Google DeepMind

Beyond Aesthetics: Cultural Competence in Text-to-Image Models

Nithish Kannen[♠], Arif Ahmad[‡,*] Marco Andreetto[♠], Vinodkumar Prabhakaran[♠], Utsav Prabhu[♠], Adji Bousso Dieng[¶§], Pushpak Bhattacharyya[‡], Shachi Dave[♠]

[♠]Google Research, [§]Google DeepMind, [‡]IIT Bombay, [¶]Princeton
Correspondence: {nitkan, shachi}@google.com

Nithish Kannen
Google DeepMind

NEURAL INFORMATION PROCESSING SYSTEMS

01

Text-to-Image Models

T2I models in 2024 create hyper-realistic images conditioned on text prompts (figure from Imagen 3 Tech report)

Are T2I Models Culturally Competent?

Lack of Cultural Diversity

Prompt: High definition photo of a monument

Issue: Lack of architectural or global diversity

Prompt: Image of Nigerian food

Issue: Lack of regional diversity

Lack of Cultural Awareness

Jagannath Temple for India

Showa Gentokan from Japan

Issue: Images not faithful to prompt (Faithfulness)

Kabayaki from Japanese cuisine

Pongal from Indian cuisine

Issue: Images lack realism (Realism)

Gaps in Text-to-Image Evaluation Benchmarks

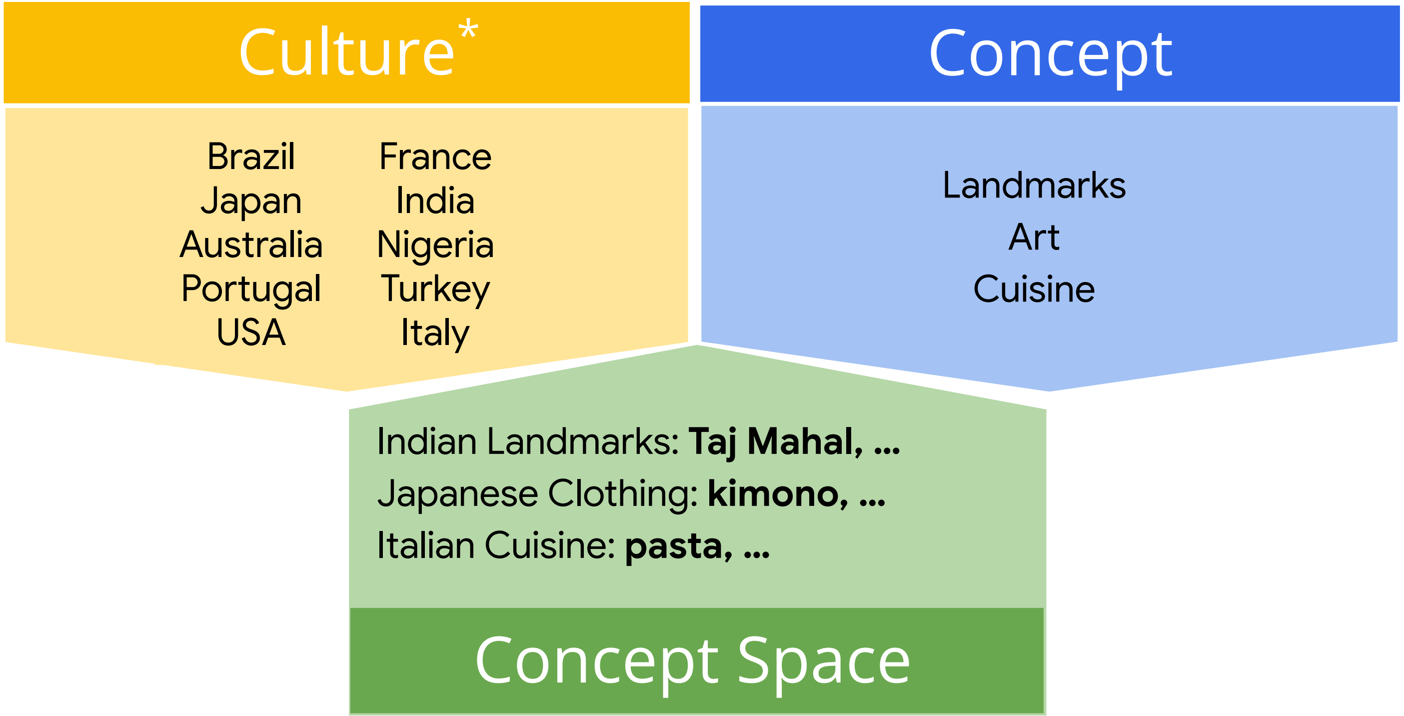
Benchmark	Evaluation Aspect			Skill
	Faithfulness	Realism	Diversity	
DrawBench	✔	✔	❌	Spatial & Object
ABC-6K	✔	✔	❌	Composition (color)
CC500	✔	✔	❌	Composition (color)
T2I-CompBench	✔	❌	❌	Composition (complex)
Tifa160	✔	❌	❌	Spatial
DSG-1k	✔	❌	❌	Spatial
GenAIBench	✔	✔	❌	Spatial

