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

Note: This paper contains examples of potentially offensive generated images.

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 language in speech and text. In this work, we take a first step towards 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 models can only translate 6% 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 translation processes (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 of "Think Global, Act Local", that tailors its ads to resonate with local cultures and experiences and

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

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

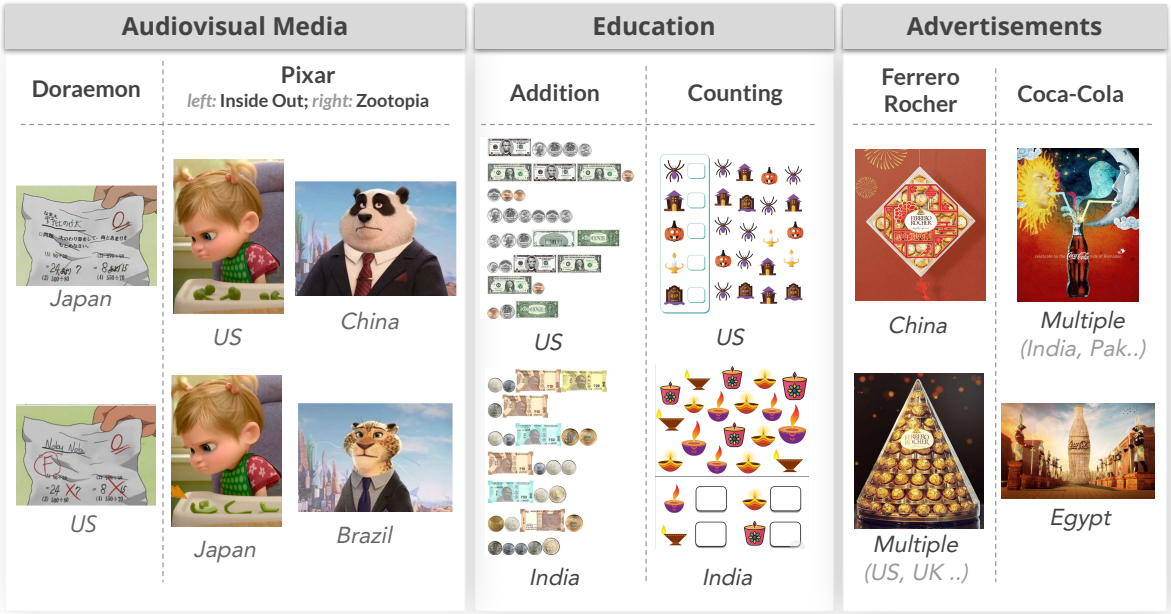


Figure 1: **Image localization** as done in various applications today: *a) Audiovisual (AV) media*: where several changes were made to adapt Doraemon to the US context like adding crosses and Fs in grade sheets, or in Inside Out, where broccoli is replaced with bell peppers in Japan as a vegetable that 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.

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 translation** 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 translating 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; and **c)** `cap-retrieve` *(caption -> LLM edit -> image retrieval)*: that uses the culturally-modified caption from `cap-edit` to retrieve a natural image instead.

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 (food, beverages, celebrations, and so on) making this dataset cross-culturally comparable. *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 translated 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 translate 6% images for Nigeria in the simpler *concept* dataset and no image translation is successful for some countries (like Brazil, Portugal) in the harder *application* dataset.

2 Pipelines for Image Translation

We introduce three pipelines for image translation, all comprising of state-of-the-art generative models.

