is because most image-editing models and retrieval-based models only work with English instructions. However, captioning and querying in languages associated with cultures the images are taken from is certainly an interesting direction for future research.

A one-to-one mapping may never exist: One may argue that a perfect substitute or equivalent of an object in another culture may never exist. While this is certainly true, we'd like to highlight that our focus here is on context-specific substitutions that convey the intended meaning within a localized setting. For example, in Figure 1, we observe that *Inside Out* substitutes broccoli with bell peppers in Japan to convey the concept of a disliked vegetable. However, in the absolute sense, bell peppers is not a substitute for broccoli when we consider other properties like taste, texture, etc. Importantly, the goal of transcreation is to, at the least, *increase* the relatability of the adapted message when compared with the original message. This is also the reason why we compare between the original and edited image's cultural relevance score in the human evaluation in §4, rather than simply looking at absolute cultural relevance values of edited images.

8 Ethical Considerations

What is the trade-off between relatability and stereotyping? Often times, models may be prone to stereotyping and only producing a small range of outputs when instructed to increase cultural relevance. We observe this a lot with InstructPix2Pix, where it randomly starts inserting sakura blossoms and Mt. Fuji peaks, out of context, to increase cultural relevance for Japan. Hence, it is essential that we build models capable pf producing a diverse range of outputs while not propagating stereotypes. Importantly, one must note that the problem itself *does not* suggest promoting stereotypes but rather an output that the audience can relate to better. We must move towards developing solutions that enable one to hit any of the multiple possible right answers in their context.

We may want to preserve the original cultural elements at times: We are also aware that many a times, the goal may be to expose the audience to diverse cultural experiences and not to localize. While we acknowledge that this is extremely important for sharing knowledge and experiences, our work is not applicable in such scenarios. It may also be that we may want to preserve certain elements, while adapt others. In the Japanese anime *Doraemon* for example, creators make some edits to adapt to the US, but preserve most of the original content which is set in the Japanese context. In future work, we'd ideally want to build a system that allows us to visit different points in the relatability/preservation spectrum, that provides for finer-grained object-level control in translation.

Using pre-existing material created for educational and literary purposes: Our application-oriented evaluation dataset is curated from content originally created to teach math concepts (education) or for children's literature. The StoryWeaver images are CC-BY-4.0 licensed, and we have been in communication with the team for simpler curation and release of data for the future. There were no licenses associated with educational worksheets. Hence, we obtain written consent to use and distribute their worksheet for non-commercial academic research purposes only. The written consent is obtained for the following task description and purpose:

Description of Task: We are assessing the capabilities of generative AI technology to edit images and make them more relevant to a particular culture. There are many concepts that are culture-specific, which people who have not been immersed in the culture may not understand or be aware of. An important end-application where something like this would be useful is education. For example, if one wants to adapt this math worksheet for children in Japan[11], they might want to replace Christmas trees with Kadomatsu (bamboo decorations used on new years). We found several such worksheets which could benefit from such local adaptation.

Purpose of Use: This is a non-commercial research project. We wish to use some of these images (complete list below), to evaluate our pipelines on cultural adaptation. We also request for permission to distribute to other researchers for non-commercial research purposes only. Please note that we are not training any model on this data and it is being used for testing purposes only. Additionally, if you find our research to be beneficial to your workflow, we would be happy to discuss long-term engagements and collaboration as well.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.

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.

Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

[11] https://www.k5learning.com/worksheets/math/data-graphing/grade-1-same-different-c.pdf

Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I Hong, and Adam Perer. 2023. Zeno: An interactive framework for behavioral evaluation of machine learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Frederic Chaume. 2018. Is audiovisual translation putting the concept of translation up against the ropes?

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

National Research Council et al. 2015. Transforming the workforce for children birth through age 8: A unifying foundation.

John Dryden. 1694. Preface to examen poeticum. In *Examen Poeticum*.

Andrea Esser, Iain Robert Smith, and Miguel Á Bernal-Merino. 2016. *Media across borders: Localising TV, film and video games*. Routledge.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

Melissa Hall, Candace Ross, Adina Williams, Nicolas Carion, Michal Drozdzal, and Adriana Romero Soriano. 2023. Dig in: Evaluating disparities in image generations with indicators for geographic diversity. *arXiv preprint arXiv:2308.06198*.

Stuart Hall. 2015. Cultural identity and diaspora. In *Colonial discourse and post-colonial theory*, pages 392–403. Routledge.

