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

Contact: tianyu.zhang@mila.quebec, xinyu.wang5@mail.mcgill.ca
luli1@sas.upenn.edu, winfred.tai@mail.utoronto.ca

Abstract

While diffusion models have revolutionized text-to-image generation with their ability to synthesize realistic and diverse scenes, they continue to struggle with generating consistent and legible text within images. This shortcoming is commonly attributed to the locality bias inherent in diffusion-based generation, which limits their capacity to model long-range spatial dependencies. In this paper, we introduce **STRICT**, a benchmark designed to systematically stress-test the ability of diffusion models to render coherent and instruction-aligned text in images. Our benchmark evaluates models across multiple dimensions: (1) the maximum length of readable text that can be generated; (2) the correctness and legibility of the generated text, and (3) the ratio of not following instructions for generating text. We assess several state-of-the-art models, including proprietary and open-source variants, and reveal persistent limitations in long-range consistency and instruction-following capabilities. Our findings provide insights into architectural bottlenecks and motivate future research directions in multimodal generative modeling. We release all our evaluation pipeline at https://github.com/tianyu-z/STRICT-Bench/.

1 Introduction

Text-to-image generation has made remarkable strides with the advent of diffusion models (Ho et al., 2020; Rombach et al., 2022; Balaji et al., 2022; Feng et al., 2023; Gal et al., 2023; Zhang et al., 2023), which can now produce highly realistic images from natural language prompts. However, the generation of accurate and coherent text within images, such as complex road signs, product labels, or blackboards, remains a major unsolved problem (Liu et al., 2023; Chen et al., 2023a). Unlike general object generation, rendering text demands strict spatial precision, character-level continuity (Fallah et al., 2025), and strong adherence to instruction semantics. Due to their iterative and local sampling nature, diffusion models often fail to maintain global coherence (Zhang et al., 2024b), leading to text that is jumbled, misspelled, or visually fragmented. These failures highlight a fundamental challenge in aligning image synthesis with structured linguistic content (Chen et al., 2025).

Recent advances, such as OpenAI's Image-4o (OpenAI, 2025), have shown promising progress in this domain, achieving near-human performance in rendering embedded text. Similarly, open-source models like HiDream-L1 (HiDream.ai, 2025) and SeedDream 3 (Gao et al., 2025) have reported comparable success in overcoming long-range dependency issues. Yet, a systematic and quantitative evaluation of these capabilities remains lacking.

In this work, we present **STRICT** (Stress-Test of Rendering Image Containing Text), a comprehensive benchmark designed to rigorously evaluate the performance of diffusion models in generating image-embedded text. Our contributions are threefold:

- We introduce a multi-lingual benchmark that tests model performance on rendering texts of varying lengths in English, Chinese, and French.
- We propose quantitative metrics for assessing (1) the maximum readable text length; (2) the correctness of the generated content, and (3) the ratio of not following instructions for generating text.
- We analyze recurring failure modes, including truncation in longer texts and the inability to follow explicit textual instructions.

Through this evaluation, we aim to expose current limitations, identify failure patterns, and guide the development of structure-aware generative mod-

*Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21149–21162
November 4-9, 2025

els capable of producing semantically and visually coherent text within images.

2 Task Design

The primary objective of our benchmark is to rigorously evaluate the capability of text-to-image generation models to render accurate and coherent text embedded within synthetic images. The task is to utilize the model to be evaluated with a ground truth content string and instructed with a natural language prompt template: *f"Produce an image of a typed document page with the following text:* **[TEXT]** *"* to generate an image containing this specific text **[TEXT]**. The characteristics of the ground truth text sample are systematically varied to probe different aspects of model performance. These variations include:

- **Text Length:** Ground truth texts range from short phrases to longer paragraphs to determine the maximum length of text a model can generate coherently. This helps assess how accuracy degrades as the quantity of text increases.
- **Language:** Texts are selected from Wikipedia (Foundation) based on different languages.

For each sample, a model generates an image based on the input prompt containing the target ground truth text and then forms a pair with its corresponding ground truth text file. We collect these generated image-text pairs for evaluating each model.

3 Evaluation Metrics

To quantify the performance of models in rendering text, we employ an OCR-based verification framework. The text content from each generated image is first extracted using an OCR engine, followed by a comparison against the original ground truth text.

3.1 OCR and Preprocessing

We utilize the Tesseract OCR engine (Ooms, 2025) for extracting text from generated images. Tesseract is a widely adopted open-source OCR tool capable of recognizing over 100 languages and offering configurable page segmentation modes (PSM) to suit different layout scenarios. In our evaluation, we primarily use English, French, and Chinese language models with PSM set to 3, which indicates fully automatic page layout analysis.

To ensure fair and robust evaluation, we perform minimal but critical preprocessing of both the OCR-extracted text and the ground truth references. Specifically, we adopt a strict text processing strategy that normalizes whitespace across both texts, collapsing all forms of whitespace (spaces, tabs, newlines) into a single space, and trims leading and trailing whitespace.

After preprocessing, we evaluate OCR performance using a suite of metrics: character-level accuracy (Character Error Rate, CER) (Radford et al., 2023; Conneau et al., 2021), word-level accuracy (Word Error Rate, WER) (Morris et al., 2004; Kim et al., 2021), normalized edit distance (NED) Fisman et al. (2022), and sequence similarity using Python package `difflib` (Python Software Foundation, 2001). For a more granular analysis, we compute these metrics under two settings: a full comparison, which compares the entire OCR output to the full ground truth, and a truncated comparison, which compares the OCR output against a truncated version of the ground truth matched to the number of tokens recognized by OCR. This dual-mode evaluation allows us to assess not only absolute performance but also the model's ability to preserve textual order and correctness under realistic generation length constraints.

Through these evaluations, we reveal systematic differences in how text-to-image models perform under varying linguistic and spatial constraints in multi-language settings.

3.2 Quantitative Metrics

Following OCR extraction and preprocessing, the recognized text is compared against the ground truth using several order-preserving metrics. These metrics assess the accuracy of the generated text at both word and character levels. We report "Full" and "Truncated" versions for each metric; the "Full" version compares against the entire ground truth, while the "Truncated" version may adapt the comparison based on the length of the shorter sequence (typically the OCR output), providing insight into partial correctness.

1. **Normalized Edit Distance (NED):** NED quantifies the dissimilarity between the ground truth and OCR output based on character-level Levenshtein distance (edit distance). We adopt the Levenshtein distance normalization proposed in Fisman et al. (2022), where the cost of inserts, deletes and swaps are

