PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

CuRe: Cultural Gaps in the Long-Tail of Text-to-Image Systems

Aniket Rege, Zinnia Nie, Mahesh Ramesh, Unmesh Raskar, Zhuoran Yu,
Aditya Kusupati, Yong Jae Lee, Ramya Korlakai Vinayak

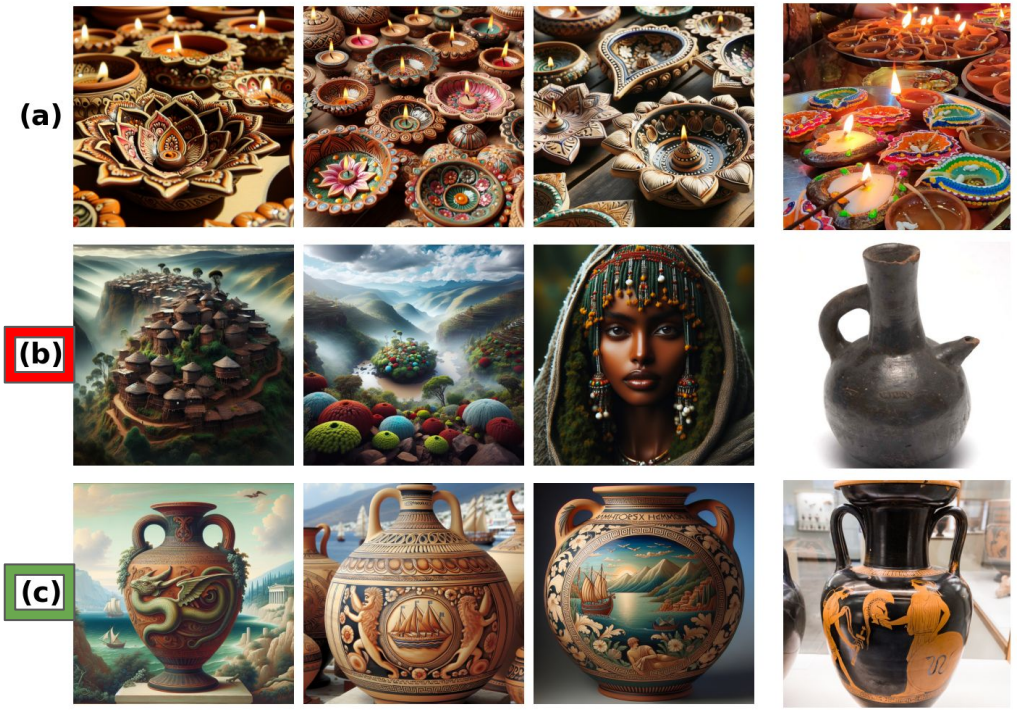







Generative Model Bias

- **Cultural Representativeness (CuRe)**: current text-to-image (T2I) systems do not represent global cultures equitably.
- America & Europe dominate pre-training data ("head"), while the “long tail” overlaps more with the Global South. How can we address this?
- Prompt Engineering? Specifying more information in prompt **can help**, but is an **unreliable method** to make the T2I system more **culturally representative**

