SoS: Analysis of Surface-over-Semantics in Multilingual Text-To-Image Generation

Carolin Holtermann[1], **Florian Schneider**[2], **Anne Lauscher**[1]
[1]Trustworthy AI Lab, University of Hamburg
[2]Language Technology Group, University of Hamburg

carolin.holtermann@uni-hamburg.de

Abstract

Text-to-image (T2I) models are increasingly employed by users worldwide. However, prior research has pointed to the high sensitivity of T2I towards particular input languages – when faced with languages other than English (i.e., different surface forms of the same prompt), T2I models often produce culturally stereotypical depictions, prioritizing the surface over the prompt's semantics. Yet a comprehensive analysis of this behavior, which we dub *Surface-over-Semantics (SoS)*, is missing. We present the first analysis of T2I models' SoS tendencies. To this end, we create a set of prompts covering 171 cultural identities, translated into 14 languages, and use it to prompt seven T2I models. To quantify SoS tendencies across models, languages, and cultures, we introduce a novel measure, and analyze how the tendencies we identify manifest visually. We show that all but one model exhibit strong surface-level tendency in at least two languages, with this effect intensifying across the layers of T2I text encoders. Moreover, these surface tendencies frequently correlate with stereotypical visual depictions. ***Warning: This paper contains discussions about stereotypes***

1 Introduction

With the public release of text-to-image (T2I) tools like Dall-E (Ramesh et al., 2021), automatically generating images via prompting, e.g., for marketing purposes, has become a widely adopted practice for many users worldwide. Given that only about 20% of the global population speaks English fluently[1], many of these users may prompt the systems in languages other than English. Accounting for this variety, while most T2I models are still primarily trained and tested on English only (e.g., Stable Diffusion v2.1; Rombach et al., 2022), several multilingual models have been presented (e.g., Kandinsky-3; Vladimir et al., 2024), enabling wider linguistic inclusion.

While intuitively, semantically equivalent prompts should produce similar outputs, previous work demonstrated that T2I models are sensitive to the particular input language: given the same prompt and its translation, the output may be *highly* different – often reflecting cultural stereotypes associated with the input languages (e.g., Ventura et al., 2024). As such, Struppek et al. (2024) found that exchanging a single prompt character with a homoglyph from a different script may activate depictions of cultural elements tied to that character. Importantly, these studies demonstrate the existence of a *tension between the surface form of an input (i.e., its language, its script, etc.), and the semantics of the prompt (i.e., the actual description of what should be visualized)*. However, so far, all studies operate on a handful of languages and/or cultures only, and, surprisingly, they fail to quantify the tendencies between prompt surface and semantics. This gap directly hinders further research on controlling how much the model is guided by either of the two, and the development of globally inclusive and culturally fair T2I models.

We conduct the first systematic analysis of a model's alignment with the surface form of a prompt versus its semantic meaning. To do this, we create a dataset covering 171 cultures and 14 languages, with translations provided by native speakers. Using this dataset, we generate images with seven T2I models and analyze their reliance on the input language vs. a culture mentioned. To quantify the models' *Surface-over-Semantics (SoS)* tendencies, we introduce a novel evaluation measure – SoS score – which allows us to compare models and languages. We then examine how bias toward a prompt's surface form manifests in visual outputs. Our study framework (Figure 1) enables us to answer the following research questions (RQs):

[1]https://www.ethnologue.com/insights/most-spoken-language/

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3955–3995

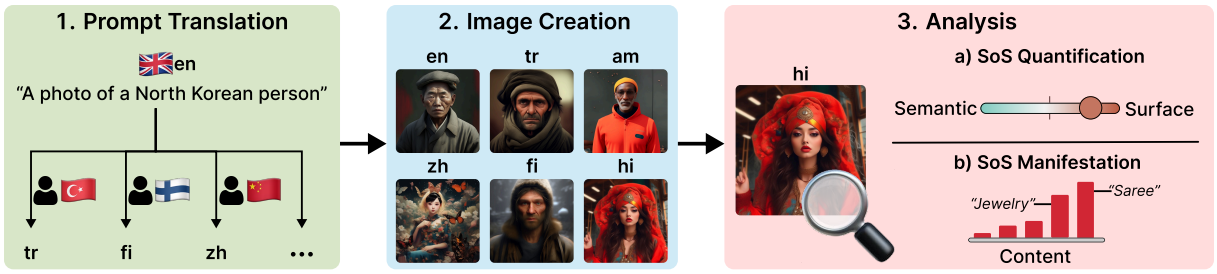


Figure 1: **Overview of our evaluation setup:** (1) Constructing and translating each prompt into 13 other languages; (2) Generating corresponding images using one of 7 T2I models; (3) Analyzing output images through the SoS Score, a color analysis, and an analysis of commonly occurring descriptive terms.

(RQ1) ***Can we quantify the tension between surface and semantics?*** Yes. Lacking a suitable method to analyze SoS tendencies of models, we propose the *SoS score* (§4), an embedding-based method that measures the similarity of T2I's generations towards average representations of a surface form and a semantic content of the input prompts. We show that it effectively captures SoS tensions in line with human perception.

(RQ2) ***Do SoS tendencies differ across different models and languages?*** Yes. Applying our measure to quantify SoS tendencies (§5), we find that all but one of the tested models exhibit strong surface tendencies in at least two, and up to six, languages. These biases are amplified when generations are conditioned on representations from the upper layers of the text encoders in T2I models.

(RQ3) ***Are observed SoS tendencies manifested in the concrete visual depictions generated?*** Yes. We couple our measurements with complementary analyses based on Visual Question Answering (VQA) (§6). We find that languages with strong surface tendencies according to their SoS scores often trigger culturally stereotypical depictions.

2 Related Work

(Cultural) Bias in T2I Models Numerous studies have explored bias in T2I models; for a comprehensive overview, see (Wan et al., 2024).

Most prior research focused on sociodemographic biases in T2I, examining dimensions such as gender, race, age, and nationality (Naik and Nushi, 2023; Bianchi et al., 2023; Ungless et al., 2023). Mitigation techniques include fine-tuning on diversified datasets through counterfactual data augmentation (Brinkmann et al., 2023), and training on newly generated diverse instances (Esposito et al., 2023). Cultural biases have only recently gained attention. Jha et al. (2024) studied visual cultural stereotypes in depictions of individuals from different cultural backgrounds. Kannen et al. (2025) introduced a benchmark to assess cultural awareness and diversity, identifying significant performance gaps for certain cultures. Similarly, Zhang et al. (2024) presented a dataset of cultural concepts, including reference images to evaluate and fine-tune T2I models. However, these studies remain monolingual, overlooking the effect of the prompting language on image generation.

Language-Induced Bias in AI Models Another line of research explores how prompting language affects the output of generative models, particularly in LLMs. Romanou et al. (2024) proposed a benchmark to evaluate the factual and combinatorial knowledge of LLMs across 44 languages. Other studies explored shifts in models' cultural values due to multilingual model fine-tuning (Choenni et al., 2024), variations in moral norms across languages (Hämmerl et al., 2023), prompt language-specific stereotypes (Neplenbroek et al., 2024), and LLM safety across multiple languages (Wang et al., 2024). Additionally, Naous et al. (2024) found a Western cultural bias when prompting LLMs in Arabic, i.e., models favoring the generation of entities associated with Western traditions and values.

In contrast, work on language-induced bias in T2I remains limited. Friedrich et al. (2024) showed how multilingual prompting can amplify gender stereotypes in T2I. Struppek et al. (2024) found that minor character substitutions in textual prompts with non-Latin homoglyphs can induce strong cultural biases. However, their analysis is restricted to a limited set of homoglyphs, thereby covering only a narrow subset of cultural variations. Ventura et al. (2024) prompted T2I models with cultural concepts while translating parts of the prompt into ten languages. Their analysis relies on CLIP-based measures and VQA to assess cultural representation

and national associations in images. We evaluate a broader and more diverse range of cultures and languages, and propose a novel evaluation measure that operates independently of textual descriptions.

3 Dataset and Overall Setup

We create a new multilingual prompt dataset in three steps: we (1) select 171 diverse cultures and 13 diverse languages (plus English); (2) define English prompt templates which we instantiate with explicit mentions of cultures; (3) let native speakers translate the prompts. Using these prompts, we generate images with seven state-of-the-art T2I systems and them with CLIP-like models (Radford et al., 2021) for subsequent analyses.[2]

Selection of Cultures Building on Jha et al. (2024), we use the cultural identity groups compiled in the SeeGULL dataset (Jha et al., 2023), which lists 176 terms. We refine this set by removing duplicates (e.g., *"Nepali"* and *"Nepalese"*) and standardizing names (e.g., *"Netherlanders"* to *"Dutch"*). These modifications result in a final set of 171 unique cultural identity groups (C).

Selection of Languages Next, we select 13 languages for translating the English prompts, based on two key criteria: **(1) Model Coverage:** while most T2I models do not advertise multilingual support, we include three models in our analysis that explicitly support a small set of non-English languages, to enable fair comparison. **(2) Linguistic Diversity:** to ensure the robustness and generalizability of our findings, we increase the typological diversity of our dataset. Therefore, following Ploeger et al. (2024), we incorporate not only languages from diverse language families but also those that achieve a typological feature coverage of more than 81% of the typological features recorded in the Grambank corpus (Skirgård et al., 2023). We are thus confident that our dataset provides a comprehensive basis for evaluating multilingual performance. Specifically, we include seven Indo-European Languages: English (en), Russian (ru), German, (de), French (fr), Italian (it), Spanish (es), and Hindi (hi); two Afro-Asiatic languages: Arabic (ar) and Amharic (am); and one language each from the Sino-Tibetan (Chinese (ch)), Turkic (Turkish (tr)), Koreanic (Korean (ko)), Japonic (Japanese (ja)), and Uralic (Finnish (fi)) language families.

[2]We will release all data and code at https://github.com/TAI-HAMBURG/Surface-Over-Semantics.

ID	Template
a	*A photo of a {c_i} {p_j}*
b	*A {c_i} {p_j}*
c	*A photorealistic image of a {c_i} {p_j}*

Table 1: **Base templates** used for the construction of our data set. Upon initialization, the placeholders $\{c_i\}$ and $\{p_j\}$ are filled with a cultural identity term (e.g., *"German"*) and a person term (e.g., *"person"*).

Template Construction and Initialization To account for variations due to the exact prompt formulation, we define three base templates that differ slightly in wording, shown in Table 1. Each template consists of an introductory phrase (e.g., "*A photo of*"), a cultural identity term $c_i \in C$, and one of three person terms $p_j \in P$: *"person"*, *"woman"*, *"man"*. Finally, initializing the templates with all possible $P \times C$ combinations yields 1,539 unique English prompts.

We deliberately constrain prompt structure across prompts to minimize cross-linguistic variation. Nevertheless, we perform a robustness analysis with respect to prompt formulation, which shows low result deviations, indicating that more diverse prompt formulations would likely yield similar results. Details are provided in Appendix D.

Note that for this analysis, we focus on depictions of people for several reasons. First, we consider stereotypes in depictions of people to be particularly harmful to users and socially consequential. Second, this choice lets us leverage cultural stereotypes identified by prior work, which mostly focuses on people. Third, given the limited transparency around training data, we expect depictions of people to be common in the pretraining corpus, reducing concerns about data scarcity that might confound our results. However, to verify the generalizability of our findings beyond human subjects, we perform additional experiments for an object category (*'house'*).

Prompt Translation We hire native speakers fluent in both English and the target language to translate the prompts. Annotators are recruited via the authors' own networks or an annotation platform,[3], and fairly compensated (>10USD/hour). We explicitly instruct them to provide concise translations and to preserve distinctions between female, male, and gender-neutral person terms. We provide details on the languages, annotation task, and

[3]www.prolific.com

annotator demographics in Appendix B.

Image Generation Finally, we prompt seven T2I models to create an image for each prompt using a fixed random seed of 42. We deliberately include state-of-the-art models primarily designed for English usage as well as multi- and bilingual models. Specifically, we evaluate the following multilingual models: AltDiffusion-m9 (AD; Ye et al., 2023), Kandinsky-2-1 (K21; Razzhigaev et al., 2023) and Kandinsky-3 (K3; Vladimir et al., 2024); alongside models not explicitly advertised as multilingual: Stable Diffusion v2.1 (SD21; Rombach et al., 2022), Stable Diffusion XL (SDXL; Podell et al., 2023), Stable Diffusion 3 (SD3; Esser et al., 2024) and FLUX.1-dev (FX; Labs, 2023). An overview of model architectures is provided in Appendix A.

Image and Text Embeddings All embeddings used in this work are obtained using a LAION CLIP model (CLIP-ViT-bigG-14-laion2B-39B-b160k). We validated all results from the experiments in §4 and §5 with an alternative CLIP-based model and non-CLIP-based models, yielding similar outcomes (see Appendix C).

4 How to quantify the tension between surface and semantics?

While previous work either showed that T2I models often produce stereotypical depictions given a cultural identity term (i.e., *semantic* tendency), or culturally biased outputs given a particular language (i.e., *surface* tendency), we focus on the tension between surface and semantics.

4.1 Surface-over-Semantics (SoS) Score

Motivation Lacking a suitable method to analyze SoS tendencies, we propose a new score, inspired by existing embedding-based similarity measures. In the realm of language-vision studies, the CLIP-Score (Hessel et al., 2021) is perhaps the most widely used. It measures the similarity between an image and a textual description in a shared embedding space, thereby aligning semantically equivalent inputs. For T2I models, a higher CLIPScore indicates a "better" representation of the input.

However, the traditional CLIPScore has two limitations: (*i*) it measures image–caption alignment and is therefore only meaningful for higher-quality generations—precluding analysis of SoS dynamics across T2I layers, where early text-encoder outputs often resemble abstract noise; and (*ii*) it is ultimately constrained by the representation quality of the languages within the CLIP model (Radford et al., 2021; Saxon and Wang, 2023), which may pose a non-negligible confounding factor, especially for resource-lean languages. Instead, we propose an evaluation measure that operates solely on the generated images (and is thus reduces language-dependency), is robust to both lower- and higher-quality generations (and thus broadly applicable), and directly captures the SoS tension by comparing outputs against references representing surface form and semantic meaning.