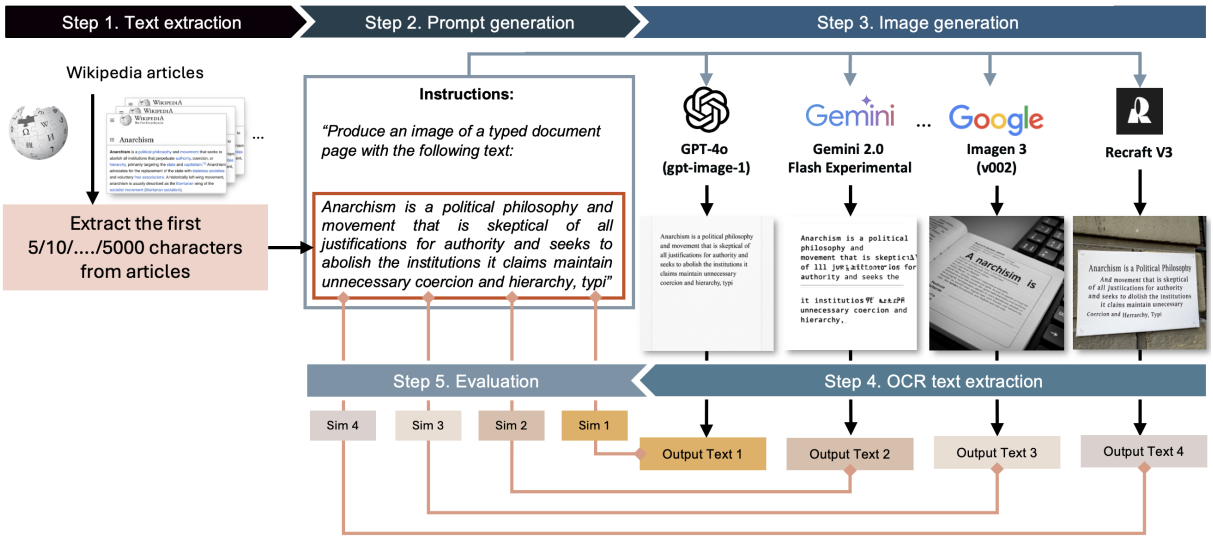

Figure 1: Illustration of the dataset creation and evaluation pipeline for **STRICT**. **Step 1:** We begin by selecting multilingual text samples from Wikipedia (Foundation), and extracting character sequences of varying lengths, ranging from 5 to 5000 characters. **Step 2:** For each text sample, we generate natural language instructions prompting models to create "a plain Word document with black text on a white background, without decorative elements," embedding the extracted text from Step 1. **Step 3:** These prompts are then passed to various text-to-image generation models to produce the corresponding output images. **Step 4:** Optical character recognition (OCR) is applied to the generated images to extract the rendered text. **Step 5:** Finally, we evaluate the quality of the rendered text by comparing the OCR output to the ground truth using similarity metrics, including normalized edit distance (NED), character error rate (CER), and word error rate (WER).

all 1, and the normalization factor is the length of the optimal edit path. This process results in a score between 0.0 (identical strings) and 1.0. Lower NED values indicate greater similarity. In our observation, some models consistently generate words with typos, making character-level metrics better reflect the actual generation performance.

2. **Character Error Rate (CER):** Other than NED, we also use two commonly used metrics in speech recognition: Character Error Rate (CER) (Radford et al., 2023; Conneau et al., 2021) and Word Error Rate (WER) (Morris et al., 2004; Kim et al., 2021). Similar to NED, CER operates at the character level. It is more sensitive to minor errors like single incorrect letters or OCR misrecognitions. CER is also normalized by the number of characters in the ground truth (for Full CER), with lower values being better. Note that the CER values do not have an upper limit.

3. **Word Error Rate (WER):** WER is a standard word-level metric commonly seen in speech recognition and OCR, calculating the minimum number of word-level insertions, deletions, and substitutions required to transform the OCR output into the ground truth text. The result is typically normalized by the number of words in the ground truth text (for Full WER). Lower WER values signify higher accuracy. Note that the WER values do not have an upper limit.

4. **Ratio of Not Following Instructions (RNFI):** RNFI measures the extent to which a model fails to follow the prompt by generating discretionary natural images instead of faithfully rendering the given text. For each sample, we compute the ratio between the number of characters extracted from the model-generated image (via OCR) and the number of characters in the ground-truth input text. We then count the number of samples where this ratio falls below 1%, indicating that the model has effectively ignored the instruction, i.e., less than 1% of the characters from the prompt are present in the generated image, **regardless of their correctness**.

For each metric, we calculate aggregate statistics, including mean, bootstrapped standard deviation for mean, and bootstrapped confidence intervals, across the entire dataset of evaluated image-text

pairs. This provides a robust overall assessment of a model's text generation capabilities under the specified evaluation conditions. The length of the ground truth text and the OCR-extracted text in characters are also recorded to provide context for the error rates.

4 Experiment Results

We evaluate a diverse set of state-of-the-art text-to-image generation models, including proprietary and open-source variants, on the **STRICT** benchmark. The models tested include: GPT-4o (`gpt-image-1`) (OpenAI, 2025), Seedream 3.0 (Gao et al., 2025), Recraft V3 (AI, 2024a), HiDream-I1-Dev (HiDream.ai, 2025), Imagen 3 (`imagen-3.0-generate-002`) (DeepMind, 2024a), FLUX 1.1 pro (Labs, 2024), Gemini 2.0 (`gemini-2.0-flash-preview-image-generation`) (DeepMind, 2024b), as well as open-source models such as Stable Diffusion 3.5 Medium (AI, 2024b), Anytext 2 (Zhao and Lian, 2024a), and the `TextDiffuser 2` (Chen et al., 2023b, 2024). We also include partial results for `Qwen-Image` (Wu et al., 2025) and `gemini-2.5-flash (nano-banana)` (DeepMind, 2025), which is released after we submit the paper. The text dataset is sourced from Wikipedia (Foundation), which is licensed under the Creative Commons Attribution-ShareAlike 3.0 (CC-BY-SA 3.0) license. The OCR processing was performed using Tesseract OCR (Ooms, 2025), which is distributed under the Apache License 2.0. Both licenses permit use for research and commercial purposes, provided the respective attribution and license terms are followed.

Overall Performance. According to Figure 3, across all evaluated languages and text length, GPT-4o and Gemini-2.0 outperform all competing models significantly by a large margin in terms of character accuracy, word accuracy, and instruction adherence. We evaluated ten models with varying capabilities. For the weakest models (Anytext 2, TextDiffuser 2, Stable Diffusion 3.5 Medium), we tested input lengths ranging from 5 to 300 characters. For moderately performing models (FLUX 1.1 pro, Seedream 3.0, HiDream-I1-Dev), the tested character lengths ranged from 50 to 2,000. Although moderate performance, the API upper limit of Recraft V3 is 1000 bytes (1,000 Latin characters or 500 Chinese characters). Thus, we test Recraft V3 from 50 to its upper limits. Moreover, the strongest models (Imagen 3, Gemini 2.0 and GPT-4o) were evaluated on inputs ranging from 50 to 5,000 characters. For both English and French, GPT-4o and Gemini maintain strong performance up to approximately 800 characters, beyond which accuracy begins to decrease. Detailed CER and WER metrics are displayed in the Appendix C for reference. For Chinese, the overall model performance remains poor. However, GPT-4o still consistently outperforms the other models. We display all the scores in the form of heatmap in Figure 2. For detailed scores with standard deviation, please check Table 1.

Ratio of Not Following Instructions. From Figure 4, we can see that some models including Flux 1.1 pro and Gemini 2.0 tend not to follow the instructions, especially when the number of characters in the given prompt becomes longer. We will discuss more in Section 5.

5 Discussion

Our study highlights both the capabilities and limitations of current text-to-image models in faithfully rendering structured textual content. While models like GPT-4o (OpenAI, 2025) demonstrate impressive gains and establish a new standard, the majority of diffusion models (Gao et al., 2025; AI, 2024a; HiDream.ai, 2025; DeepMind, 2024a; Labs, 2024; AI, 2024b; Chen et al., 2023b, 2024) still face significant challenges in instruction-following, text alignment, and multi-lingual generalization.

Performance Degradation with Text Length. Most diffusion models demonstrate a marked decline in performance as the input text length exceeds approximately 200 characters. This threshold aligns with the 77-token limit of the CLIP text encoder, which is used in HiDream, stable diffusion series and other diffusion models. Taking a step back, as mentioned in (Zhang et al., 2024a), the real effective token length of CLIP is 20 rather than 77. Considering the above two points together, it is likely that the model's ability to capture and condition on longer instructions is restricted. Moreover, existing training corpora rarely include such long and instruction-heavy prompts, further exacerbating generalization challenges in these regimes.