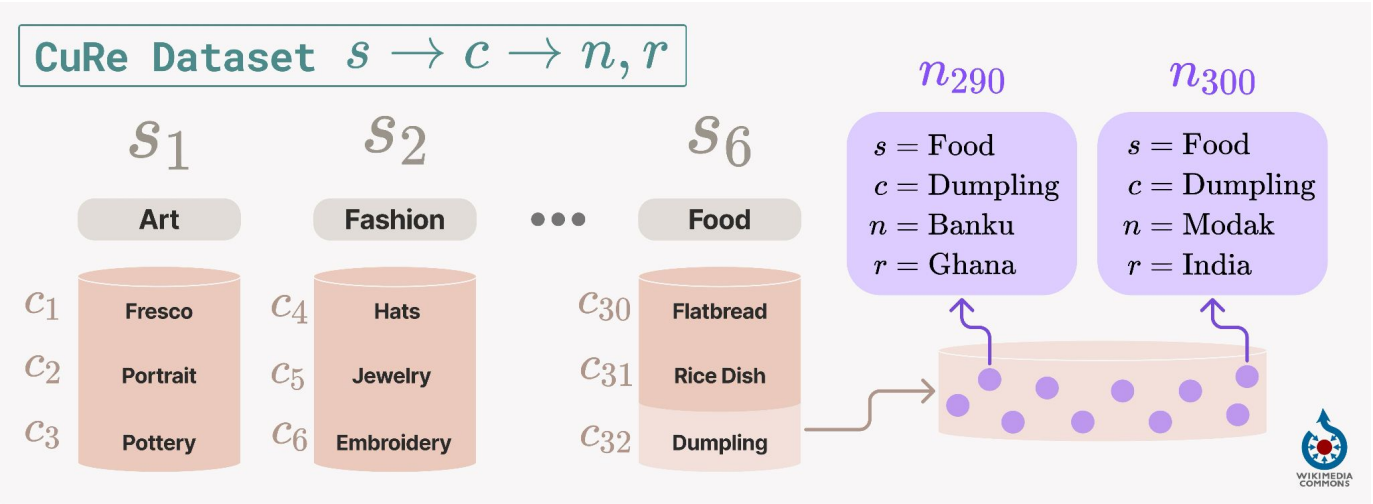


a) “Ceramic Diyas.”
b) “Jebena, from Ethiopia.”
c) “Amphora of Hermonax, a type of pottery from Greece.”

Images generated with Dall-E 3

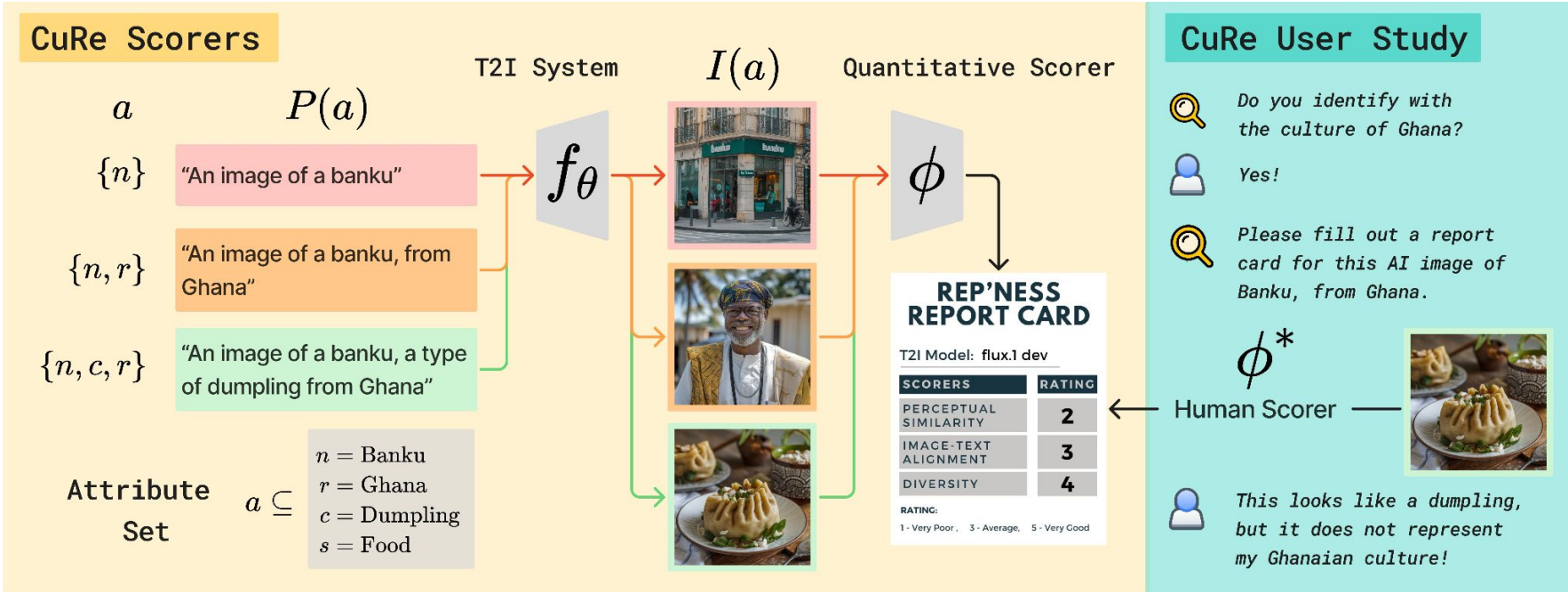
CuRe: a Benchmark and Scoring Suite

- Before solving for T2I system long-tail bias, we need to **accurately measure it.**
- **Ideal**: query humans from each culture to rate outputs (expensive, not scalable).
- **Our insight:** approximate CuRe by measuring how T2I systems respond to marginally increasing the attributes (**MIA**) in the text prompt!
- **New benchmark dataset** from Wikimedia with a *hierarchy of attributes* we can increase!



