## STRICT: Stress-Test of Rendering Image Containing Text

Tianyu Zhang<sup>1\*</sup>, Xinyu Wang<sup>2\*</sup>, Lu Li<sup>3\*</sup>, Zhenghan Tai<sup>4</sup>, Jijun Chi<sup>4</sup>, Jingrui Tian<sup>5</sup>, Hailin He<sup>6</sup>, Suyuchen Wang<sup>1</sup>

Imagen 3 (FR) - 0.81 | 0.51 | 0.33 | 0.35 | 0.46 | 0.67 | 0.69 | 0.68 | 0.78 | 0.77 | 0.78 | 0.83 | 0.82 | 0.84 | 0.92 | 0.88

Imagen 3 (ZH) - 0.98 | 0.92 | 0.93 | 0.94 | 0.94 | 0.95 | 0.96 | 0.96 | 0.95 | 0.96 | 0.96 | 0.96 | 0.97 | 0.98 | 0.97 | 0.98

Nano Banana (EN) - 0.96 | 0.78 | 0.81 | 0.87 | 0.80 | 0.82 | 0.67 | 0.67 | 0.60 | 0.71 | 0.64 | 0.63 | 0.78 | 0.78 | 0.81

Nano Banana (FR) - 0.96 | 0.86 | 0.87 | 0.83 | 0.73 | 0.63 | 0.87 | 0.63 | 0.74 | 0.61 | 0.74 | 0.69 | 0.69 | 0.77 | 0.82

Nano Banana (ZH) - 0.95 | 1.00 | 0.99 | 0.95 | 0.96 | 0.93 | 0.94 | 0.96 | 0.95 | 0.96 | 0.96 | 0.95 | 0.94 | 0.96 | 0.96

Recraft V3 (EN) - 0.96 0.81 0.83 0.39 0.28 0.27 0.35 0.44 0.59 0.73

Recraft V3 (FR) - 0.99 0.90 0.79 0.38 0.36 0.30 0.43 0.51 0.64 0.75

Recraft V3 (ZH) - 0.98 | 0.97 | 0.99 | 0.95 | 0.97 | 0.98 | 0.99 | 0.92 | / / /

Seedream 3.0 (EN) - 0.93 | 0.67 | 0.72 | 0.74 | 0.80 | 0.76 | 0.74 | 0.73 | 0.76 | 0.79 | 0.80 | 0.81 | 0.81

Seedream 3.0 (FR) - 0.93 | 0.75 | 0.75 | 0.77 | 0.73 | 0.71 | 0.70 | 0.69 | 0.77 | 0.75 | 0.77 | 0.80 | 0.86

Seedream 3.0 (ZH) - 0.94 | 0.94 | 0.93 | 0.94 | 0.95 | 0.96 | 0.95 | 0.96 | 0.98 | 0.97 | 0.96 | 0.97 | 0.98

Qwen-Image (EN) - 0.86 | 0.77 | 0.77 | 0.64 | 0.59 | 0.67 | 0.65 | 0.62 | 0.69 | 0.72 | 0.75 | 0.79 | 0.76 | 0.79 | 0.79 | 0.81

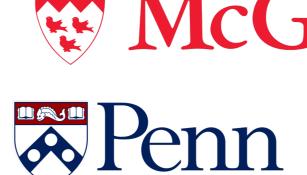
Qwen-Image (FR) - 0.91 | 0.82 | 0.73 | 0.74 | 0.60 | 0.82 | 0.78 | 0.79 | 0.77 | 0.81 | 0.81 | 0.82 | 0.84 | 0.84 | 0.86 | 0.87

Qwen-Image (ZH) - 0.97 | 0.93 | 0.91 | 0.97 | 0.98 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.98 | 0.98 | 0.98 | 0.99 | 0.98 | 0.97