Linda Hammond, Channa Flook, Cook-Harvey, Bridgid Barron, and David Osher. 2020. Implications for educational practice of the science of learning and development. *Applied Developmental Science*, 24(2):97–140.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

George Ho. 2016. Translating advertisements across heterogeneous cultures. In *Key Debates in the Translation of Advertising Material*, pages 221–243. Routledge.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189.

Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. Underspecification in scene description-to-depiction tasks. *arXiv preprint arXiv:2210.05815*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Roman Jakobson. 1959. On linguistic aspects of translation. *Harvard Educational Review*, 29(1):232–239.

Jerome. 384. Letter to pammachius. Translated in Kelly, J. N. (Ed.) (2009). Jerome: Letters (Vol. 1). Oxford University Press.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan K. Reddy, and Sunipa Dev. 2024. Visage: A global-scale analysis of visual stereotypes in text-to-image generation.

Heidi Keinonen. 2016. Cultural negotiation in an early programme format: the finnish adaptation of romper room. *New Patterns in Global Television Formats. Bristol: Intellect*, pages 95–108.

Mary Ritchie Key and editors Bernard Comrie. 2015. Ids. *Max Planck Institute for Evolutionary Anthropology, Leipzig*.

Ibn Khaldun. 1377. *The Muqaddimah: An introduction to history*.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. *arXiv preprint arXiv:2210.14712*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Albert Moran. 2009. Global franchising, local customizing: The cultural economy of tv program formats. *Continuum*, 23(2):115–125.

Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808.

Eugene A. Nida. 1964. *Principles of correspondence in translating*. Summer Institute of Linguistics.

Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. 2021. Few-shot image generation via cross-domain correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10743–10752.

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.

Nathalie Ramière. 2010. Are you" lost in translation"(when watching a foreign film)? towards an alternative approach to judging audiovisual translation. *Australian Journal of French Studies*, 47(1):100–115.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Juan José Martínez Sierra. 2008. *Humor y traducción: Los Simpson cruzan la frontera*. 15. Universitat Jaume I.

Jeanette Steemers and Alessandro D'Arma. 2012. Evaluating and regulating the role of public broadcasters in the children's media ecology: The case of home-grown television content. *International Journal of Media & Cultural Politics*, 8(1):67–85.

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*.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2023. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *arXiv preprint arXiv:2310.01929*.

A Example Outputs

Here, we include sample outputs from the pipelines for select images. All pipelines have their own set of limitations, indicating that we have a long way to go before we can solve this task. Patterns observed for each pipeline can be found below:

A.1 `e2e-instruct`: Instruction-based editing

The models seem to associate flags and colors in them with a particular country/culture and includes these features in the edited images irrespective of the objects mentioned in the caption prompts. Some examples can be seen in Figure 16, where the American flag colors are applied over the Burger to make it relevant to the United States. Similarly, Figure 19 includes Brazil map and flag as part of the editing process. The code to run this pipeline is here.[12] We simply pass in the original image with the instruction *make this image culturally relevant to* `COUNTRY`.

A.2 `cap-edit`: Caption, Text-edit, Image-edit

Spatial dimensions are highly preserved in this pipeline as can be seen in Figure 25. This can sometimes lead to undesirable outcomes or the outputs to look unnatural, as shown below. The code to run this pipeline can be found here.[13]. The exact prompts used for captioning and LLM editing can be found in Table D.2.

(a) original (b) `cap-edit`

Figure 8: Example of how preserving the spatial layout of the original image can lead to unnatural looking outputs. Here, the final image shows *a cup of chai*, but a typical cup of chai looks different in India.

[12]https://anonymous.4open.science/r/image-translation-6980/src/pipelines/e2e-instruct.py

[13]https://anonymous.4open.science/r/image-translation-6980/src/pipelines/caption-llm_edit.py

A.3 `cap-retrieve`: Caption, Text-edit, Retrieval

The obtained images through the retrieval pipeline seem to be noisy with a low precision but high recall. Some of the images are better representatives of that country's culture compared to the other two pipelines, given that they are real images. However, this pipeline also suffers from failure cases of retrieving images which may be too different from the source image or retrieving irrelevant outputs. Examples are shown below:

(a) original (b) `cap-retrieve`

Figure 9: Example of how the retrieved output may at times look completely different from the original image.