A Lens of Information Gain

Measure CuRe by evaluating how a **marginal change in information** explicitly provided to a T2I system (via text prompt) **changes its behavior**.



Quantitative Scorers

Perceptual Similarity φ_{PS}

Info Change: visual closeness of artifact → category (img-img)

Backbones: SigLIP 2 [12], DINOv2 [13], AIM v2 [14]

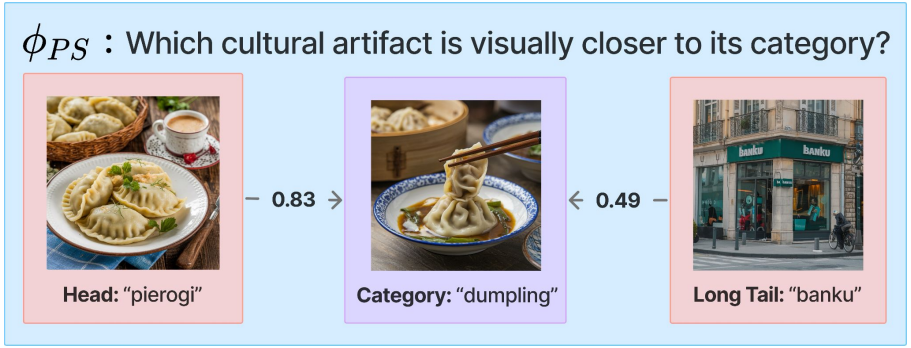
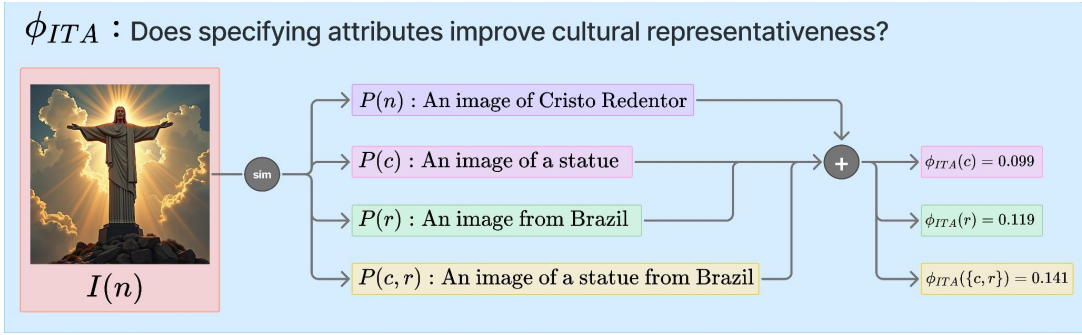


Image-Text Alignment φ_{ITA}

Info Change: textual closeness of an artifact to a change in attributes (img-txt)

Backbones: OpenCLIP [15], SigLIP 2

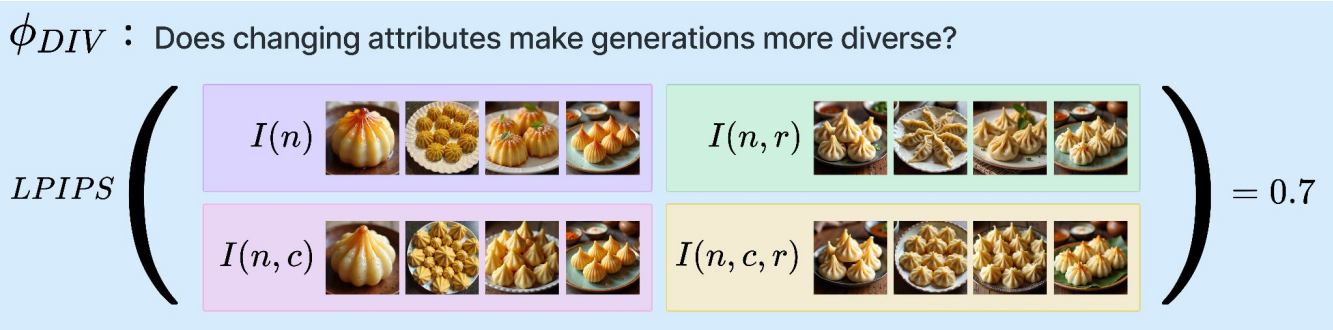


Quantitative Scorers (Cont.)

Diversity φ_{PS}

Info Change: visual diversity with changing attributes, e.g. Banku → Banku, a type of dumpling (img-img)

Backbones: LPIPS [16], Vendi Scores [17]



A Large Scale User Study

How do we know our scorers are good?

