TextInVision: Text and Prompt Complexity Driven Visual Text Generation Benchmark

Forouzan Fallah♠* Maitreya Patel♠ Agneet Chatterjee♠
Vlad I. Morariu◊ Chitta Baral♠ Yezhou Yang♠
♠Arizona State University ◊Adobe Research

Abstract

Generating images with embedded text is crucial for the automatic production of visual and multimodal documents, such as educational materials and advertisements. However, existing diffusion-based text-to-image models often struggle to accurately embed text within images, facing challenges in spelling accuracy, contextual relevance, and visual coherence. Evaluating the ability of such models to embed text within a generated image is complicated due to the lack of comprehensive benchmarks. In this work, we introduce TextInVision, a large-scale, text and prompt complexity driven benchmark designed to evaluate the ability of diffusion models to effectively integrate visual text into images. We crafted a diverse set of prompts and texts that consider various attributes and text characteristics. Additionally, we prepared an image dataset to test Variational Autoencoder (VAE) models across different character representations, highlighting that VAE architectures can also pose challenges in text generation within diffusion frameworks. Through an extensive analysis of multiple models, we identify common errors and highlight issues such as spelling inaccuracies and contextual mismatches. By pinpointing the failure points across different prompts and texts, our research lays the foundation for future advancements in AI-generated multimodal content. Dataset and benchmark can be found at: `https://github.com/TextinVision`

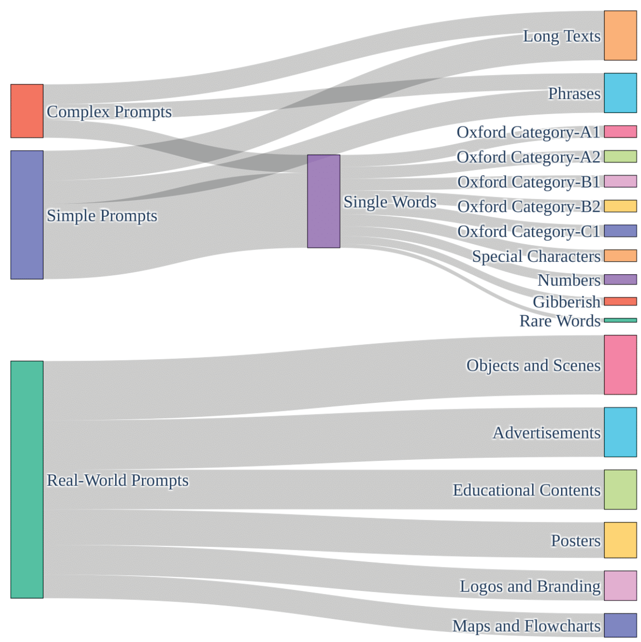

Figure 1. A comprehensive benchmark that enables separate analysis of prompt complexity and text attributes in evaluating T2I models. The benchmark features diverse prompt types—including Simple prompts (e.g., "a paper with the word 'text' written on it"), Complex prompts (e.g., "A sunny day at a theme park, a roller coaster with cars ascending the track, each car labeled with the text spelling out 'text', excited riders waving their hands"). It also incorporates a range of text attributes from single words to gibberish, numbers, special characters, phrases, and long texts.

1. Introduction

Denoising diffusion probabilistic models [7, 21, 24, 26] have significantly boosted the development of general text-to-image (T2I) generation, demonstrating the capability of generating surprisingly high-quality images over the past few years. Despite their impressive achievements, these models exhibit a notable limitation: difficulty in generating images that include specific visual text. Image generation with embedded text is a critical task with wide-ranging applications, from advertising, where the precise rendering of brand names on products can alter consumer perception, to educational resources, where accurate depictions of text in diagrams or illustrations can significantly impact learning outcomes. However, the State-Of-The-Art (SOTA) diffusion models often struggle with this task [22].