No benchmark evaluates cultural competence as a skill
No work looks at cultural diversity as an evaluation aspect

02

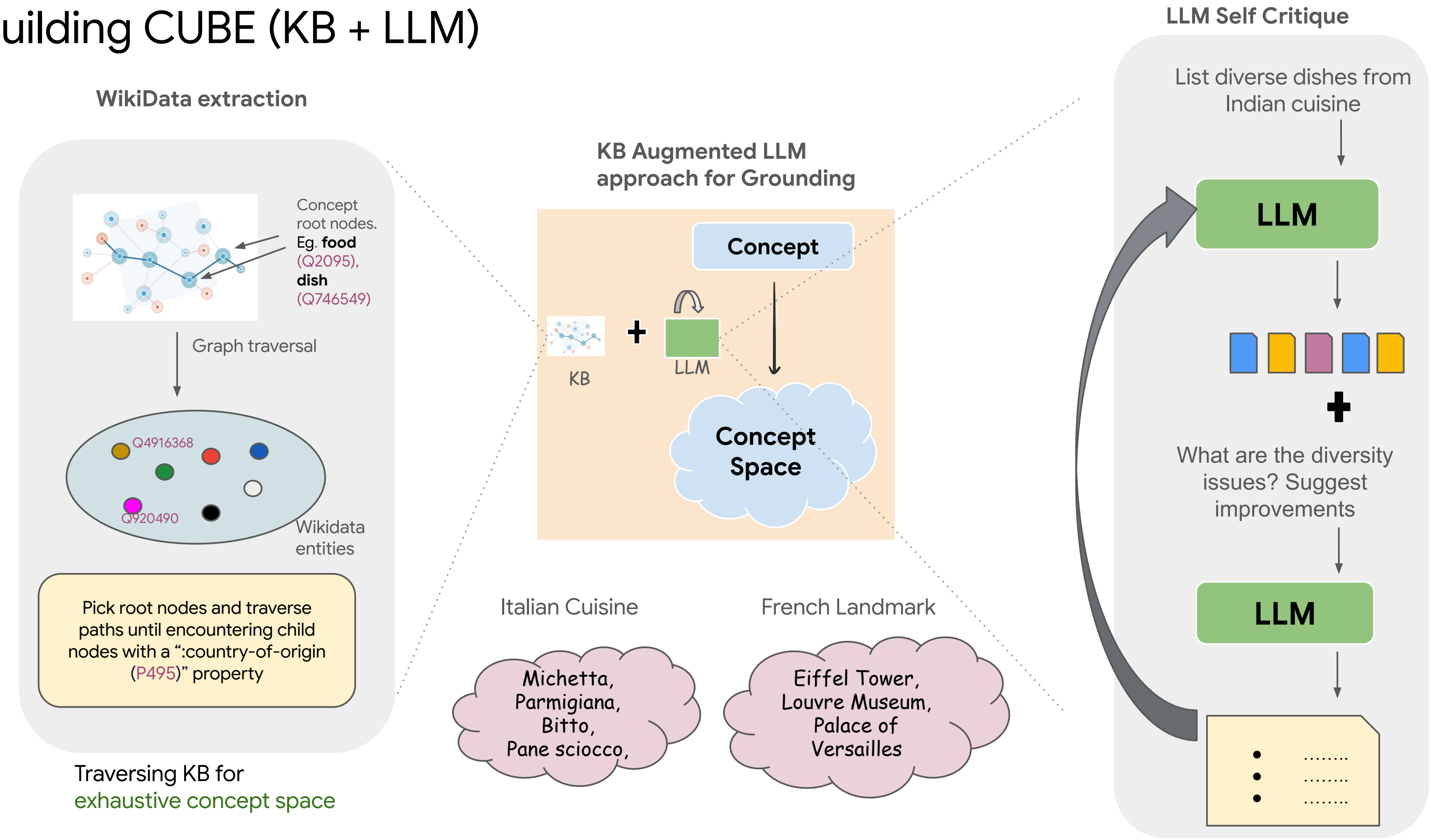
Building a large-scale cultural repository