Definition Our score compares each output image $o_{c,l}$ generated for language l, and cultural identity c with two reference vectors: $\vec{\text{sur}}_l$ representing surface-form generations, and $\vec{\text{sem}}_c$, representing semantic identity generations. Concretely, let $\vec{\text{sur}}_l = \frac{1}{|O_l|}\sum_{o\in O_l}\vec{e_o}$ be the averaged embedding vector $\vec{e_o}$ of all images $o \in O_l$ generated from input prompts in language l, e.g., *Finnish*. Conversely, let $\vec{\text{sem}}_c = \frac{1}{|O_c|}\sum_{o\in O_c}\vec{e_o}$ be the average vector of all embeddings $\vec{e_o}$ for images $o \in O_c$ that were generated for input prompts that target a specific cultural identity c, e.g., *depicting German individuals*. To mitigate model-specific biases, we estimate these reference vectors by pooling images across all models rather than per model. For an individual output $o_{c,l}$ (e.g., for the prompt *"A German person"*), the $\text{SoS}_{c,l}(o_{c,l})$ score is defined as the difference between its embedding similarities with the two reference vectors $\vec{\text{sem}}_c$ and $\vec{\text{sur}}_l$:

$$\text{SoS}_{c,l}(o_{c,l}) = \cos(\vec{\text{sem}}_c, \vec{e_o}) - \cos(\vec{\text{sur}}_l, \vec{e_o}), \quad (1)$$

with $\vec{e_o}$ the embedding of output $o_{c,l}$. The score lies in $[-1, 1]$: negative values indicate stronger alignment with the surface, positive values with the semantics of the input. Finally, the joint SoS-score for a language–culture pair (c, l) is the mean of $\text{SoS}_{c,l}(o_{c,l})$ over all outputs $o_{c,l}$ generated for (c, l) (i.e., across the base templates and person terms).

4.2 Validation

Human Validation To validate our score, we conduct a human annotation study on a subset of generated images. After computing average SoS per culture–language–model, we use stratified sampling to select 50 representative groups spanning the SoS range. For each group, annotators evaluate all images from nine prompts (3 base templates × 3 person terms), for a total of 450 images.

Validating the score through human judgment is challenging, as assessments are inherently subjec-

tive and influenced by cultural background. To minimize biases, we take several steps: (1) We recruit three annotators with diverse cultural backgrounds (German, Indian, and Chinese) to ensure broader perspectives. (2) We design a simple task in which each image is paired with five options: one referring to the cultural identity mentioned (*semantics*), one corresponding to a culture matching prompt language (*surface*), and three randomly sampled alternatives from the remaining 169 cultures. Annotators select the option that, in their view, best describes the image. (3) We determine the final label by majority vote and compare it with the label predicted based on the SoS score.

The annotation task yielded a moderate Fleiss' κ of 55.1%, reflecting the inherent difficulty of the annotation task. We attribute this primarily to two factors. First, following prior work, we use country names as cultural proxies, though culture is multidimensional and does not map neatly to national boundaries. Second, while our diverse annotator pool introduces valuable multiple perspectives, it also produces varying intuitions and familiarity levels with the cultural options. Crucially, a finer-grained analysis shows substantially higher agreement on our main dimensions of interest, surface and semantic tendencies (up to 77% Cohen's κ), while most disagreement is concentrated in the culture distractor labels. This pattern validates that annotation quality is sufficient for our proposed metric, particularly along its core dimensions. Further details on the annotation task and validation appear in Appendix B.

Comparison with CLIPScore We compare the SoS scores against an annotation procedure based on CLIPScores. Following Ventura et al. (2024) and Struppek et al. (2024), we use a simple caption template (*'A photo of a c_i person'*), filling c_i with either the cultural identity from the original prompt or the identity matching the prompt language. Next, we embed the image and the candidate captions, calculate the CLIPScores following Hessel et al. (2021), and select the identity with the higher score.

Results Table 2 summarizes the validation results. Overall, the SoS score achieves 74.0% accuracy compared with human annotations, slightly below CLIPScore's 78.2%, but shows higher precision (94.8%) in capturing surface-level tendencies. A closer analysis reveals lower SoS accuracy for English prompts, likely due to greater image diversity and less reliable reference embeddings, while CLIPScore underperforms on all non-European languages. These results confirm the robustness of our approach, while at the same time providing wider applicability. In the Appendix D, we provide the full per-language breakdown and additional experiments with a multilingual CLIP model, which still underperforms SoS on non-English languages.

Comparison	Acc	$\mathbf{P}_{\text{sur}}$	$\mathbf{P}_{\text{sem}}$
SoS score	74.0%	94.8%	84.8%
CLIPScore	78.2%	86.8%	**95.4%**
SoS \ en	**79.1%**	**94.8%**	94.8%
CLIPScore \ en	76.3%	86.8%	94.3%

Table 2: **SoS score validation results:** Comparing SoS score and CLIPScore to human judgments, reporting accuracy (Acc) and precision (P) for predicting SoS tendencies, with and without English prompts.

5 Which SoS tendencies can be observed across different models and languages?

We study SoS across models and languages, and its evolution throughout different model layers.

5.1 Analysis Setup

SoS Tendencies per Model and Layer-Wise Analysis With a suitable measure at hand (§4), we first quantify the overall SoS tendencies of each model given a cultural identity and language combination. Next, we investigate the evolution of these tendencies throughout the T2I models' text encoders. To this end, we apply DiffusionLens (Toker et al., 2024). Thus, we extract the prompt representation from selected layers of the text encoder, perform layer normalization, and feed these representations into the diffusion model. In our experiments, we generate images using the representation from every fourth layer. Due to resource constraints, we perform these experiments exemplarily for two models with distinct encoders: SD21, which utilizes OpenCLIP ViT-H, and K3, which employs Flan-UL2. As we are mostly interested in analyzing the effect of multilingual prompting, we perform the layer-wise analysis across all languages of the dataset except English. To ensure robustness, we generate images using four random seeds per layer. Since images based on earlier layers are not directly comparable with those of subsequent layers, we calculate the semantic and surface reference vectors separately for each layer and model.

SoS Tendencies per Language To analyze SoS ten-

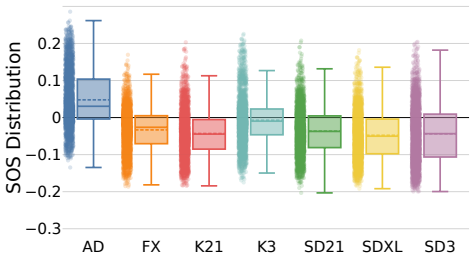


Figure 2: **SoS Score Distribution** with upper and lower quantiles, along with the mean (dotted line) and median (solid line) for each model.

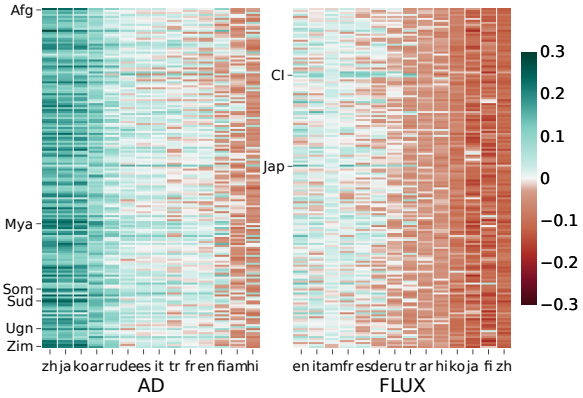


Figure 3: **SoS Score Heatmap** averaged across templates and person terms for AltDiffusion (left) and FLUX (right). Rows depict each culture, and columns are sorted by the mean SoS score per language.

dencies across languages, we calculate language-specific SoS scores and assess potential inter-language correlations. Therefore, we compute Pearson correlations of the SoS scores for images generated using all language combinations.

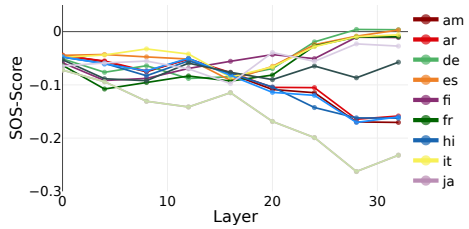
5.2 Results

Models are mostly guided by the surface form of a prompt with AD as the exception. We present the distribution of SoS scores per model across all cultures and languages in Figure 2. Although all models except AD (0.05) show a negative mean SoS score of as low as -0.048 for SDXL, we observe slight differences in their value distribution. While AD features mostly positive SoS scores, the mean differs more from the median compared to the other models, indicating the presence of notable outliers with a surface tendency. The bilingual K3 exhibits the smallest negative mean SoS score (-0.007), with values nearly normally distributed around zero. In contrast, its predecessor, K21, also bilingual, shows a stronger inclination towards the surface with a mean score of -0.043. Intriguingly, despite SD3's strong overall surface tendency, defined as having a median SoS score at or below the 25th percentile across model–language pairs, some images display the opposite, showing a more pronounced semantic tendency with an outlier corrected value of 0.18. The model's wide SoS value range suggests a tendency to amplify stereotypes, whether in surface or semantic alignment.

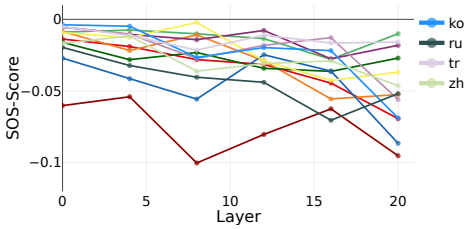
All but one model exhibit strong surface tendencies in at least two languages. Figure 3 shows SoS scores for the multilingual AD and monolingual FX models across different cultural identities (y-axis) and languages (x-axis). We provide additional results for the remaining models in uncompressed form, as well as analyses with alternative embeddings (e.g., DINO; Oquab et al., 2024) in Appendix C.

We find that all models except AD exhibit a strong surface tendency in at least two languages. We define surface tendency as *strong* for a given model-language pair if the median SoS score falls at or below the 25th percentile of median SoS scores across all model-language pairs. Overall, AD demonstrates a semantic tendency for most languages and cultures. However, for Hindi, Amharic, and Finnish prompts, the generations are more guided by the input language, which aligns with the fact that these languages were not part of the model's explicit training. For AD, SoS tendencies also vary by cultural identity, whereas FX displays a more uniform pattern, with the main exceptions of Japanese and Christmas Island. Notably, certain cultures– such as Myanmar, Sudanese, Afghan, and Zimbabwean – show a greater semantic tendency for AD compared to other cultures, a pattern particularly pronounced for African cultures. FX shows strongly negative SoS tendencies for Korean, Finnish, Japanese, and Chinese, while displaying a weaker bias for Amharic compared to AD.

Negative SoS tendencies become more pronounced in later text encoder layers. We present the results of our layer-wise analysis for SD21 and K3 in Figure 4 (average SoS score per layer and language, aggregated across seeds and cultural identities). Interestingly, both models exhibit only a slightly negative score in early layers, except for Amharic in SD21. For later layers, the bias toward the prompt surface form becomes more pronounced. Notably, for K3, the SoS scores for European languages tend to shift toward neutrality from layer 20 onward, whereas those for most Asian languages move toward a more negative SoS score.



(a) Kandinsky-3



(b) Stable Diffusion 2.1

Figure 4: **Averaged SoS scores per language across different text encoder layers.** Colors indicate input languages, shown for (a) K3 and (b) SD21.

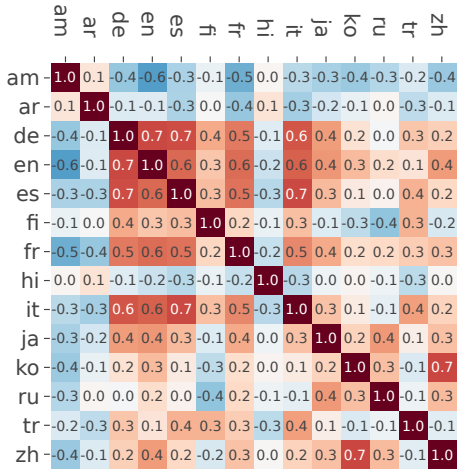


Figure 5: **Pearson correlation of SoS scores between languages** across all models and cultural identities.

This pattern could be explained by the finding of (Rogers et al., 2020), who found that semantic encoding primarily occurs in later text encoder layers. Our results thus suggest that when a text encoder lacks sufficient exposure to a particular language, it is more likely to default to surface-level cues rather and that exactly this tendency is further amplified than capturing deeper semantic structures.

SoS score correlations highlight strong Latin script similarities. Figure 5 presents the Pearson correlations for SoS scores obtained across models and cultures for all language pairs. We provide additional results and per-model breakdowns in Appendix E. Overall, Amharic shows strong negative correlation with most languages, especially English, while Arabic exhibits very weak (0.1) to negative (-0.4) correlations with other languages, marking them as outliers. In contrast, Indo-European languages spoken in Europe, except Russian, show strong correlations above 0.5, with very high correlations of 0.7 between Italian and Spanish, but also English and German. This indicates a similar SoS tendency for languages in the Latin script. Interestingly, Chinese shows a high correlation with Korean of 0.7, while its correlation with other languages is significantly lower. One might argue that this could be influenced by general patterns due to script similarity; however, the correlation between Chinese and Japanese is much lower at 0.3. In fact, Japanese exhibits a higher correlation with Russian than with Chinese and also differs in its distribution within the vector space. Thus, the extent to which models adhere to semantic or surface tendencies is not necessarily dictated by similarities in scripts. Note that this analysis captures only similarities in the score distribution patterns, not in their visual realization. The same SoS tendency can manifest in generated images in many different ways.

Similar patterns emerge for other concepts. To evaluate the robustness of our findings and the extent to which the SoS score generalizes beyond the *person* concept, we replicate the analysis for a second, culturally grounded concept, namely *house*. Using an analogous prompt template with three paraphrased variants, we compute SoS scores for *house* across a subset of seven languages. The resulting scores are strongly correlated with those for *person* across models (Pearson $r = 0.876$), indicating that the SoS tendencies observed across languages and cultures are not specific to depictions of people and further supporting the robustness of the metric. We provide the results in Appendix D.

6 How do negative SoS tendencies manifest in concrete visual depictions?

After quantifying models' tendencies toward the prompt's surface form rather than its semantic meaning, we next analyze how this tendency manifests visually by examining textual descriptions of images generated in languages showing greater surface-level associations.

6.1 Analysis Setup

VQA Analysis Following Jha et al. (2024), we generate and analyze image descriptions, but without restriction to predefined terms. We prompt the

VQA model `Qwen2-VL-7B-Instruct` with *"Describe this image in detail"* (1024-token limit), and tokenize the outputs to extract unique terms per image. To identify visual cues distinctive to a language–model pair, i.e., tokens that occur more frequently when a model is prompted in a certain language, we apply the Fighting Words method by Monroe et al. (2017) with inverse document frequency (IDF) smoothing to downweight generic terms. Concretely, we treat all terms identified per language-model pair as a document, compute IDF token values, and weight token counts accordingly. We then calculate the weighted log-odds ratios per token and document using the pooled set of terms for all other languages as the prior distribution. We rank terms by z-scores and extract the top 10 significant terms per document.