Instruction-Following Failures. With longer text prompts, diffusion models increasingly fail to adhere to instruction semantics. Instead of generating a document-like image containing the target

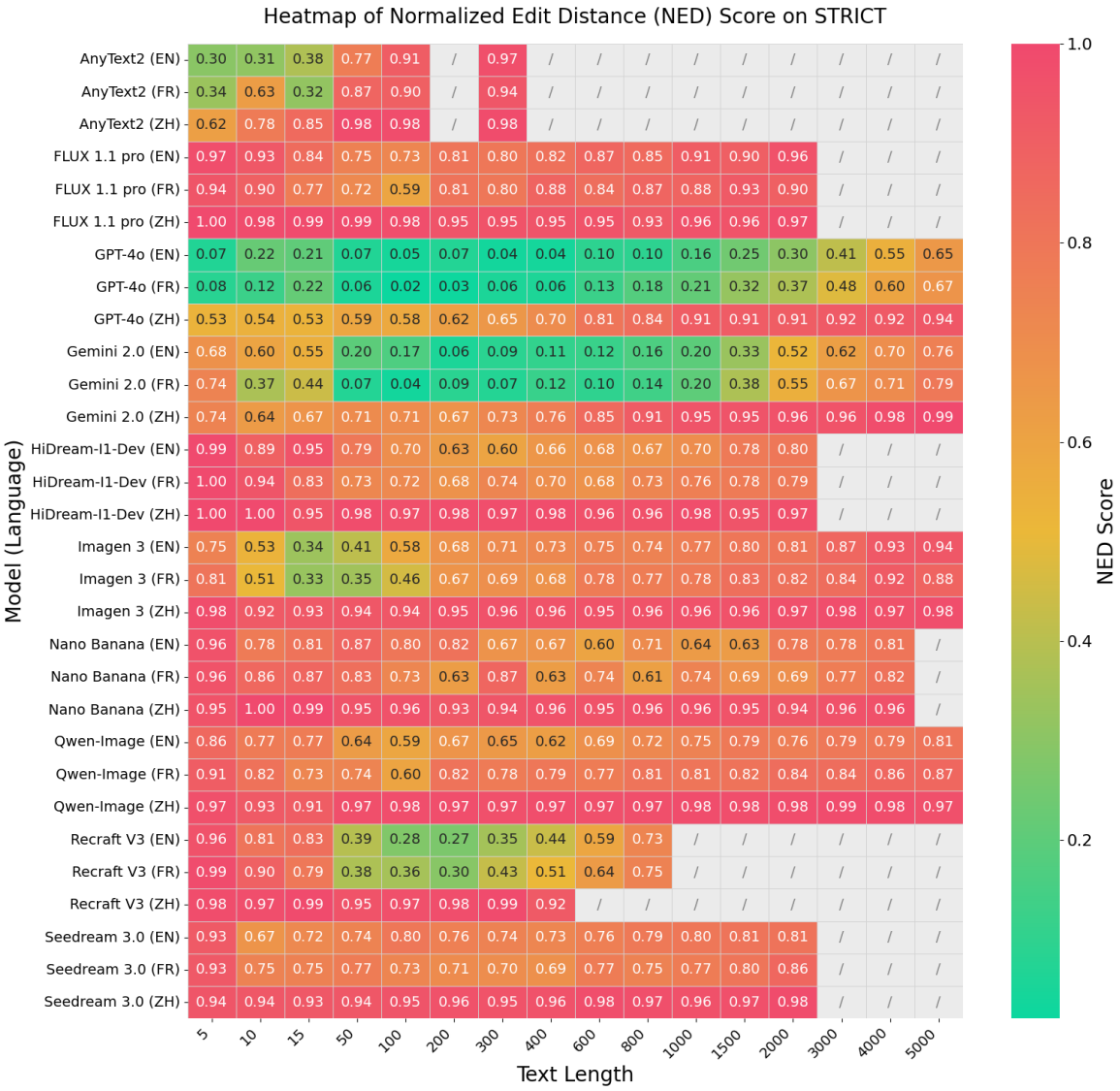

Figure 2: Heatmap of Normalized Edit Distance (NED) scores on the **STRICT** benchmark. Models are evaluated across three languages (EN: English, FR: French, ZH: Chinese) and varying text lengths. Lower NED scores (green) indicate better performance (higher text rendering accuracy). Grey cells denote untested lengths, as models were evaluated on ranges corresponding to their capabilities or API limitations.

text, many models instead synthesize a naturalistic image related to the subject matter of the prompt, completely omitting any embedded text. For example, in Figure 5, when instructed to render a document stating "Asia (,) is the largest continent in the world by both land area and population....", some models, especially Flux 1.1 pro, return an illustration of an Asian map instead of the text itself as the number of characters increase. We hypothesize that this failure arises from three main causes: (1) the dilution of the instruction signal in long prompts, which reduces the model's focus on the "document should contain the text:" instruction; (2) the limited capacity of CLIP-based or similar text encoders to capture long-range logical structure in extended inputs. These encoders may process such inputs more like a bag-of-words and thus compromise instruction fidelity; (3) the use of certain models, such as HiDream-L1, which incorporate LLMs (e.g., Llama-3.1-8B (Grattafiori et al., 2024)) as encoders, although LLMs are primarily designed as decoders and are not well-suited for encoding tasks. Prior work (Huang et al., 2024; BehnamGhader et al., 2024) has highlighted that LLMs generally underperform in encoder roles, although some modifications can improve their effectiveness.

Cross-Lingual Variation. We observe consistent differences in performance across languages.

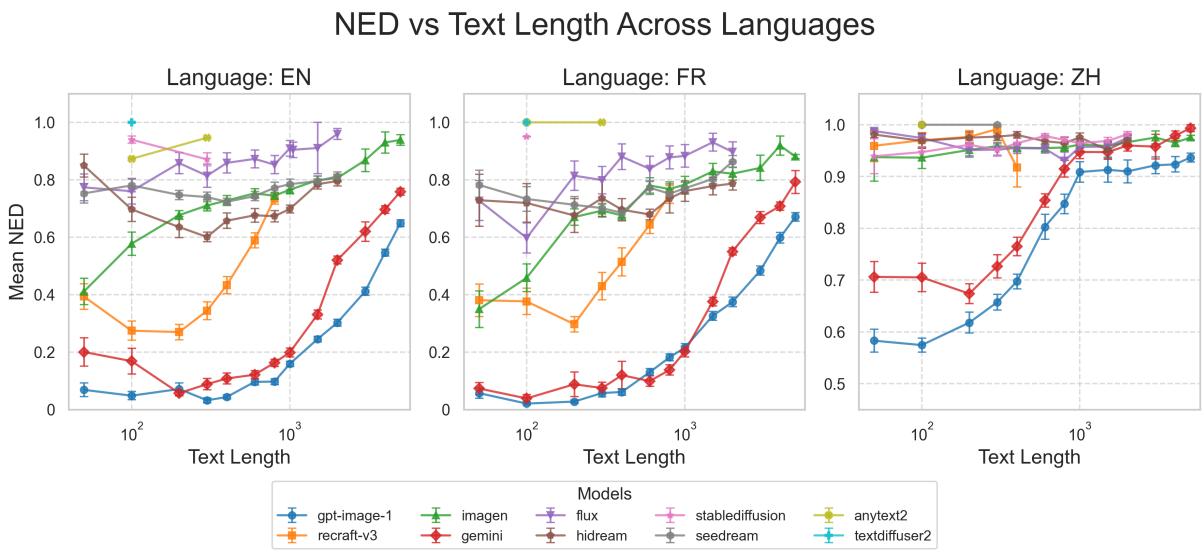

Figure 3: **Normalized Edit Distance (NED) vs. Text Length across Languages.** We evaluate ten state-of-the-art text-to-image generation models on multilingual text rendering using English (EN), French (FR), and Chinese (ZH) excerpts sampled from Wikipedia, with input lengths ranging from 5 to 5000 characters. Each model is prompted with identical semantic content across varying lengths, and OCR is applied to the generated images to compute character-level NED scores. Higher-performing models such as GPT-4o, Gemini 2.0, and Imagen 3 are evaluated up to 5000 characters, while Stable Diffusion 3.5, AnyText2, and TextDiffuser2 are evaluated up to 300 characters, and the remaining models up to 2000. Lower NED scores indicate better text fidelity and layout consistency.

