TrICR: Translating Images for Cultural Relevance (Group 27)

Bhuvan Jhamb [*1] **Simran Khanuja** [*1] **Bao Nguyen** [*1] **Long Vân Tran Ha** [*1]

Abstract

We address the novel challenge of *cultural transcreation*, which extends usual image editing to semantic change for cultural relevance conditioned by text. This task introduces additional constraints not addressed by off-the-shelf image-editing models, on which we first improve with modular explainable pipelines. To design unified models tailored for our task, we focus on cultural food translation, collecting a curated dataset of food images from two cultures. Leveraging our high-quality data and heuristic pairing, we fine-tune different models specifically adapted for our task, finally paving the way for text-controlled culturally-aware semantic image-editing models. We conduct thorough result analysis of our final approaches and make our code publicly available at https://github.com/simran-khanuja/vlr-project.

1. Motivation

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. However, machine translation systems remain confined to language in speech and text. With increased distribution of multimedia content, scholars in translation studies [5, 25, 28] 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 [7]. The term *transcreation* has been coined to refer to the adaptation of content across several modalities to preserve its intended meaning.

Figure 1 presents different real-world transcreations from popular real-world culture, which are currently manually edited by culturally-aware humans. For example, 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.[1] Our work is thus motivated by the need for diverse, relevant, faithful and automated transcreations. This is also a new task in the current image-editing paradigm, with unique constrains: semantic change for cultural relevance, proximity with the original image, diversity, and fine-grained text conditioning. Current baseline methods do not handle all these constrains, and we pave the way for more general semantic image editing.

2. Prior Work

Image-generating models have reached impressive realistic results, from the initial graphical models like VAE [16] to the generator-discriminator paradigm in GANs [10] and the rediscovered diffusion models [11], along with sophisticated variants of these models. **Image-editing models** were derived in parallel to evolve from being capable of single editing tasks like style transfer [8, 9, 14]. For example, StyleGAN enables fine-grained image editing on particular directions in the latent space, but finding and optimizing the translation vector for a specific change is tedious. More recent models handle multiple such tasks in one model [13, 6, 12, 19]. Today, their capabilities range from performing *targeted* editing that preserves spatial layout [32], local in-painting [18], and most notably, following natural language instructions [3]. Notably, InstructPix2Pix [3] and Imagic [15] allow users to give natural language instructions, as opposed to other models requiring text labels, captions, segmentation masks, example output images and so on. The models are 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*.[2,3] These models could be used as a baseline for us, with limited results to improve.