(a) original (b) `cap-retrieve`

Figure 10: Example of how the retrieved output may be irrelevant/noisy. Here, we can see it behaving like a bag-of-words since the llm-edit used to prompt for retrieval is: *A sunflower stands tall against the backdrop of a clear blue sky in India.*

A.4 GPT4-o + GPT-4 + DALLE-3

We use the GPT-4 family of models for this pipeline. Since DALLE-3 works with detailed prompts (Betker et al., 2023), we prompt GPT4-o

Figure 11: Screenshot of how one instance looks like for human evaluation on the Zeno platform.

to give detailed captions for images. We use GPT4 to edit these captions and prompt DALLE-3 to generate images. To make the images look natural, we add "photo, photograph, raw photo, analog photo, 4k, fujifilm photograph" to the prompt.[14] Even then, the images do have a distinct style. Qualitatively, we observe that the captions and caption-edits capture fine-grained details which shorter captions in the previous two pipelines cannot. The overall pipeline can be found in Figure 12. All visualizations can be found in the released code repository. Note that GPT4-o + DALLE-3 outputs could not be human evaluated since their APIs were released on May 13, 2024. Further, the images' distinct style defeats the purpose of randomizing pipeline outputs for human evaluation.

B Annotation Instructions

Our annotation and human evaluation instructions are as follows. We host our data on the Zeno[15] (Cabrera et al., 2023) platform and hire people on Prolific[16] to do the annotation and evaluation. Each worker is paid in the range of 10-15 USD per hour for the job. This work underwent IRB screening prior to conducting the evaluation.

[14] https://www.reddit.com/r/dalle/comments/1au10g6/generate_realistic_pictures_with_dalle/

[15] https://zenoml.com/

[16] https://www.prolific.com/

B.1 Part-1: Concept Collection

This task is part of a research study conducted by *[name]* at *[place]*. In this research, we aim to create AI models that can generate images that are appropriate for different target audiences, such as people who live in different countries.

You will be given a set of universal categories that cover a diverse range of objects and events. These categories include things like bird, food, clothing, celebrations etc. You have to give Wikipedia links for 5 salient concepts for each category, that are most prevalent in your country and culture, for each of these categories.

The two key requirements are for the concepts to be: **a)** commonly seen or representative of the speaking population of your country; **b)** ideally, to be physical and concrete.

You have to make sure that the concept you select can be represented visually, i.e., an image can be used to represent the concept.

A few examples for the food category for United States are given below:

- https://en.wikipedia.org/wiki/Hamburger
- https://en.wikipedia.org/wiki/Hot_dog

Note: Links to wikipedia pages in English is preferred, but you can even provide a link to other languages if the concept is not present on English Wikipedia.

The categories are as follows: Bird, Mammal, Food, Beverages, Clothing, Houses, Flower, Fruit, Vegetable, Agriculture, Utensil/Tool, Sport, Celebrations, Education, Music, Visual Arts, Religion.

B.2 Part-1: Image Collection

This task is part of a research study conducted by *[name]* at *[place]*. In this research, we aim to create AI models that can generate images that are appropriate for different target audiences, such as people who live in different countries.

You will be given a set of universal categories that cover a diverse range of objects and events. These categories include things like bird, food, clothing, celebrations etc. You will also be given 5 concepts in each category that are highly relevant to your culture.

Your task is to give us an image for each concept such that it reflects how it appears in your culture and native surroundings. Ideally this can be a wikipedia or wikimedia image itself. However,

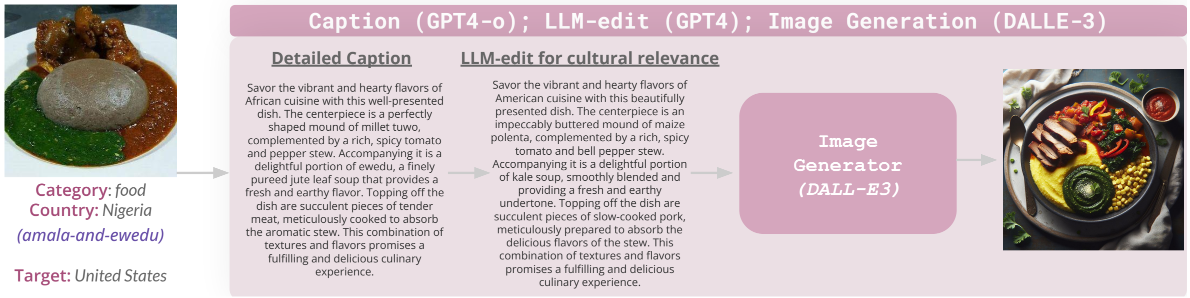


Figure 12: Pipeline for GPT4-based experiments.

if you feel the wikipedia image is not appropriate, please provide us with a CC-licensed image from google image search. To filter for CC-licensing, look at the screenshot below.