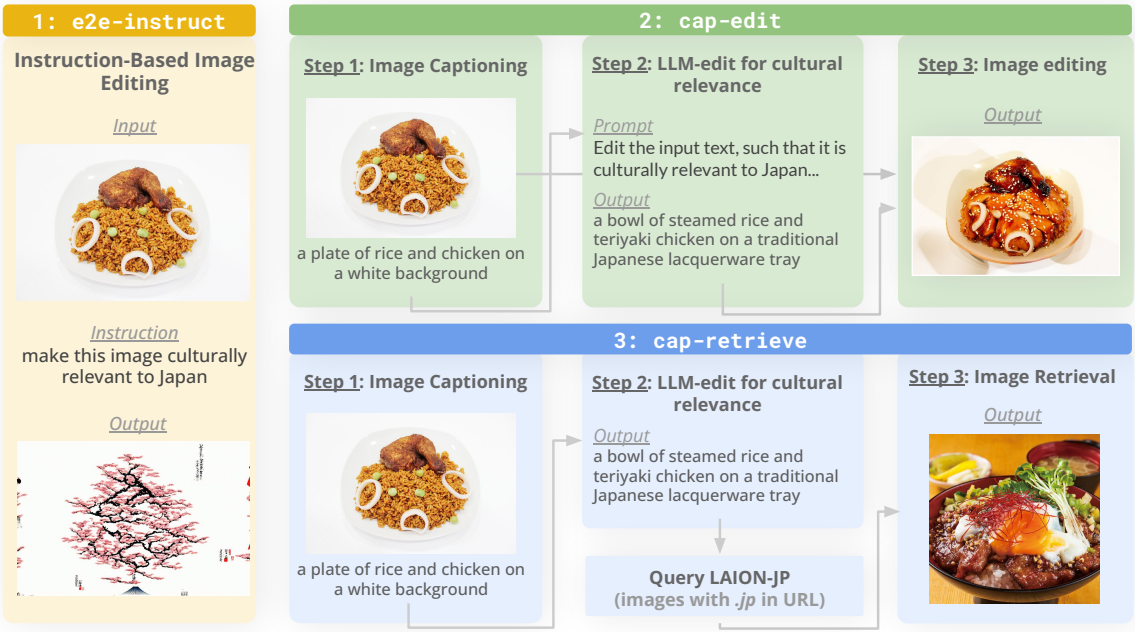


Figure 2: *Pipelines to translate images:* `e2e-instruct` takes as input the original image and a natural language instruction; `cap-edit` is a modular approach which first captions the image, uses a LLM to edit the caption for cultural relevance, and edits the original image using the LLM-edit as instruction; `cap-retrieve` similarly captions the image and LLM-edits the caption, but retrieves from a country-specific image dataset instead.

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.

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). Large language models have been trained on trillions of tokens of text (Touvron et al., 2023; Achiam et al., 2023), and have been shown to exhibit at least a certain degree of cultural awareness (Arora et al., 2022). Concretely, we adopt a method that first performs image captioning, translates the caption for cultural relevance using an LLM, and then edits the image using a caption-based image editing model. In experiments, we use BLIP2-FlanT5-XXL[3] (Li et al., 2023) as the image captioner, GPT-3.5[4] for caption transformation, and PlugnPlay as the image editing model (Tumanyan et al., 2023).

2.3 `cap-retrieve`: Caption, Text-edit, Image Retrieval

In *Step-3* of `cap-edit`, we edit the original image, which can lead to the final output not being reflective of how the concept naturally appears in the target country (§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://huggingface.co/Salesforce/blip2-flan-t5-xxl
[4] 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 part (*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

The first part contains images collected for the same universal categories, across seven countries, as shown in Figure 3. We follow the annotation protocol of MaRVL (Liu et al., 2021) which involves people local to a region driving the entire annotation process. This approach ensures that the data accurately captures their lived experiences. Concretely, the 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, machine learning 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, aligning with established practices in related fields.

Category Selection: Ideally, datasets for different cultures should reflect the most salient concepts and their typical visual denotations, while retaining some thematic coherence for comparability. Hence, following Liu et al. (2021), 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 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).

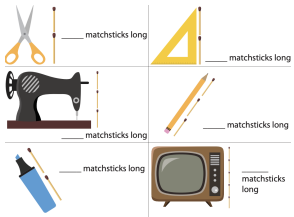
Post-Filtering: The selected concepts and images are additionally verified by 3 native speakers, and those without a majority voting (< 2) are filtered out. We obtain 85 images per country, which become roughly 580 images overall, post-filtering.