Several research efforts have attempted to address visual text rendering in image generation [1, 4, 6, 15, 27, 31]. Al-

*Corresponding author: `ffallah@asu.edu`

(a) Sample images generated using simple prompts and single words.

(b) Sample images generated using simple prompts and long texts.

(c) Sample images generated using complex prompts and single words.

Figure 2. Generated images from multiple T2I models demonstrating common limitations in accuracy, clarity, and legibility.

though these models significantly improve the fidelity of generated visual text, they generally fall short in rendering longer textual elements[2, 9, 30]. Figure 2 illustrates the challenges of accurate text rendering in different models. This limitation not only restricts the utility of these models but also highlights a gap in our understanding of how to effectively integrate textual and visual information in generated content. Moreover, existing models tend to default to generating common words or phrases, substituting user-specified unique terms, names, or phrases—often unique or novel—in the generated images. This behavior detracts from the personalized experience users seek and limits the applicability of these models in domains requiring high specificity and accuracy.

A significant obstacle in advancing this area is the absence of comprehensive benchmarks specifically designed to evaluate the ability of T2I models to generate images containing accurate and contextually relevant text. In this paper, we address this gap by introducing TextInVision, a large-scale benchmark tailored to evaluate the integration of visual text in image generation models. Our key contributions are as follows:

1. We present TextInVision, the first-of-its-kind visual text generation benchmark that mirrors real-world use cases, driven by word and prompt complexity. This benchmark emphasizes practical applications and challenges encountered outside controlled experimental settings, establishing a new standard for evaluating image generative models in dynamic environments.
2. We craft a diverse set of texts and prompts that consider various attributes and characteristics. Our methodology involves a carefully curated selection, systematically categorized to ensure a thorough examination of models' performance across a wide range of textual inputs. By independently varying prompt complexity and text attributes, this benchmark provides a detailed assessment of how well T2I models embed readable and meaningful text within different visual contexts.
3. We prepare an image dataset designed to test Variational Autoencoder (VAE) models across different character representations. This highlights challenges posed by VAE architectures in text generation within diffusion frameworks, an area less explored in existing literature.
4. Through comprehensive experiments using multiple open-source models, we rigorously dissect the T2I translation process to identify the origins of errors and evaluate their impact. Our key findings are: a) Word frequency impacts

performance less than previously assumed. b) The visual encoder is a major contributor to error. c) Rare words present substantial challenges to T2I models, with gibberish prompts unexpectedly outperforming rare words in preserving translation quality. d) Word length critically influences performance: models generate multiple short words more effectively than a single lengthy word.

2. Related work

Visual Text Generation Models Recent T2I diffusion models have significantly enhanced their text generation capabilities [3, 8, 10, 11, 16, 22, 32, 34]. In visual text generation, these models have been adapted to focus on generating more readable and visually clear text. For instance, Stable Diffusion (SD) trained with data and a powerful CLIP text encoder is proficient at generating high-quality images but often fail to generate precise and accurate visual text [20]. Building upon SD, SDXL [22] improves Stable Diffusion by using a 3× larger U-Net and a second text encoder and introducing another refinement model to improve image quality. According to [15], incorporating a Character-aware text encoder results in significantly fewer spelling errors compared to using Character-blind encoders. GlyphControl [31] improves the performance of SD models in generating accurate visual text by using additional glyph conditional information. In TextDiffuser [6], the process starts by extracting keywords from the text prompts to create a layout, which is then used to generate images by model from U-Net [23]. In TextDiffuser-2 [4], a language model for layout planning improves spelling accuracy through the use of fine-grained tokenizers, such as character and position tokens from [15]. Diff-Text [33] introduces a training-free framework for generating multilingual visual text images. It utilizes attention constraints and image-level constraint within the U-Net's cross-attention layer to improve the accuracy of visual text. AnyText [27] comprises a text-control diffusion pipeline with an auxiliary latent module and a text embedding module, which can generate multilingual visual text. These developments demonstrate the ongoing efforts to improve visual text generation within diffusion models. However, effectively evaluating these models requires robust benchmarks that can accurately assess their capabilities.