A few examples for the food category for United States are given below:

1. **Concept (given to you)**: Hamburger (https://en.wikipedia.org/wiki/Hamburger). **Image link (you have to provide)**: https://upload.wikimedia.org/wikipedia/commons/c/ce/McDonald%27s_Quarter_Pounder_with_Cheese%2C_United_States.jpg

2. **Concept (given to you)**: Hotdog (https://en.wikipedia.org/wiki/Hot_dog) **Image link (you have to provide)**: https://upload.wikimedia.org/wikipedia/commons/thumb/b/b1/Hot_dog_with_mustard.png/220px-Hot_dog_with_mustard.png

Ensure that the images are clear and provide a good representation of the concept as it is experienced or seen in your culture and surroundings.

B.3 Human Evaluation

This task is part of a research study conducted by *[name]* at *[place]*. In this research, we aim to create AI models that can generate images that are appropriate for different target audiences, such as people who live in different countries. You need to be native to one of the following countries, and aware of its culture, to complete the task: Brazil, India, Japan, Nigeria, Portugal, Turkey, United States.

In this evaluation, you will be shown 4 images, as shown in the Figure 11. The top-most image (*Image-1*) is sourced from the internet, from a diverse set of domains like agriculture, food, birds, education etc. This image is being edited to make it culturally relevant to your country and culture, using three state-of-the-art generative AI technologies (*Image-2, Image-3, Image-4*).

You will be asked whether you agree with six questions or statements about each of the images, from **5 (strongly agree)** to **1 (strongly disagree)**:

C0) There are visual changes in the generated image, when compared with the source (top-most) image (*1 → no visual change; 5 → high visual changes*).

C1) The image contains similar content as the source image. For example, if the source is a food item, the target must also be a food item. Use the label to see which domain the source image is from (*1 → dissimilar category; 5 → same category*).

C2) The image maintains the spatial layout of the source image (this can be thought in terms of shapes and overall structure and placement of objects etc.) (*1 → different layout; 5 → same layout*).

C3) The image seems like it came from your country or is representative of your culture (*1 → not culturally relevant; 5 → culturally relevant*).

C4) The image reflects naturally occurring scenes/objects (it does not look unnaturally edited and is something you can expect to see in the real world) (*1 → unnatural; 5 → natural*).

C5) This image is offensive to you, or is likely offensive to someone from your culture (*1 → not offensive; 5 → offensive*).

Stories
S1) The image would match the text of the story in a children's storybook, as shown in the label.
S2) The image seems like it came from your country or is representative of your culture.

Education
E1) The image can be used to teach the concept of the original worksheet, as shown in the label.
E2) The image seems like it came from your country or is representative of your culture.

[Optional]: We would appreciate if you can share observations of certain patterns you found while doing the evaluation, post the study. For example, a few things we noticed are as follows:

1. Some models insert the flag or flag colors in the image, without any context, to increase the cultural relevance of it.

2. Some models exhibit color biases, like making things red/black, when asked to edit an image to make it culturally relevant to Japan.

3. Some models start inserting culturally prominent objects to increase relevance. For example, they commonly insert Mt. Fuji peaks, or cherry blossoms, to make an image culturally relevant to Japan.

B.4 Observations as noted by human evaluators

This is the feedback received for the optional comments in the human evaluation as asked for above. Almost everyone found outputs to be semantically incoherent with random insertions of colors, cultural entities, flag elements and so on, uncovering several biases and gaps that these models have today.

B.4.1 Brazil

- *Overall, I noticed that the colors of Brazil's flag were extensively used in various contexts, creating an unnatural effect on the subject of the pictures. I cannot precisely articulate why, but I felt that these images gave me an impression of Africa rather than Brazil, even though Brazil is an extremely diverse country with a significant African influence. Additionally, I observed numerous abstract representations where only the basic shape from the original picture was retained.*

- *Some images had the colors of the Brazilian flag as if "superimposed" on the objects and images, without making sense with the figure itself*

B.4.2 Japan

- *There are not enough variations to represent Japan. Commonly used subjects - cherry blossoms, pine trees, Mt.Fuji*

- *Characters in Japanese children's picture books tend to have American-leaning faces, making Japanese faces look more adult-oriented*

B.4.3 India

- *Models have put some improper Indian images with only cultural costume and also found many bad generated faces*

B.4.4 Nigeria

- *Some models just changed the pictures to green in an attempt to make it look Nigerian. Images did not match the description.*

- *Models has a lot of black scary images that did not fit the context and doesn't make it culturally relevant to Nigeria. Images generated did not match the original image neither was it relevant to the Nigerian culture.*

B.4.5 Portugal

- *In the math worksheets, for so many times it was generated a picture that would add, random parts of the portugese flag or colors making no sense at all and sometimes it looks like Morocco*

- *Some problems are not related to mathematics: such as the question of associating what each "element" can carry on its back*