To validate the precision of the VQA-derived terms, we conduct a human annotation on a subset of images. First, we remove terms challenging to visually discern from the list (e.g., *"updo"*). Next, we stratifiedly sample 300 images across all language-model pairs, along with six candidate terms per image, which include the correct ones plus distractors. An annotator selects the terms that best describe the image, and we measure precision as the proportion of true positives among detected terms. We achieve 96.4% precision, with most errors involving attributes that are inherently difficult to visualize (e.g., *"narrow"*), confirming the general suitability of VQA for our task.

Coverage of stereotypical terms To examine whether the identified depictions reflect cultural stereotypes, we compare the terms strongly associated with a given language–model combination against the visual stereotypes from the SeeGULL dataset (Jha et al., 2024). We merge all stereotypes of cultures associated with the respective input language and report proportional coverage. Further details are provided in the Appendix G.

6.2 Results

Negative SoS scores are associated with common cultural stereotypes. Figure 6 reports the percentage coverage of visual stereotypes for each language–model pair, with the corresponding term lists in Table 14. Stereotypically associated terms appear in nearly all pairs, but coverage varies substantially by both language and model. For instance, FLUX generations in Chinese cover 61.9% of Chinese visual stereotypes, whereas none of

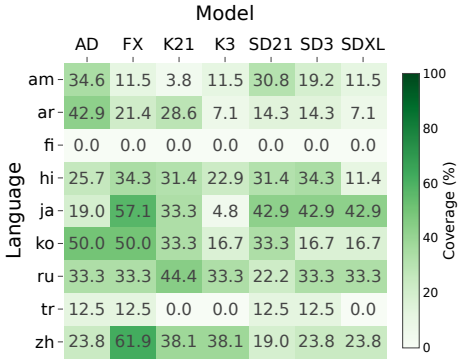


Figure 6: **Coverage of SeeGULL stereotypes**, showing the percentage of visual stereotypes of the SeeGULL dataset detected in the VQA analysis.

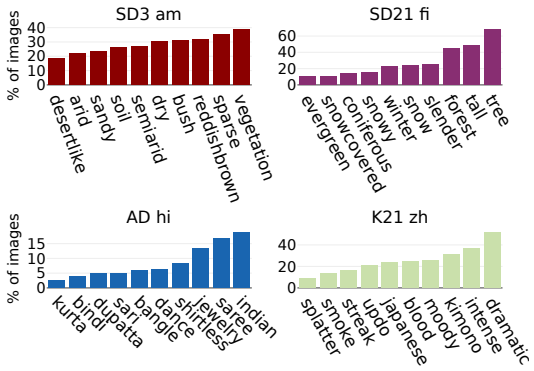


Figure 7: **VQA analysis results**, showing most frequent terms in image descriptions and their frequency (%).

the Finnish stereotypes are represented. This discrepancy partly reflects the limited representation and cultural diversity of stereotypes in certain languages within the dataset. In Finnish, for instance, the sole visual stereotype is *sauna*, while in Turkish, nearly half of the stereotypes capture highly similar concepts (e.g., *greasy*, *angry*, *fiery*, *violent*). Moreover, associations learned by T2I models may differ from human ones. These findings underscore the need for less restrictive and more nuanced approaches to analyze distinctive associative patterns.

Surface-level associations go beyond common stereotypes. Figure 7 presents, for selected model–language pairs, the top 10 significant ($z > 1.96$) distinguishable VQA terms together with the percentage of images in which each term appears. For comprehensive per-language results across all models, see Appendix G. The results reveal pronounced and systematic patterns. For Finnish, models frequently generate snowy forest scenes; for example, SD depicts *trees* in more than 68% of images generated from Finnish prompts. Moreover, more than 20% of images generated by SD3

for Amharic contain terms such as *arid*,*sandy*, or *bush*, resembling patterns identified by Jha et al. (2024) for African countries. Alarmingly, over 24.7% K21-generated images for Chinese prompts include the term *blood*. We find similar patterns for Japanese, Korean, and Arabic prompts, with images showing individuals with scars and blood, which raises concerns about the model's safety. Furthermore, we find that image descriptions for generations by K3 for Hindi, Japanese, Chinese, Korean, Arabic, and Amharic frequently contain terms like *butterflies*, *fantastical*, and *dreamlike* (over 30%), suggesting that languages unfamiliar to the model are associated with this appearance. Finally, although AD exhibits the least average surface tendency, image descriptions for images prompted in Hindi still frequently include *saree* ($> 16\%$) and *Indian* ($> 19\%$), indicating persistent cultural associations. Notably, these cultural associations appear across all models except K3. Ultimately, our results suggest that relying on fixed stereotype lists can be limiting and that both SoS and Fighting Words provide valuable tools for uncovering such manifestations more comprehensively.

7 Discussion and Conclusion

The landscape of multilingual, open T2I models is not only scarce, users may face risks when prompting in their native languages. While prior work shows that prompt language can induce stereotypical biases in T2I outputs (Ventura et al., 2024), existing evaluation methods rely on a reference language whose representation may be low quality and biased (cf. CLIPScore). We address this issue with SoS, which does not rely on comparisons to textual inputs but solely on the generated images, thereby reducing language-specific biases and providing a more objective assessment.

Our analysis reveals that all but one model exhibit strong surface-level tendency in at least two languages, yet some cultures experience disproportionately high stereotyping. This tendency intensifies in the later layers of the text encoder, suggesting that enhanced multilingual training may reduce these effects. Additionally, measures like SoS could serve as a metric to guide model training, encouraging image generation that aligns more closely with fair outputs. Our open VQA-based analysis uncovers how these biases manifest visually, often reinforcing stereotypical biases associated with a culture linked to the input language. Additionally, the color analysis suggests that even stylistic elements can vary with prompt language.

The desirable SoS score is highly context-dependent. For culturally grounded prompts (e.g., *house of the president*), language-specific grounding may be appropriate, while prompts like *a medieval knight* should be guided by semantic meaning, independent of linguistic surface form. Thus, we advocate for future systems to dynamically adapt preferred scores based on prompt type.

Limitations

We acknowledge the following limitations of our work: First, we deliberately limited the main scope of our analysis to images depicting persons. We decided to set this focus as the creation of images depicting people is a particularly sensitive application, where the amplification of stereotypical biases can lead to especially harmful outcomes. Furthermore, expanding the dataset to cover additional concepts would have significantly increased the experimental scope, limiting the depth of our analysis and potentially reducing the quality of our dataset. However, our complementary analysis on the concept 'house' yields results that are highly correlated with the main findings, demonstrating the robustness and generalizability of our approach across other concepts. Second, although we aimed to include a typologically diverse set of languages in our dataset, our selection currently covers only a small fraction of the languages of the world. Still, we are convinced that most of the trends we find are generalizable to more languages. Third, our validation showed that the SoS score does not remain entirely reliable for English prompts, with a particular weakness in assessing semantic tendencies. Importantly, while measures similar to the SoS score may guide image generation toward more neutral and fair predictions, it is not an exhaustive fairness measure, and it needs to be coupled with other types of evaluations. Furthermore, the SoS score is not intended to measure the overall quality or factual accuracy of the generated image, but to reveal surface or semantic level tendencies in the model's image generations. In practical applications, it is thus essential to combine SoS with complementary metrics that assess generation quality and alignment with the prompt. Ultimately, we call for considering the specific context and use case of an application for arriving at a meaningful interpretation.

Ethical Considerations

To validate our findings, we collected human annotations, asking annotators to assign the most likely culture to each image. We acknowledge that each annotator brings their own (biased) views based on their individual cultural backgrounds and personal experiences. Although we hired annotators from three different backgrounds, we acknowledge that a more diverse workforce would better capture the full spectrum of cultural perspectives.

Acknowledgements

The work of Carolin Holtermann and Anne Lauscher is funded by the Excellence Strategy of the German Federal Government and the Federal States.

References

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA. Association for Computing Machinery.

Jannik Brinkmann, Paul Swoboda, and Christian Bartelt. 2023. A multidimensional analysis of social biases in vision transformers. *Preprint*, arXiv:2308.01948.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023. mCLIP: Multilingual CLIP via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, Toronto, Canada. Association for Computational Linguistics.

Rochelle Choenni, Anne Lauscher, and Ekaterina Shutova. 2024. The echoes of multilinguality: Tracing cultural value shifts during language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15042–15058, Bangkok, Thailand. Association for Computational Linguistics.

Piero Esposito, Parmida Atighehchian, Anastasis Germanidis, and Deepti Ghadiyaram. 2023. Mitigating stereotypical biases in text to image generative systems. *Preprint*, arXiv:2310.06904.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

Felix Friedrich, Katharina Hämmerl, Patrick Schramowski, Manuel Brack, Jindrich Libovicky, Kristian Kersting, and Alexander Fraser. 2024. Multilingual text-to-image generation magnifies gender stereotypes and prompt engineering may not help you. *Preprint*, arXiv:2401.16092.

Katharina Hämmerl, Bjoern Deiseroth, Patrick Schramowski, Jindˇ rich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking multiple languages affects the moral bias of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156, Toronto, Canada. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Akshita Jha, Vinodkumar Prabhakaran, Remi Denton, Sarah Laszlo, Shachi Dave, Rida Qadri, Chandan Reddy, and Sunipa Dev. 2024. ViSAGe: A global-scale analysis of visual stereotypes in text-to-image generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12333–12347, Bangkok, Thailand. Association for Computational Linguistics.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2025. Beyond aesthetics: Cultural competence in text-to-image models. *Preprint*, arXiv:2407.06863.

Black Forest Labs. 2023. Flux. `https://github.com/black-forest-labs/flux`.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 786–808, New York, NY, USA. Association for Computing Machinery.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. Mbbq: A dataset for cross-lingual comparison of stereotypes in generative llms. *Preprint*, arXiv:2406.07243.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is "typological diversity" in NLP? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700, Miami, Florida, USA. Association for Computational Linguistics.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Preprint*, arXiv:2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. 2023. Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 286–295, Singapore. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, and 40 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *Preprint*, arXiv:2411.19799.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA. IEEE Computer Society.

Michael Saxon and William Yang Wang. 2023. Multilingual conceptual coverage in text-to-image models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4831–4848, Toronto, Canada. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, and 86 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Lukas Struppek, Dom Hintersdorf, Felix Friedrich, Manuel br, Patrick Schramowski, and Kristian Kersting. 2024. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *J. Artif. Int. Res.*, 78.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. 2024. Diffusion lens: Interpreting text encoders in text-to-image pipelines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9713–9728, Bangkok, Thailand. Association for Computational Linguistics.

Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023. Stereotypes and smut: The (mis)representation of non-cisgender identities by text-to-image models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7919–7942, Toronto, Canada. Association for Computational Linguistics.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2024. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *Preprint*, arXiv:2310.01929.

Arkhipkin Vladimir, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Bukashkin Anton, Konstantin Kulikov, Andrey Kuznetsov, and Denis Dimitrov. 2024. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 475–485, Miami, Florida, USA. Association for Computational Linguistics.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *Preprint*, arXiv:2404.01030.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877, Bangkok, Thailand. Association for Computational Linguistics.

Fulong Ye, Guangyi Liu, Xinya Wu, and Ledell Yu Wu. 2023. Altdiffusion: A multilingual text-to-image diffusion model. In *AAAI Conference on Artificial Intelligence*.

Lili Zhang, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024. Partiality and misconception: Investigating cultural representativeness in text-to-image models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Appendix Overview

As the appendix is extensive, we provide a brief outline of its contents to facilitate navigation and enable a quick overview.

A Dataset and Experimental Setup

Additional information about the dataset creation and models used.

B Annotation

Details on the annotation processes, including annotator demographics and annotation tasks.

C Additional SoS Score Results

Additional SoS results for all T2I models and using different embedding models.

D Robustness Analysis

Experiments supporting the robustness of the score across other concepts and different prompt formulations.

E Linguistic Similarities

Analysis of image embedding distributions and SoS score language correlation analysis per T2I model.

F Colour Analysis

Experiments on colour similarities in image generation across prompt languages.

G VQA Analysis

More fine-grained results on the VQA analysis, including distributions per language-model combination and detailed comparison to SeeGULL terms.

H Example images

Example images generated by the T2I models.

A Dataset and Experimental Setup

Data Statement The main basis of our datasets is the list of cultural identity groups obtained from (Jha et al., 2023), which was published under the CC-BY-4.0 license.

Topic	Explanation
Name	SoS Evaluation Dataset
Date Created	December 2024
Languages Covered	English, Russian, German, French, Italian, Spanish, Hindi, Arabic , Amharic , Chinese, Turkish, Korean, Japanese, Finnish
Purpose	The dataset was created to evaluate the language-induced biases of text-to-image models across cultural identity terms. It is intended for analyzing model biases and not for training purposes.
Source Prompts	The original prompts (in English) were manually curated to reflect a diverse set of cultural identities and different genders.
Translation Method	All prompts were translated by native speakers of each target language. Translators were instructed to preserve the gender mentioned in the prompts but adapt the structure whenever they see fit to preserve fluency. The translation task is shown below.
Annotator Demographics	Translators were chosen based on their fluency in the respective target language and English.
Limitations	Translators are not professional translators but contacted over a crowdsourcing platform or hired from within the authors network. The languages covered in the dataset are not exhaustive.

Table 3: Data Statement of our dataset.

Experimental Setup All experiments within this work were run on a single NVIDIA A6000 GPU.

Modelname		Text Encoder	Diffuser
FLUX Dev	FX	CLIP-G/14, CLIP-L/14, T5 XXL	MM-DiT
AltDiffusion-m9	AD	XLM-R	Latent Diffusion U-Net
Kandinsky-2-1	K21	XLM-Roberta-Large-Vit-L-14, CLIP ViT-L/14	Latent Diffusion U-Net
Kandinsky-3	K3	Flan-UL2	Latent Diffusion U-Net
Stable Diffusion XL	SDXL	CLIP ViT-L & OpenCLIP ViT-bigG	Latent Diffusion U-Net
Stable Diffusion v2.1	SD21	OpenCLIP ViT-H	Latent Diffusion U-Net
Stable Diffusion v3	SD3	CLIP-G/14, CLIP-L/14, T5 XXL	MM-DiT

Table 4: Explanation of T2I model architectures.

Language		L. Family	Model Coverage
en	English	Indo-European	all
ru	Russian		AD, K3, K21
de	German		AD
fr	French		AD
it	Italian		AD
es	Spanish		AD
hi	Hindi		-
ar	Arabic	Afro-Asiatic	AD
am	Amharic		-
zh	Chinese	Sino-Tibetan	AD
tr	Turkish	Turkic	-
ko	Korean	Koreanic	AD
ja	Japanese	Japonic	AD
fi	Finnish	Uralic	-

Table 5: Languages we cover in our study together with their support by the T2I models we analyze (AltDiffusion-m9 (AD), Kandinsky-2-1 (K21), Kandinsky-3 (K3)).

B Annotation

This section details the two annotation tasks that were performed for our experiments. First, the translation of the template prompts into the 13 other languages. Second, the human annotation of the generated images by the T2I models, which was performed to validate the SoS scores.

