Exploiting Cultural Biases via Homoglyphsin Text-to-Image Synthesis

Lukas Struppek [1, 4] Dominik Hintersdorf [1, 4] Felix Friedrich [1, 3] Manuel Brack [1, 4] Patrick Schramowski [1, 3, 4, 5] Kristian Kersting [1, 2, 3, 4]

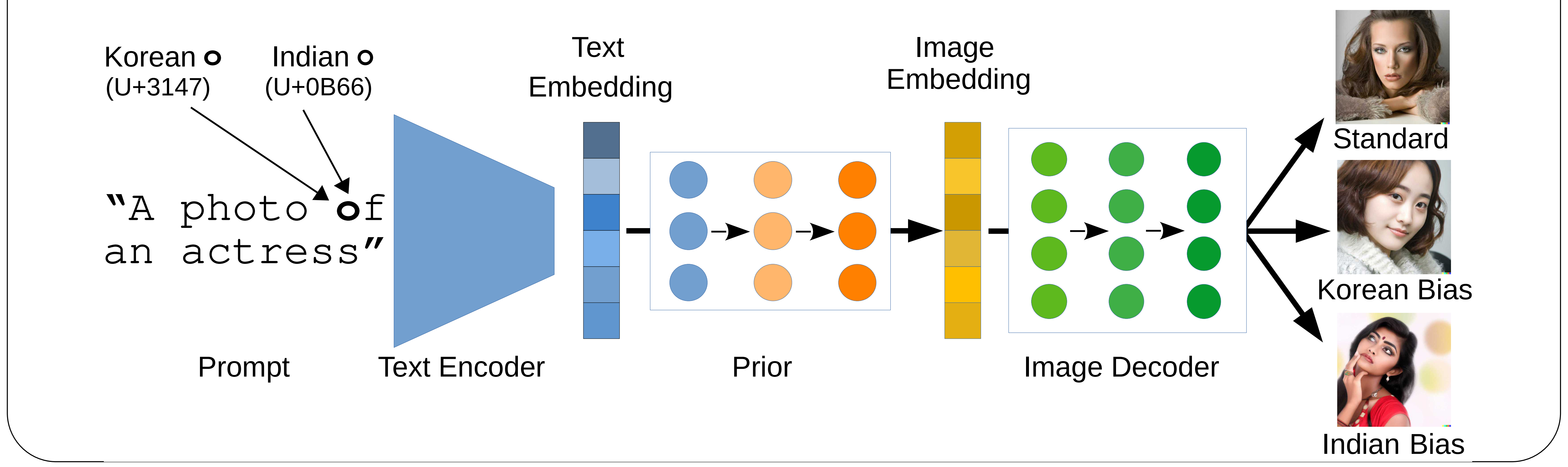
[1] Technical University of Darmstadt [2] Centre for Cognitive Science [3] Hessian Center for AI (hessian.AI) [4] German Research Center for AI (DFKI) [5] Ontocord

At a Glance

- Text-to-image synthesis systems react sensitively to character encodings in the input prompts.
- Generated images reflect cultural biases and stereotypes when inserting non-Latin characters.
- Inserting characters from native language scripts allows users to tailor the images to their cultural background.
- But this behavior can also be exploited to create racist stereotypes by replacing characters with homoglyphs.

Homoglyph Manipulations

Replacing **single characters** with similarly-looking characters from non-Latin scripts, so-called homoglyphs, leads to images reflecting cultural stereotypes and influences.