B.4.6 Turkey

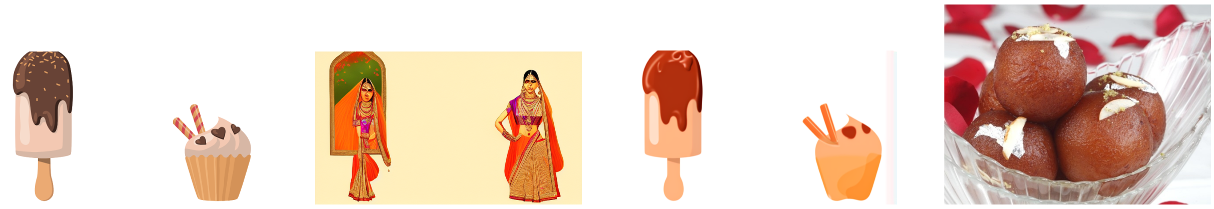
- *Observed that a lot of the edited images included turkey (the animal) illustrations, and also some of the edited images included Turkish flag, mosques, Turkish food, Turkish tea and some clothing styles that were mostly used in ancient times. Some of the edited images were only consisting of the colors of the Turkish flag, which are red and white.*

- *In some instances, where there was a person of color or a person with a different ethnicity in the topmost image, the skin color of the person was changed in the edited images and sometimes beards were added on men, and headscarves were added on women*

B.4.7 USA

- *I did notice that in the majority of images with people/faces, that the AI image rearranged/disoriented the facial features*

- *The AI images related to plants, food and nature seem to be more natural in the edits and effects and way more natural than when applying the same change of effects on people*



(a) original (b) `e2e-instruct` (c) `cap-edit` (d) `cap-retrieve`

Figure 13: *Application:* Education; *Target:* India — *Task*: Pick the largest one among the two icecreams; *InstructBLIP caption*: a cupcake and an ice cream pop on a white background; *LLM-edited caption*: a gulab jamun and a kulfi on a white background. `e2e-instruct` inserts women in traditional indian clothing not relevant to the task, the LLM makes a pretty good edit but the image-editing model in `cap-edit` probably doesn't understand indian sweets like gulab jamun and kulfi, and the retriever in `cap-retrieve` only retrieves one item of two.

(a) original (b) `e2e-instruct` (c) `cap-edit` (d) `cap-retrieve`

Figure 14: *Source:* Japan; *Target:* Brazil — *BLIP caption*: a bowl of ramen with meat and vegetables; *LLM-edited caption*: a bowl of feijoada with beef and vegetables. `e2e-instruct` simply inserts flag colors, `cap-edit` highly preserves structural layout, `cap-retrieve` retrieves a natural image but is structurally different from the source.

(a) original (b) `e2e-instruct` (c) `cap-edit` (d) `cap-retrieve`

Figure 15: *Source:* India; *Target:* Japan — *BLIP caption*: a field of cotton plants; *LLM-edited caption*: a rice paddy field. `e2e-instruct` inserts sakura blossoms and multiple Mt. Fuji peaks in the background, `cap-edit` highly preserves structural layout but looks pretty realistic here.

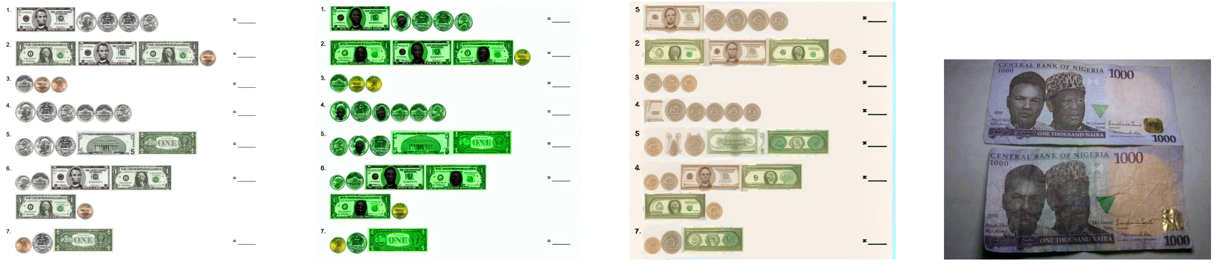
(a) original (b) `e2e-instruct` (c) `cap-edit` (d) `cap-retrieve`

Figure 16: *Source:* USA; *Target:* USA — *BLIP caption*: a hamburger with cheese and pickles on a white background; *LLM-edited caption*: a cheeseburger with pickles on a white bun. `e2e-instruct` heavily inserts flag colors, in `cap-edit` the LLM makes the bun white, `cap-retrieve` works well. Ideally, we do not want any change to be made in this case.