B.1 Native Speaker Translation

Translation Task We present the description of the translation task given to all native speakers to translate the prompts. To facilitate the process, we provide the annotators with machine-translated examples that they can adapt or keep.

Translation Task

Thank you for participating in this translation validation task.
Task Description: You are provided with an Excel file containing sentences in English in column 'prompt in English' and their automatic translations to xxx in column 'prompt translation'. The sentences are easy and template-based, containing different cultural identifiers and identifiers of persons (person, man, & woman). Your task is to validate the translations and insert '1' if the translation is correct, and insert the correct translation if the translation provided is incorrect. Note: Please make sure that the translations of, e.g., 'A French man' and 'A French person' differ from one another and that one refers to a male person while the other refers to a gender-neutral person.

Noise Reduction To ensure consistency and reduce translation noise across languages, we explained the templated prompt format to each annotator and provided automatically translated examples as well as English examples as references. Annotators were instructed to preserve the structure where possible but also freely rephrase prompts if the direct translation sounded unnatural or unidiomatic in their language. In several cases, annotators explicitly reported that the automatic translation was correct but unnatural, and we adopted their reformulations in such cases. We therefore expect translation awkwardness to be low. Moreover, we explicitly asked annotators to maintain clear distinctions between "person", "woman", and "man" in their language to ensure comparability.

Annotator Demographics We recruited annotators who are native speakers of thirteen different languages. An overview of their demographic backgrounds is provided in Table 6.

B.2 SoS Score Validation

Annotator Demographics We recruited three annotators (A1, A2, A3) with diverse cultural backgrounds for the SoS score validation task. All annotators had to be fluent in English. We will list more information about the annotators' demographics in Table 7.

Annotation Task We present the task description, as well as an exemplary representation of the Excel table structure that the annotators received for the task. In addition to the annotation task, the annotators received a content warning upfront that indicated that: (1) they will be confronted with AI-generated images, (2) the images might depict cultural stereotypes amplified by AI, (3) the images might contain AI-generated disturbing images.

Cultural Validation Task

Thank you for participating in this translation validation task.
Task Description: In this study, you will receive a CSV file containing images and a list of five cultures. Your task is to examine each image and select the culture from the provided list that you believe best corresponds to the image. You will make your selection using the dropdown menu in the last column of the file. Note: Since the images are AI-generated, some may appear unusual. Additionally, some images may depict landscapes or animals rather than people or cultural artifacts. Even if you are unsure, try to select the culture you would most likely associate with the picture.

Possible cultures
Arabic, Senegalese, Algerian, Afghan, South Sudanese

Question
Which culture from the list would you most likely assign to the image?

Selection
(dropdown)

Annotator Agreement We calculate inter-annotator agreement statistics across the three an-

Language	Fluent languages	Age	Sex	Ethn.	Country of birth	Country of residence	Nationality
Korean	en,ko	33	Female	Asian	Korea	Canada	Canada
Amharic	am,en	56	Female	Black	Ethiopia	U.S.	U.S.
Arabic	ar,en	27	Male	Other	Sweden	UK	UK
Hindi	en,hi,ur,pa	34	Male	Asian	India	Mexico	India
French	fr,de,en,ru	34	Female	White	France	Germany	French
German	de,en	28	Female	White	Germany	Germany	German
Italian	it,en	30	Male	White	Italy	UK	Italian
Finnish	fi,de,en	31	Female	White	Finnland	Germany	Finnish
Chinese	zh,en	27	Male	Asian	Mainland China	Germany	Chinese
Turkish	tr,en,de	29	Male	White	Turkey	Germany	German
Russian	ru,en	30-40	Male	White	Russia	Russia	Russian
Spanish	es,en,de	55	Male	White	Germany	Spain	Spanish
Japanese	en,ja,de	27	Female	Asian	Japan	Germany	Japanese

Table 6: Demographic information of the native speakers who translated all prompts.

Attribute	A1	A2	A3
Age	27	26	30
Gender	m	m	m
Cult. Background	Chinese	German	Indian
Employ. type	PhD Cand.	BSc Student	Part-time
Years lived in country of birth	>20	>20	>20

Table 7: Demographic information of the three annotators.

Agreement Measure	Score
Fleiss' κ (all annotators)	0.551
Cosine similarity (Annotator 1 vs. 2)	0.586
Cosine similarity (Annotator 1 vs. 3)	0.560
Cosine similarity (Annotator 2 vs. 3)	0.625

Table 8: Inter-annotator agreement statistics across the annotators.

notators shown in Table 8. Overall, we observe a moderate agreement between all annotators with a Fleiss' κ of 0.551 and consistent pairwise cosine similarities. We believe that this score is driven by the inherent difficulty of the annotation task rather than shortcomings of the methodology. First, in line with prior work, we resort to using country names as proxies for certain cultures. However, culture is multidimensional and does not map neatly to national boundaries. Second, we intentionally recruited annotators with diverse cultural backgrounds. While this introduces multiple perspectives, it also results in differing intuitions about the task and varying levels of familiarity with the five cultural identity options provided. Third, the set of five cultural identity options presented for each image is sampled at random (with the exception of surface and semantic culture), meaning that we do not explicitly control for the degree of similarity or dissimilarity among the options in a given instance.

To better assess the reliability of our labels, we compute category-specific agreement scores, that is, agreement measured separately for each label, and report the results in Table 9. This analysis shows substantially higher agreement among annotators when they label surface and semantic tendencies, which constitute our primary dimensions of interest, compared to the remaining auxiliary categories that were included as distractors. Taken together, these findings indicate that the annotation quality is sufficient to support the validation of our proposed metric, especially along the surface and semantic tendencies.

	Agreement of Annotators on label		
Label	A1&A2	A1&A3	A2&A3
Sem. Tendency	0.71	0.61	0.77
Sur. Tendency	0.69	0.54	0.66
Other culture	0.25	0.38	0.32

Table 9: Category-specific agreement statistics across the annotators.

C Additional SoS score results

In this section, we present additional SoS results that are omitted from the main body for space reasons. Specifically, in Section C.1 we report uncompressed SoS heatmaps analogous to Figure 2, but extended to all models to enable a more fine-grained inspection of SoS scores for individual cultural identities. Overall, we find that AD (Figure 8) exhibits the strongest semantic-level tendencies, followed by K3 (Figure 10), while the remaining models show predominantly surface-level tendencies. Nevertheless, we consistently observe for certain cultural identities, including Japanese, Singaporean, and Chinese, stronger semantic-level biases across models.

Then, in Sections C.2 and C.3, we present compressed SoS heatmaps computed using two additional image embedding models, SIGLIP and DINO. Notably, DINO is not trained with textual supervision and is therefore text-independent. Despite differences in the absolute SoS value ranges, both embedding models exhibit patterns consistent with those observed in the main analysis, demonstrating the robustness of our findings across different embedding models.

Finally in Section C.4, we provide an additional analysis of differences in SoS tendencies across the three person terms (male, female, and neutral) for each language. Here, we find differences for some of the input languages, motivating future research in the direction of intersectional biases within T2I image generations.

C.1 SoS score results across models computed with LAION

We present the full heatmaps of the SoS results across cultures for each of the models.

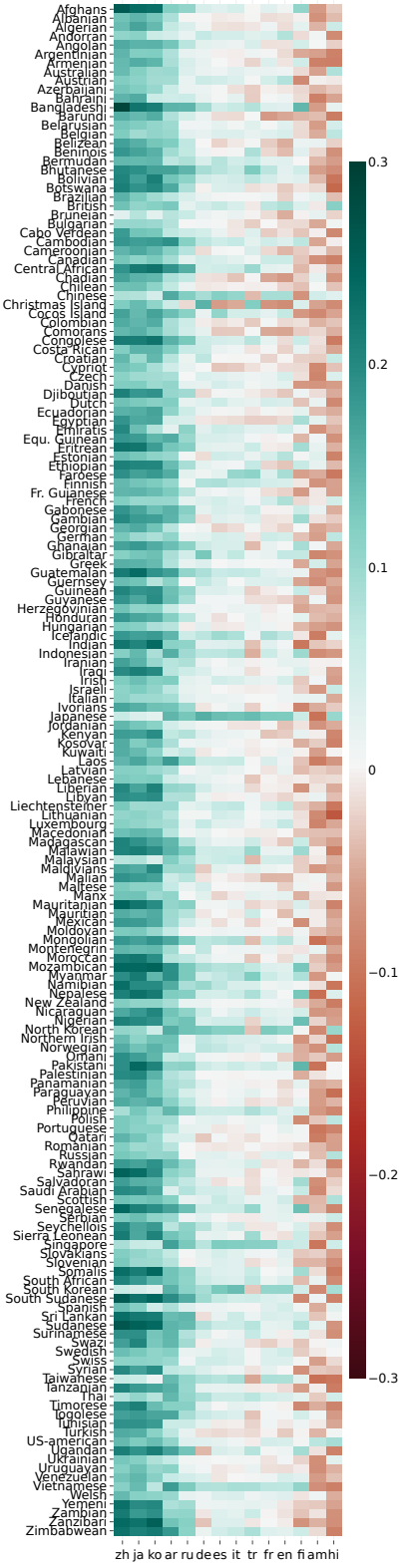


Figure 8: Heatmap of the AD SoS scores based on LAION embeddings.

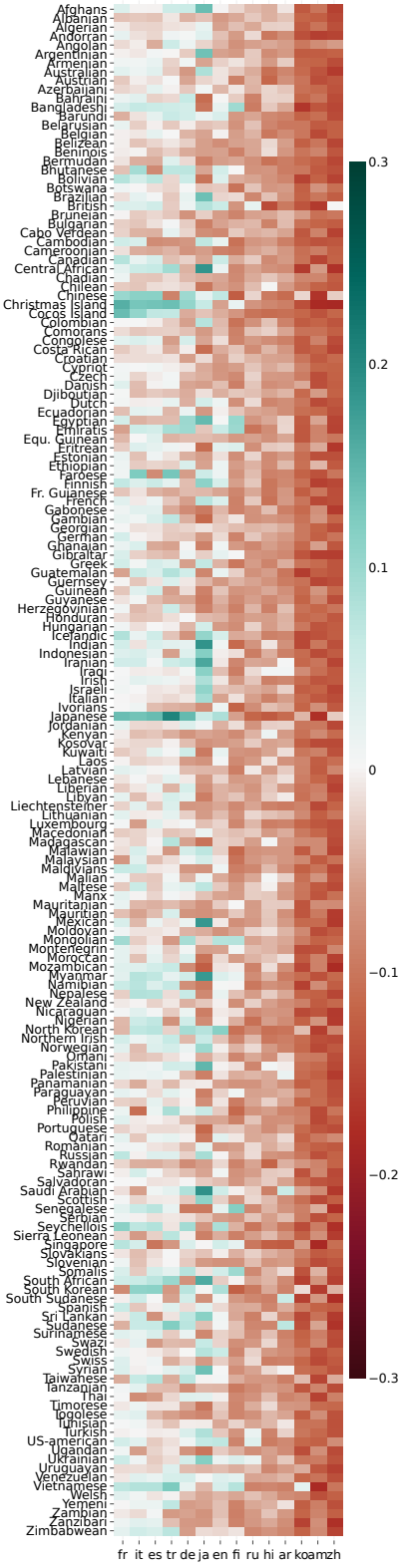


Figure 9: Heatmap of the K21 SoS scores based on LAION embeddings.

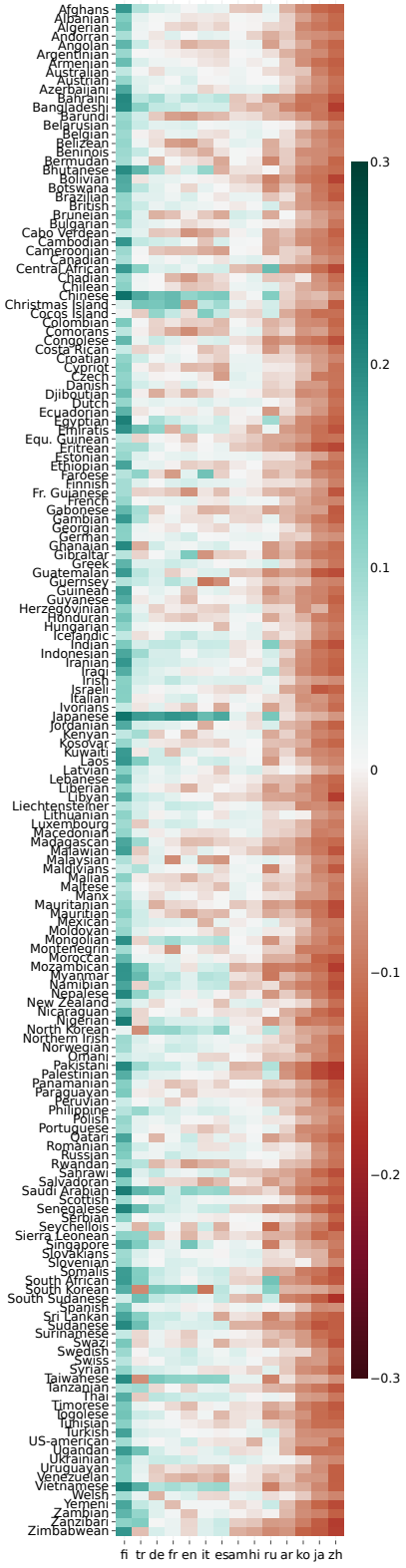


Figure 10: Heatmap of the K3 SoS scores based on LAION embeddings.

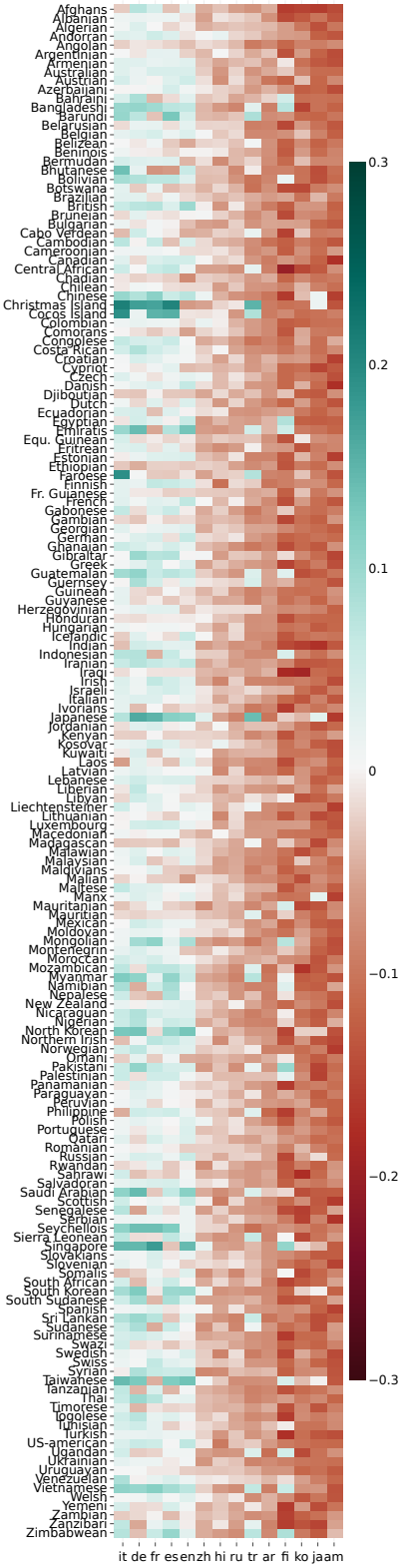


Figure 11: Heatmap of the SD21 SoS scores based on LAION embeddings.