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. Hence, we source worksheets from two websites: K5 Learning[5] and NCERT.[6] K5 Learning is US-based while NCERT is specific to India. The translation 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 translation, 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 StoryWeaver,[7] a digital library of stories for children. A story is sequential, consisting of at least 14-15 images with text, on this website. However, dealing with consistency for sequential data is a hard problem in the image-generation space and out-of-scope of the current work. Further, models today struggle to translate a single image alone, as noted in Section 4. Hence, we simply collect the image used alongside the title of the story (Figure 5).

3.3 Why the two-part dataset?

Even though our eventual goal is to translate 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.

Figure 5: *Title of story:* Grandma's glasses

With *concept*, we build a prototype which has the following features: **a)** *diverse*: images are collected across 7 geographically spread-out countries; **b)** *cross-culturally comparable*: concepts selected from a universal set of semantic categories (e.g., food, agriculture), hence enabling cross-cultural comparisons; **c)** *single concept or object per image*: making it easier to analyse model errors when one image represents a concept in isolation; **d)** *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.[8] While image-editing focuses on image quality and how closely the edit follows the instruction, image-translation 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. Evaluators are shown the source image and the three pipeline outputs in a single instance, (Figure 8). This ensures that scores capture relative differences across pipelines. Further, the order of pipeline outputs is randomized so as to not bias the ratings.

All images are run through all pipelines, separately for each of the 7 chosen countries as target. We ask six questions per instance for *concept*, and two per instance for *application*, as detailed

[5] 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.

[6] https://kirandul.kvs.ac.in/school-academics-overview/elementary-education

[7] https://storyweaver.org.in/

[8] 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				
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 title of the story in a children's storybook?	`story-title`	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).

below. To put all of this in perspective, 12,500 questions have been given a 5-point rating across 2,800 images, by each of the 14 participants (two per country) in the evaluation.

4.1 *Questions and Findings*: Concept

The translation here is open-ended; our end goal is to make an edit that doesn't change the universal category, and increases cultural relevance. However, some applications have additional constraints like maintaining the spatial layout (ex: AV media like Doraemon). Hence, our evaluation is multi-dimensional, analysing the strengths and weaknesses of each pipeline, which can help inform its appropriateness for an end-application. We ask human evaluators to give ratings on a 5-point scale on six questions as shown in Table 1. We summarize key findings below, and a detailed analysis for all questions is in §D.

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 `1`s.

C1: `semantic-equivalence` – First, we filter out images with scores `1` or `2` on **C0** since here we want to capture that **if** an edit has been made, 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. For example, it has higher performance for India and US but low for Nigeria. Image generation models have been shown to be reflective of cultures in the US followed by India for under-specified inputs (Basu et al., 2023), which indicates higher quality model representations and training data for these countries, explaining why `cap-retrieve` would perform better for them.

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

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 6% (Nigeria) to 30% (India).

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.

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 6% for countries like Nigeria, indicating that this task is far from solved.

Quantitative Metrics for image-editing 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 17 and 18. 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 **C0:** `visual-change` and **C2:** `spatial layout`. For the *country-relevance score*, we observe a similar trend as that for **C3:** `cultural-relevance`.

4.2 *Questions and Findings*: Application

Here, we must make sure that: **a)** *conditioned on the task*, the "meaning" of the original image is preserved; **b)** the cultural relevance increases. Our observations here are as follows:

E1: `education-task` and **L1:** `story-title` – 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 title of the story (for stories), harder than simply matching a category.

E1+E2 and **S1+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 Brazil and Portugal in education, no pipeline is able to translate any image successfully.

