GEORGE MASON UNIVERSITY

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Crossroads of Continents

Automated Artifact Extraction for Cultural Adaptation with Large Multimodal Models

Anjishnu Mukherjee
amukher6@gmu.edu

Ziwei Zhu
zzhu20@gmu.edu

Antonios Anastasopoulos
antonis@gmu.edu

Motivation: Adapting content across cultures

An image of a front door in Greece

- {Greece} + {India}

An image of a front door in India

This work

Cultural Awareness → Cultural Artifacts → Cultural Adaptation

Guess the country

Greece

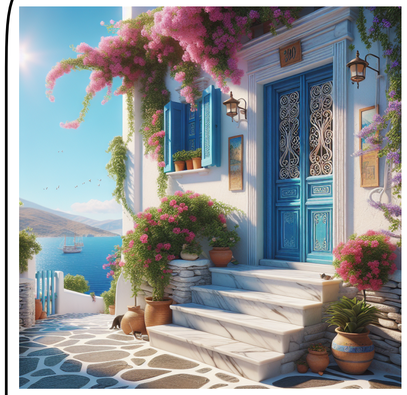
ouzo, grapevines, olive oil

India

saree, tea, dupatta, cow, spices

Country-Artifact associations

- {Greece}



+ {India}

Adapted artifacts

Step 1: Where do models stand wrt cultural awareness?

Better than human baseline

Existing datasets do not cover {country, concepts} in a balanced way

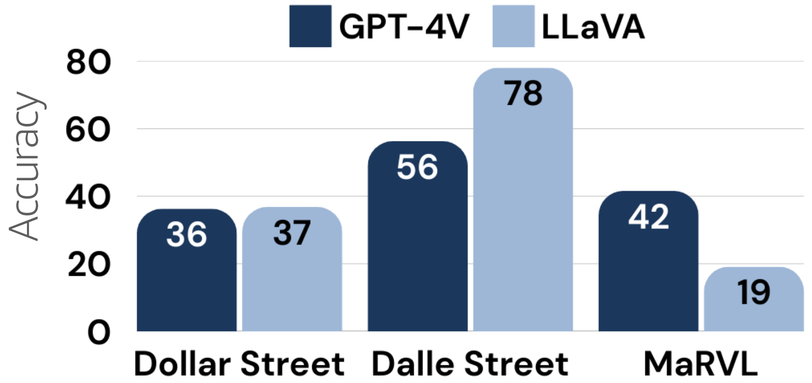
NEW Dataset

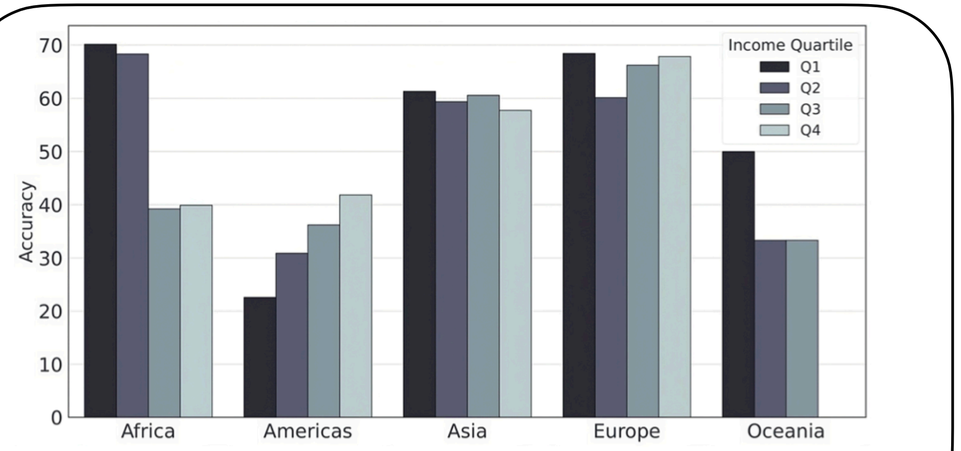
We create a large synthetic dataset with DALL-E 3:
- 67 countries (5 geographical regions)
- 10 culturally relevant concepts

NEW Benchmark

- open-source and closed-source models
- human evals on both data quality and task performance

Human accuracy on a subset of Dalle Street is 48%.





Accuracy for each normalized economic quartile

Step 2: How do models understand culture?