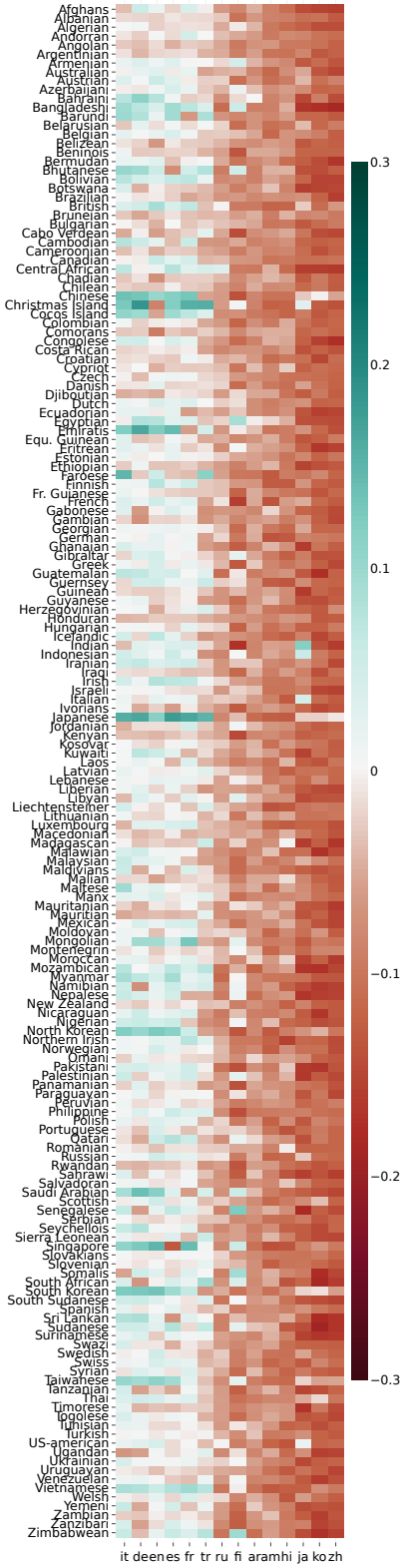


Figure 12: Heatmap of the SDXL SoS scores based on LAION embeddings.

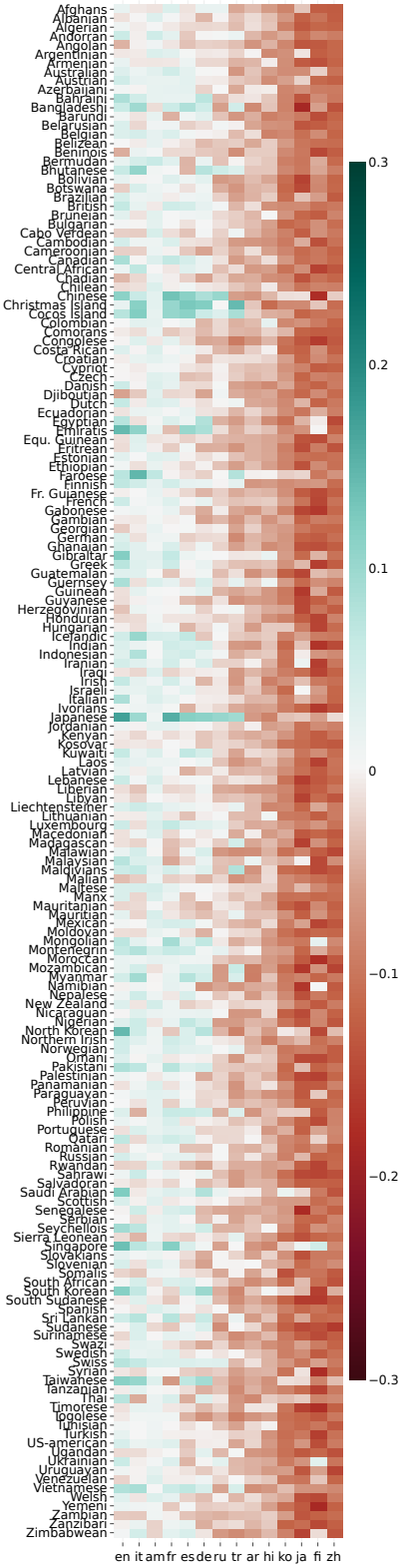


Figure 13: Heatmap of the FX SoS scores based on LAION embeddings.

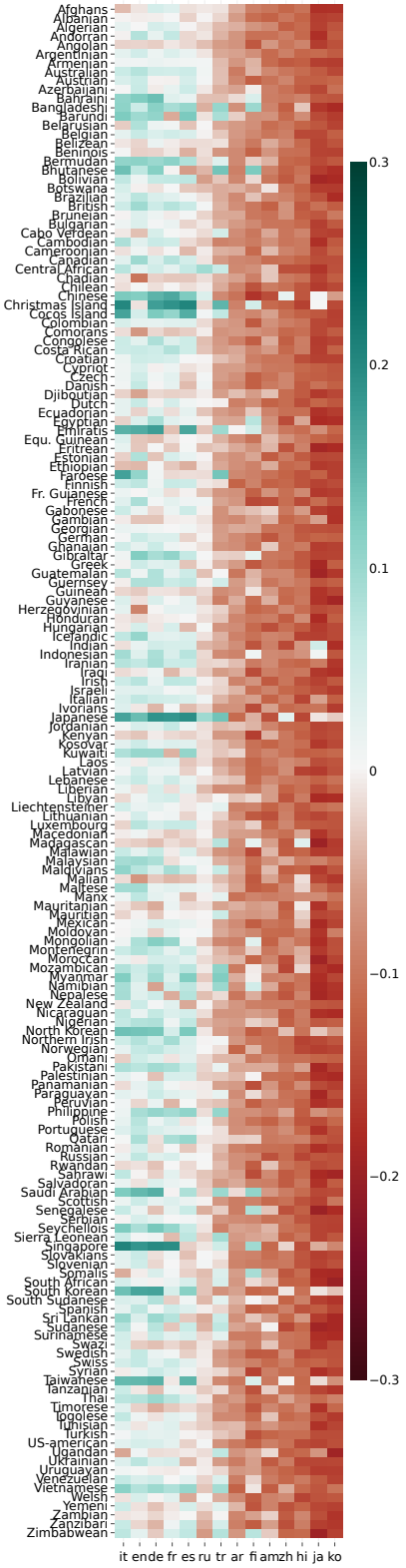


Figure 14: Heatmap of the SD3 SoS scores based on LAION embeddings.

C.2 SoS score validation results with SIGLIP

Using the image embedding model SIGLIP SIGLIP-SO400M-PATCH14-384 to embed the images, we obtain smaller SoS score values but see the same patterns per language and model. This proves the robustness of our findings.

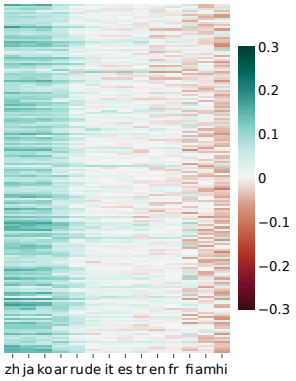


Figure 15: Heatmap of the AD SoS scores based on SIGLIP embeddings.

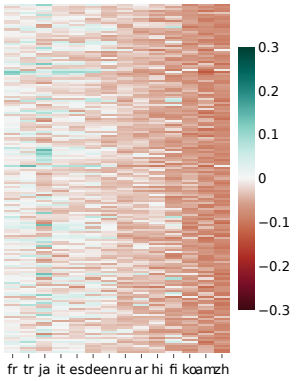


Figure 16: Heatmap of the K21 SoS scores based on SIGLIP embeddings.

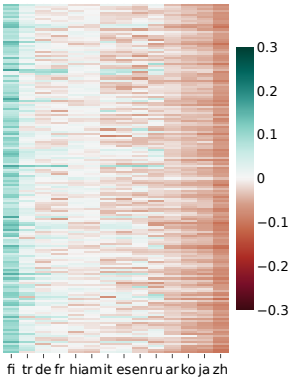


Figure 17: Heatmap of the K3 SoS scores based on SIGLIP embeddings.

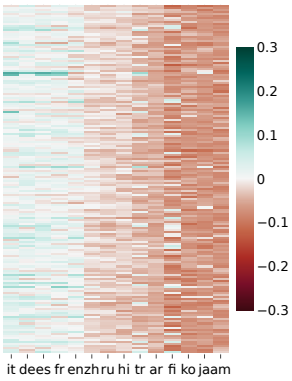


Figure 18: Heatmap of the SD21 SoS scores based on SIGLIP embeddings.

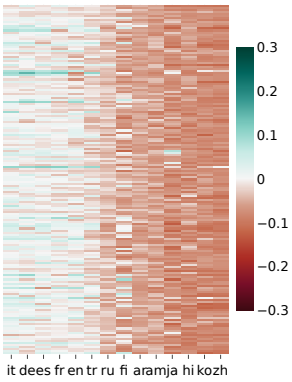


Figure 19: Heatmap of the SXL SoS scores based on SIGLIP embeddings.

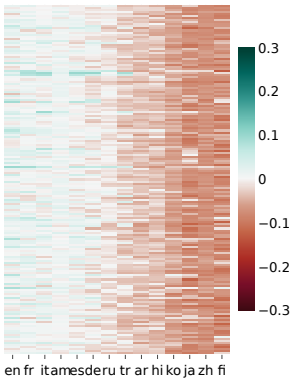


Figure 20: Heatmap of the FX SoS scores based on SIGLIP embeddings.

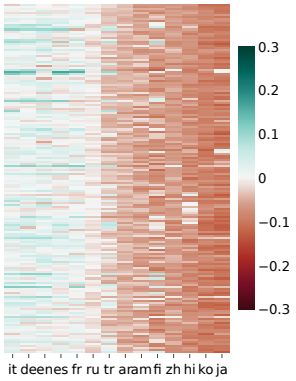


Figure 21: Heatmap of the SD3 SoS score based on SIGLIP embeddings.

C.3 SoS score validation results with DINO

Even though we are aiming to eliminate language biases by purely relying on the visual encodings of the CLIP model, without using any textual input to compare against, the CLIP model is inherently biased due to its training on a combination of image and text with predominantly English textual inputs. To account for this, we validate our results using the DINO image encoder (Oquab et al., 2024), which was trained without textual supervision. We present all results created using the `facebook/dinov2-with-registers-giant` model below:

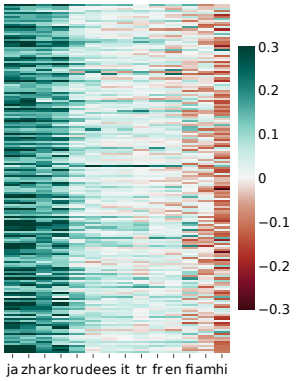


Figure 22: Heatmap of the AD SoS scores based on DINO embeddings.

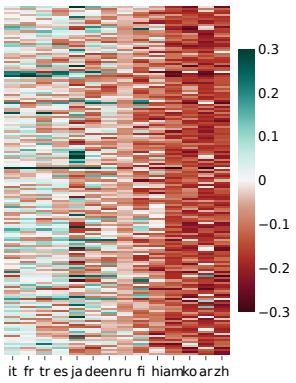


Figure 23: Heatmap of the K21 SoS scores based on DINO embeddings.

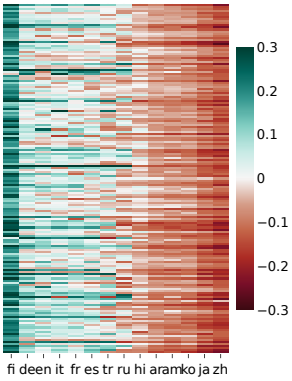


Figure 24: Heatmap of the K3 SoS scores based on DINO embeddings.

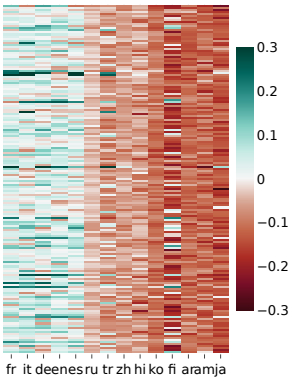


Figure 25: Heatmap of the SD21 SoS scores based on DINO embeddings.

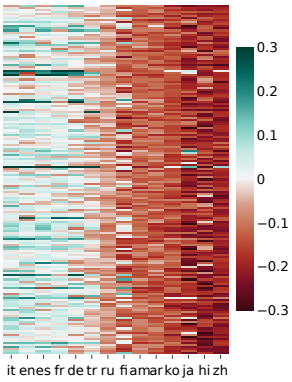


Figure 26: Heatmap of the SDXL SoS scores based on DINO embeddings.

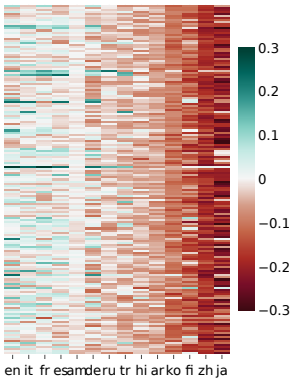


Figure 27: Heatmap of the FX SoS scores based on DINO embeddings.

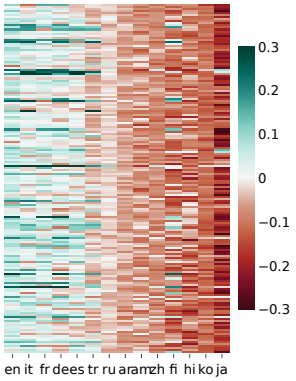


Figure 28: Heatmap of the SD3 SoS scores based on DINO embeddings.