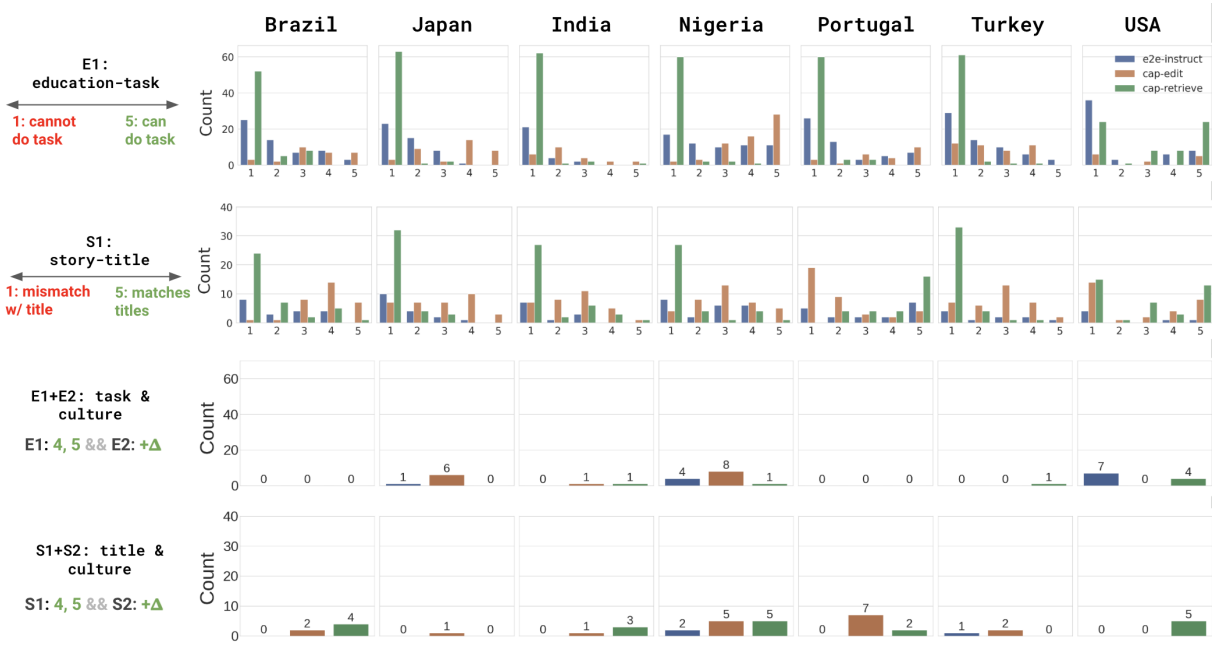


Figure 7: *Human ratings for the application dataset*: Our goal is to test whether the edited image can be used for the application as before (**E/S1**), and whether it increases cultural relevance (**E/S2**). We plot the count of images that can do both above (**E/S1+E/S2**), and observe that even the best pipeline cannot translate any image successfully in some cases, like for Brazil and Portugal in education.

For other countries, the best pipeline is able to translate 10-15% of total images.

5 Related Work

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, and most notably, following natural language instructions (Brooks et al., 2023). We choose, InstructPix2Pix (Brooks et al., 2023) to experiment with since their model allows users to give natural language instructions, as opposed to other models requiring text labels, captions, segmentation masks, example output images and so on. The model is capable of following a wide variety of instructions, ranging from concrete ones like *swap sunflowers with roses* to abstract ones like *make it Paris* or *make it the 1900s*.[9] From a practical standpoint, this model is consistently one of the most downloaded image-editing models on HuggingFace,[10] indicating its wide usage in the community.

6 Conclusion

In this paper, we introduce the task of **image translation** 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 6% images for select countries (like Nigeria) in the easier *concept* dataset and no image translation is successful for some countries (like Brazil, Portugal) in the harder *application* dataset. Our code and data will be released to facilitate future work in this new, exciting line of research.

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

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

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*.

Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. *arXiv preprint arXiv:2305.11080*.

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.

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?

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.

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.

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.

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*.

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.

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.