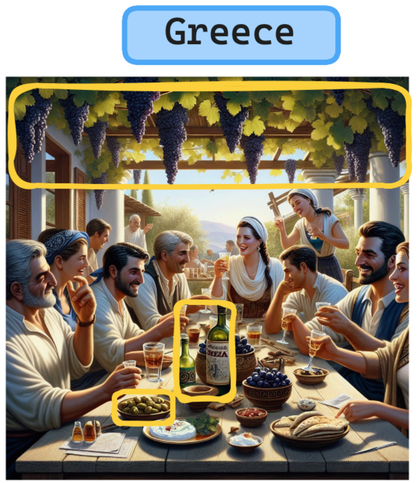
By associating countries with cultural artifacts

We use open-set object detection capabilities of GPT to extract a list of objects it thinks exists in the image

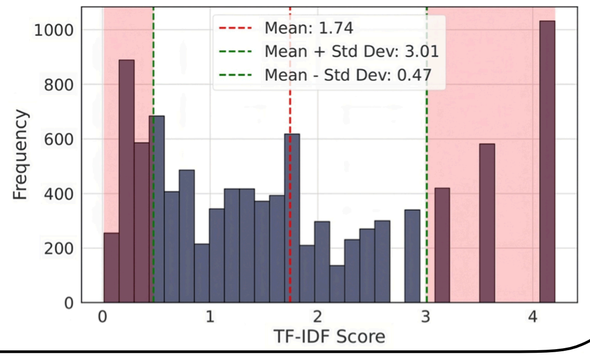
- Filter for frequently occurring items at country level
- Ground using traditional object detector
- 18k+ cultural artifacts in total

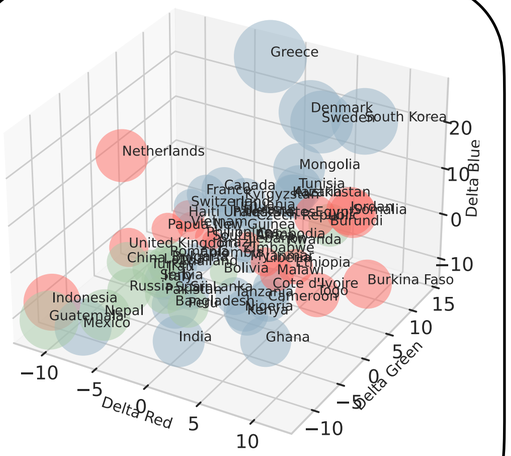
Qualitative human evals:
- hallucination of artifacts detected
- cultural relevance

Additional analysis:
- color associations with countries
- number of people generated in images

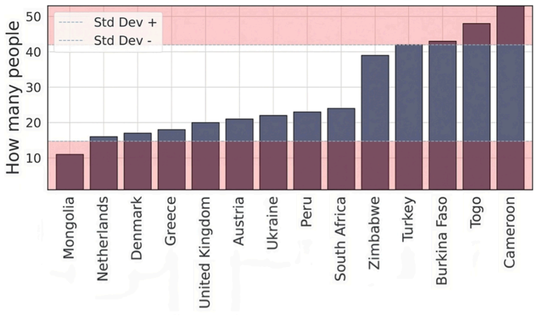


ouzo, grapevines, olive oil





Color associations



People-count associations

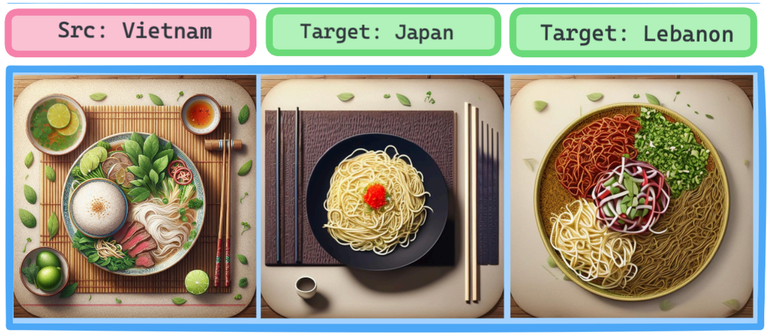
Step 3: Cultural adaptation of artifacts