Number of Char: 100

Number of Char: 50



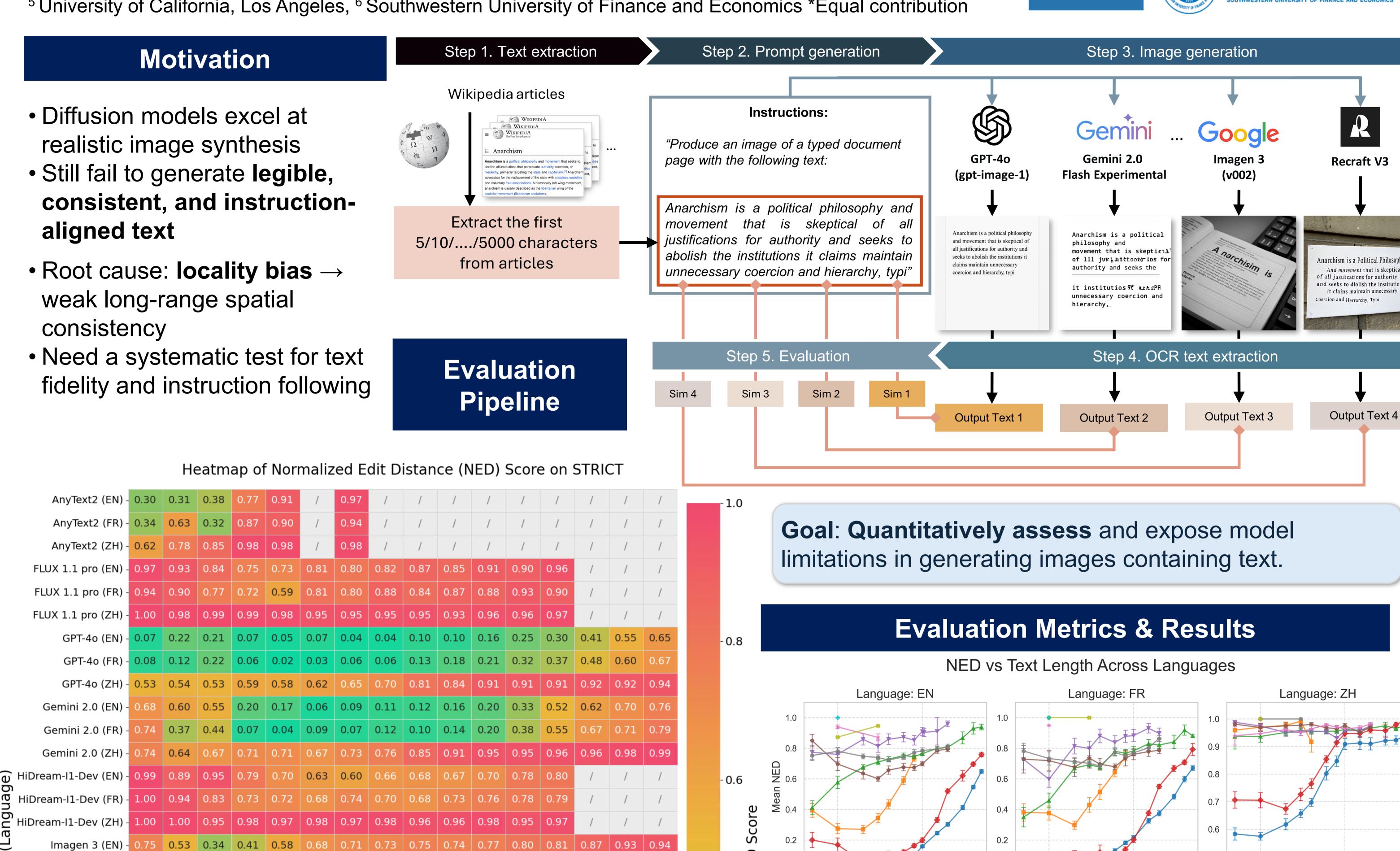




<sup>1</sup> Mila, University of Montreal, <sup>2</sup> McGill University, <sup>3</sup> University of Pennsylvania, <sup>4</sup> University of Toronto,

<sup>5</sup> University of California, Los Angeles, <sup>6</sup> Southwestern University of Finance and Economics \*Equal contribution