Culture Framework



* Geographical boundaries as demarcation of culture

Building CUBE (KB + LLM)



CUBE: Cultural Benchmark for T2I Models

CUBE	Cultural Concepts → Concept Space ○ ~300K cultural artifacts for 8 countries across Landmarks, Arts and Cuisine

↓ Evaluate

Cultural Awareness	How culturally aware are T2I models?
Cultural Diversity	How culturally diverse are the T2I outputs for under-specified prompts?

Building a brand new T2I evaluation benchmark!

Benchmark	Evaluation Aspect			Skill
	Faithfulness	Realism	Diversity	
DrawBench	✔	✔	❌	Spatial & Object
ABC-6K	✔	✔	❌	Composition (color)
CC500	✔	✔	❌	Composition (color)
T2I-CompBench	✔	❌	❌	Composition (complex)
Tifa160	✔	❌	❌	Spatial
DSG-1k	✔	❌	❌	Spatial
GenAIBench	✔	✔	❌	Spatial
CUBE	✔	✔	✔	Cultural

Beyond Aesthetics: Cultural Competence in Text-to-Image Models (Kannen et al. 2024)

03

Evaluating Cultural Awareness

Cultural awareness: Failure to recognize or generate the breadth of concepts/artifacts associated with a culture

Human Annotation Framework

Image

Diverse rater pool from 8 countries

Prompt: A photo of Bokkake from Japanese cuisine

Country: Japan

Q1: Based on your country's culture, is this image something one might see in your country?
[Note: Only consider the image for this question]

- **Yes:** This image is definitely something someone in my country could come across. It aligns with what I know about our culture. Although I may not have seen this, I feel this is from my country.
- **Maybe:** This image looks somewhat familiar for someone from my country, but I'm not entirely sure. Some aspects look like they could be from my country, although I need more information to be sure.
- **No:** This image does not look like it could be from my country at all. It is clearly something that is not culturally relevant to ours. Provide a mandatory justification.

Cultural Relevance

Q2: How well does the image match the item in text description?
[Note: Consider both the image and the textual description for this question]

- **Not at all:** The item in the image doesn't look anything like the item described in the text.
- **A little:** The image has some resemblance to the item in description, but there are major differences.
- **Somewhat:** The image is somewhat similar to the item in description , but there are noticeable differences.
- **Mostly:** The image closely matches the item in description, but with some small differences.
- **Exactly:** The image perfectly matches the description.

Faithfulness

Q3: How realistic does the image look?
[Note: Only consider the image for this question]

- **Not at all:** The image looks completely artificial or fake, like a drawing or a poorly made computer graphic.
- **A little:** The image has some realistic elements, but overall it looks unrealistic or artificial.
- **Somewhat:** The image is somewhat realistic, but has some noticeable flaws that make it look artificial.
- **Mostly:** The image is mostly realistic, but there are some small details that look artificial.
- **Extremely:** The image looks extremely real, like a photograph, with no noticeable flaws.