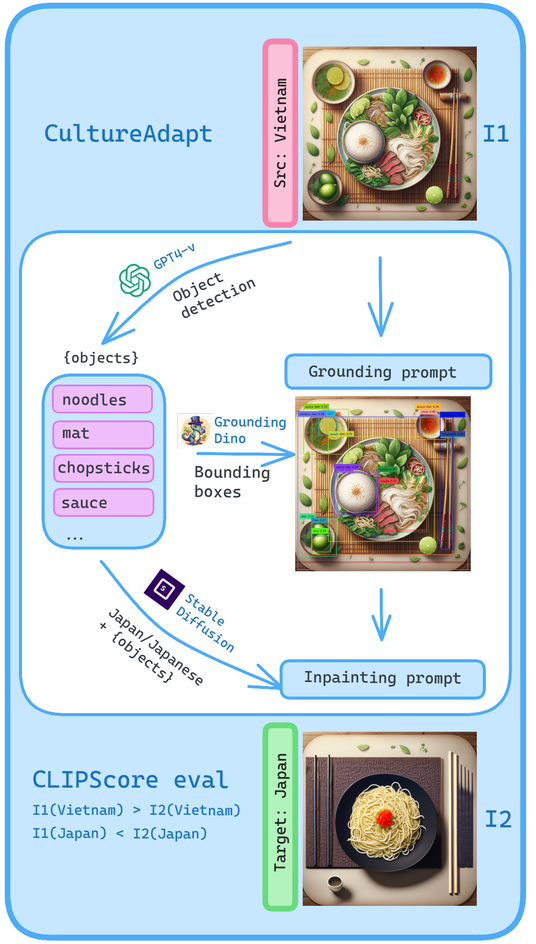
We mask the detected objects and use inpainting conditioned on the target country to adapt the image

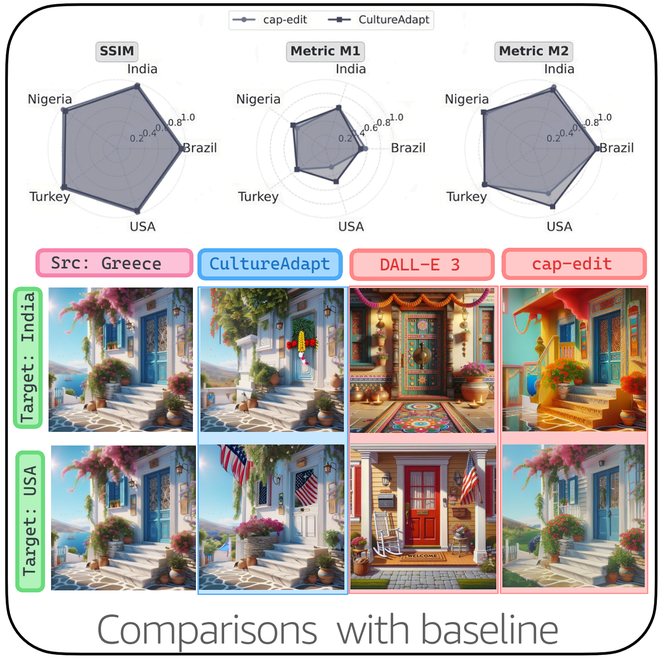
Metrics:
- CLIP similarity deltas b/w src, target images and country names
- Structural similarities between/w src/target images

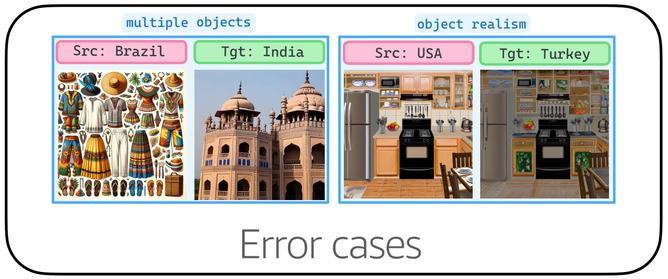
Baseline comparisons:
- ours is as good as or better always
- ours maintains high structural similarity

Qualitative analysis:
- cultural relevance of target image
- structural similarity of src/target images
- common error cases









Baseline comparisons

Our results appear to be culturally relevant while making minimal changes to structure

Other approaches:
- change structure too much
- only change the color
- sometimes make no changes
- do not edit relevant artifacts

Future work

- Our approach is a feasible simple approach for cultural adaptation
- But it does struggle in some cases for example with human figures and/or adapting multiple objects at the same time
- These are easy to improve with additional guidance for face/pose/etc for humans and iterating over object masks one at a time

Doing yoga

- {Algeria} + {India}

Doing yoga

Thank you!

Questions?

https://github.com/iamshnoo/crossroads

amukher6@gmu.edu