C.4 Analysis of SoS score differences per person term

Gender-related biases in T2I models vary across languages, with distinct regional patterns affecting SoS tendencies. To analyze the intersectional bias between gender and culture and find out whether the SoS tendency is more salient for a certain gender, we aggregate the SoS score per gender across all models. Figure 29 shows the mean SoS score per language and gender with confidence intervals. For some cultures, such as Amharic, Arabic, and Korean, we observe no significant differences in SoS scores across gender identities. However, in other cultures, notable patterns emerge. Our aggregation shows that for Chinese, Japanese, and Hindi (i.e., Asian languages), images associated with female identities exhibit a significantly more negative SoS score compared to male and neutral identities, indicating a stronger reliance on surface-level linguistic features. Conversely, for Russian, Turkish, and Finnish (i.e., Eastern European languages), we observe the opposite effect: female identities are less biased toward the surface form of the prompt compared to their male and neutral counterparts. These findings suggest that gender-related biases in T2I models are not uniform across languages but instead follow distinct regional patterns, reinforcing the need for a more nuanced evaluation of intersectional biases in T2I models.

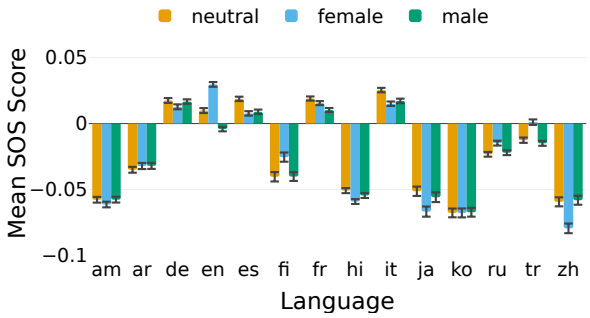


Figure 29: SoS score differences across genders. We present the SoS score mean along with the confidence interval for each gender and language, averaged over all models.

D Robustness Analysis

In this section, we present additional experiments that further support the robustness of our findings. Specifically, we demonstrate that our main results are robust to variations in terms of variations due to prompt formulation and generalize beyond a single concept. To this end, we (1) report additional SoS results for a second concept (*house*), which exhibit high correlation with the results obtained for the primary concept (*person*);(2) assess consistency across different prompt formulations; (3) provide per-language performance comparisons between SoS and CLIPScore, revealing substantial gains for multiple languages; and (4) report additional results using mCLIP, which, while improving over CLIPScore, does not match the performance of SoS.

D.1 Concept Generalization

To demonstrate the robustness of our findings and the effectiveness of the SoS-score across different concepts, we extend our analysis to another culturally specific concept, i.e., the concept "house". Using a prompt template analogous to the main results – "A photo of a house a c_i person lives in" – with three paraphrased formulations per prompt, we compute the corresponding SoS-scores across a subset of languages. For the concept house, these scores show a very strong alignment with those obtained for the concept "person", yielding a Pearson correlation coefficient of 87.6%. Figure 30 presents the detailed SoS-scores across cultures, languages, and models as heatmaps.

D.2 Prompt Formulation

To assess the robustness of SoS with respect to prompt formulation, we compute the mean absolute deviation (MAD) and Pearson correlation coefficient (PCC) between the SoS score distributions obtained from the three different prompt formulations (a, b, and c), which we report in Table 10. The high correlations and low deviations indicate that our SoS results are stable across these prompt variants.

While our prompts are synthetically constructed and may not fully reflect real-world user inputs, this design choice is deliberate, as it minimizes variation across languages and allows observed differences to be more reliably attributed to language–culture effects rather than to prompt noise. However, given the high correlation between SoS score distributions, we do not expect substantial variation under more diverse prompt formulations, though future work could explore more naturalistic prompts drawn from real usage scenarios.

Prompt Formulation	a & b	a & c	b & c
MAD	0.0245	0.0250	0.0272
PCC	0.9020	0.9040	0.8830

Table 10: Variation across the three prompt formulations (a, b, and c). We report the mean absolute deviation (MAD) and the Pearson correlation coefficient (PCC) between the resulting SoS score distributions for images generated by each prompt formulation.

Language	SOS	CLIPScore
arabic	**0.78**	0.73
chinese	**0.96**	**0.96**
english	0.51	**0.86**
finnish	0.75	**0.81**
french	**0.53**	0.49
german	0.76	**0.80**
hindi	**0.67**	0.56
italian	**0.74**	0.70
japanese	**0.85**	0.74
korean	**0.67**	**0.67**
russian	**0.81**	0.78
spanish	**1.0**	**1.0**
turkish	**0.96**	0.87

Table 11: Proportion of correctly categorized surface and semantic-level tendencies using SoS and CLIPScore.

D.3 Comparison of SoS and CLIPScore across languages

The overall improvements on accuracy of SoS over the ClipScore are modest for non-English languages. However, our primary objective extends beyond raw accuracy to the precise identification of surface-level and semantic-level tendencies, where SoS exhibits clearer and more consistent gains. To additionally support the results presented in Figure 2, we provide a detailed per-language comparison of classification accuracy for both approaches in Table 11. In particular, SoS substantially outperforms CLIPScore for Japanese, Turkish, and Hindi, while achieving comparable performance for several other languages.

D.4 Comparison of SoS and CLIPScore based on mCLIP

To provide evidence that the limited performance of the CLIPScore on assessing semantic and surface-level tendencies is not due to its multilingual capa-

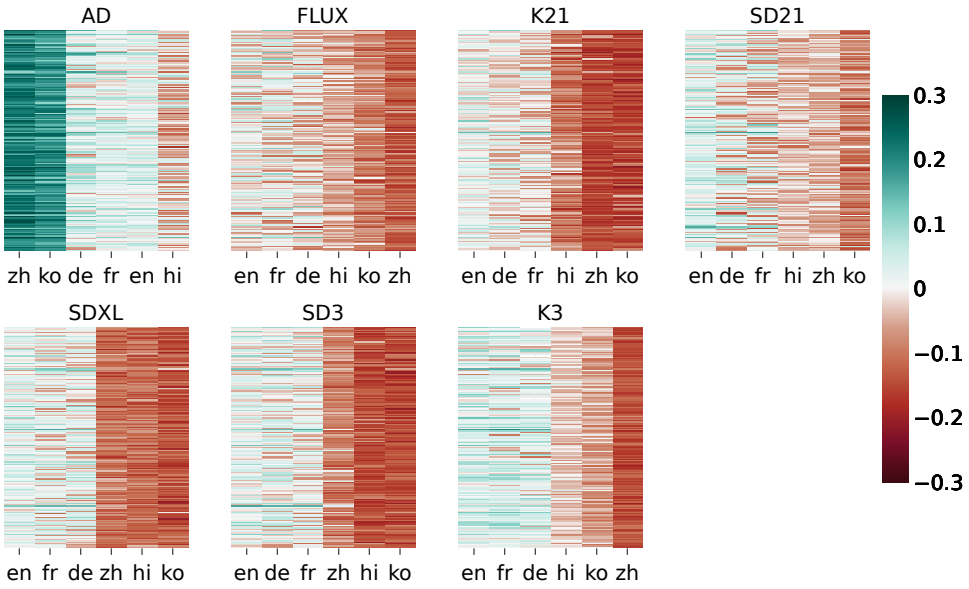


Figure 30: Heatmap of SoS-scores based on LAION embeddings for the concept “house”.

	Acc	$\mathbf{P_{sur}}$	$\mathbf{P_{sem}}$
mCLIP	77.6%	88.7%	84.8%
SoS score	74.0%	94.8%	84.8%
mCLIP \en	**75.9**	88.7	92.2
SoS \ en	**79.1**%	**94.8**%	94.8%

Table 12: Comparison of mCLIP-based scores with SoS scores on the subset of human labelled instances.

bilities, we additionally report results obtained with mCLIP (Chen et al., 2023) in Table 12. We find that while mCLIP slightly improves performance for non-English languages compared to CLIP, it does not surpass the best results achieved by SoS.

E Linguistic Similarities

E.1 Image Embedding Distribution

We examine the distribution of images created for different input languages when embedded in 2D. To this end, we perform a principal component analysis (PCA) (Maćkiewicz and Ratajczak, 1993) with a random seed of 42, reducing to 2 components.

European Languages Cluster together for all models. Also, multilingual models exhibit language clusters. Two things have already become very clear here. In contrast to all other models, AltDiffusion shows a more diverse distribution of images according to the prompted cultures and no clustering. Despite this, smaller clusters can be recognized for languages such as Amharic and Finnish, but they are not quite as clearly delineated. The plot of SD3, on the other hand, paints a very different picture. While the higher resource European languages such as English, Italian, Spanish, German, and French are also nicely distributed across each other in a cross-lingual space, you can see clearly separated clusters for some languages such as Arabic, Amharic, Hindi, and Turkish. Interestingly, you can even see that the languages Japanese, Korean and Chinese cluster on top of each other in a separate cluster, which already suggests a greater similarity of these images to each other in contrast to the other images. We also see similar patterns for the other non-explicitly multilingually trained models.

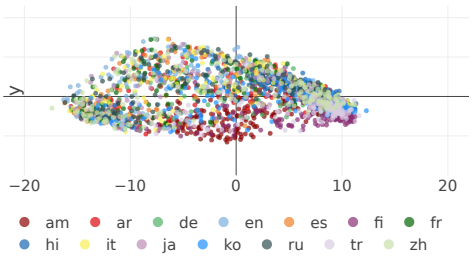


Figure 31: Distribution of image embeddings for images generated by AD.

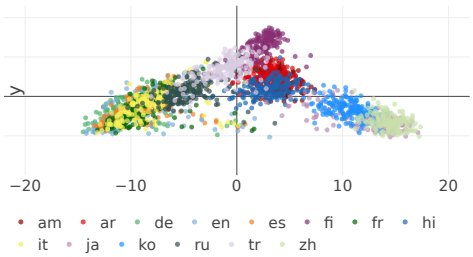


Figure 32: Distribution of image embeddings for images generated by FX.

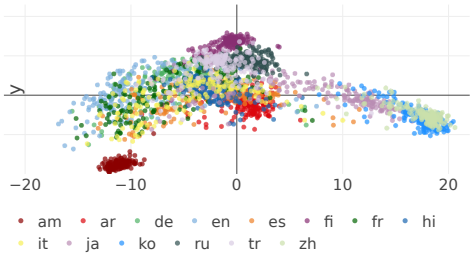


Figure 33: Distribution of image embeddings for images generated by SD21.

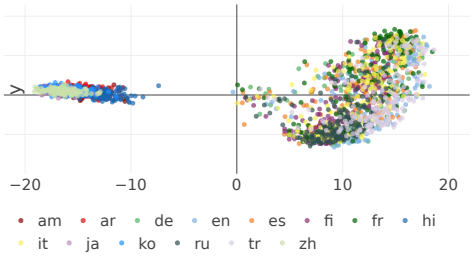


Figure 34: Distribution of image embeddings for images generated by K3.

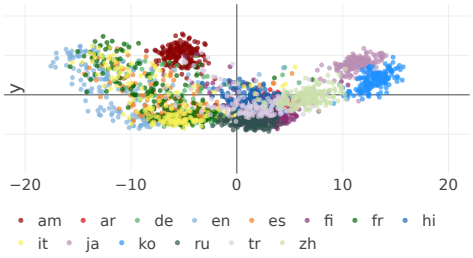


Figure 35: Distribution of image embeddings for images generated by SD21.

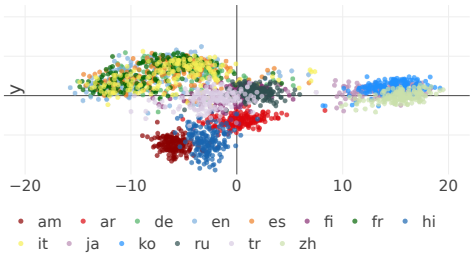


Figure 36: Distribution of image embeddings for images generated by SDXL.

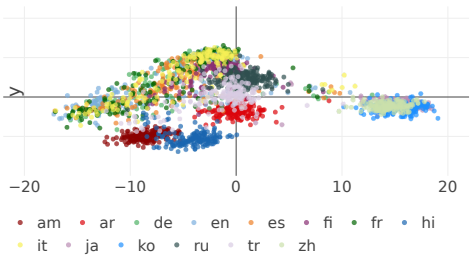


Figure 37: Distribution of image embeddings for images generated by SD3.

E.2 Language Correlations

We present the correlation analysis for all language pairs for each of the evaluated models.

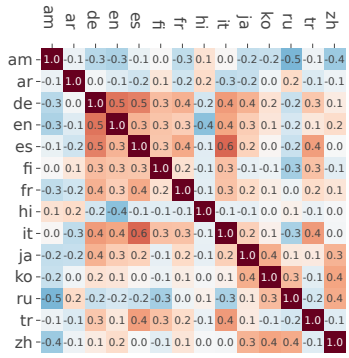


Figure 38: Per language Correlation of the images obtained by prompting SD21.

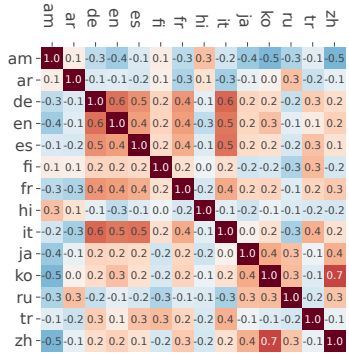


Figure 39: Per language Correlation of the images obtained by prompting SDXL.

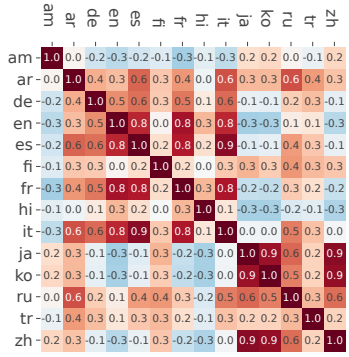


Figure 40: Per language Correlation of the images obtained by prompting AD.

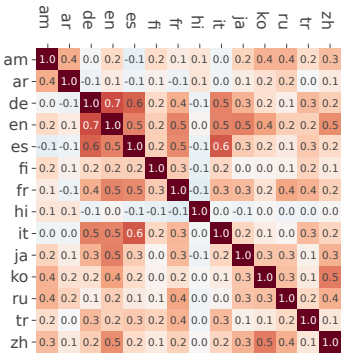


Figure 41: Per language Correlation of the images obtained by prompting FX.

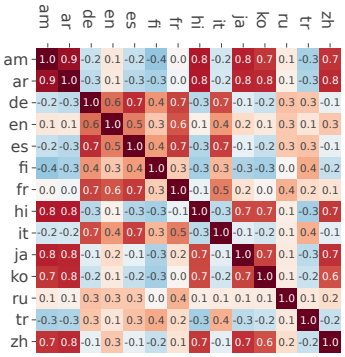


Figure 42: Per language Correlation of the images obtained by prompting K3.

