STRICT: Stress-Test of Rendering Image Containing Text

Tianyu Zhang[1*], Xinyu Wang[2*], Lu Li[3*], Zhenghan Tai[4], Jijun Chi[4], Jingrui Tian[5], Hailin He[6], Suyuchen Wang[1]

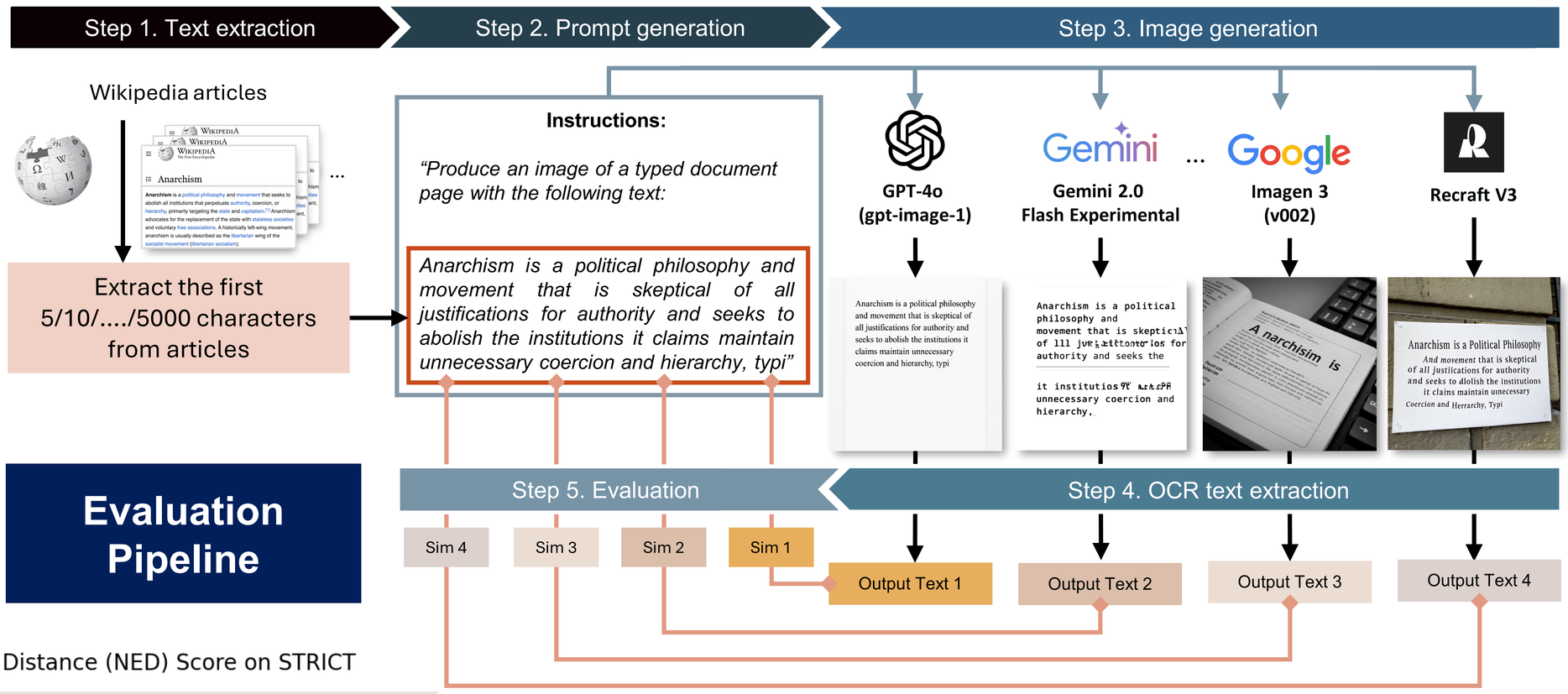
[1] Mila, University of Montreal, [2] McGill University, [3] University of Pennsylvania, [4] University of Toronto, [5] University of California, Los Angeles, [6] Southwestern University of Finance and Economics *Equal contribution