English generally yields the highest accuracy, followed by French, while Chinese exhibits the lowest performance. This is likely attributable to insufficient training data for Chinese, rather than intrinsic difficulties in rendering Chinese characters. Structurally, we do not think there is a fundamental reason showing Chinese characters should be harder to generate than Latin alphabet. The performance gap underscores a need for broader multilingual data inclusion in models' pertaining.

6 Related Work

6.1 Diffusion Models

Diffusion models have emerged as a dominant paradigm in text-to-image synthesis, surpassing traditional generative models like GANs and VAEs in generating high-fidelity and diverse images. Foundational works such as DDPM (Ho et al., 2020) and Latent Diffusion Models (Rombach et al., 2022) laid the groundwork for this progress. Subsequent models like EDIFI (Balaji et al., 2022), ERNIE-ViLG 2.0 (Feng et al., 2023), and Textual Inversion (Gal et al., 2023) have further enhanced the alignment between generated images and textual prompts.

Recent advancements, Stable Diffusion series (Rombach et al., 2022; Esser et al., 2024), have integrated large language models like T5 (Raffel et al., 2020) to better encode textual information. Despite these improvements, challenges remain in rendering complex, multi-line, or structured text within images (Zhao et al., 2023).

6.2 Autoregressive Models

Autoregressive models offer an alternative approach to image generation by modeling images as sequences of discrete tokens. Recent developments have focused on enhancing spatial consistency and instruction adherence. VAR (Tian et al., 2024) employs multi-resolution next-token prediction to improve long-range coherence. InstructCV (Gan et al., 2024) frames various visual tasks within a unified text-guided generation framework using multi-modal prompts. Models like SeedX (Ge et al., 2024) and Chameleon (Team, 2024) unify text and image sequences within the same autoregressive framework, enhancing fluency and cross-modal alignment. Additionally, LlamaGen (Lu et al., 2024) adopts refined tokenization and pretrained language modeling to narrow the performance gap with diffusion-based models.

6.3 Models on Text Rendering

Rendering accurate and structured text within images remains a core challenge for genera-

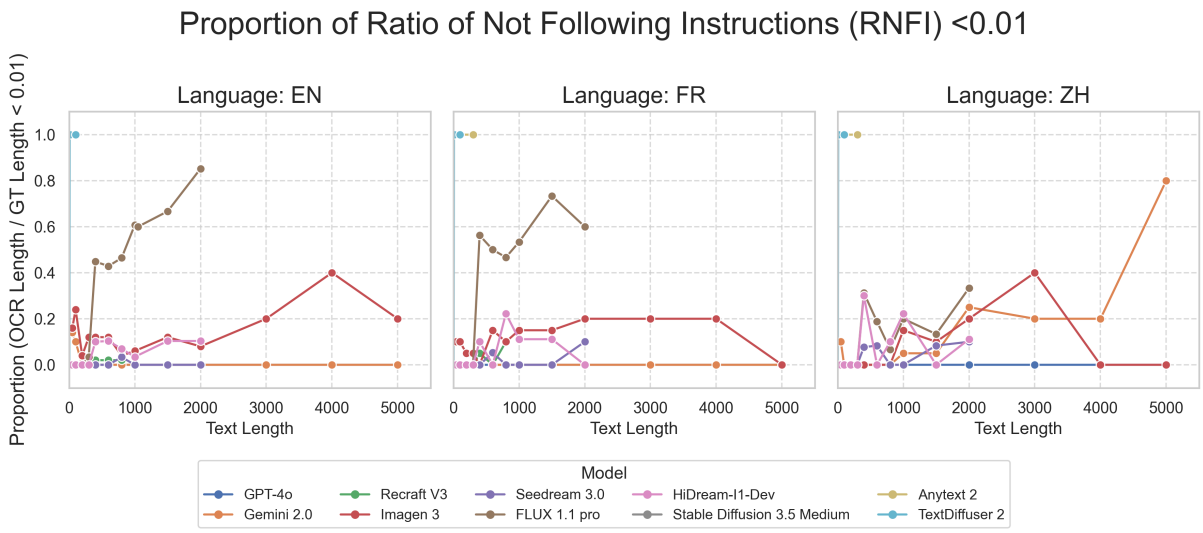

Figure 4: **Instruction Following Failure Rate across Text Lengths.** We plot the Ratio of Not Following Instructions (RNFI), defined as the percentage of samples where a model fails to render the input text. A failure is recorded if the ratio of characters in the generated image (measured via OCR) to characters in the input text is less than 1%. This metric captures catastrophic failures, such as generating a natural image instead of text, and does not penalize minor rendering errors. Evaluations use multilingual text (EN, FR, ZH) from Wikipedia with input lengths from 5 to 5,000 characters. Lower values indicate better robustness and instruction adherence.

tive models. To address spatial precision and layout constraints, GlyphControl (Yang et al., 2023) enables user-guided glyph placement, while GlyphDraw (Ma and Zhao, 2023) and TextDiffuser (Chen et al., 2023b) adopt keyword-driven generation and layout masks for structured rendering. TextDiffuser-2 (Chen et al., 2024) further incorporates layout planning and line-level encoding for improved diversity.

Models such as Recraft and the AnyText series (Zhao and Lian, 2023, 2024a) enhance multilingual and stylistic versatility by supporting diverse languages and font styles. Meanwhile, methods such as Glyph-SDXL (Zhao and Lian, 2024b), Glyph-SDXL-v2 (Zhao and Lian, 2024c), and GlyphDraw2 (Ma and Zhao, 2024) use OCR-guided glyph representations to improve layout accuracy and visual coherence.

Character-level approaches such as DiffSTE (Zhang and Lian, 2024b), UDiffText (Zhao and Lian, 2024d), and Brush Your Text (Zhang and Lian, 2024a) refine alignment through attention-based interventions.

Recent models have further advanced text rendering capabilities. Recraft V3 demonstrates proficiency in generating images with long texts and diverse styles (AI, 2024a). HiDream-I1-Dev, an open-source model with 17B parameters, achieves high-quality image generation with prompt adherence (HiDream.ai, 2025). Imagen 3, Google's latest model, offers improved detail and text rendering (DeepMind, 2024a). FLUX 1.1 pro delivers enhanced composition and artistic fidelity (Labs, 2024). Gemini 2.0 integrates multimodal inputs for native image generation (DeepMind, 2024b). Stable Diffusion 3.5 introduces a Multimodal Diffusion Transformer architecture, improving typography and complex prompt understanding (AI, 2024b). SeedDream 3.0 (Gao et al., 2025), a strong open-source diffusion model, demonstrates competitive accuracy and layout consistency in rendering multi-line and structured text.

6.4 Text-to-Image Benchmarks

Recent work has introduced specialized benchmarks to systematically evaluate T2I models' abilities to render readable, instruction-aligned, and multilingual text. TIFA (Hu et al., 2023) assesses semantic faithfulness via QA-based probing, while TypeScore (Sampaio et al., 2024) evaluates OCR-based text fidelity and instruction following.