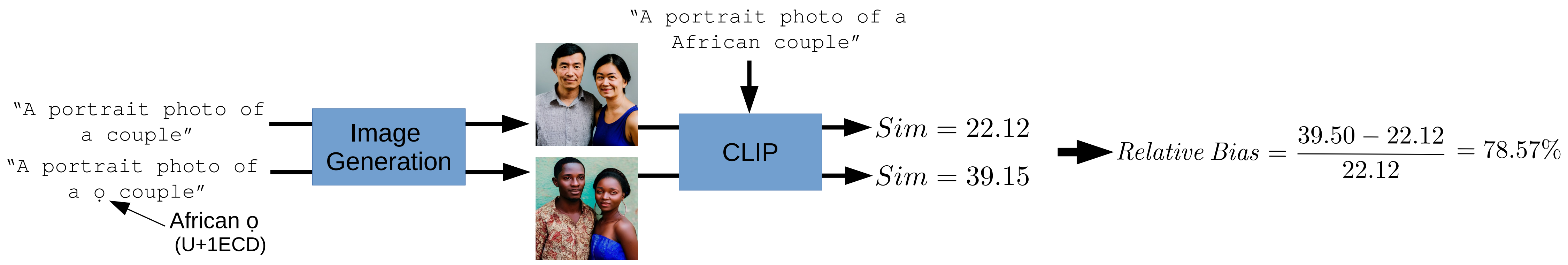


Code & Paper

www.github.com/LukasStruppek/Exploiting-Cultural-Biases-via-Homoglyphs

Measuring Cultural Biases

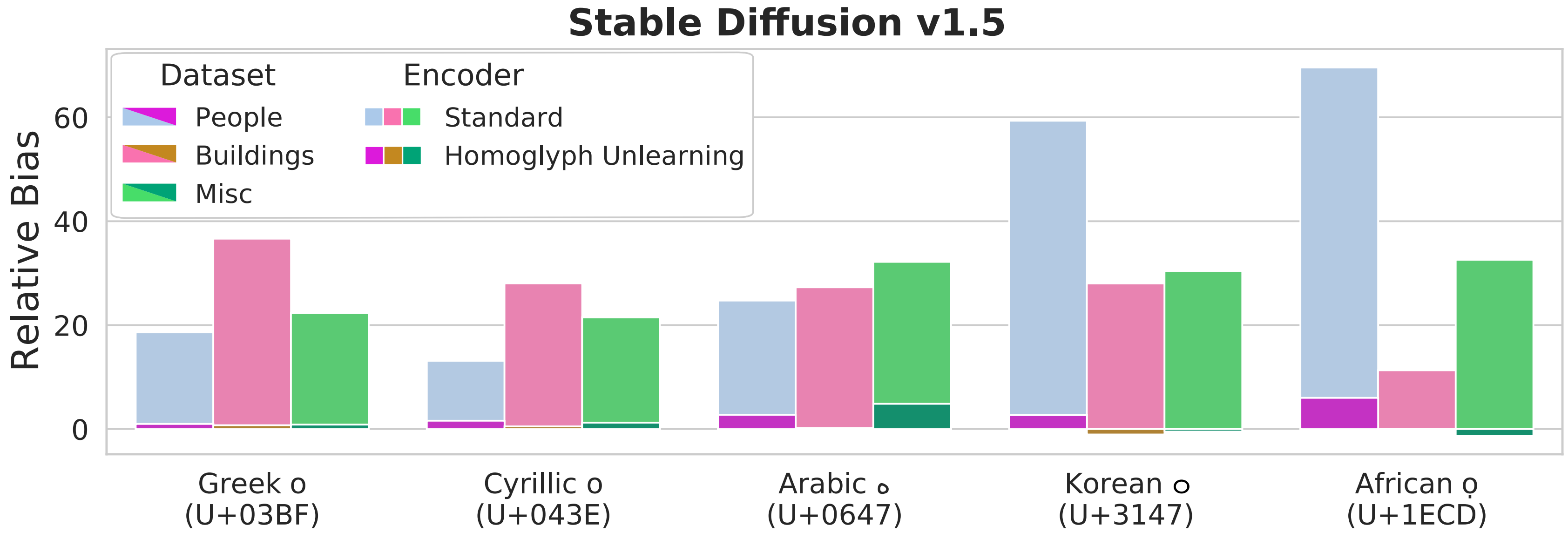
The Relative Bias induced by non-Latin characters is measured by **comparing the images generated with and without a non-Latin character**. The higher the CLIP similarity with a script’s associated culture, the stronger the induced cultural bias.