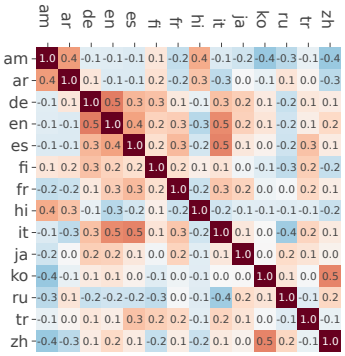


Figure 43: Per language Correlation of the images obtained by prompting K21.

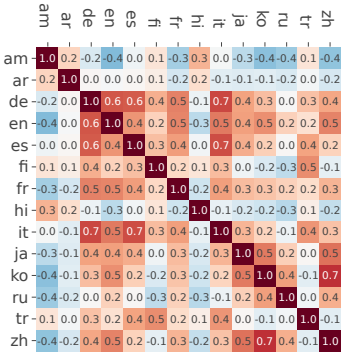


Figure 44: Per language Correlation of the images obtained by prompting SD3.

F Colour Analysis

In this section, we provide additional results on the color analysis. First, we provide the full results for the cultural identity group *german* across languages and models. Second, we provide more examples across several cultural identities for images when prompting in German and images when prompting in Chinese. Finally, we provide distribution plots of the 'value' in the colors HSV codes. The value represents the brightness of a color, ranging from 0 (complete darkness) to 1 (full brightness). We observe notable differences in brightness distribution across models. Specifically, images generated for European languages tend to feature darker tones compared to those produced for languages like Chinese and Hindi.

F.1 Full results for one cultural identity

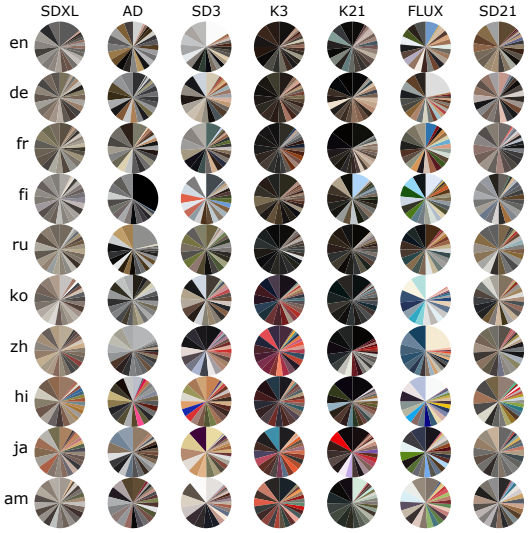


Figure 45: Most prominent colors across languages for the cultural identity *German*.

F.2 Results for multiple cultural identities for prompts in German and Chinese

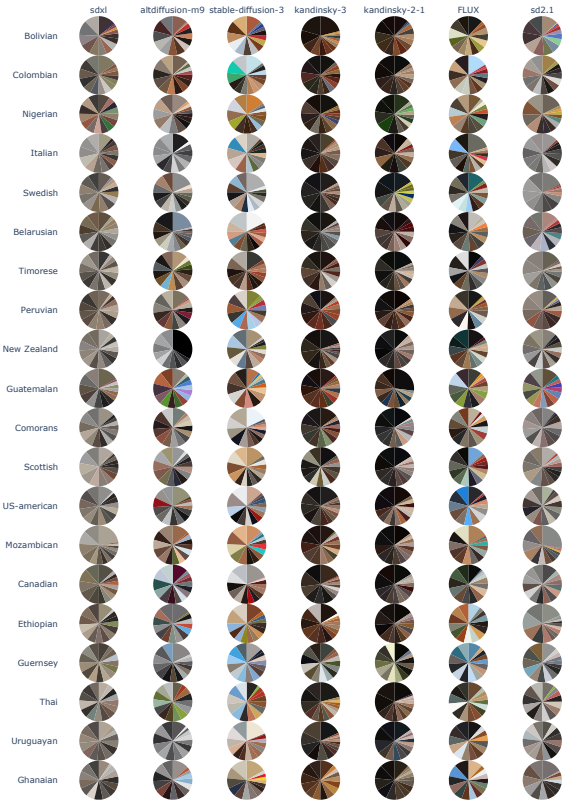


Figure 46: Prominent colors across 20 random Cultures when prompting in *German*.

Figure 47: Prominent colors across 20 random Cultures when prompting in *Chinese*.

F.3 Distribution of HSV-Values

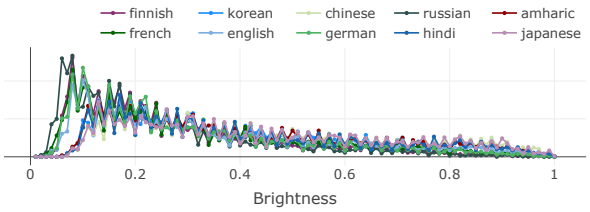


Figure 48: Value distribution of HSV values across the 8 most prominent color clusters across images for different languages obtained by prompting K3.

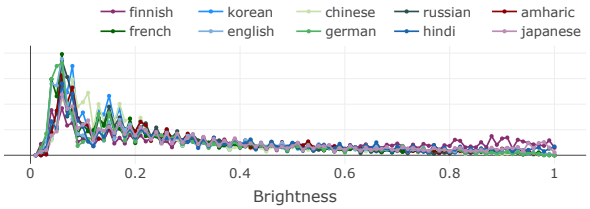


Figure 49: Value distribution of HSV values across the 8 most prominent color clusters across images for different languages obtained by prompting K21.

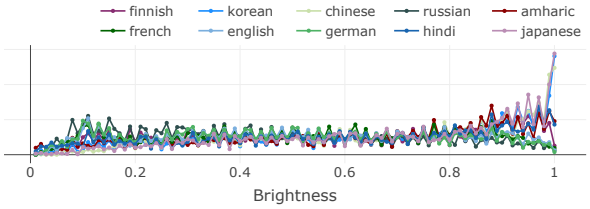


Figure 50: Value distribution of HSV values across the 8 most prominent color clusters across images for different languages obtained by prompting FX.

G VQA Analysis

G.1 VQA Analysis using Fighting Words Method

Category	Terms
Generic image describing terms	item, product, object, hardship, abandonment, character, couple, text, word, font, title, subject, depiction, complexity, texture, photograph, uneven, person, contours, picturesque, population, highrise, collage, room, portrait, individual, illustration, figure, abstract, forehead, mood, outfit, hair, face, wall, pose, shoulder, people, shirt, head, wrap, group, pack, mute
Quantitative and temporal terms	20th, 19th, century, archival, digital, historical
Generic adjectives	richly, tall, dynamic, dapple, official, blackandwhite, updo, minimalist, welllit, scenic, simple, chaotic, fantastical, bright, vibrant, grand, unsettling, classical, clear, long, short, intense, densely, rough, neutral, plain, undisturbed, gridlike, casual, peaceful, tightly, contemplative, sparse, richness, heavy, overcast, dark, soft, brown, detailed, traditional
Directional and relational terms	north, south, east, underneath, outermost, central, directly, subject, title, fourth, second
Generic verbs	elaborate, crash, wear, stand, enchanting, help, desolate, use, neglect, highlight, stamp
Pronouns and person identifier	man, mans

Table 13: List of filtered terms that were not validated by human annotators due to lack of visual appearance.

G.2 VQA Analysis Results

We present the results of the VQA analysis for all language-model combinations that exhibited surface-level tension. We further present example images for each of them.

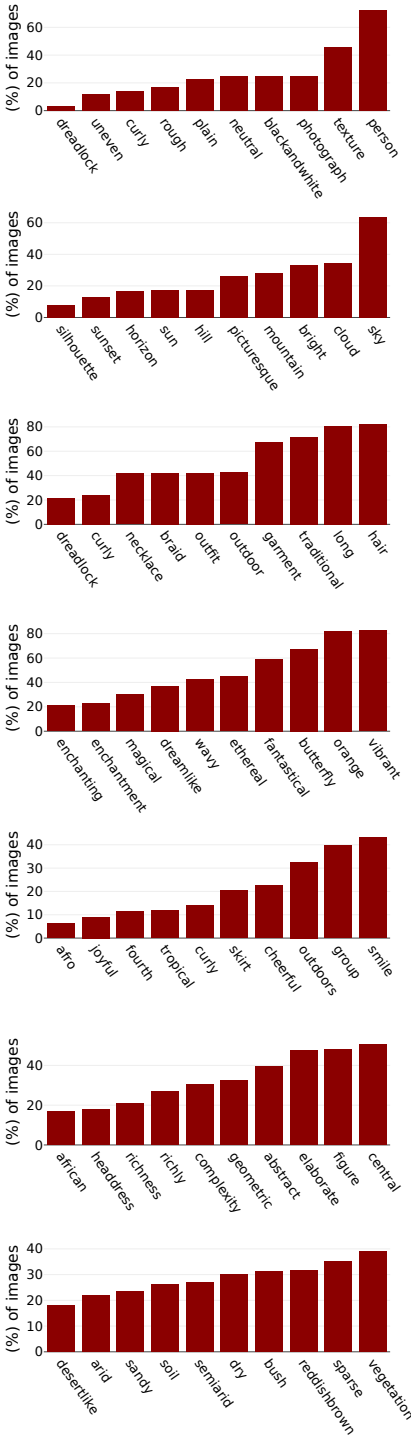


Figure 51: VQA results: We present the top 15 terms for images generated using input prompts in Amharic.

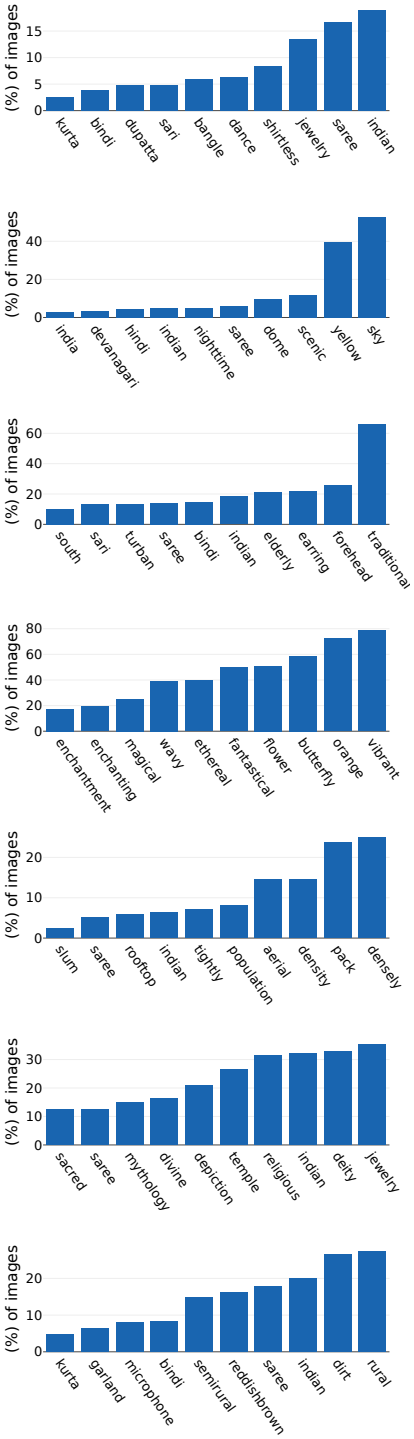


Figure 52: VQA results: We present the top 15 terms for images generated using input prompts in Hindi.

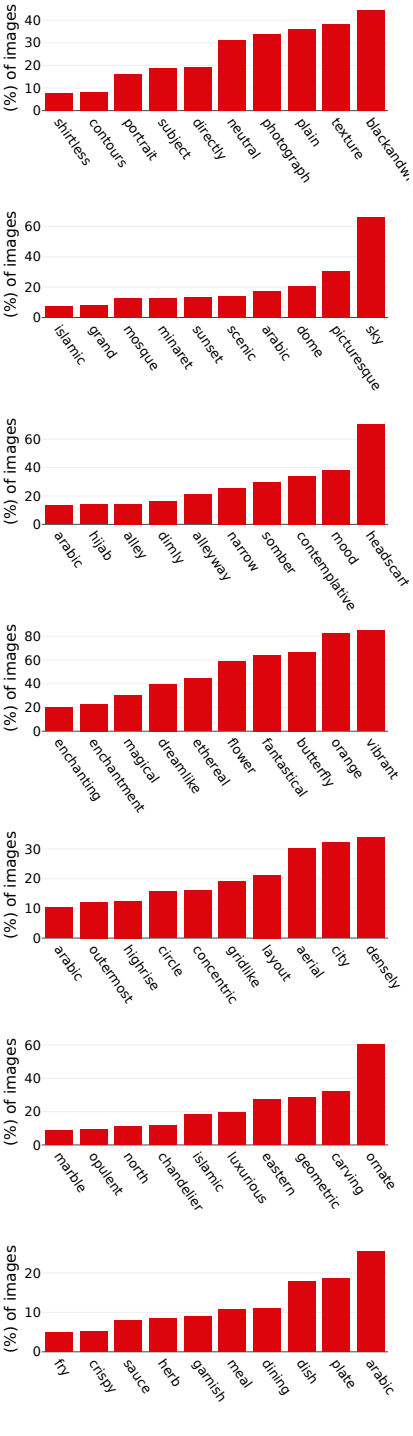


Figure 53: VQA results: We present the top 15 terms for images generated using input prompts in Arabic.

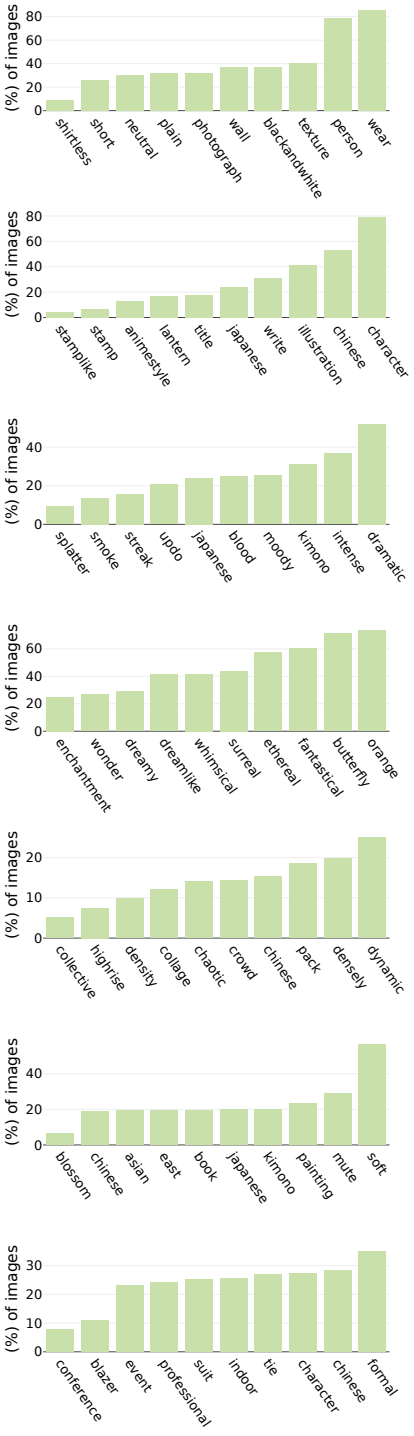


Figure 54: VQA results: We present the top 15 terms for images generated using input prompts in Chinese.