A Example Outputs

Here, we include sample outputs from the pipelines for select images. All three pipelines have their own set of limitations, indicating that we have a long way to go before we can solve this task. Limitations and 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 at Figure 12, where the American flag colors are applied over the Burger to make it relevant to the United States. Similarly, Figure 16 includes Brazil map and flag as part of the editing process.

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

Both the image editing models preserve spatial dimensions, as can be seen in Figure 21

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.

B Annotation Instructions

Our annotation and human evaluation instructions are as follows. We host our data on the Zeno[12] (Cabrera et al., 2023) platform and hire people on Upwork[13] 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.

B.1 Concept and image collection for the *concept* evaluation dataset

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.

[12] https://zenoml.com/

[13] https://www.upwork.com/

Figure 8: Screenshot of how one instance looks like for human evaluation on the Zeno platform.

These categories include things like bird, food, clothing, celebrations etc. You have to give 5 salient concepts for each category, that are most prevalent in your country and culture, for each of these categories.

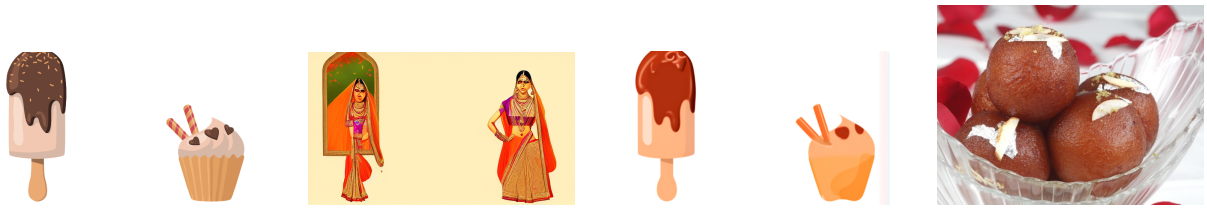
The two key requirements are for the concepts to be: **a) commonly seen** or **representative** of the speaking population of your country; b) ideally, to be **physical and concrete**.

Additionally, you have to query the web and get one image for each concept you list. The image should be of high-quality, clearly displaying the concept as it would appear in your culture or surroundings.

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 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.



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

Figure 9: *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 10: *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`

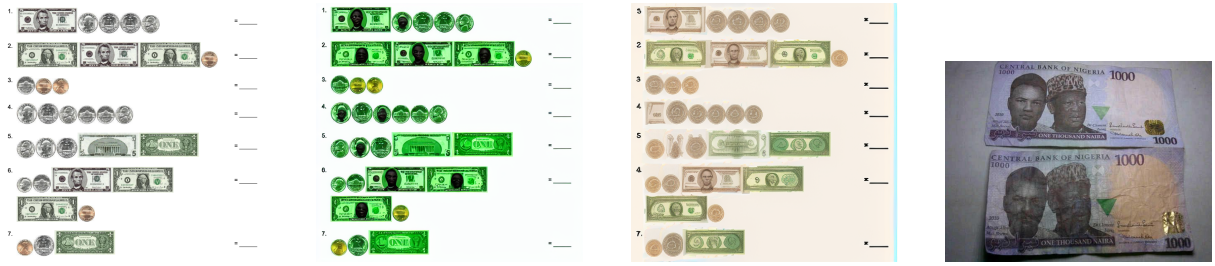
Figure 11: *Source:* India; *Target:* Japan — *BLIP caption*: a field of cotton plants; *LLM-edited caption*: a rice paddy field. `e2e-instruct` inserts sakura blossoms and multiple Mt. Fuji peaks in the background, `cap-edit` highly preserves structural layout but looks pretty realistic here.

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

Figure 12: *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.

