Non-Archival Abstract

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

Carolin Holtermann[1], Florian Schneider[2],
[1]Trustworthy AI Lab, University of Hamburg
[2]Cohere
Correspondence: carolin.holtermann@uni-hamburg.de

Note: This extended abstract describes work published in (Holtermann et al., 2026).

Introduction Text-to-image (T2I) models are increasingly adopted by users worldwide, yet most are trained primarily on English data. Since only approximately 20% of the global population speaks English fluently[1], a substantial share of users may instead interact with these systems in their native language. While semantically equivalent prompts should ideally yield comparable outputs regardless of the input language, prior work has shown that T2I models are sensitive to the surface form of a prompt, meaning its language and script (Ventura et al., 2025; Struppek et al., 2024), rather than to its underlying semantic content. This tension, which we term Surface over Semantics (SoS), has so far resisted systematic quantification, limiting progress toward culturally fair and globally inclusive generative systems. We present the first comprehensive analysis of SoS tendencies in multilingual T2I generation, with three core contributions.

Dataset We construct a multilingual prompt dataset covering 171 cultural identities and 14 languages drawn from typologically diverse families, including Indo-European, Afro-Asiatic, Sino-Tibetan, Turkic, Koreanic, Japonic, and Uralic. Prompts in all languages other than English are validated by native speakers. Using this dataset, we generate images with seven state of the art T2I models, encompassing both multilingual or bilingual models and systems primarily trained in English.

SoS Score We introduce a novel evaluation measure that quantifies the tension between surface and semantic tendencies:

$$\mathrm{SoS}_{c,l}(o_{c,l}) = \cos\big(\overrightarrow{\mathrm{sem}}_c,\, \vec{e}_o\big) - \cos\big(\overrightarrow{\mathrm{sur}}_l,\, \vec{e}_o\big)\,,$$

comparing each generated image embedding $\vec{e}_o$ against averaged semantic ($\overrightarrow{\mathrm{sem}}_c$) and surface ($\overrightarrow{\mathrm{sur}}_l$) reference embeddings. Operating solely on images avoids language specific confounds and increases robustness to low quality generations over CLIPScore (Hessel et al., 2021), achieving 94.8% precision for surface level tendency detection and outperforming CLIPScore for several non-European languages, including Japanese, Turkish, and Hindi.

Analysis Applying our measure, we find that all but one of the evaluated models exhibit strong surface level tendencies in at least two languages. An analysis across text encoder layers using DiffusionLens (Toker et al., 2024) reveals that these tendencies intensify in the upper layers, suggesting limited multilingual training leads models to default to surface cues. A correlation analysis across language pairs further shows that SoS patterns cluster by script family, with Latin script European languages displaying strong correlation, while Amharic and Arabic emerge as outliers. Complementary Visual Question Answering analyses, conducted using the Fighting Words method (Monroe et al., 2017), reveal how surface biases manifest visually: languages with strong surface tendencies frequently elicit stereotypical depictions, including depictions of *saree* and *bindi* for Hindi, *snowy forest* scenes for Finnish. Coverage of SeeGULL (Jha et al., 2023) visual stereotypes reaches up to 61.9% for certain language and model combinations.

Conclusion Our findings show that users prompting T2I models in different languages may receive biased or stereotypical outputs, with this risk distributed unevenly across linguistic communities. We argue that the SoS score can serve as a meaningful signal to guide model training toward fairer multilingual generation, while acknowledging that the appropriate balance between surface and semantic grounding remains context dependent.

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

Acknowledgements The work of Carolin Holtermann is funded by the Excellence Strategy of the German Federal Government and the Federal States. The work of Florian Schneider was done while he was with the Language Technology Group at the University of Hamburg.

References

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.

Carolin Holtermann, Florian Schneider, and Anne Lauscher. 2026. SoS: Analysis of surface over semantics in multilingual text-to-image generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3955–3995, Rabat, Morocco. 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.

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.

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.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2025. Navigating cultural chasms: Exploring and unlocking the cultural POV of text-to-image models. *Transactions of the Association for Computational Linguistics*, 13:142–166.