EMNLP 2025 Suzhou, China 中国苏州

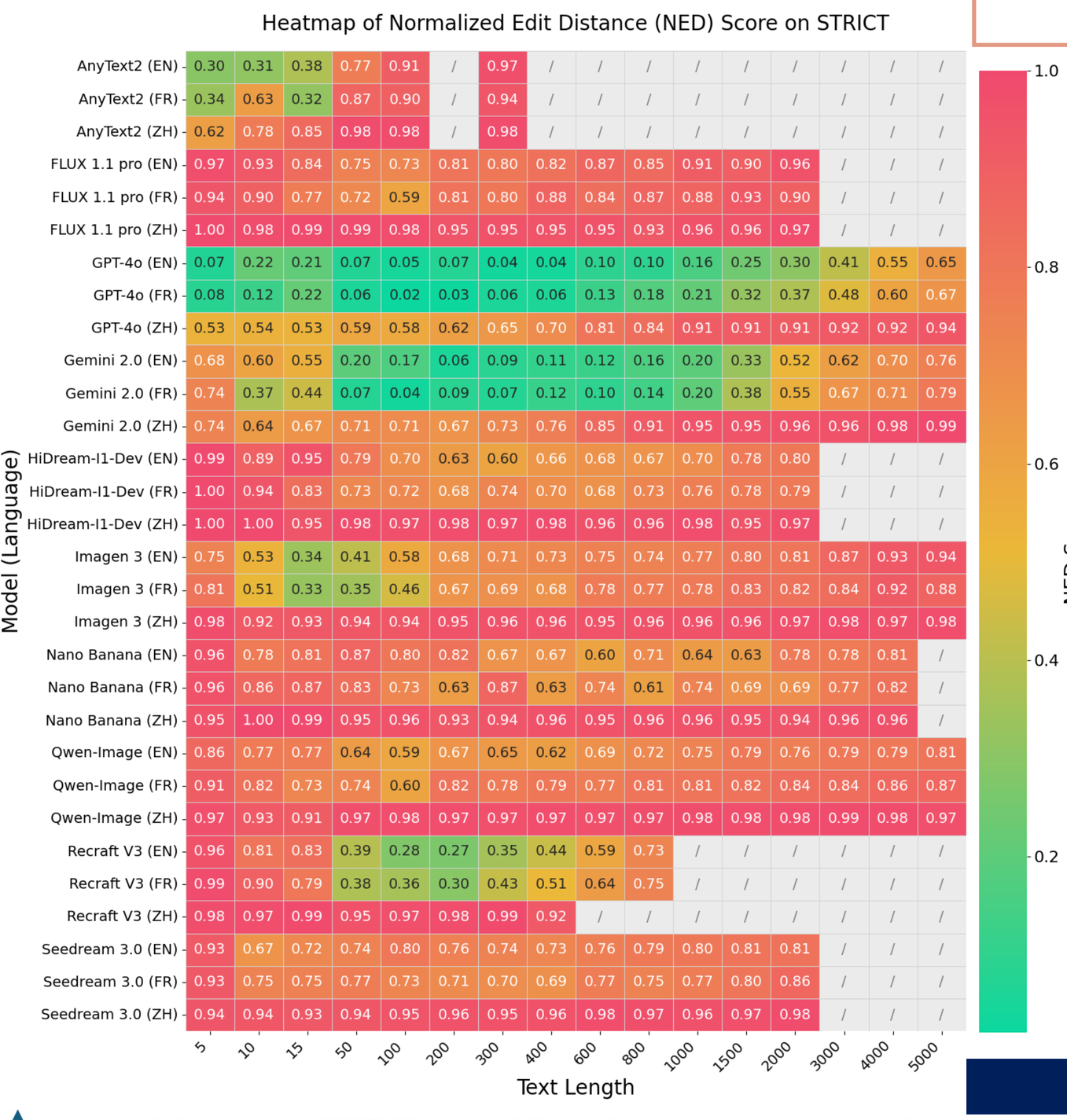
Motivation

- Diffusion models excel at realistic image synthesis
- Still fail to generate **legible, consistent, and instruction-aligned text**
- Root cause: **locality bias** → weak long-range spatial consistency
- Need a systematic test for text fidelity and instruction following

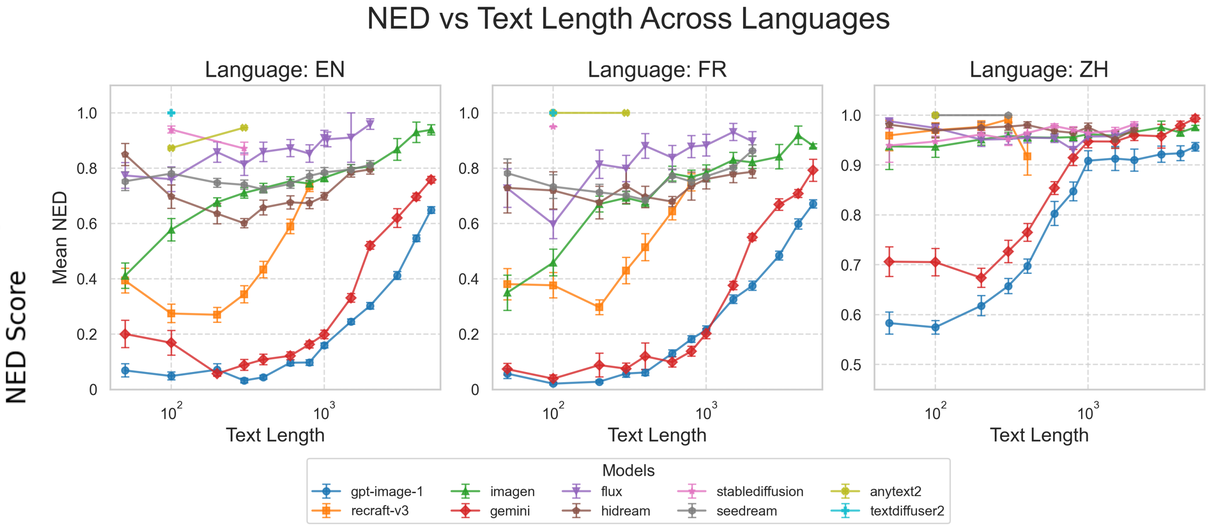
Evaluation Pipeline



Goal: **Quantitatively assess** and expose model limitations in generating images containing text.