“A portrait photo of a couple”
“A portrait photo of a ọ couple”
African ọ (U+1ECD)
Image Generation
“A portrait photo of a African couple”
CLIP

$Sim = 22.12$

$Sim = 39.15$

$$Relative\ Bias = \frac{39.50 - 22.12}{22.12} = 78.57\%$$



Contact

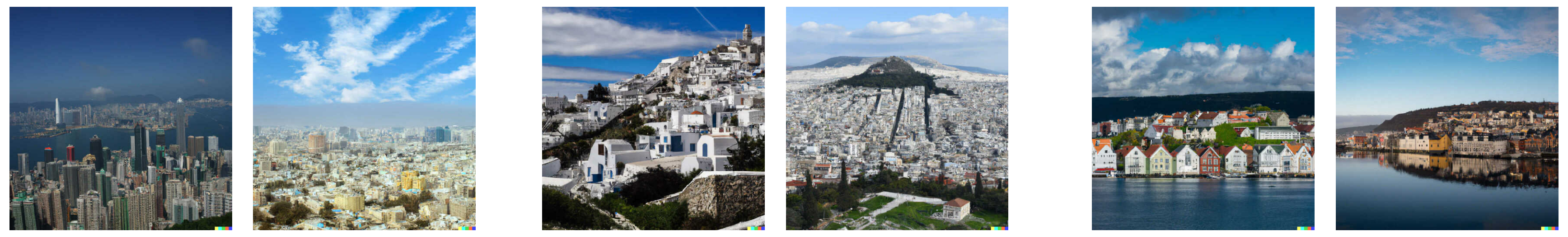
Please feel free to reach out to us!

Lukas Struppek
Technical University of Darmstadt
✉ struppek@cs.tu-darmstadt.de
𝕏 @LukasStruppek
🌐 lukasstruppek.github.io

Inducing Cultural Biases by Single Characters

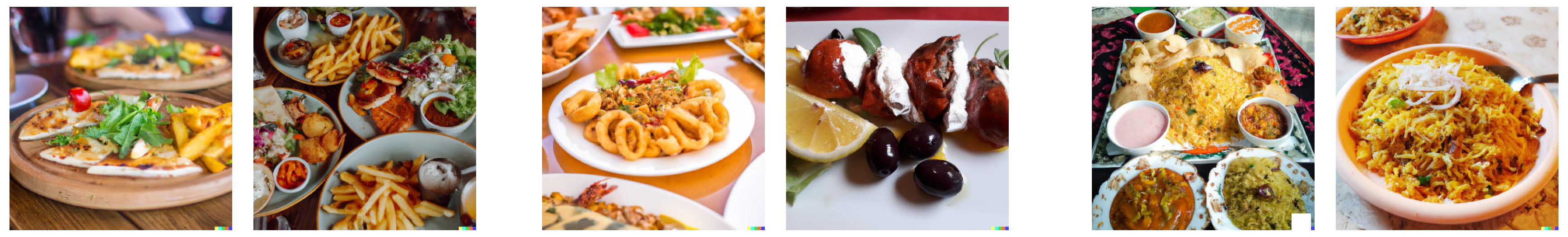
Characters from a wide range of scripts induce various cultural biases, including the appearance of **architecture, food, and people‘s visual appearance**, among many more domains.

DALL-E 2:
“A city in bright sunshine”



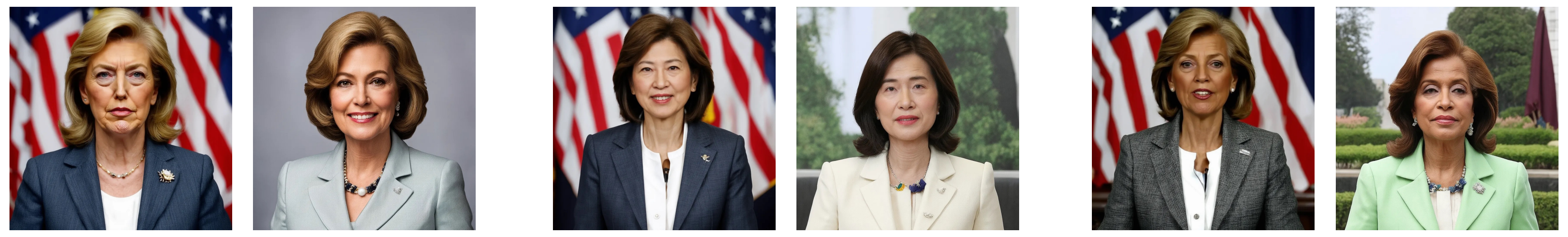
Latin A (U+0041) Greek Α (U+0391) Scandinavian Å (U+00C5)