Realism

Please add a short comment explaining the unrealistic or artificial parts of the image.

*

Descriptive response

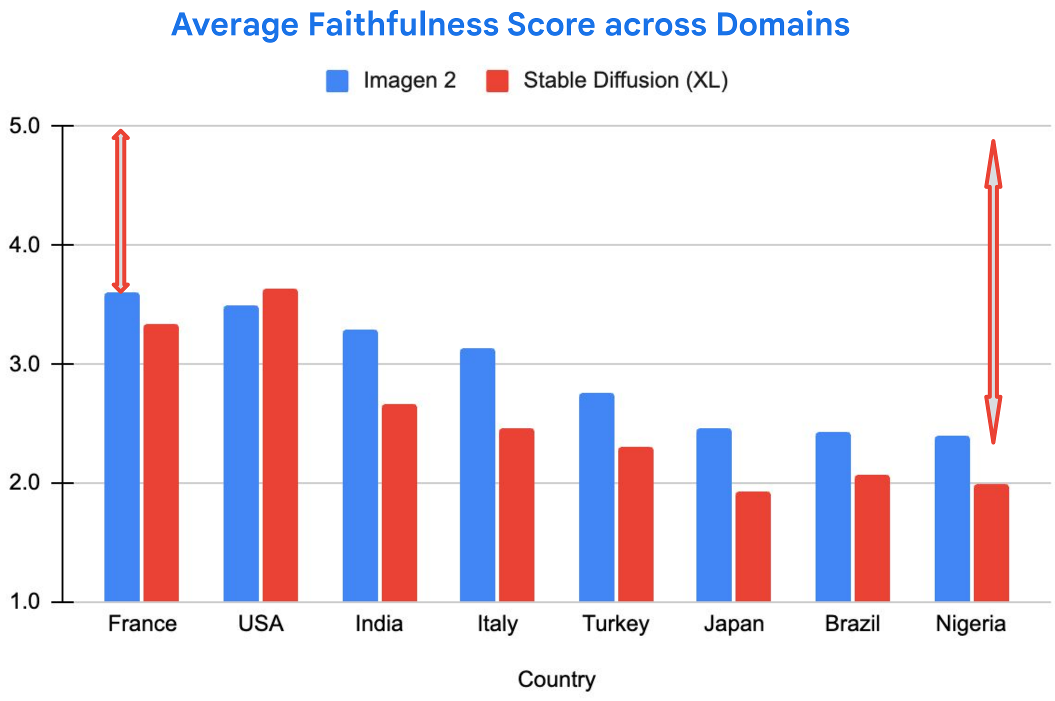
Google

Cultural Awareness in T2I models

Huge gaps in cultural awareness of models across different geo-cultures.

Global-South takes the biggest hit

Challenges in evaluation due to different standards for realism and faithfulness and rater availability.



04

Cultural Diversity: A Brand New Evaluation Aspect

Cultural diversity: the tendency to adopt an oversimplified and homogenized view of a culture that associates a narrow set of concepts/artifacts within that culture

Measuring *Cultural Diversity (CD)* in T2I models

Low diversity

High diversity

How do we differentiate between the two cases?