Visual Text Generation Evaluation The evaluation of T2I models is a crucial step in understanding their effectiveness and identifying areas for improvement. To thoroughly evaluate T2I models, the choice of the prompts is key, as it determines which abilities or skills of the models are being evaluated [30]. The MARIO-Eval [6] is a benchmark designed to assess text rendering quality in images, including a collection of prompts sourced from the MARIO-10M test set and other datasets. The LAION-Glyph [31] introduces two evaluation benchmarks, SimpleBench and CreativeBench. Words for these prompts are selected from Wikipedia and categorized into four frequency-based buckets. Gecko [30], although useful, is limited because it is not designed specifically for the text rendering task and lacks a comprehensive set of prompts to thoroughly evaluate models. LenCom-EVAL [13] is designed to assess the robustness of T2I models by challenging them with prompts that include special characters, numbers, and lengthy text. However, its effectiveness is constrained by a maximum prompt length of 17 words, and it is not publicly available, which limits broader access and comparative analysis.

Concurrently, TextAtlasEval [29] has introduced a long-text rendering dataset and benchmark, underscoring the growing emphasis on evaluating model performance in generating images that incorporate long-form text.Despite recent efforts, current benchmarks for T2I models often rely on unstructured and arbitrary selections of prompts, failing to systematically evaluate models capability on visual text generation. To address this gap, we introduce TextInVision, a comprehensive benchmark that evaluates models by separately analyzing varying levels of prompt complexity and a wide range of text attributes. By including an image dataset specifically designed to evaluate the VAE components of T2I models, TextInVision enables a comprehensive assessment of the models for text fidelity.

3. TextInVision benchmark

In this section, we introduce a novel benchmark for evaluating visual text generation, designed to evaluate their performance across varying levels of complexity from simple tasks to real-world scenarios. Our benchmark addresses practical applications and challenges beyond controlled settings, establishing a new standard for assessing these models. Table 1 presents a comparison between TextInVision attributes and other existing benchmarks used to evaluate the visual text rendering task. To create meaningful evaluation prompts, we gathered diverse real-world scenarios such as advertisements, promotional materials, and educational content. Recognizing that current models often struggle with complex prompts and integration issues, we included both intricate and simpler prompts in our benchmark. This approach allows us to incrementally track model performance and identify specific areas for improvement in model design and fine-tuning. TextInVision comprises over 50,000 methodically designed prompts, assessing models' capabilities across a broad spectrum of scenarios. By independently varying prompt complexity and text attributes, our benchmark provides a detailed assessment of how well T2I models embed readable and meaningful text within different visual contexts. Appendix A contains further details and multiple examples.

Table 1. Comparison of benchmark characteristics for evaluating visual text rendering tasks.

Features	MARIO-Eval	GlyphControl	LenCom-Eval	Gecko	TextInVision (Ours)
Number of Prompts	5414	400+	3000	1500	50000+
Visual Complexity	✗	✓	✓	✓	✓
Text Style	✗	✓	✓	✓	✓
Frequency in COCO & LAION Dataset	✗	✗	✗	✗	✓
Special Characters & Numbers	✗	✗	✓	✓	✓
Gibberish Text	✗	✗	✓	✓	✓
Text Complexity Category	✗	✗	✗	✗	✓
Text Length Category	✗	✗	✓	✗	✓
Public Availability	✓	✓	✗	✓	✓

3.1. Prompt selection

We employ a systematic approach to derive prompts from real-world scenarios, ensuring comprehensive testing across various text-in-image applications. Our selection process involves categorizing images into distinct scenarios, each reflecting typical contexts where text-inclusive images occur. Specifically, we define categories such as professional settings, commercial environments, and educational contexts. For instance, professional settings are represented by items like corporate reports and official documents, which assess the models' precision and formality in text presentation. Commercial signage, including store signs and safety notices, evaluates clarity and impact in retail and safety-critical environments. Educational materials test the models' ability to render informative text in diagrams and instructional images. In addition to these diverse and complex scenarios, we incorporate a selection of very simple prompts. By including simpler prompts, such as single-word labels or basic phrases, we evaluate the models' effectiveness in accurately rendering text with clarity and precision in less demanding contexts. This comprehensive approach ensures that we can pinpoint the models' strengths and weaknesses across different application domains.