TextInVision (Fallah et al., 2025) addresses structural challenges by varying prompt lengths and complexities to assess how diffusion models handle diverse textual inputs. MARIO-Eval (Chen et al., 2023c), built upon the extensive MARIO-10M dataset, offers a large-scale OCR benchmark for evaluating text rendering quality. For multilin-

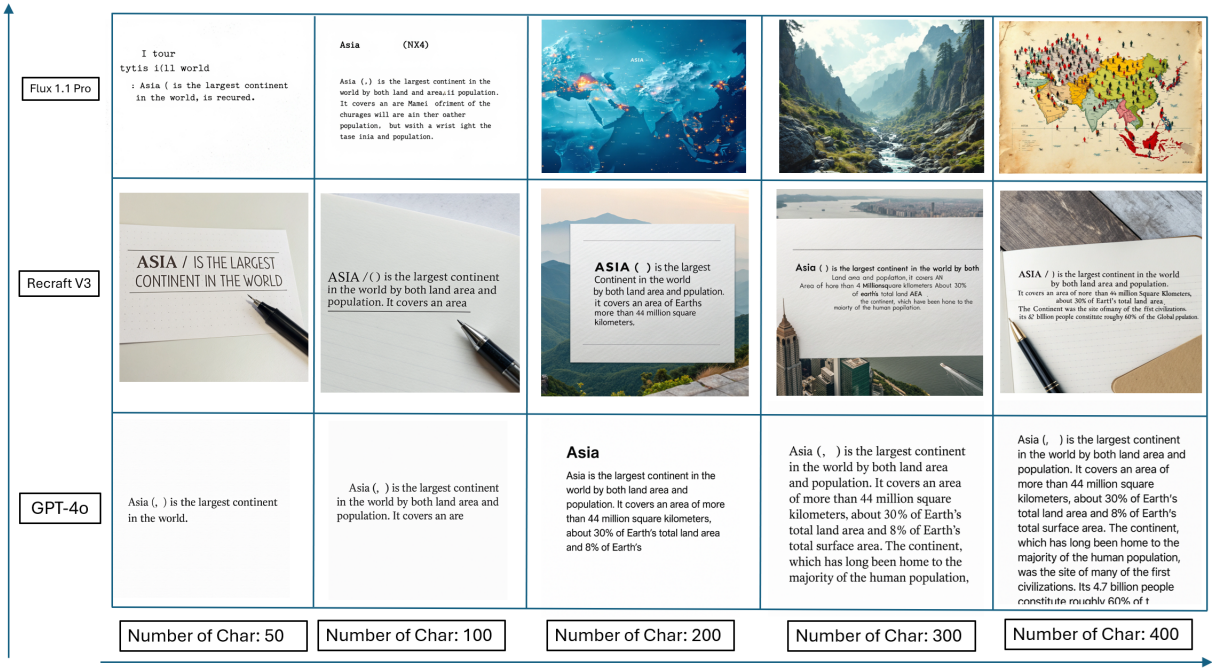

Figure 5: Case study on instruction-following failures. The three rows correspond to Flux 1.1 Pro, Recraft V3, and GPT-4o. GPT-4o consistently adheres to the given instructions, whereas Flux 1.1 Pro increasingly ignores them as the character count grows. Recraft V3 continues to generate text within the image but introduces background elements that were not requested in the prompt.

gual evaluation, AnyText (Tuo et al., 2024) introduces a dataset and metrics that encompass various languages and font styles.

Comprehensive benchmark suites like HEIM (Lee et al., 2023) and HRS-Bench (Bakr et al., 2023) jointly assess image quality, text rendering, and prompt adherence. LenCom-Eval (Lakhanpal et al., 2025) focuses on long-form prompts to expose generation degradation.

Recent architectural evaluations show LLM-grounded generation (Lian et al., 2025) significantly improves prompt alignment and cross-lingual generalization. TextMatch (Luo et al., 2024) refines outputs through multimodal feedback from VQA and LLMs. Collectively, these efforts lay a solid foundation for benchmarking advanced T2I systems under realistic, structured, and semantically rich prompts.

7 Conclusion

We introduced **STRICT**, a comprehensive benchmark for evaluating the ability of text-to-image models to render accurate, instruction-aligned, and multilingual text. Our evaluation reveals that while recent models like GPT-4o and Gemini 2.0 show strong performance, most open-source diffusion models still struggle with long-range coherence and instruction fidelity. We highlight key failure modes such as instruction neglect and language-specific disparities. By providing standardized tasks and metrics, **STRICT** enables targeted diagnosis and guides future improvements in multimodal generation systems.

Limitations

Firstly, despite nearly two decades of continuous development, the Tesseract OCR engine (Ooms, 2025) still encounters failure cases in which humans can easily recognize the text. These limitations remain challenging until we can fully overcome the drawbacks of OCR technologies.

Furthermore, if **STRICT** becomes a widely adopted benchmark, there is a risk that future models may be fine-tuned or hard-coded to perform well specifically on the dataset's structure and instructions. This undermines the benchmark's utility as an unbiased generalization test and could lead to inflated leaderboard results without corresponding real-world gains.

References

Recraft AI. 2024a. Recraft v3 text to image. https://fal.ai/models/fal-ai/recraft/v3/text-to-image.

Stability AI. 2024b. Introducing stable diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5.

Omar Bakr and 1 others. 2023. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12345–12354.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *CoRR*, abs/2211.01324.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv: 2404.05961*.

Hang Chen, Qian Xiang, Jiaxin Hu, Meilin Ye, Chao Yu, Hao Cheng, and Lei Zhang. 2025. Comprehensive exploration of diffusion models in image generation: a survey. *Artificial Intelligence Review*, 58(99):1–35.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023a. Textdiffuser: Diffusion models as text painters. *Preprint*, arXiv:2305.10855.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023b. Textdiffuser: Diffusion models as text painters. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023c. Textdiffuser: Diffusion models as text painters. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2024. Textdiffuser-2: Unleashing the power of language models for text rendering. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 2426–2430. ISCA.

DeepMind. 2025. Master the nano banana prompt: A beginner's guide on how to use nano banana for stunning ai images. Nano Banana AI Blog.

Google DeepMind. 2024a. Imagen 3. https://deepmind.google/technologies/imagen-3/.

Google DeepMind. 2024b. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

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, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*. OpenReview.net.

Forouzan Fallah, Maitreya Patel, Agneet Chatterjee, Vlad I. Morariu, Chitta Baral, and Yezhou Yang. 2025. Textinvision: Text and prompt complexity driven visual text generation benchmark. *arXiv preprint arXiv:2503.13730*.

Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, Li Chen, Hao Tian, Hua Wu, and Haifeng Wang. 2023. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *CVPR*, pages 10135–10145. IEEE.

D. Fisman, Joshua Grogin, Oded Margalit, and Gera Weiss. 2022. The normalized edit distance with uniform operation costs is a metric. *Annual Symposium on Combinatorial Pattern Matching*.

Wikimedia Foundation. Wikimedia downloads.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*. OpenReview.net.

Yulu Gan, Sungwoo Park, Alexander Schubert, Anthony Philippakis, and Ahmed M. Alaa. 2024. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. In *ICLR*. OpenReview.net.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xuanda Wang, Xun Wang, and 12 others. 2025. Seedream 3.0 technical report. *arXiv preprint arXiv: 2504.11346*.

Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. 2024. SEED-X: multimodal models with unified multi-granularity comprehension and generation. *CoRR*, abs/2404.14396.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv: 2407.21783*.

HiDream.ai. 2025. Hidream-i1: Open-source image generative foundation model. https://github.com/HiDream-ai/HiDream-I1. Accessed: 2025-05-20.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)*.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Chunyu Wang, Xiyang Dai, Dongdong Chen, Chong Luo, and Lili Qiu. 2024. Llm2clip: Powerful language model unlocks richer visual representation. *arXiv preprint arXiv: 2411.04997*.

Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. 2021. Semantic distance: A new metric for ASR performance analysis towards spoken language understanding. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 1977–1981. ISCA.

Black Forest Labs. 2024. Flux1.1 [pro] text to image. https://fal.ai/models/fal-ai/flux-pro/v1.1.