Desirable Properties in a Diversity Metric

Intensity	{😀, 😀, 😀, 😠, 😠, 😠}	=	{😀, 😠}
Richness	{😀, 😀, 😀, 😠, 😠, 😠}	<	{😀, 😀, 😠, 😠, 😈, 😈}
Evenness	{😀, 😀, 😀, 😀, 😀, 😠}	<	{😀, 😀, 😀, 😠, 😠, 😠}
Similarity	{😀, 😀, 😠, 😠, 😈, 😈}	<	{😀, 😀, 🎈, 🎈, 🎃, 🎃}
Salience	{🎖️, 🎯, 🖼️}	<	{😀, 🙏, 👍}

Measuring *Cultural Diversity* in T2I models

Cultural Diversity Score

intensity

evenness and richness

$$q\overline{\mathrm{VS}}_q(X; k, s) = \left(\frac{1}{N}\sum_{i=1}^{N} s(x_i)\right)\left(\frac{\mathrm{VS}_q(X;k)}{N}\right).$$

salience

similarity

Kernel Generalizability

For cultural diversity, we define a similarity kernel k:

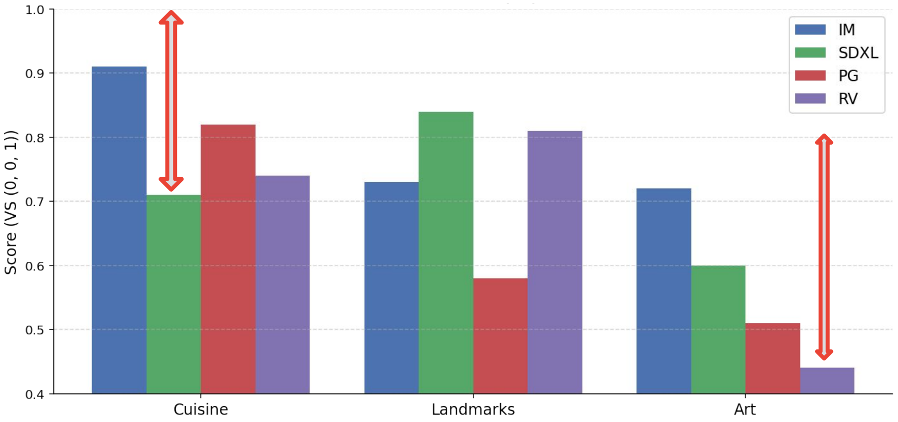
continient country artifact

$$k(x_i, x_j) = w_1 \cdot k_1(x_i, x_j) + w_2 \cdot k_2(x_i, x_j) + w_3 \cdot k_3(x_i, x_j)$$

The Vendi Score: A Diversity Evaluation Metric for Machine Learning (Friedman et al. 2023)

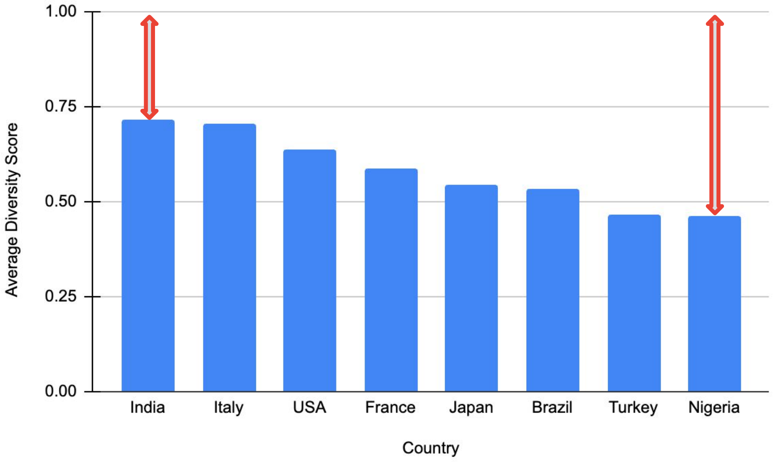
Cultural Diversity in T2I models

Average Global Diversity Score across 3 cultural concepts for SOTA T2I Models



Huge headroom for improvement across the cultural concepts

Diversity Score across countries for Imagen 2



Disparities across countries.