DALL-E 2:
“Delicious food on a table”



Latin o (U+006F) Greek ο (U+03BF) Indian । (U+0964)

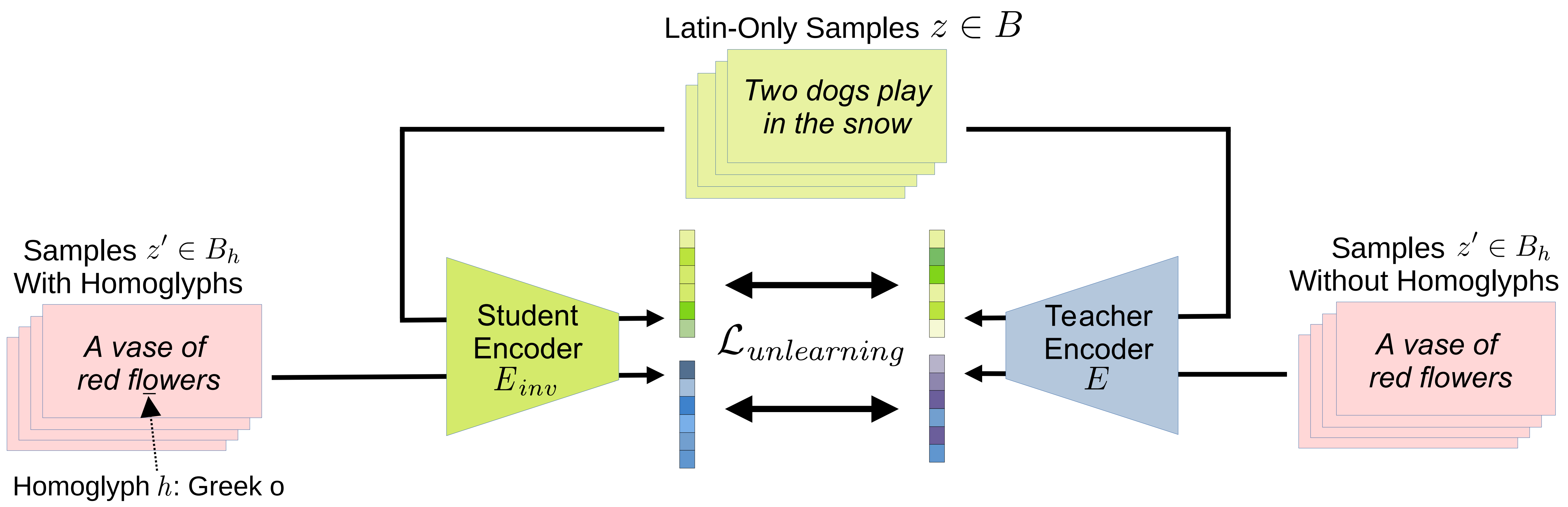
Stable Diffusion 3
“A photo of a female president”



Latin o (U+006F) Korean ㅇ (U+3147) Arabic ه (U+0647)

Homoglyph Unlearning

We fine-tune the text encoder to **map homoglyphs to their Latin counterparts**, making the model invariant to these characters. This invariance is achieved using an English text dataset and the original encoder as a teacher model for training signals.


Latin-Only Samples $z \in B$
Two dogs play in the snow
Samples $z' \in B_h$ With Homoglyphs
A vase of red flowers
Homoglyph h: Greek o
Student Encoder E_{inv}
$\mathcal{L}_{unlearning}$
Teacher Encoder E
Samples $z' \in B_h$ Without Homoglyphs
A vase of red flowers