Sanyam Lakhanpal, Shivang Chopra, Vinija Jain, Aman Chadha, and Man Luo. 2025. Refining text-to-image generation: Towards accurate training-free glyph-enhanced image generation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. 2023. Holistic evaluation of text-to-image models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2025. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. In *Proc. International Conference on Learning Representations (ICLR)*.

Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. 2023. Character-aware models improve visual text rendering. *Preprint*, arXiv:2212.10562.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *CVPR*, pages 26429–26445. IEEE.

Yucong Luo, Mingyue Cheng, Jie Ouyang, Xiaoyu Tao, and Qi Liu. 2024. Textmatch: Enhancing image-text consistency through multimodal optimization. *arXiv preprint arXiv:2410.17746*.

Zhao Ma and Yiming Zhao. 2023. Glyphdraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM)*.

Zhao Ma and Yiming Zhao. 2024. Glyphdraw2: Automatic generation of complex glyph posters with diffusion models and large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768.

Jeroen Ooms. 2025. *tesseract: Open Source OCR Engine*. R package version 5.2.3.

OpenAI. 2025. Introducing 4o image generation. Accessed: 2025-05-20.

Python Software Foundation. 2001. difflib – helpers for computing deltas. https://docs.python.org/3/library/difflib.html. Accessed: 2025-05-19.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE.

Georgia Gabriela Sampaio, Ruixiang Zhang, Shuangfei Zhai, Jiatao Gu, Josh Susskind, Navdeep Jaitly, and Yizhe Zhang. 2024. Typescore: A text fidelity metric for text-to-image generative models. *arXiv preprint arXiv:2411.02437*.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *CoRR*, abs/2405.09818.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *NeurIPS*.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. 2024. Anytext: Multilingual visual text generation and editing. In *Proc. International Conference on Learning Representations (ICLR)*.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, and 20 others. 2025. Qwen-image technical report. *Preprint*, arXiv:2508.02324.

Yang Yang, Yiming Zhao, and Zhouhui Lian. 2023. Glyphcontrol: Glyph conditional control for visual text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Beichen Zhang, Pan Zhang, Xiao wen Dong, Yuhang Zang, and Jiaqi Wang. 2024a. Long-clip: Unlocking the long-text capability of clip. *European Conference on Computer Vision*.

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. 2023. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

Jianyi Zhang, Yufan Zhou, Jiuxiang Gu, Curtis Wigington, Tong Yu, Yiran Chen, Tong Sun, and Ruiyi Zhang. 2024b. Artist: Improving the generation of text-rich images with disentangled diffusion models and large language models. *arXiv preprint arXiv:2406.12044*.

Yiming Zhang and Zhouhui Lian. 2024a. Brush your text: Enhancing text rendering in diffusion models with attention map interventions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yiming Zhang and Zhouhui Lian. 2024b. Diffste: Improving diffusion models for scene text editing with dual encoders. In *International Conference on Learning Representations (ICLR)*.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. 2023. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *NeurIPS*.

Yiming Zhao and Zhouhui Lian. 2023. Anytext: Multilingual visual text generation and editing. *CoRR*, abs/2311.03054.

Yiming Zhao and Zhouhui Lian. 2024a. Anytext2: Visual text generation and editing with customizable attributes. *CoRR*, abs/2411.15245.

Yiming Zhao and Zhouhui Lian. 2024b. Glyph-byt5: A customized text encoder for accurate visual text rendering. *CoRR*, abs/2403.09622.

Yiming Zhao and Zhouhui Lian. 2024c. Glyph-byt5-v2: A strong aesthetic baseline for accurate multilingual visual text rendering. *CoRR*, abs/2406.10208.

Yiming Zhao and Zhouhui Lian. 2024d. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

A Appendix: Prompt Variants

To test the robustness of our benchmark against prompt phrasing, we experimented with a set of diverse but semantically equivalent instructions for generating document-style images. These prompt variants yielded consistent performance trends across multiple models. To standardize our evaluation pipeline, we selected a single representative prompt (highlighted in red below) for all metric-based experiments:

Prompt Variants Explored

- *Generate a scanned document image with following text:* ***[TEXT]***
- *Create a mockup of a scanned document containing the text:* ***[TEXT]***
- *Design a sample document scan with the following text:* ***[TEXT]***
- *Generate an image of printed note that include these text:* ***[TEXT]***
- *Produce an image of a typed document page with the following text:* ***[TEXT]***
- *Generate a document scan visualization showing this text:* ***[TEXT]***
- *Produce a sample of how a scanned memo might look with this text:* ***[TEXT]***
- *Generate an image of a plain Word document with black text on white background without decorative elements, document should contain the text:* ***[TEXT]***

We observed no significant variation in performance across these prompts, reinforcing the robustness of our task design. For all reported experiments, we standardized on the red-highlighted prompt.

B Appendix: Detailed NED Scores in table

We present the detailed NED scores and corresponding standard deviation in table 1.

C Appendix: CER and WER Figure

Please check Figure 6 and Figure 7 for Character Error Rate (CER) and Word Error Rate (WER).

D Use of AI Tools in Manuscript Preparation

In the preparation of this manuscript, we utilized a large language model as an assistive tool. The LLM's role was confined to improving the grammatical structure and clarity of our writing. Furthermore, it was used to assist in debugging code snippets and generating routine documentation such as docstrings. The core research ideas, experimental design, analysis, and conclusions were conceived and executed entirely by the authors. All LLM-generated outputs were critically reviewed and edited by the authors, who take full responsibility for the final content of this paper.