C Quantitative metrics

We find a linear correlation between image-image similarity scores and human evaluation ratings on **C0:** `visual-change`. This helps us determine a threshold beyond which, on average, images get a visual-change score of 1 or 2 (1 means no visual change). A correlation plot for one of the countries is shown in Figure 22.

For the application-oriented evaluation, we simply ask whether the edited image can be used to

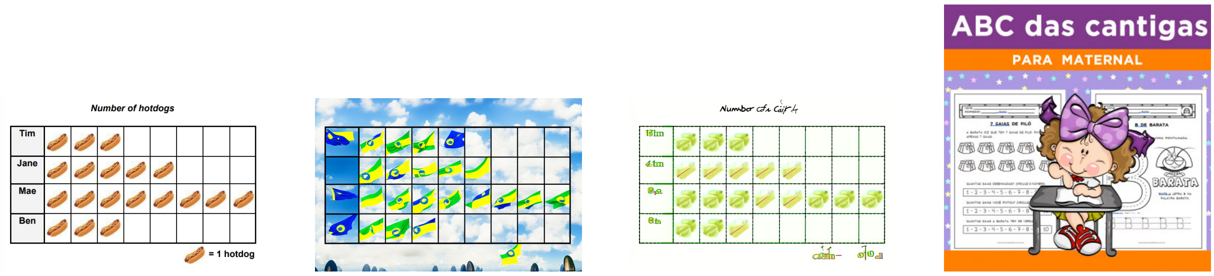


(a) original (b) `e2e-instruct` (c) `cap-edit` (d) `cap-retrieve`

Figure 17: *Application:* Education; *Target:* Nigeria — *Task*: Add the US currency notes; *InstructBLIP Caption*: a math worksheet with coins and notes on it *LLM-edit Caption*: a math worksheet with Naira coins and notes on it. We see the pipelines exhibiting strong color bias both for the notes and the background itself.

(a) original (b) `e2e-instruct` (c) `cap-edit` (d) `cap-retrieve`

Figure 18: *Source:* United States; *Target:* Turkey — *BLIP caption*: a coca cola bottle with a red lid; *LLM-edited caption*: a bottle of coca cola with a red cap in Turkey. `e2e-instruct` doesn't know that coca-cola is black, and makes it red for Turkey, `cap-edit` adds flag details to the logo and the LLM also simply adds "turkey" in the caption while `cap-retrieval` just produces an irrelevant output.



(a) original (b) `e2e-instruct` (c) `cap-edit` (d) `cap-retrieve`

Figure 19: *Application:* Story; *Target:* Brazil — *Task*: Count the number of hotdogs. Here, we see a strong tendency to output elements of the map and flag colors in these models.

solve the same task (in education) or whether it matches the title of the story (for stories). However, if the image is not edited at all, pipelines would still score high on this question, thus biasing our analysis. Since we notice a linear correlation in image-similarity and human ratings for the same question in *concept* evaluation, we determine a threshhold in image similarity beyond which humans give a rating of 1 or 2 to the image (1 means no visual change). This threshhold typically hovers around 0.95-0.97 for each country.

For E1 and S1 application plots in Figure 23, we employ these thresholds to filter images that haven't been edited at all. Images whose image-

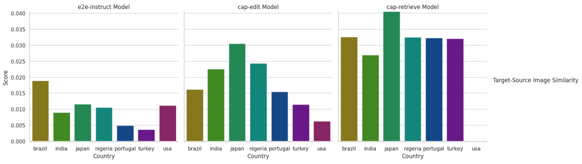


Figure 20: `target-source similarity`, capturing the difference in image-text similarity scores between target and source

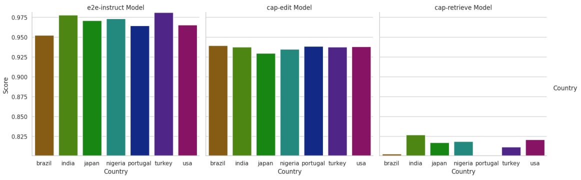


Figure 21: `image similarity difference`, capturing the difference in image similarity scores between target and source

similarity scores greater than the thresholds calculated are filtered out, ensuring that only those images that have been edited are considered for further analysis.