3.2. Text selection

We choose different words and phrases from real images that contain text to ensure our prompts are realistic and relevant. We categorize the selected text into single words, phrases, and long text, with lengths varying from one word to several sentences. To assess the models' ability to handle text of varying complexity, we utilize the Oxford 5000 by CEFR level [19]. By including words across all language proficiency levels, from basic (A1) to advanced (C1), we test the models' performance on text that ranges in difficulty and importance. In addition to text complexity, we consider word frequency by drawing data from the COCO [14] and LAION [25] datasets. By selecting words that occur at high, medium, and low frequencies, we evaluate whether the models exhibit any bias toward commonly seen words or can handle rare words with equal proficiency. This analysis helps us understand how the frequency of words used during training influences model performance.

To further challenge the models, we incorporate inputs beyond plain alphabetic characters. This includes gibberish text, intentional misspellings, special characters (e.g., @, #, $), numbers, and long rare words. By including such varied text elements, we test whether the models can accurately reproduce the provided text exactly as given, including any errors or unconventional characters. This aspect is crucial for assessing the models' ability to handle precise input, which is important for applications requiring high fidelity and accuracy.

Finally, we vary the length of the text, measured by the number of letters or words in phrases or sentences. This variability allows us to test the models' performance as a function of different lengths. Short text demands precise placement and clarity in small spaces, while longer text challenges the models to maintain readability and aesthetic integration over larger areas. By ensuring that the models can effectively render text of any length, we address practical application needs where text length can vary significantly.

Having described the TextInVision benchmark, we transition to conducting experiments and analyses that leverage it to assess and comprehend the performance of T2I models.

4. Experimental design and results

This section details the experiments and analysis performed with the TextInVision Benchmark. We report results on end-to-end image generation methodologies, explore the role of the VAE component in setting performance upper bounds for diffusion models and evaluate the performance of visual text image generation.

4.1. Real-world scenarios prompts

We initiated our evaluation by utilizing real-world scenario prompts from six main categories (see Figure 3) and their respective subcategories. For detailed information, please refer to Appendix A.

To assess model performance, we analyzed the average edit distance of generated texts, employing OCR models (see Appendix B for details on the OCR and algorithms used). Categories with more diverse and complex prompts, such as advertisements, exhibit higher average edit distances compared to other categories like logos and branding.

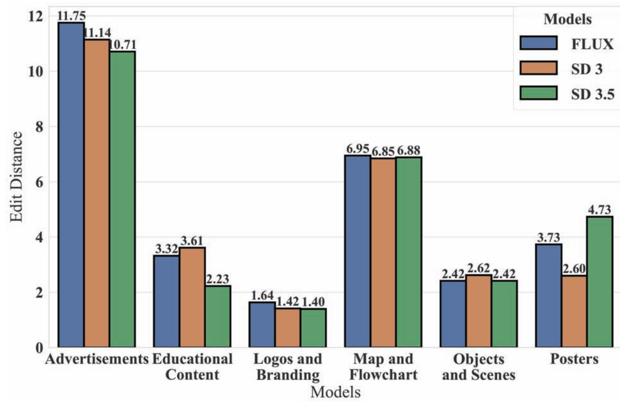

Figure 3. Average edit distance using real-world prompts.

The poor performance observed in the three SOTA T2I models prompted us to conduct a more granular analysis. We divided our analysis into two main areas: text and prompt. This involved testing single words, phrases, and long texts, alongside two sets of simple and complex prompts, to individually assess the models and identify their specific failure points. The subsequent sections delve deeper into these findings.