- For whole dataset, query 3 workers from each country who identify with its culture (2700+ workers from Prolific)
- T2I systems evaluated: Stable Diffusion 1.5, Stable Diffusion 3.5, FLUX.1 [dev]

Workers rate from 1 to 5:

1. **Cultural Representativeness** – "Could this image plausibly be found in your country?"
2. **Perceptual Similarity** – visual similarity to four real reference images from Wikimedia.
3. **Ground-Truth Likelihood** – correctness of the image's label (e.g. "Is this an image of spaghetti?").

We find: stronger rank correlation of our scorers to real human perceptions!

Results: a Teaser

	Images		Metric	ϕ
n_1			[43]	0.67
			[63]	0.79
			ϕ^*_{CuRe}	0.83
			ϕ^*_{PS}	0.75
			$\phi_{PS} \downarrow$	0.49
n_2			[43]	0.62
			[63]	0.71
			ϕ^*_{CuRe}	0.31
			ϕ^*_{PS}	0.44
			$\phi_{PS} \downarrow$	0.65
	AI : $I(n)$	**Real :** $G(n)$		

(a) Perceptual Similarity Scorer (ϕ_{PS}). n_1: "Chicken Biryani", n_2: "Omurice". Images were generated with SD 3.5 Large.

	Images		Metric	ϕ
			[31]	0.13
			[65]	0.11
			ϕ^*_{CuRe}	1.00
			ϕ^*_{PS}	0.75
			$\phi_{ITA} \uparrow$	0.14
			[31]	0.11
			[65]	0.09
			ϕ^*_{CuRe}	0.17
			ϕ^*_{PS}	0.15
			$\phi_{ITA} \uparrow$	0.01
	AI : $I(n)$	**Real :** $G(n)$		

(b) Image-Text Alignment Scorer (ϕ_{ITA}). n_1: "Sombrero", n_2: "Toquilla". Images were generated with FLUX.1 [dev].

	Images		Metric	ϕ
			LPIPS(n)	0.72
			VS(c)	0.24
			ϕ^*_{CuRe}	0.93
			ϕ^*_{PS}	0.66
			$\phi_{DIV} \downarrow$	0.57
			LPIPS(n)	0.70
			VS(c)	0.24
			ϕ^*_{CuRe}	0.46
			ϕ^*_{PS}	0.40
			$\phi_{DIV} \downarrow$	0.79
	AI : $I(n)$	**Real :** $G(n)$		

(c) Diversity Scorer (ϕ_{DIV}). n_1: "Spaghetti and meatballs", n_2: "Saimin". Images were generated with FLUX.1 [dev].

When humans judgments of T2I performance differ (φ^*_{CuRe} & φ^*_{PS}), our scorers are able to differentiate them (φ_{PS}, φ_{ITA}, φ_{DIV}) , but baselines cannot!

Conclusion

Takeaways

- Just like T2I systems, scorers do not work equally well across global cultures
- Lens of marginally increasing attributes provides valuable signal to approximate human judgments of cultural representativeness and perceptual similarity to ground-truth

More in our Paper:

- Which scorer configurations (backbone, attribute subset) most closely matches human judgments → Spearman Correlation plots for each scorer
- User study UI, questions, design choices
- How do strong multimodal LLMs perform?
- Long tail analysis in LAION, worker disagreement, scorers as long-tail predictors
- Lots of qualitative examples & discussion

Q&A

Thank you for attending!

Get in touch with me - happy to chat and open to collaborations!

Twitter: @wregss

Website: https://aniketrege.github.io/

Email: aniketr@cs.wisc.edu

CuRe Project Page

References

[1] Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

[2] Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[3] Li, Yuheng, et al. "Gligen: Open-set grounded text-to-image generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[4] Li, Yuheng, et al. "Generate Anything Anywhere in Any Scene." arXiv preprint arXiv:2306.17154 (2023).

[5] Mohammad Norouzi, William Chan, Jonathan Ho, Chitwan Saharia, Shayaan Abdullah, Jenny Lei, and Jacob Lu. ideogram.ai. https://ideogram.ai/, 2023.

[6] Schuhmann, Christoph, et al. "Laion-5b: An open large-scale dataset for training next generation image-text models." Advances in Neural Information Processing Systems 35 (2022): 25278-25294.

References

[7] Yu, Jiahui, et al. "Scaling autoregressive models for content-rich text-to-image generation." arXiv preprint arXiv:2206.10789 2.3 (2022): 5.