Text Length	GPT-4o	Gemini 2.0	Recraft V3	Imagen 3	Seedream 3.0	FLUX 1.1 pro	HiDream-I1-Dev	Stable Diffusion 3.5 Medium	Anytext 2	TextDiffuser 2	Qwen-Image	nano-banana
						EN						
5	0.07 ± 0.03	0.68 ± 0.06	0.96 ± 0.01	0.75 ± 0.05	0.93 ± 0.03	0.97 ± 0.04	0.99 ± 0.00	1.00 ± 0.00	0.17 ± 0.00	1.00 ± 0.00	0.86 ± 0.06	0.96 ± 0.03
10	0.22 ± 0.06	0.60 ± 0.06	0.81 ± 0.05	0.53 ± 0.06	0.67 ± 0.06	0.93 ± 0.07	0.89 ± 0.05	0.99 ± 0.03	0.08 ± 0.00	1.00 ± 0.00	0.77 ± 0.06	0.78 ± 0.07
15	0.21 ± 0.05	0.55 ± 0.05	0.83 ± 0.04	0.34 ± 0.04	0.72 ± 0.05	0.84 ± 0.05	0.95 ± 0.02	0.99 ± 0.00	0.16 ± 0.00	1.00 ± 0.00	0.77 ± 0.06	0.81 ± 0.06
50	0.07 ± 0.17	0.20 ± 0.35	0.39 ± 0.31	0.41 ± 0.32	0.75 ± 0.24	0.77 ± 0.23	0.85 ± 0.18	–	–	–	0.64 ± 0.05	0.87 ± 0.05
100	0.05 ± 0.10	0.17 ± 0.31	0.27 ± 0.24	0.58 ± 0.29	0.78 ± 0.15	0.76 ± 0.24	0.70 ± 0.23	0.94 ± 0.04	0.87 ± 0.00	1.00 ± 0.00	0.59 ± 0.05	0.80 ± 0.06
200	0.07 ± 0.15	0.06 ± 0.06	0.27 ± 0.19	0.68 ± 0.11	0.75 ± 0.11	0.86 ± 0.19	0.64 ± 0.20	–	–	–	0.67 ± 0.03	0.82 ± 0.06
300	0.03 ± 0.06	0.09 ± 0.14	0.34 ± 0.22	0.71 ± 0.13	0.74 ± 0.12	0.81 ± 0.22	0.60 ± 0.09	0.87 ± 0.08	0.95 ± 0.00	1.00 ± 0.00	0.65 ± 0.04	0.67 ± 0.07
400	0.04 ± 0.05	0.11 ± 0.13	0.43 ± 0.21	0.73 ± 0.12	0.72 ± 0.09	0.86 ± 0.19	0.66 ± 0.15	–	–	–	0.62 ± 0.04	0.67 ± 0.07
600	0.10 ± 0.08	0.12 ± 0.10	0.59 ± 0.18	0.75 ± 0.12	0.74 ± 0.09	0.87 ± 0.16	0.68 ± 0.13	–	–	–	0.69 ± 0.02	0.60 ± 0.06
800	0.10 ± 0.07	0.16 ± 0.09	0.73 ± 0.13	0.74 ± 0.08	0.77 ± 0.11	0.85 ± 0.16	0.67 ± 0.10	–	–	–	0.72 ± 0.01	0.71 ± 0.05
1000	0.16 ± 0.07	0.20 ± 0.11	–	0.77 ± 0.10	0.78 ± 0.10	0.91 ± 0.15	0.70 ± 0.08	–	–	–	0.75 ± 0.01	0.64 ± 0.05
1500	0.25 ± 0.07	0.33 ± 0.11	–	0.80 ± 0.10	0.80 ± 0.09	0.91 ± 0.15	0.78 ± 0.09	–	–	–	0.79 ± 0.02	0.63 ± 0.04
2000	0.30 ± 0.08	0.52 ± 0.10	–	0.81 ± 0.09	0.81 ± 0.08	0.96 ± 0.10	0.80 ± 0.09	–	–	–	0.76 ± 0.01	0.78 ± 0.04
3000	0.41 ± 0.10	0.62 ± 0.08	–	0.87 ± 0.09	–	–	–	–	–	–	0.79 ± 0.02	0.78 ± 0.03
4000	0.55 ± 0.08	0.70 ± 0.03	–	0.93 ± 0.08	–	–	–	–	–	–	0.79 ± 0.02	0.81 ± 0.03
5000	0.65 ± 0.08	0.76 ± 0.03	–	0.94 ± 0.04	–	–	–	–	–	–	0.81 ± 0.01	–
						FR						
5	0.08 ± 0.04	0.74 ± 0.08	0.99 ± 0.01	0.81 ± 0.07	0.93 ± 0.04	0.94 ± 0.05	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.91 ± 0.04	0.96 ± 0.04
10	0.12 ± 0.06	0.37 ± 0.09	0.90 ± 0.05	0.51 ± 0.09	0.75 ± 0.07	0.90 ± 0.04	0.94 ± 0.07	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.82 ± 0.07	0.86 ± 0.07
15	0.22 ± 0.07	0.33 ± 0.08	0.79 ± 0.06	0.44 ± 0.09	0.75 ± 0.07	0.77 ± 0.05	0.83 ± 0.09	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.73 ± 0.07	0.87 ± 0.05
50	0.06 ± 0.08	0.07 ± 0.09	0.38 ± 0.25	0.35 ± 0.28	0.78 ± 0.23	0.73 ± 0.27	0.73 ± 0.29	–	–	–	0.74 ± 0.05	0.83 ± 0.08
100	0.02 ± 0.02	0.04 ± 0.06	0.38 ± 0.21	0.46 ± 0.22	0.73 ± 0.19	0.60 ± 0.21	0.72 ± 0.21	0.95 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.60 ± 0.06	0.73 ± 0.08
200	0.03 ± 0.02	0.09 ± 0.19	0.30 ± 0.12	0.67 ± 0.13	0.71 ± 0.13	0.81 ± 0.21	0.68 ± 0.19	–	–	–	0.82 ± 0.03	0.63 ± 0.08
300	0.06 ± 0.06	0.07 ± 0.09	0.43 ± 0.21	0.69 ± 0.10	0.70 ± 0.07	0.80 ± 0.19	0.74 ± 0.20	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.78 ± 0.04	0.87 ± 0.05
400	0.06 ± 0.05	0.12 ± 0.22	0.51 ± 0.22	0.68 ± 0.07	0.68 ± 0.07	0.88 ± 0.18	0.70 ± 0.12	–	–	–	0.79 ± 0.04	0.63 ± 0.09
600	0.13 ± 0.06	0.10 ± 0.08	0.64 ± 0.14	0.78 ± 0.12	0.77 ± 0.10	0.84 ± 0.17	0.68 ± 0.06	–	–	–	0.77 ± 0.02	0.74 ± 0.07
800	0.18 ± 0.05	0.14 ± 0.08	0.75 ± 0.14	0.77 ± 0.11	0.75 ± 0.07	0.88 ± 0.16	0.73 ± 0.15	–	–	–	0.81 ± 0.02	0.61 ± 0.07
1000	0.21 ± 0.07	0.20 ± 0.08	–	0.78 ± 0.13	0.77 ± 0.09	0.88 ± 0.15	0.76 ± 0.11	–	–	–	0.81 ± 0.02	0.74 ± 0.07
1500	0.33 ± 0.07	0.38 ± 0.07	–	0.83 ± 0.12	0.80 ± 0.07	0.93 ± 0.12	0.78 ± 0.09	–	–	–	0.82 ± 0.02	0.69 ± 0.05
2000	0.37 ± 0.07	0.55 ± 0.06	–	0.82 ± 0.10	0.86 ± 0.09	0.90 ± 0.13	0.79 ± 0.07	–	–	–	0.78 ± 0.04	0.69 ± 0.04
3000	0.48 ± 0.07	0.67 ± 0.05	–	0.84 ± 0.10	–	–	–	–	–	–	0.84 ± 0.02	0.77 ± 0.03
4000	0.60 ± 0.08	0.71 ± 0.03	–	0.92 ± 0.07	–	–	–	–	–	–	0.86 ± 0.02	0.82 ± 0.02
5000	0.67 ± 0.07	0.79 ± 0.09	–	0.88 ± 0.02	–	–	–	–	–	–	0.87 ± 0.02	–
						ZH						
5	0.53 ± 0.04	0.74 ± 0.05	0.98 ± 0.01	0.98 ± 0.01	0.94 ± 0.02	1.00 ± 0.00	1.00 ± 0.07	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.02	0.95 ± 0.02
10	0.54 ± 0.02	0.64 ± 0.03	0.97 ± 0.01	0.92 ± 0.04	0.94 ± 0.01	0.98 ± 0.00	1.00 ± 0.06	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.93 ± 0.03	1.00 ± 0.00
15	0.53 ± 0.02	0.67 ± 0.03	0.99 ± 0.00	0.93 ± 0.03	0.93 ± 0.02	0.99 ± 0.00	0.95 ± 0.10	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.91 ± 0.05	0.99 ± 0.00
50	0.58 ± 0.10	0.71 ± 0.13	0.96 ± 0.14	0.94 ± 0.21	0.94 ± 0.14	0.99 ± 0.01	0.98 ± 0.02	–	–	–	0.97 ± 0.01	0.95 ± 0.02
100	0.57 ± 0.06	0.71 ± 0.12	0.97 ± 0.07	0.94 ± 0.09	0.95 ± 0.04	0.97 ± 0.03	0.97 ± 0.05	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.98 ± 0.01	0.96 ± 0.02
200	0.62 ± 0.09	0.67 ± 0.08	0.98 ± 0.05	0.95 ± 0.05	0.96 ± 0.03	0.95 ± 0.05	0.97 ± 0.04	–	–	–	0.97 ± 0.01	0.93 ± 0.02
300	0.66 ± 0.07	0.73 ± 0.10	0.99 ± 0.02	0.96 ± 0.03	0.95 ± 0.05	0.95 ± 0.04	0.98 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.00	0.94 ± 0.01
400	0.70 ± 0.06	0.76 ± 0.08	0.92 ± 0.05	0.96 ± 0.03	0.96 ± 0.05	0.95 ± 0.04	0.98 ± 0.02	–	–	–	0.97 ± 0.01	0.96 ± 0.01
600	0.80 ± 0.11	0.85 ± 0.06	–	0.95 ± 0.03	0.98 ± 0.02	0.95 ± 0.03	0.97 ± 0.03	–	–	–	0.97 ± 0.01	0.95 ± 0.02
800	0.85 ± 0.08	0.91 ± 0.07	–	0.96 ± 0.03	0.97 ± 0.02	0.93 ± 0.07	0.96 ± 0.03	–	–	–	0.97 ± 0.01	0.96 ± 0.01
1000	0.91 ± 0.09	0.95 ± 0.06	–	0.96 ± 0.03	0.96 ± 0.02	0.96 ± 0.03	0.98 ± 0.03	–	–	–	0.98 ± 0.01	0.96 ± 0.01
1500	0.91 ± 0.10	0.95 ± 0.06	–	0.96 ± 0.04	0.97 ± 0.02	0.96 ± 0.03	0.95 ± 0.05	–	–	–	0.98 ± 0.01	0.95 ± 0.01
2000	0.91 ± 0.10	0.96 ± 0.05	–	0.97 ± 0.04	0.98 ± 0.02	0.97 ± 0.03	0.97 ± 0.03	–	–	–	0.98 ± 0.01	0.94 ± 0.02
3000	0.92 ± 0.07	0.96 ± 0.05	–	0.98 ± 0.03	–	–	–	–	–	–	0.99 ± 0.00	0.96 ± 0.01
4000	0.92 ± 0.06	0.98 ± 0.02	–	0.97 ± 0.02	–	–	–	–	–	–	0.98 ± 0.01	0.96 ± 0.01
5000	0.94 ± 0.04	0.99 ± 0.02	–	0.98 ± 0.01	–	–	–	–	–	–	0.97 ± 0.01	–