4.2. Impact of text & prompt complexity

We benchmark seven SOTA T2I models: Flux [3], SD 3 [8], SD 3.5 [8], SD-XL [18], DeepFloyd [12], AnyText [27], and Glyph Control [31]. We assess the models' capabilities in generating images with accurately integrated text from prompts. To this end, we randomly selected texts and prompts, covering various evaluation tasks. Preliminary tests with other models, such as SD 2.1 [20], Pixart-α [5], and TextDiffuser [6], yielded unsatisfactory results, and these were excluded from further analysis. The selected models demonstrated decent performance in text embedding, setting a clear benchmark for future use.

Word/text complexity Using the Oxford Dictionary's word difficulty classifications (A1 to C1), we analyzed the average edit distance across different word complexity groups using two sets of simple and complex prompts. The results show that although there is a slight increase in average edit distance from simpler words (A1) to more complex ones (C1), the differences are minimal. The observed pattern (see Figure 4) follows the sequence C1 > B2 > B1 > A2 > A1, indicating a subtle gradation in complexity. These findings suggest that the performance of the T2I models is not significantly affected by the single word complexity as defined by the Oxford classifications.

Building upon this, we extended our analysis to compare simple prompts with complex prompts (with single words) that include multiple descriptors, actions, or contexts. In this comparison, we observed a clear decrease in performance as the level of detail in the prompts increased. The added details in these prompts may introduce ambiguities or conflicting instructions, making it challenging for the model to determine which elements to prioritize in the generated image. Although T2I models are trained on vast datasets, they may still struggle with intricate linguistic structures or less common vocabulary found in complex prompts, leading to misinterpretations.

To ensure that the models were following the prompts, we tested the CLIP scores over complex prompts. As shown in Table 2, there are only slight differences between the CLIP scores of the models, even though they exhibit significant variations in terms of edit distances. This suggests that while the models may capture the general content of the prompts to a similar extent, the precision and quality of the generated images, as reflected by the edit distances, can differ substantially.

Consequently, this performance drop highlights the current limitations of T2I models in handling increased prompt complexity and underscores areas for future improvement in model training and architecture.

Word frequency analysis We hypothesized that model performance on text generation might correlate with the frequency of words in the training dataset. One of our hypotheses was to determine whether the model can recall, understand, or remember specific spellings and words from its training dataset. We compiled a list of frequent words from the COCO and LAION caption datasets to test this. We evaluated the models against this list to explore any potential relationship between word frequency and edit distances.

Our analysis revealed a slight correlation between word frequency and edit distances, as shown in Table 3. We observed a small negative correlation, indicating that edit distances tend to decrease slightly as word frequency increases. However, the correlation coefficient is small, suggesting this effect is not statistically significant. Thus, while a slight trend indicates better performance for more frequent words, word frequency alone is not a reliable predictor of a model's ability to generate text in images accurately.

Word/text length Given that text complexity, importance, and frequency do not significantly affect model performance, we focused on word length, analyzing words ranging from 3 to 16 characters (see Figure 5) and phrases and sentences spanning 2 to 70 words (see Figure 6).

Our results indicate that model performance degrades as word length increases. Interestingly, models performed better with two words totaling 16 characters than with a single 16-character word, suggesting that shorter segments are easier to process. Longer words introduce more opportunities for errors, as each additional character increases the likelihood of deviations from the expected text.