[8] Saharia, Chitwan, et al. "Photorealistic text-to-image diffusion models with deep language understanding." Advances in Neural Information Processing Systems 35 (2022): 36479-36494.

[9] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." International Conference on Machine Learning. PMLR, 2021.

[10] Sharma, Piyush, et al. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.

[11] Thomee, Bart, et al. "YFCC100M: The new data in multimedia research." Communications of the ACM 59.2 (2016): 64-73.

[12] Tschannen, Michael, et al. "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features." arXiv preprint arXiv:2502.14786 (2025).

[13] Oquab, Maxime, et al. "DINOv2: Learning Robust Visual Features without Supervision." Transactions on Machine Learning Research Journal (2024).

References

[14] Fini, Enrico, et al. "Multimodal autoregressive pre-training of large vision encoders." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

[15] Cherti, Mehdi, et al. "Reproducible scaling laws for contrastive language-image learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[16] Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[17] Senthilkumar, Nithish Kannen, et al. "Beyond aesthetics: Cultural competence in text-to-image models." Advances in Neural Information Processing Systems 37 (2024): 13716-13747.

Bonus Slides

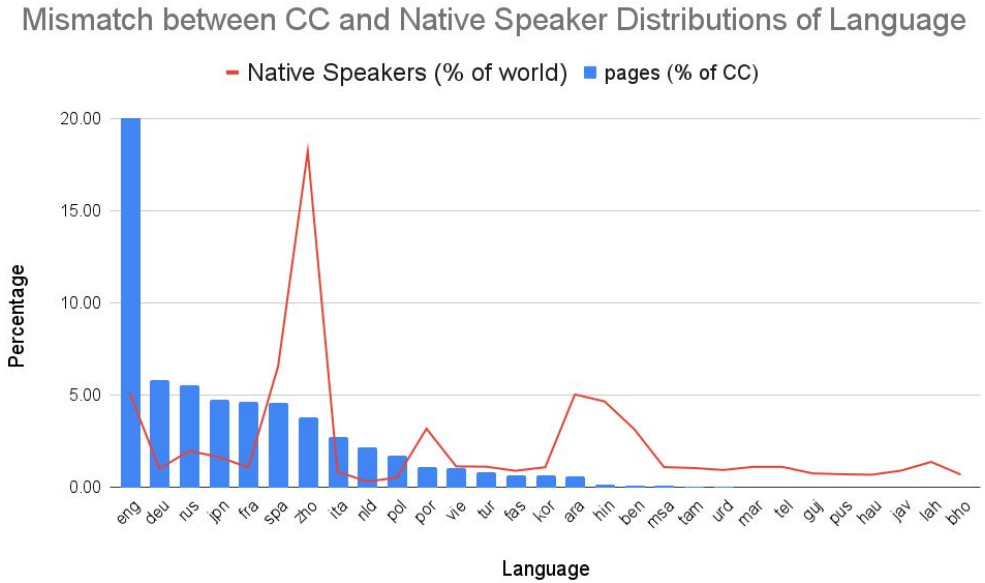
Where Does the Bias Come From?

What do we pretrain large T2I systems on?

→ Filtering Common Crawl

Common Recipe: LAION [6] subset + internal data

- Parti [7] → LAION-400M + FIT400M + JFT-4B
- Imagen [8] → LAION-400M + Internal 460M
- Dall-E [9] → CoCa [10] + YFCC100M [11] subset (~ 1B scale)



46% of CC is in English, but only 5% of world are native English speakers, while Chinese is natively spoken by 18% of the world, but is only 4% of CC.

Common Crawl language statistics from:
https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html

Image Synthesis

- **Goal:** Imagination to Image Translation
- Many existing works towards controllable image generation
  - Prompt-to-Prompt [1]
  - InstructPix2Pix [2]
  - GLIGEN [3], PACGen [4]
- **Problems:**
  - We can't fully specify what we imagine through text
  - To give grounding info (bbox, edges), we need an image of what we want a priori

Humans create art iteratively

Francesco del Giocondo
Draw my wife please.

Da Vinci
How's this?

Francesco del Giocondo
... no

Da Vinci
sigh... nobody appreciates genius

Image Credits : https://www.pinterest.com/pin/311029917987716070/

Imagination

What if Sir Floofy was a news anchor?

Describe the image in my head with text

"Square photograph of a feline news anchor at a studio, microphone ready, delving into the topic of whether it's preferable to be inside or outside. Behind the cat, the background showcases a split screen of a warm living room and a sunny park with the headline 'Breaking Debate: Inside vs. Outside"

DALL·E 3

Image

Keep modifying the description until we like the generated Image