D Continued analysis of human evaluation

We continue analysis of questions asked in Table 1 below:

C0: `visual-change` – First we ask whether the image has been edited at all, to help understand if the edits make sense in the questions that follow. Across all countries, `cap-retrieve` maximally edits images, with roughly 90% scoring 5 (Figure 6). This is expected since here the original image is not input at all in producing the final image. `e2e-instruct` on the other hand makes no edit sometimes, with 40-60% images being given a score of 1. For countries like Brazil and US, this pipeline overwhelmingly paints the image with the flag or flag colors (§A.1), explaining the relatively lower number of 1s.

C1: `semantic-equivalence` – Here, we ask that if an edit is made (**C0**<3) is it a meaningful one? In Figure 6, we observe that `cap-edit` scores highest, while `cap-retrieve's` performance varies based on the country (lower for countries with low digital presence).

C2: `spatial-layout` – For `e2e-instruct` and `cap-retrieve`, we observe similar trends as those observed in **Q1**). For `cap-edit`, while it scores mid to high on visual changes, it surprisingly maintains spatial layout, performing similar to `e2e-instruct`. This signifies that even though `cap-edit` makes visual edits, it does so while preserving spatial layout, helpful for audiovisual translation like in Doraemon, Inside Out and so on.

C3: `culture-concept` – Each original image's cultural relevance score may be different to begin with. Hence, here we plot the delta in scores, relative to the original image. If $\text{score}_{\text{edited}} < \text{score}_{\text{original}}$, we bucket it into $-\Delta$ (*negative change*); if $\text{score}_{\text{edited}} = \text{score}_{\text{original}}$, we bucket it into `0` (*no change*), and if $\text{score}_{\text{edited}} > \text{score}_{\text{original}}$, we bucket it into $+\Delta$ (*positive change*). We observe that `cap-retrieve` performs best across all countries, followed by `cap-edit` and finally `e2e-instruct`. This indicates that while end-to-end image-editing models still have a long way to go in understanding cultural relevance, LLMs can take the responsibility of cultural translation and provide them with concrete instructions for editing or retrieval.

C4: `naturalness` – `cap-retrieve` receives highest scores here since these are natural images retrieved from the internet. `cap-edit` receives a significant number of 4s, because it doesn't look as natural as retrieved images, but probably natural enough, as discussed in §A.2.

C5: `offensiveness` – Almost no images are found to be offensive, which is encouraging.

C1+C3: `meaningful-edit` – We plot counts of pipelines that score above 3 on `semantic-equivalence` and have a positive change in `culture-concept` score ($+\Delta$). These images have been edited such that they increase cultural relevance while staying with bounds of the universal category, which is our end-goal for *concept*. From Figure 6, we can see that performance of the best pipeline is as low as 5% for countries like

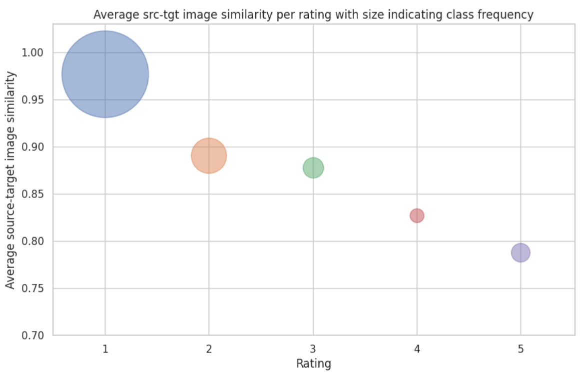


Figure 22: `correlation plot`, capturing linear correlation between human and machine evaluation for Brazil

Nigeria, indicating that this task is far from solved.

D.1 Application Dataset

E1: `education-task` and **S1:** `story-text` – Our observations are similar to what we observe for **C1:** `semantic-equivalence` in *concept*. The retrieval pipeline is especially noisy, given that the requirement of "equivalence" here is that the edited image must be able to teach the same concept (for education) or match the text of the story (for stories), harder than simply matching a category.

E/S1+E/S2: `meaningful-edit` – Similar to **C1+C3**, the count of images that increase cultural relevance, while preserving meaning as required by the end-application, is very low. For countries like Portugal, no pipeline is able to translate any image successfully. For some other countries, the best pipeline is able to translate 10-15% of total images.

D.2 Quantitative Metrics

For image-editing, these typically capture how closely the edited image matches – (i) the original image; and (ii) the edit instruction. Following suit, we calculate two metrics: **a)** *image-similarity*: we embed the original image and each of the generated images using DiNO-ViT (Caron et al., 2021) and measure their cosine similarity; and **b)** *country-relevance*: we embed the text – `This image is culturally relevant to {COUNTRY}`, and the edited images using CLIP (Radford et al., 2021) and calculate their cosine similarity. We present results for both metrics in Figures 20 and 21. A discussion on correlation of these metrics with human evaluation is in §C.