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

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

[3] https://imagic-editing.github.io

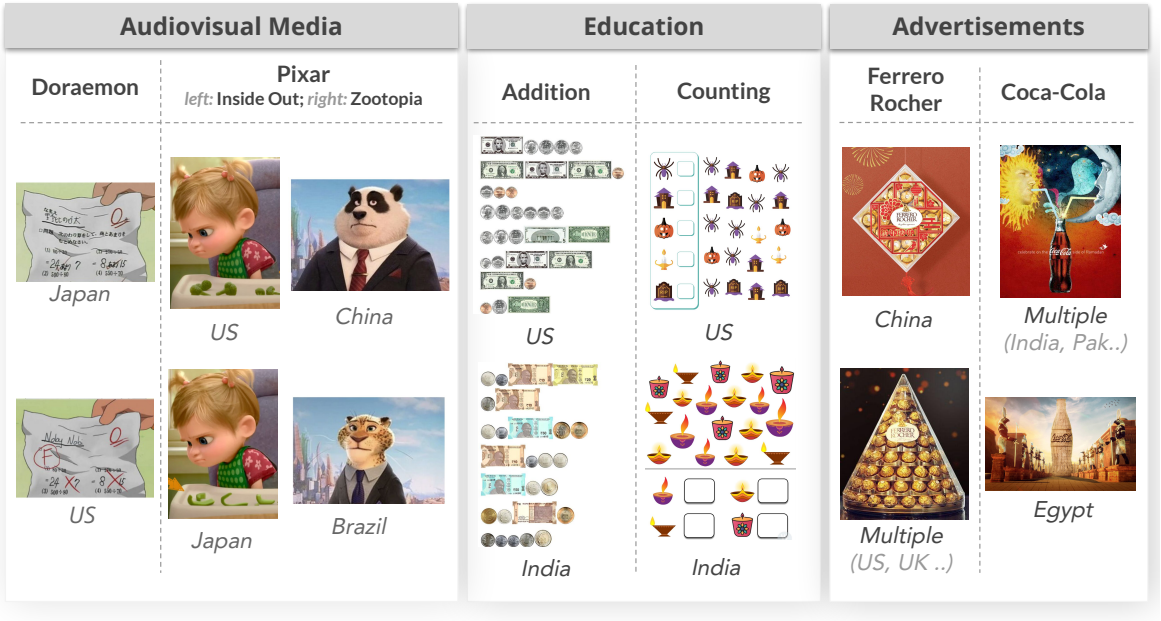

Figure 1. **Image localization** as done in various applications today: *a) Audiovisual (AV) media*: where several changes were made to adapt Doraemon to the US context like adding crosses and Fs in grade sheets, or in Inside Out, where broccoli is replaced with bell peppers in Japan as a vegetable that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

To the best of our knowledge, no prior work has studied the performance of such image-editing models for the task of **cultural transcreation**. This distinctive task diverges from conventional image translation studies, where a straightforward one-to-one mapping between the original image and its translated counterpart is typically pursued, involving elementary changes (such as style or color) [20]. Cultural transcreation, on the other hand, demands the preservation of the original image's style while necessitating an external grasp of cultural nuances to discern what aspects of the image should be modified — a task inherently lacking of systematic guidelines. The modifications requires a delicate balance: they should be culturally relevant and varied, devoid of biases toward any specific cultural motif or cliché, while simultaneously minimizing changes to maintain the style and meaning of the original image. These three constraints constitute competing objectives, necessitating thorough evaluation through both human assessment and automated metrics.

While numerous extensions of GANs have been proposed for image translation tasks, including those tailored for unpaired data like CycleGAN [34], the latest most promising image-editing models are diffusion models which are widely used for image generation, potentially conditioned on text [24]. However, existing off-the-shelf variants of these image-editing models lack an intrinsic understanding of cultural context and fidelity to the original image, often resorting to clichéd representations (see the preliminary results). ControlNet [32] introduces a convenient way to fine-tune diffusion models for text-controlled image editing, which stays close to the structure of the original image. However, we do not want to enforce such a strict structure preservation, so this only constitutes a baseline for us. Another line of work for image editing is inpainting [31], where we first must find a mask for the zone to edit to enhance the cultural relevance and then provide some input examples to edit this patch conditioned on a target image. However, given that current image-editing models can still locate things in the image without masking supervision, we rather focus on one-step editing pipeline that do not require masking supervision.

In short, faithfulness, cultural relevance, and diversity remain a non-trivial challenge unaddressed by conventional image translation methodologies. Our contributions are the following:

1. We characterize the socially common task of image *transcreation* in the current multimodal image-editing paradigm and evidence the current gaps from off-the-shelf baselines.

2. We introduce modular pipelines that better address our specific tasks, notably leveraging culturally-aware editing in the natural language space, and perform analysis on them.

3. We focus on the task of cultural food image translation between two cultures, Japan and India, for which we gather and curate a dataset, extended with different captioning techniques. We provide a heuristic pairing between images for each culture, providing a general framework for data collection for the task.

4. We leverage this data by fine-tuning different models. Notably, we propose a new weakly supervised adversarial setting leveraging paired and unpaired data, when the generator is a one-step diffusion-model, fine-tuned for a specific culturally-aware task with text-conditioning. Error analysis and further promising ideas are provided.

3. Methodology

3.1. Dataset and metrics

Here, we describe the dataset and metrics we plan on using for training and evaluation:

Dataset: LAION [26] dataset consists of 5 billion image-text pairs across multiple languages. We mine a subset of LAION for our task. We first associate each image-text pair to an country/culture based on the URLs (for example, URLs with ".jp" are put in the Japan subset, ".in" are put in the India subset and so on).

We generate paired data for food images from India and Japan. To do this, we get open-clip embeddings of the India and Japan subsets of LAION and create autofaiss indices of these embeddings for efficient retrieval. We use the clip-retrieval[4]. We prompt for "photo of authentic Indian/Japanese food, high quality" and retrieve top 4000-5000 images. Next, we synthetically create parallel pairs from this retrieved data in two ways:

Data-pairing approach 1: Here, we pair Indian food images and Japan food images based on clip and Dino similarity. For each Indian food image in our LAION subset, we calculate a weighted sum of clip and Dino similarity with each Japanese food image. Based on these scores, we can pair it with the best match or the top k matches. This method yields satisfactory pairing, as illustrated in Figure 8. We can also reverse the process to generate matches for each Japanese food image.