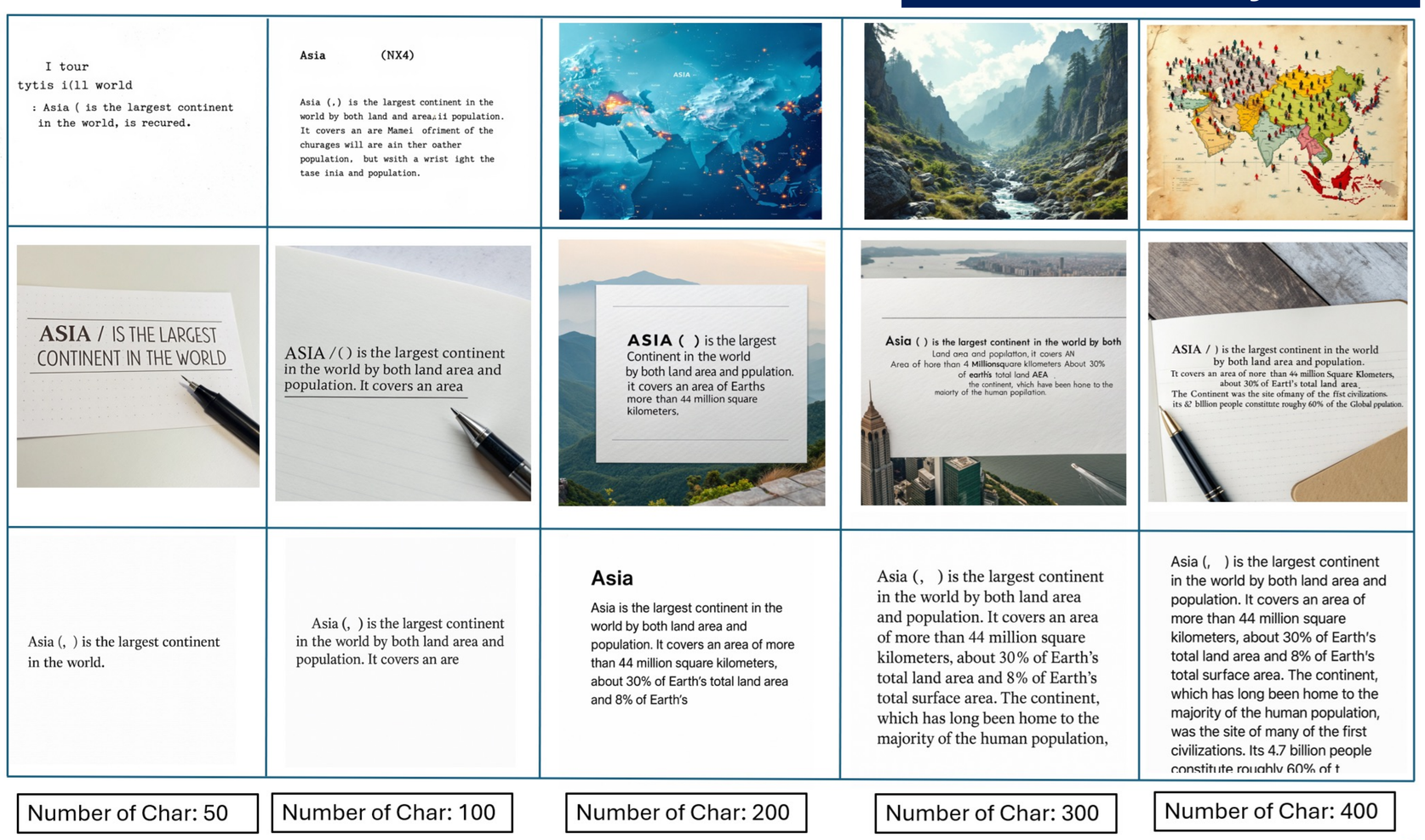


Evaluation Metrics & Results



- **Normalized Edit Distance (NED)** – character-level dissimilarity
- **Performance vs Text Length**
  - Accuracy drops **sharply beyond 200–300 characters**
  - GPT-4o maintains best NED (<0.3 up to 2K chars)
  - Open-source models degrade to **NED > 0.8** early
  - Long-text rendering remains a universal failure case

Case Study



Takeaways

- Diffusion models still **cannot render image containing texts reliably**
- Exposes **instruction-following and consistency gaps**
- **STRICT** provides a **flexible and controllable synthetic benchmark** to evaluate the abilities of models to generate images containing texts

More Info

Paper

Dataset

GitHub