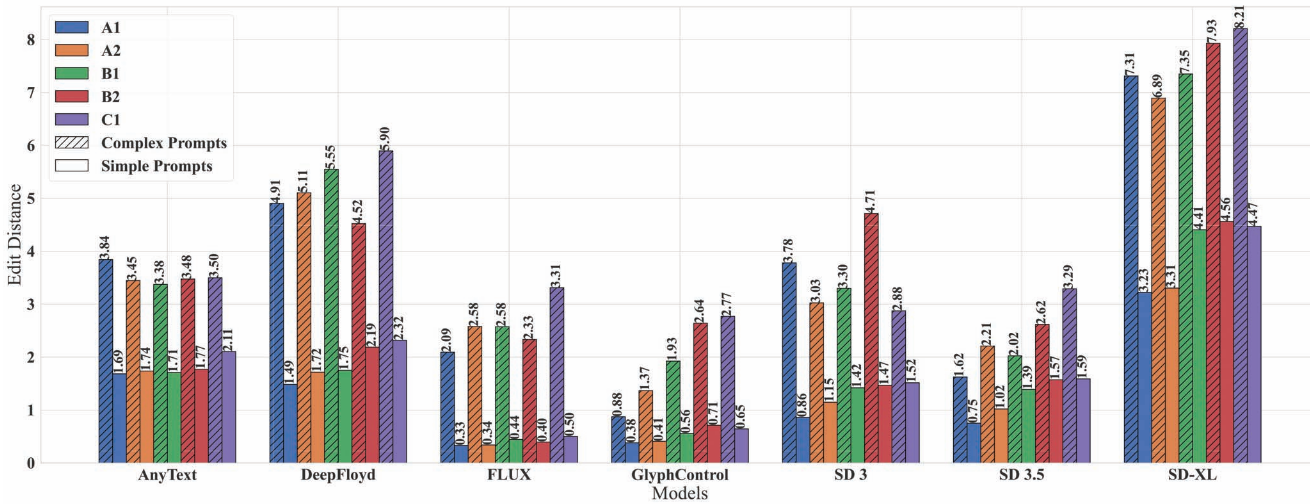

Figure 4. Average edit distance using different types of prompts and single words, compared across multiple models.

Table 2. CLIP score analysis among multiple T2I models utilizing complex prompts and single words.

Models	AnyText	DeepFloyd	FLUX	GlyphControl	SD 3	SD 3.5	SD-XL
CLIP Score	0.334	0.338	0.337	0.338	0.336	0.345	0.330

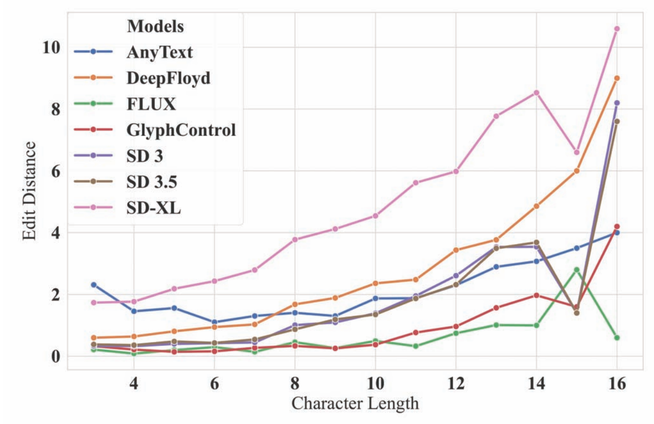

Figure 5. Longer input texts tend to have a higher average edit distance and lead to lower image quality or legibility, further compounding the difficulty in accurately extracting the text.

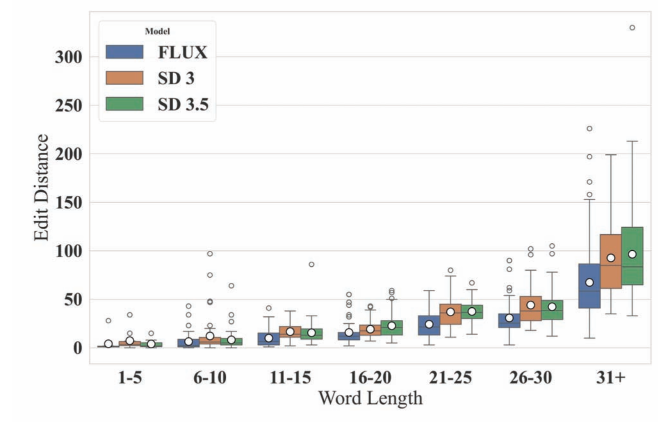

Figure 6. Average edit distance of best models across various text lengths, evaluated using simple prompts and long texts.