In this evaluation, you will be shown 4 images, as shown in the Figure 8. 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,

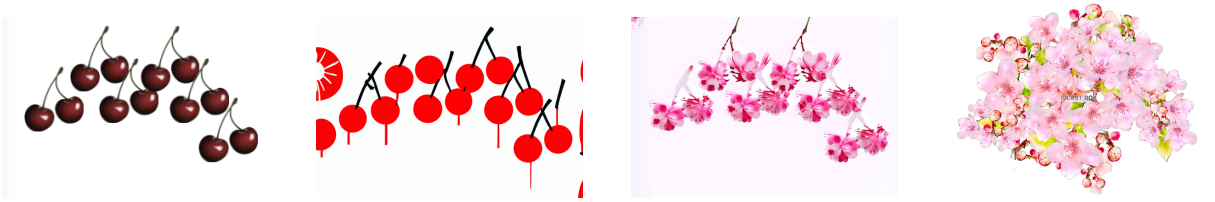


(a) original (b) `e2e-instruct` (c) `cap-edit` (d) `cap-retrieve`

Figure 13: *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 14: *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` doesn't changes english to turkish-looking text 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 15: *Application:* Education; *Target:* Japan — *Task*: count the number of cherries. Here, even though there is a semantic drift from cherries to flowers, the output of `cap-edit` can be still be used to solve the task.

from **5 (strongly agree)** to **1 (strongly disagree)**:

C0) There is a visual change in the generated image, when compared with the source (top-most) image.

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.

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.).

C3) The image seems like it came from your country or is representative of your culture.

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).

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

Figure 16: *Application:* Story; *Target:* Brazil — *Title*: DIVE!. Here, we see a strong tendency to output elements of the map and flag colors in these models.

C5) This image is offensive to you, or is likely offensive to someone from your culture.

Stories
S1) The image would match the title 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.3 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.3.1 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.3.2 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.3.3 India

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

B.3.4 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.3.5 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.3.6 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.3.7 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*

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 19.

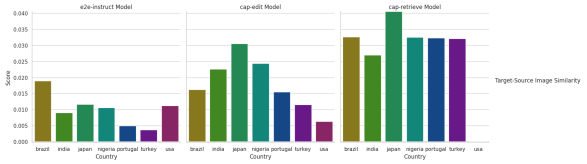


Figure 17: `target-source similarity`, capturing the difference in image-text similarity scores between target and source

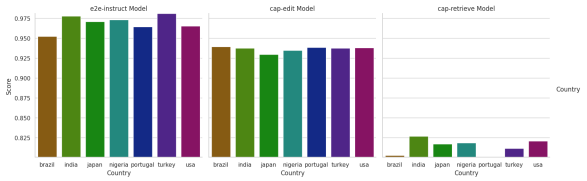


Figure 18: `image similarity difference`, capturing the difference in image similarity scores between target and source

For the application-oriented evaluation, we simply ask whether the edited image can be used to 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 7, we employ these thresholds to filter images that haven't been edited at all. Images whose image-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:

Q3: `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 trans-

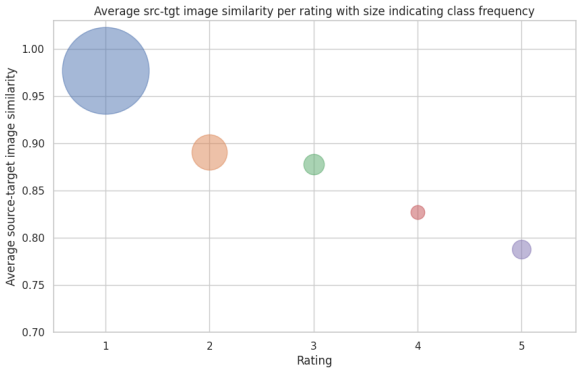


Figure 19: `correlation plot`, capturing linear correlation between human and machine evaluation for Brazil

lation like in Doraemon, Inside Out and so on.

Q5: `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.

Q6: `offensiveness` – Almost no images are found to be offensive, which is encouraging.