Data-pairing approach 2: The above approach heavily relies on structural similarity during pairing. To achieve a more semantic pairing, we employ another method. Here we generate captions for each Indian food image, and edit the caption for cultural relevance using LLMs (more details below). We use these LLM-edits to retrieve top-100 images with a closest match to the text query and utilize Dino and clip similarity to find the top match among the retrieved images.

[4] https://github.com/rom1504/clip-retrieval

Figure 2. Examples of data paired by our approach. Top row contains Food Images from India, bottom row contains corresponding Japanese food images retrieved by our approach

Metrics: Quantitative Metrics for image-editing typically capture how closely the edited image matches the original image and the edit instruction. We can use following metrics:

1. **image-similarity:** Methods that only capture structural similarity (like SSIM, PSNR etc.) would not be useful. Thus we would need some way to calculate perceptual similarity like similarity of Dino-ViT embeddings [4] or LPIPS score [33]. We chose Dino-ViT.

2. **country-relevance:** We can embed the text – "This image is culturally relevant to {country}", and the edited images using CLIP [23] and calculate their cosine similarity.

We backup these automated evaluations by human judgment since cultural questions are often subjective, and report our error analysis in the next section.

3.2. Baselines

`InstructPix2Pix`: We use out-of-the-box instruction-based image editing models to translate the image in one pass. Specifically, we use InstructPix2Pix [3], a model enabling users to define edits using natural language. The original image is fed into the model with an instruction to *make the image culturally relevant to* `COUNTRY`, following a similar prompt format as that used for training the model. Although this pipeline is simple and flexible, it heavily depends on the image models' ability to perform culturally relevant edits, which they may currently lack as we see below.

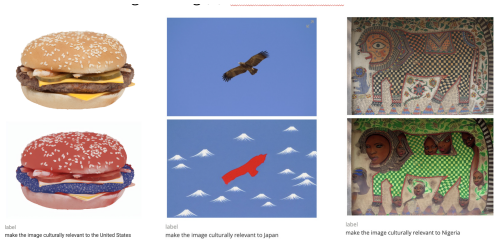

Figure 3. Instruct Pix2Pix (baseline) results. 2nd row shows results when we prompt the model to make images in 1st row culturally relevant to USA, Japan and Nigeria respectively

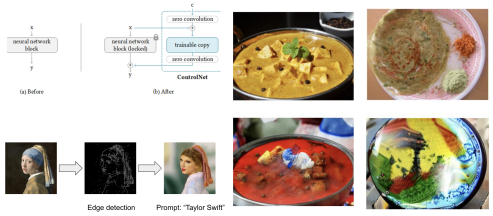

Figure 4. Controlnet(baseline) results. We prompt the model to convert Indian food images to French food images

`ControlNet`: [32] leverages the encoding layers of large diffusion models and add a diverse set of conditional controls (edges, depth, ...) with additional adapters layers during fine-tuning with low-regime data. Here, we chose the conditional control by the edges, i.e. the edited image based on the text prompt will have the same structure in terms of edges as the original one.

3.3. Modular approaches leveraging LLMs

In the above examples, we clearly see that the model lacks cultural understanding. They simply paste the flag colors or culturally salient objects like the Eiffel tower for France, Mt. Fuji for Japan etc., to make it culturally relevant. Large language models have been trained on trillions of tokens of text [29, 1], and have been shown to exhibit at least a certain degree of cultural awareness [2]. Hence, we leverage them in the loop to give image-editing models more concrete instructions.

`Caption, LLM-edit, Image-edit`: Concretely, we first caption the image, prompt a LLM to edit the caption to make it more culturally relevant, and then edit the image using a caption-based image editing model.

`Caption, LLM-edit, Image-retrieval`: Here, we observe that image editing models sometimes make unnatural edits as shown above. Hence, we try retrieving images from LAION instead, given the LLM-edit as a text query.

Figure 5. caption edit (modular pipeline) results. 2nd row shows results when we prompt the model to make images in 1st row culturally relevant to USA, Japan, USA and Japan respectively

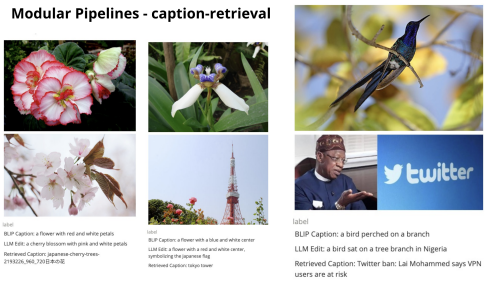

Figure 6. Caption Retrieval (modular pipeline) results. 2nd row shows results when we prompt the model to make images in 1st row culturally relevant to Japan, Japan and Nigeria respectively

In experiments, we use BLIP2-FlanT5-XXL[5] [17] as the image captioner, GPT-3.5[6] for caption transformation, and PlugnPlay as the image editing model [30].