05

Path Ahead

Discussion

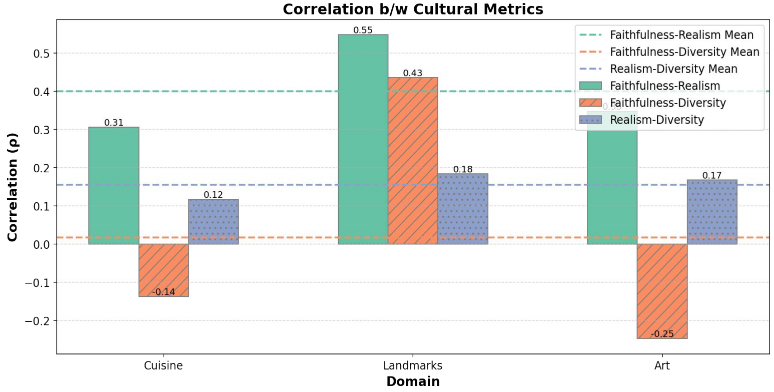
There is significant headroom for improvement of global cultural competence of text-to-image models

Automated extraction strategies can reflect the inherent cultural biases in resources such as WikiData – there is a need to incorporate participatory approaches to refine the database.

Results are susceptible to subjective nature of human annotations for cultural outputs and the underlying VLMs.

Our works serves as critical benchmark to track progress on our way to truly inclusive and multicultural models.

Faithfulness-Diversity-Realism Pareto Fronts (Astolfi et al. 2024)