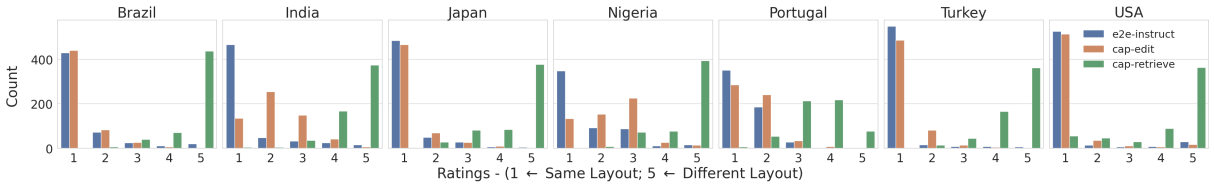


Figure 20: **Q3**: `spatial-layout`, capturing if the structure of the original image is maintained.

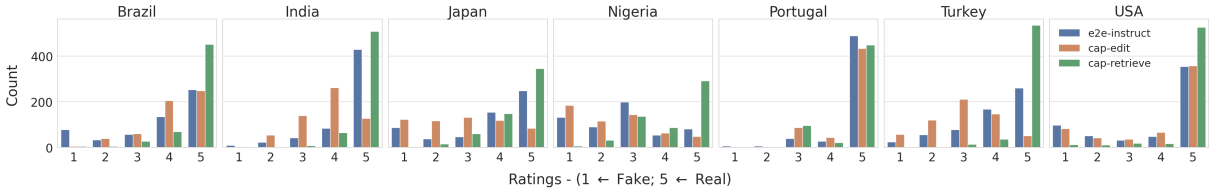


Figure 21: **Q5**: `naturalness` capturing the naturalness of the edited or retrieved image.

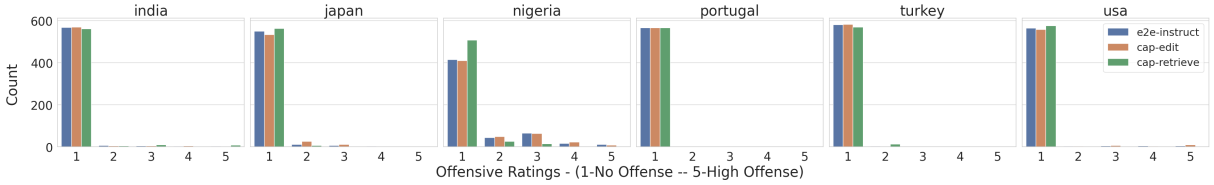


Figure 22: **Q6**: `offensiveness` capturing how offensive each pipeline is

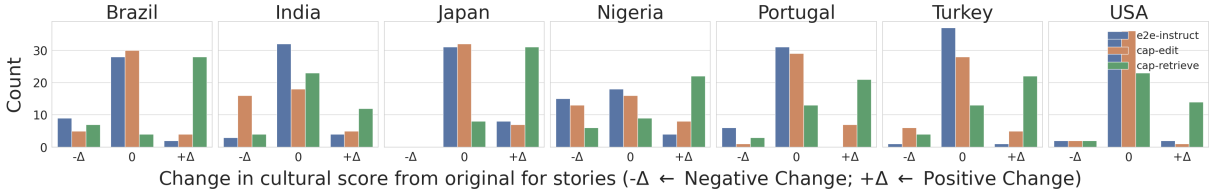


Figure 23: Cultural Score - Story

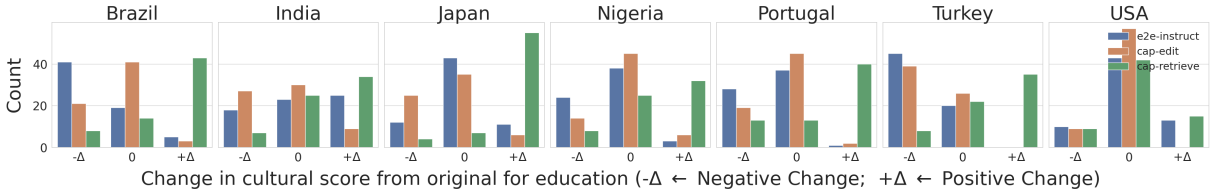


Figure 24: Cultural Score - Education