Table 1: Normalized Edit Distance (NED) scores for multilingual text rendering across various text lengths. Each model is prompted to generate an image embedding ground truth text sampled from Wikipedia (Foundation) in English, French, or Chinese. OCR is applied to the generated images, and NED is computed between the OCR output and the ground-truth text. Lower scores indicate higher fidelity in character-level text rendering.

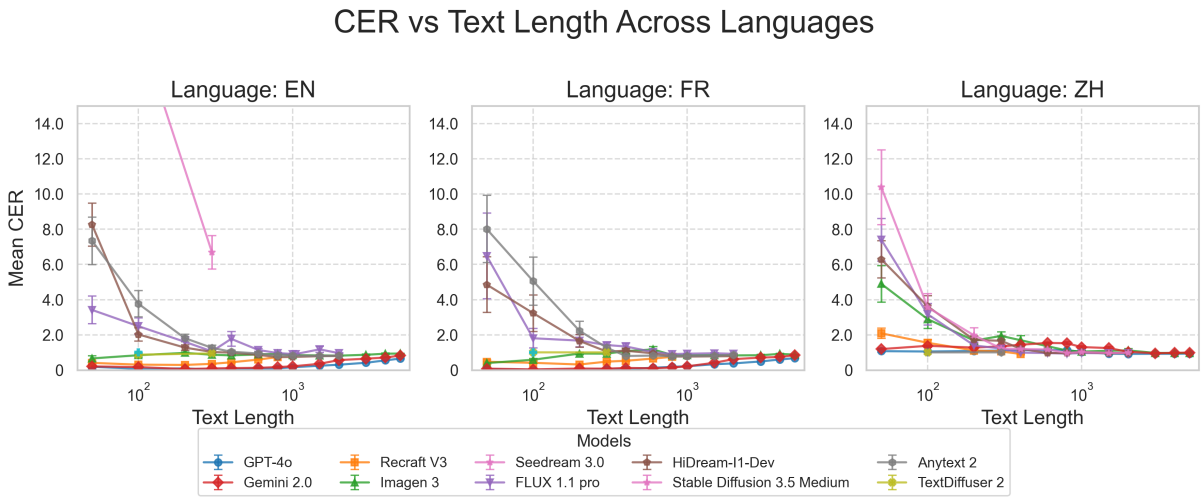

Figure 6: Character Error Rate (CER) vs. Text Length across Languages. We evaluate ten state-of-the-art text-to-image generation models on multilingual text rendering using English (EN), French (FR), and Chinese (ZH) excerpts sampled from Wikipedia, with input lengths ranging from 5 to 5000 characters. Each model is prompted with identical semantic content across varying lengths, and OCR is applied to the generated images to compute Character Error Rate (CER). Higher-performing models such as GPT-4o, Gemini 2.0, and Imagen 3 are evaluated up to 5000 characters, while Stable Diffusion 3.5, AnyText2, and TextDiffuser2 are evaluated up to 300 characters, and the remaining models up to 2000. Lower CER scores indicate better text fidelity and layout consistency.

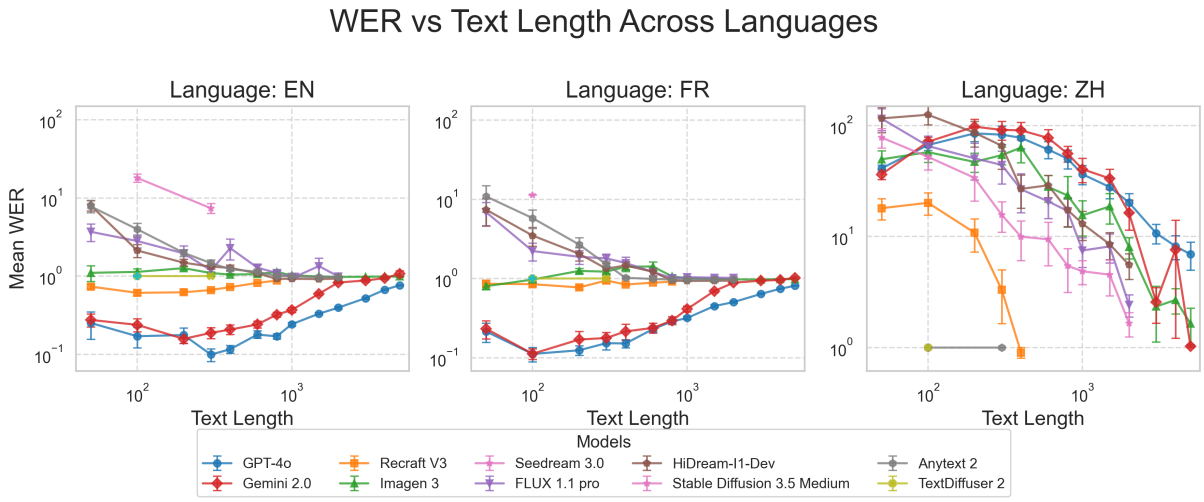

Figure 7: Word Error Rate (WER) vs. Text Length across Languages. We evaluate ten state-of-the-art text-to-image generation models on multilingual text rendering using English (EN), French (FR), and Chinese (ZH) excerpts sampled from Wikipedia, with input lengths ranging from 5 to 5000 characters. Each model is prompted with identical semantic content across varying lengths, and OCR is applied to the generated images to compute Word Error Rate (WER). Higher-performing models such as GPT-4o, Gemini 2.0, and Imagen 3 are evaluated up to 5000 characters, while Stable Diffusion 3.5, AnyText2, and TextDiffuser2 are evaluated up to 300 characters, and the remaining models up to 2000. Lower WER scores indicate better text fidelity and layout consistency.