3.4. Fine-tuning end-to-end image editing models

`pix2pix-turbo` and `CycleGAN-Turbo`: [21] introduces two models, pix2pix-turbo and CycleGAN-Turbo, using one-step diffusion models to perform various image-to-image translation tasks, both in supervised (paired data) and unsupervised (unpaired data) settings respectively. Given our dataset's characteristics, which include both paired and unpaired data, we aim to leverage these models effectively for culturally relevant image transcreation tasks.

Unsupervised and supervised fine-tuning for image translation:

To go beyond the baselines and leverage our unpaired food data from India/Japan and our heuristic pairing, we tweaked and fine-tuned two models, inspired [22]. We fine-tune both CycleGAN-Turbo and Pix2Pix-Turbo using our food image data. For the former, we simply need unsupervised

[5] https://huggingface.co/Salesforce/blip2-flan-t5-xxl

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

Figure 7. CycleGAN Turbo Finetuning Results. We finetune the model to convert Indian food images to Japanese food images

Figure 8. Pix2Pix Turbo Finetuning results. We finetune the model to convert Indian food images to Japanese food images

images from both domains. For the latter, we pair the data as described in Section 3.1, and fine-tune Pix2Pix-Turbo. However, simple fine-tuning using losses introduced in the paper doesn't work well for us since our goals here differ from traditional image-editing tasks. For example, we don't want to retain all of the original structure because that loses out on the naturalness of the target image. Here are how we adapted the models from [22] for our requirements:

- removing the skip connections because we want to deviate more from the identity mapping unlike traditional night-to-day image translation,
- during training, conditioning on specific captions given by a captioning model and their culturally-edited versions given by a LLM (unlike fix text-conditioning previously),
- We tweak the different losses previously used, since we have a one-to-many mapping and we cannot and should not expect an exact L1-reconstruction. In [22], the similarity between two images the sum of the L1-loss and the perceptual loss [33], which less focuses on exact reconstruction. In our weighting, we prioritize the latter versus the exact reconstruction. Besides, we tried to get rid of the identity loss (see below for the loss details), but we realized it was essential for the training stability, especially in the semi-supervised setting.

Semi-supervised (unsupervised with small amounts of paired data) Leveraging insights from [27], which explored semi-supervised techniques for image-to-image translation, we aim to bridge the gap between supervised and unsupervised settings by integrating pix2pix-turbo and CycleGAN-Turbo losses in one formulation. More precisely, for a batch of size 1, we have:

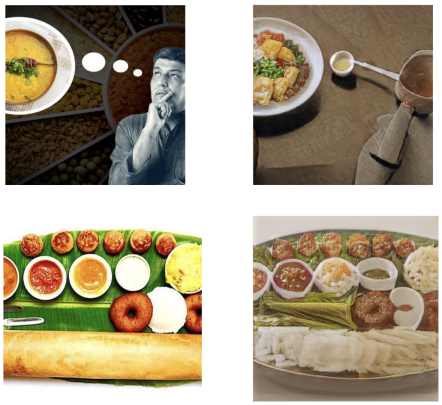

Figure 9. Pix2Pix Turbo Finetuning failure cases

- x_A, x_B the images from domain A and domain B, and their original captions c_{x_A}, c_{x_B} and edited captions c_{x_Aed}, c_{x_Bed},
- y_A, y_B the target images for x_A, x_B and their captions c_{y_A}, c_{y_B}.

Then our final loss is a weighted sum of:

- $L_{cycle} = L_{rec}(G(G(x_A, c_{x_Aed}), c_{x_A}), x_A) + L_{rec}(G(G(x_B, c_{x_Bed}), c_{B_A}), x_B)$,
- L_{GAN}, L_{idt}, similarly adapted from [22] with our edited captions,
- a supervised loss $L_{sup} = L_{rec}(G(x_A, c_{y_A}), y_A) + L_{rec}(G(x_B, c_{y_B}), y_B)$.