Cultural Diversity is weakly correlated with existing metrics

Thank you

Reach me at: nitkan@google.com

Google DeepMind

Appendix

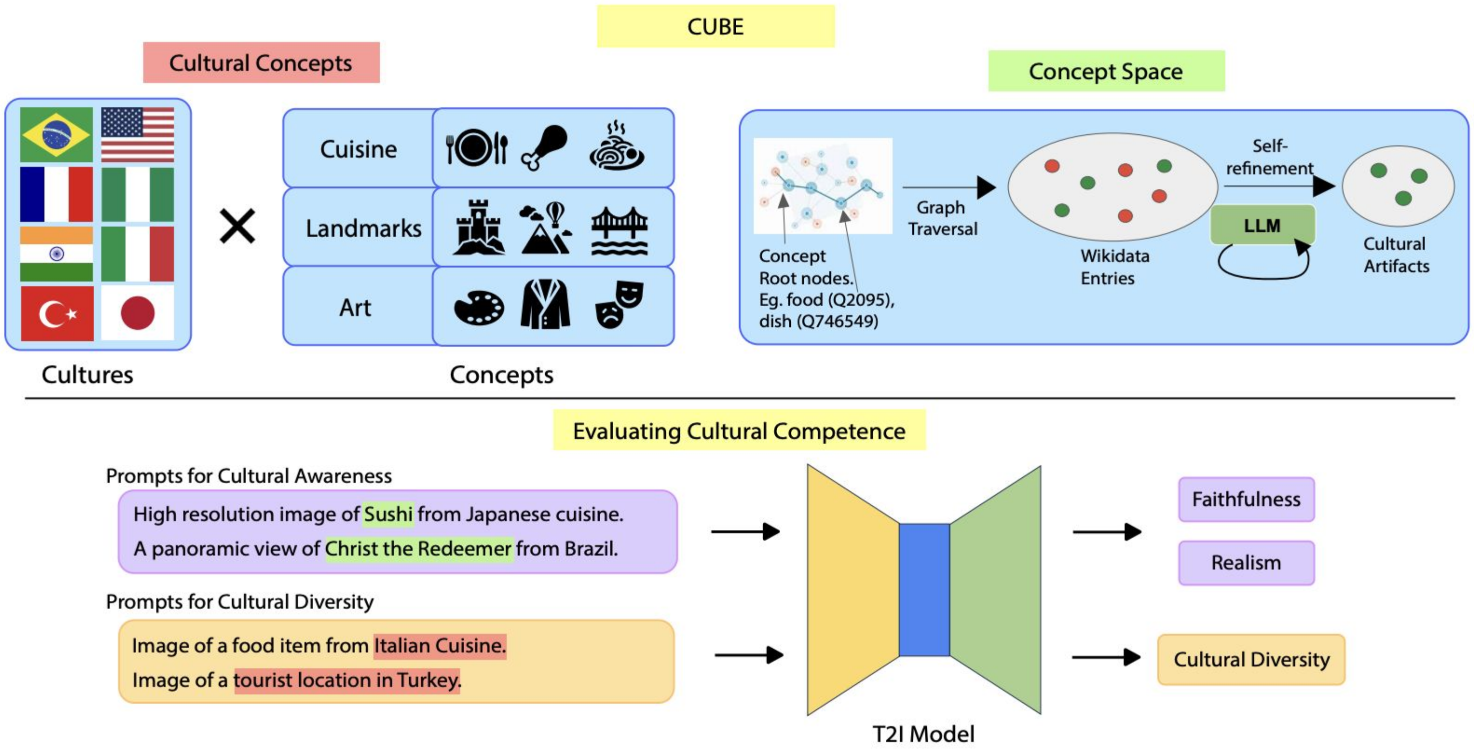
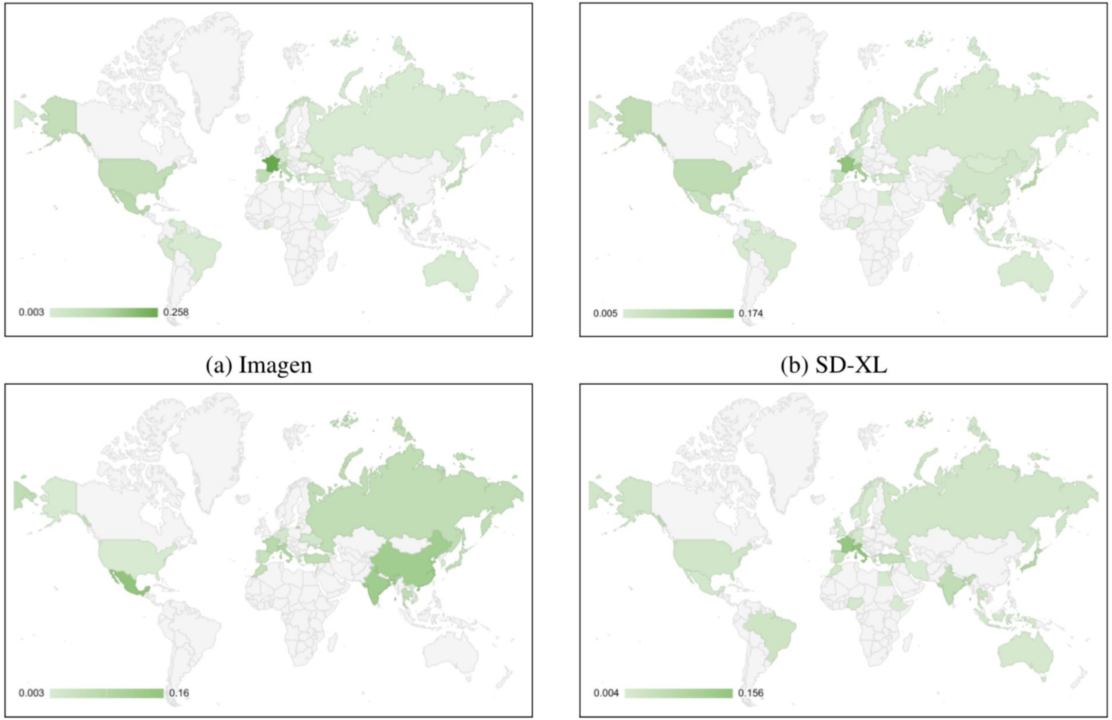


Figure 2: **Framework for evaluating cultural competence in T2I models**. The top subfigure shows the definition of *cultural concepts* and the extraction of *concept space* from KB + LLM. The bottom shows example task prompts to probe the model for cultural awareness and cultural diversity.

Geographical Inclination of models for under-specified prompts

Cuisine Domain

prompt: "images of food dishes"



(a) Imagen (b) SD-XL (c) Playground (d) RealVis

Skewed representation of global cuisines across models