- 0.4

- 0.2

Number of Char: 300

Normalized Edit Distance (NED) – character-level dissimilarity

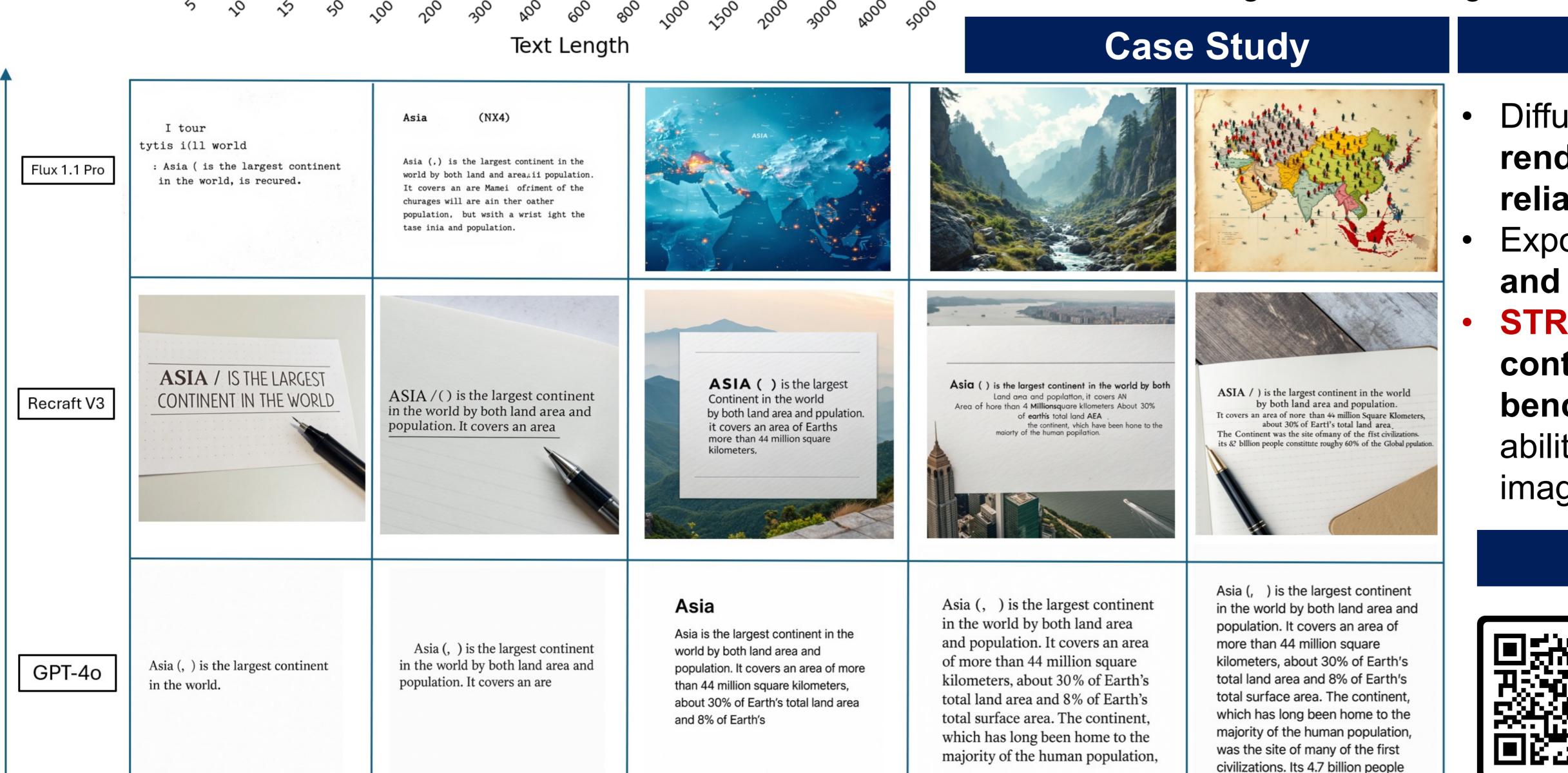
Text Length

Performance vs Text Length

constitute roughly 60% of t

Number of Char: 400

- Accuracy drops sharply beyond 200–300 characters
- GPT-40 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



Number of Char: 200

## **Takeaways**

Text Length

- 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