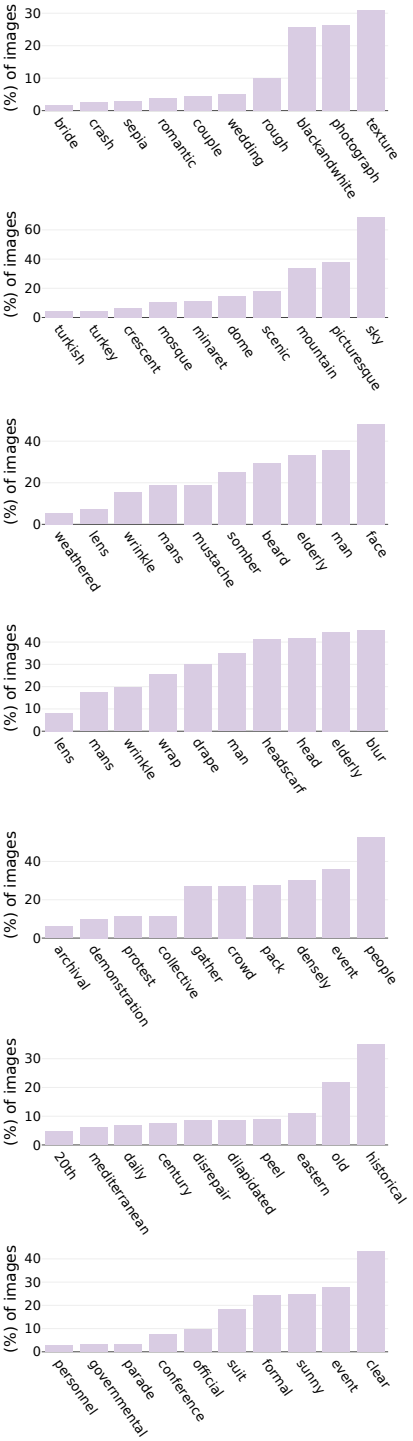


Figure 55: VQA results: We present the top 15 terms for images generated using input prompts in Turkish.

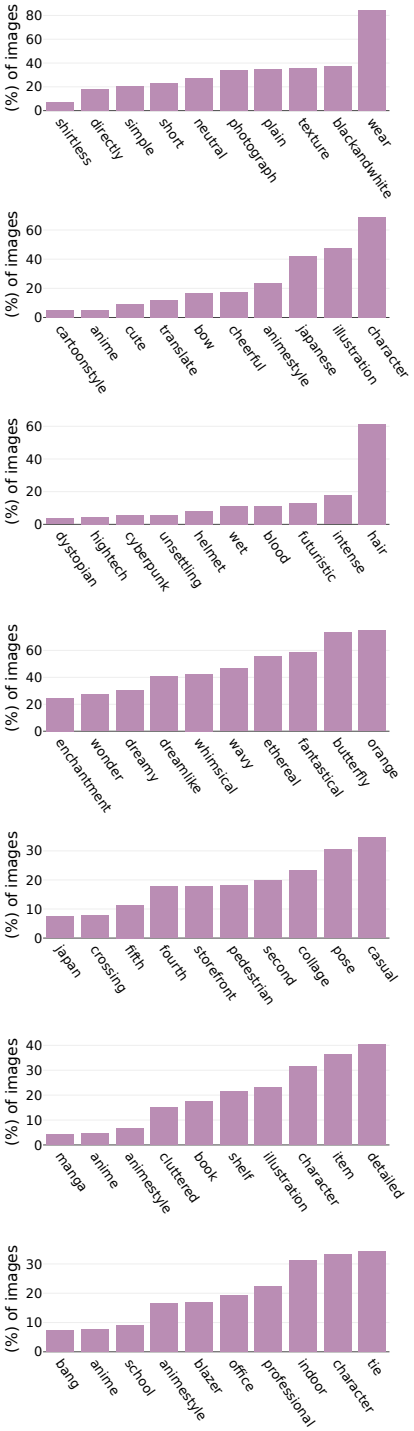


Figure 56: VQA results: We present the top 15 terms for images generated using input prompts in Japanese.

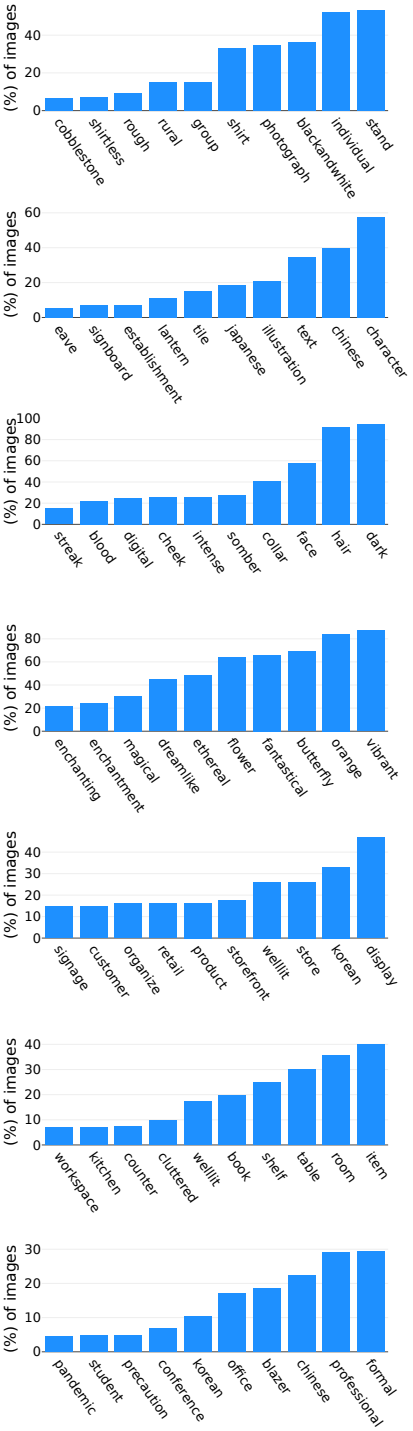


Figure 57: VQA results: We present the top 15 terms for images generated using input prompts in Korean.

G.3 Validation Correlation of SeeGULL Terms

Lang.	Associated Culture(s)	#	Visual Stereotypes
zh	Chinese	22	red, strong, china, yellow, rice, tire, bright, clear, angry, thin, short, blonde, tea, girl, beautiful, brunette, chinese, untidy, left, drug, dirty
fi	Finnish	1	sauna
ko	North Korean, South Korean	6	athletic, taekwondo, thin, kimchi, yellow, short
ja	Japanese	21	attractive, sushi, sad, dog, rice, sex, tire, sick, tsunami, japanese, samurai, thin, short, car, monkey, nervous, robot, ninja, organize, white, fish
hi	Indian	35	athletic, cow, attract, sad, watch, thick, red, unorganized, light, cricket, hockey, violent, elephant, open, snake, brown, tall, black, indian, angry, dark, organized, thin, old, beautiful, sweaty, nervous, blonde, curry, coast, untidy, vegetarian, white, colorful, dirty
ru	Russian	9	sauce, blonde, snow, vodka, russian, drinking, malnourished, strong, unwashed
ar	Bahraini, Egyptian, Jordanian, Kuwaiti, Lebanese, Libyan, Mauritanian, Omani, Palestinian, Qatari, Saudi Arabian, Syrian, Tunisian, Emiratis, Yemeni	14	kidnapping, oil, dark, malnourished, violent, unorganized, pyramid, tall, hairy, sexy, barren, wet, dirty, beautiful
tr	Turkish	8	greasy, angry, fiery, violent, untidy, organized, tall, sexy
am	Ethiopian	26	attractive, zoo, thick, torture, strong, die, coffee, bald, tall, sick, black, child, malnourished, dark, poor, underweight, thin, village, short, cannibal, blonde, poverty, undernourished, malnutrition, scrawny, dirty

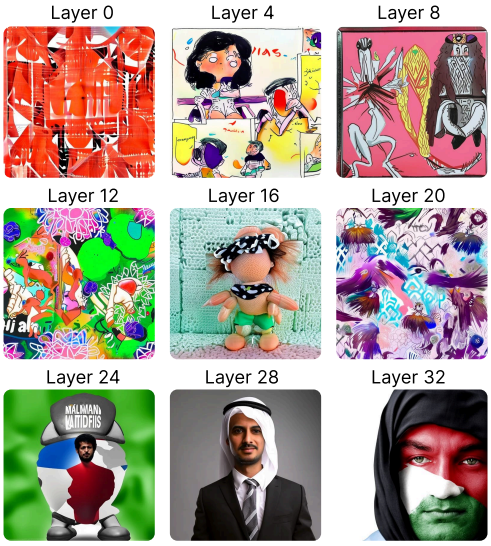
Table 14: Visual stereotypes of the SeeGULL dataset. We merge all cultural stereotypes of countries where the language is the official language, remove multi-token examples and non visually depictable adjectives. Highlighted red terms are terms that are never matched within our analysis.

H Example Images

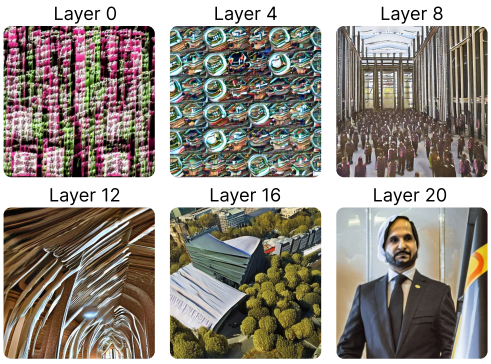
In this section, we present example images generated by the T2I models. The first subsection shows images generated from different text-encoder representations of an input prompt. The second subsection presents examples spanning multiple languages, models, and prompted cultural identities.

H.1 Examples of layer-wise images

One limitation we set out to address with SoS, relative to CLIPScore, is the reliance on a textual anchor. Text-based matching presupposes both high-quality images and faithful realization of the captioned concepts. These assumptions break down for generations guided by early layers of the text encoder, which often yield low-quality outputs or random visual noise. In contrast, SoS operates directly on images without requiring a caption anchor. To illustrate the issue, we show two example generations produced from different layers of the text encoder for the concept "Baharini man". The early layer outputs lack coherent semantics, so image-to-caption alignment is uninformative, whereas SoS remains applicable.



(a) Per layer images generated by K3.



(b) Per layer images generated by SD21.

Figure 58: Example images generated from successive text-encoder layers by K3 (a) and SD21 (b) for the prompt "Baharini man". For both T2I models, early-layer representations exhibit low visual quality and weak semantic coherence.

H.2 Example images generated by T2I models

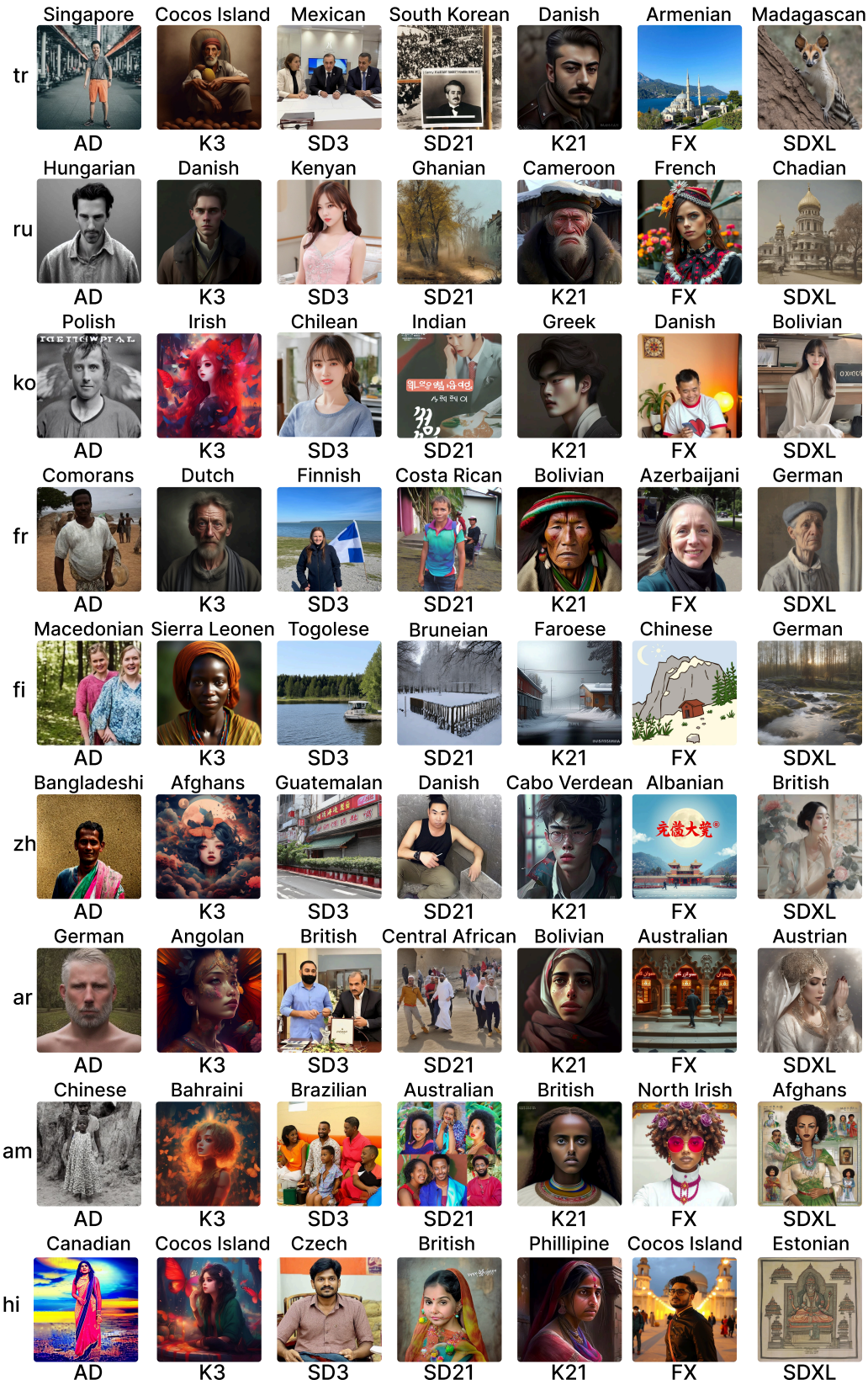


Figure 59: Example images across languages and models. Above each image, we indicate the target cultural identity to be generated; below each image, the T2I model used; and on the left, the language code of the input prompt.