4. Experiments and Results

4.1. Visual results

For baselines and modular approaches, we selected some images across many categories like food, beverages, animals and so on, from the Indian culture. We mostly focused on two target cultures: France, and the US - so that all team members can evaluate cultural references (two team members are from India, two are from France). Please find some relevant links on the collaborative platform Zeno where some results where uploaded:

- India → France with `ControlNet`

- India → France with `InstructPix2Pix`
- India → the US with `InstructPix2Pix`
- India → the US with `Caption`, `LLM-edit`, `Image-edit`

4.2. Qualitative Observations

Here are some high-level takeaways:

4.2.1. Baselines

`InstructPix2Pix` often fails to edit the original image many times. When it does change the image it usually makes some stylistic and unnatural changes. This might also be an artifact of the model being trained on synthetic data.

`ControlNet` is slightly better than InstructPix2Pix. While it rigidly maintains the original image's layout, ensuring semantic similarity, it also effectively incorporates textual instructions to modify the image within the existing framework. Though occasionally resulting in somewhat eerie output due to strict adherence to edges, and sometimes deviating towards excessive alterations, it consistently achieves cultural translation.

Both baselines exhibit a strong tendency to rely on clichés when integrating elements from the target culture, often incorporating national flags or iconic landmarks (e.g., the Eiffel Tower for France). They also just paste this on top of the image which doesn't make any meaningful sense and often leads to comical outputs.

4.2.2. Modular approaches leveraging LLMs

`Caption, LLM-edit, Image-edit/retrieval` tends to make more substantial modifications to the images, typically guided by external language and learning models adept at culturally translating captions. However, despite this advantage, the overall results are underwhelming due to limitations with the image-editing model conditioned on the new caption. Further they strongly preserve the original structure which makes images look unnatural, which the retrieval part alleviates to a certain extent but retrieval also has a lot of noisy outputs as discussed before.

4.2.3. Fine-tuning image-editing models

`CycleGAN-Turbo` The model does reasonably well job in translating food images from India to Japan. However it seems to put too much emphasis on the structure of starting image and tries to match it, as visible in figure 7. This leads to unnatural edits sometimes. [7]. We note that in the purely unsupervised setting, the modifications are still marginal and too close to the identity function, and include some clichés modifications linked to how our discriminators perceive the culture through our unpaired data: adding salmon for Japan, adding Naan for India. We suspect the model needs supervision to learn more semantic edits. The semi-supervised setting is very promising but hard to tune, but yielded more edits, and we have further ideas to improve that.

`Pix2Pix-Turbo` Qualitatively, the results with this pipeline look most promising. They make sufficient number of edits which resemble the target domain. However, the quality is still not the best and there are failure cases as we observe in Figure 9. For future work, we want to experiment with unfreezing the text encoders as well since we leverage rich captions and LLM-edits for text conditioning, not just a task instruction. We try finetuning with one-to-one paired data and one-to-many paired data. The wandb links for both can be found here.[8]

Across all pipelines, performance tends to be slightly better for translations targeted at the US compared to France. This discrepancy is likely attributable to pre-trained models having encountered more examples from American culture, thus possessing a richer repository for translation.

4.3. Quantitative metrics

Let A denote the original test images of Indian food, B the original test images of Japanese food, and C the translated images produced by the models from dataset A.

CLIP similarity
To gauge the quality of translation, we employ CLIP similarity to measure the proximity between the images and the prompt "Japanese food". The baseline mean CLIP similarity between dataset A of Indian food and the prompt "Japanese food" is established at 0.12. A higher CLIP similarity between the translated images C and the prompt indicates a successful transformation closer to "Japanese food".

Dino similarity Dino similarity provides insights into the semantic similarity between two images. Here, we compute the mean Dino similarity between the translated dataset C and the original Indian dataset A. This evaluation allows us to assess the fidelity of visual translation achieved by the models, specifically in terms of preserving the visual prop-

[7] https://wandb.ai/manhbao/sup_final/runs/h2jvwa6w?nw=nwusermanhbaon and https://wandb.ai/manhbao/unsup_final/runs/89g8ft9w?nw=nwusermanhbaon

[8] https://wandb.ai/skhanuja/pix2pix_turbo_in-jp?nw=nwuserskhanuja and https://wandb.ai/skhanuja/pix2pix_turbo_in-jp-many-1?nw=nwuserskhanuja

erties of the original images while changing their content.