We find that overall for *image-similarity*, `e2e-instruct` scores highest, closely followed by `cap-edit`, while `cap-retrieve` lags behind, consistent with human ratings. For the *country-relevance score*, we observe a similar trend as that for **C3:** `cultural-relevance`.

Figure 23: *Human ratings for the application dataset*: Our goal is to test whether the edited image can be used for the application as before (**E/S1**), and whether it increases cultural relevance (**E/S2**). We plot the count of images that can do both above (**E/S1+E/S2**), and observe that even the best pipeline cannot transcreate any image successfully in some cases, like for Portugal.

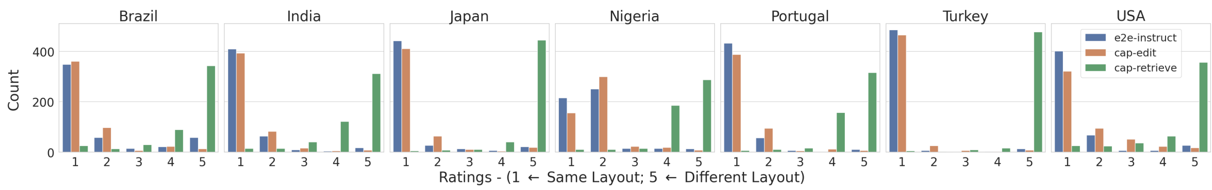


Figure 24: **Q3**: `spatial-layout`, capturing if the structure of the original image is maintained.

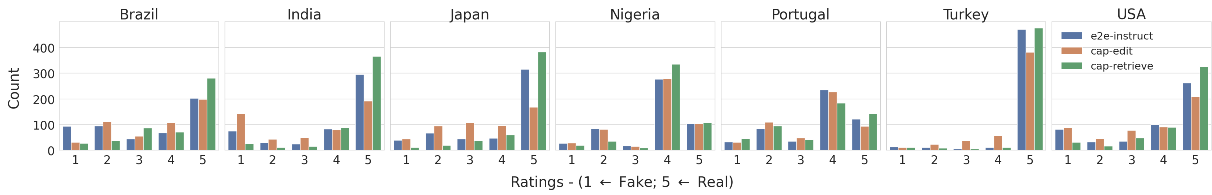


Figure 25: **Q5**: `naturalness` capturing the naturalness of the edited or retrieved image.

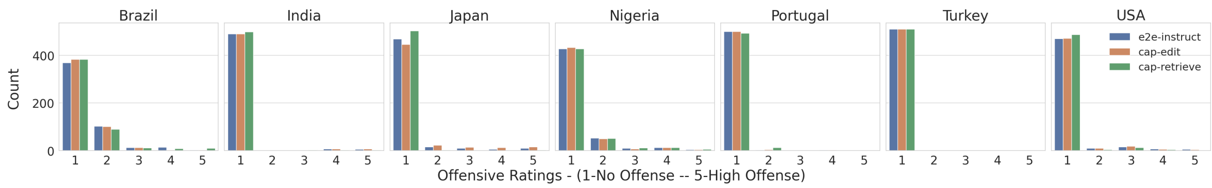


Figure 26: **Q6**: `offensiveness` capturing how offensive each pipeline is

Prompts used for the pipelines described in Section 2.

InstructBLIP Prompt (Captioning)

Concept Dataset

"A short image description:"

Application Dataset (Education)

"This image is from a math worksheet titled: TASK. Describe the image such that it talks about details relevant to the task of the worksheet. The output should be ONLY ONE sentence long."

Application Dataset (Stories)

"This image is from a storybook for children. Caption the image such that it describes details relevant to the story."

GPT3.5 Prompt (LLM-editing)

Concept Dataset

"Edit the input text, such that it is culturally relevant to COUNTRY. Keep the output text of a similar length as the input text. If it is already culturally relevant to COUNTRY, no need to make any edits. The output text must be in English only.
Input: "

Application Dataset (Education)

"Edit the input text, such that it is culturally relevant to COUNTRY. The text describes an image in a math worksheet titled: TASK. Hence, make sure the edit preserves the intent of the task in the worksheet. Keep the output text to be of a similar length as the input text. If it is already culturally relevant to COUNTRY, there is no need to make any edits. The output text must be in English only.
Input: "

Application Dataset (Stories)

"Edit the input text, such that it is culturally relevant to COUNTRY. The text describes an image in a storybook for children. Make sure the edit preserves the meaning of the story. Keep the output text to be of a similar length as the input text. If it is already culturally relevant to COUNTRY, there is no need to make any edits. The output text must be in English only.
Input: "