The table below summarizes the quantitative metrics obtained from different methods:

Method	CLIP Similarity	Dino Similarity
Instruct Pix2Pix	0.188	0.895
ControlNet	0.188	0.896
pix2pix-turbo	0.187	**0.939**
CycleGAN-Turbo	0.185	0.905

Table 1. Comparison of different methods, with the baseline at 0.12 for CLIP

All methods exhibit similar CLIP scores that surpass the established baseline of 0.12 for CLIP. Notably, the new methods pix2pix-turbo and CycleGAN-Turbo demonstrate higher Dino similarity compared to the other methods, indicating superior fidelity in image translation. These results suggest successful training and translation, as the models have acquired the ability to capture both cultural nuances and semantic similarities between the two image domains. Among the methods, pix2pix-turbo stands out as the most proficient in translating and preserving visual fidelity. This outcome is coherent with its supervised nature, as it is trained on paired data, enabling it to better capture and retain the visual characteristics during translation.

4.4. Limitations and Future Work

Our proposed approach demonstrates promising outcomes in converting Indian food images to Japanese food images. However, there are several limitations we aim to address in future investigations. Initially, our model fine-tunes solely on a dataset comprising around 5000 image pairs. Upscaling the training data should definitely help us push the performance. Furthermore, once we have access to more extensive datasets, we intend to experiment with unfreezing clip encoders, and adding a text-image alignment in the semi-supervised setting, as it is done in pix2pix-turbo. Another significant avenue for future research involves extending our architecture beyond the food category and achieving generalization across diverse cultural contexts beyond India and France, which can happen by tweaking the current dual adversarial setting.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[2] Arora, A., Kaffee, L.A., Augenstein, I.: Probing pre-trained language models for cross-cultural differences in values. arXiv preprint arXiv:2203.13722 (2022)

[3] Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)

[4] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)

[5] Chaume, F.: Is audiovisual translation putting the concept of translation up against the ropes? (2018)

[6] Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)

[7] Esser, A., Smith, I.R., Bernal-Merino, M.Á.: Media across borders: Localising TV, film and video games. Routledge (2016)

[8] Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)

[9] Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)

[10] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks (2014)

[11] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)

[12] Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)

[13] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)

[14] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR **abs/1812.04948** (2018), http://arxiv.org/abs/1812.04948

[15] Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)

[16] Kingma, D.P., Welling, M.: Auto-encoding variational bayes (2022)

[17] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)

[18] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Gool, L.V.: Repaint: Inpainting using denoising diffusion probabilistic models (2022)

[19] Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10743–10752 (2021)

[20] Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: Methods and applications (2021)

[21] Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models. arXiv preprint arXiv:2403.12036 (2024)

[22] Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models (2024)

[23] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

[24] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. CoRR **abs/2102.12092** (2021), https://arxiv.org/abs/2102.12092

[25] Ramière, N.: 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 (2010)

[26] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)

[27] Shukla, S., Van Gool, L., Timofte, R.: Extremely weak supervised image-to-image translation for semantic segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3368–3377. IEEE (2019)

[28] Sierra, J.J.M.: Humor y traducción: Los Simpson cruzan la frontera. No. 15, Universitat Jaume I (2008)

[29] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[30] Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: 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. pp. 1921–1930 (2023)

[31] Wang, S., Saharia, C., Montgomery, C., Pont-Tuset, J., Noy, S., Pellegrini, S., Onoe, Y., Laszlo, S., Fleet, D.J., Soricut, R., Baldridge, J., Norouzi, M., Anderson, P., Chan, W.: Imagen editor and editbench: Advancing and evaluating text-guided image inpainting (2023)

[32] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)

[33] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)

[34] Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)

A. Appendix

Here, we present examples of image translation across cultures beyond India and Japan, encompassing various categories such as animals, food etc.

(a) Original image

(b) Instruct Pix2Pix

(c) cap-edit

Figure 10. Translation of a rhinoceros (India) to the American culture. We see that e2e-instruct does not change much of the original image if only the general tone color. However, cap-edit is able to condition the editing on a LLM-translated caption converting the rhinoceros to a bison.

(a) Original image

(b) Instruct Pix2Pix

(c) ControlNet

Figure 11. Translation of a Lychee fruit (India, or at least Asian culture) to the French culture. InstructPix2Pix (middle) does not modify the fruit but simply adds Eiffel towers in the background which is uncanny, and does not modify the main cultural marker. On the other hand, ControlNet is able to preserve the shape/function of the image without adding uncanny cliché details. However, the shape is strictly the same, which is too conservative and the results is unclear to interpret, even though it looks 'more' French - close to nuts.