Although Figure 5 shows that Anytext and GlyphControl appear to outperform SD 3 and SD 3.5, this observation is misleading due to limitations in the evaluation metric. Anytext and GlyphControl often fail to generate long text, as a result, the edit distance metric may suggest better performance because shorter or incomplete outputs inherently result in a smaller edit distance, despite lacking semantic accuracy. For the analysis of longer texts, we focused on the three best-performing models, SD 3, SD 3.5, and Flux, that consistently generated coherent long texts to ensure a fair and meaningful evaluation of model performance on longer text generation tasks.

Misspelled text, numbers, and special characters To evaluate the robustness of T2I models, we crafted prompts that incorporated misspelled words, gibberish text, long rare words, special characters, and numbers. This experiment aimed to assess the models' ability to interpret and depict intended text despite deviations from standard spelling and syntax, thereby gaining insights into whether the models could recognize and correct inaccuracies or generate the text exactly as provided. Our results (see Figure 7) indicate that rare words pose significant challenges for the models. Al-

Table 3. The correlation between edit distance and word frequency across models.

Models	AnyText	DeepFloyd	FLUX	GlyphControl	SD 3	SD 3.5	SD-XL
COCO Dataset	-0.04	-0.07	-0.01	-0.05	-0.09	-0.08	-0.05
LAION Dataset	-0.08	-0.19	-0.04	-0.12	-0.19	-0.22	-0.20

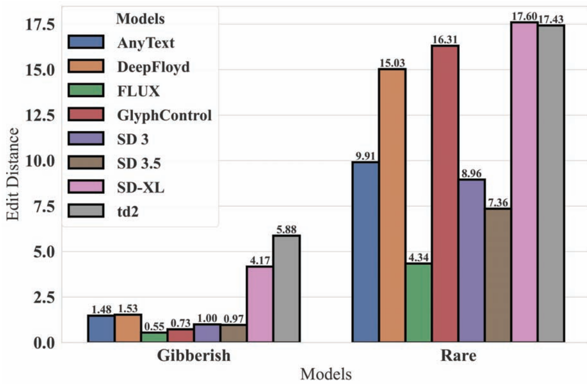

Figure 7. Average edit distance for categories rare and gibberish text, compared across multiple models.

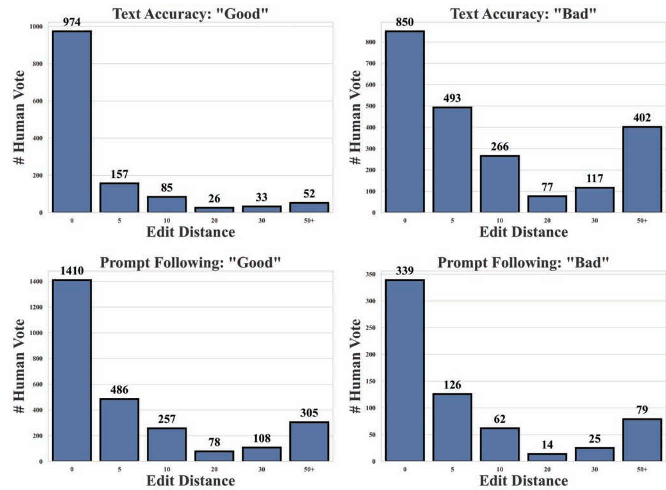

Figure 8. Aligning edit distance scores with human assessment standards.

though the models recognize these rare words as meaningful and attempt to generate them, they often fail to do so accurately. Conversely, models performed better with gibberish text, possibly because the lack of semantic pressure allows them to focus on visual coherence. These insights highlight the complexity of T2I models and the nuanced challenges models face with different types of textual input, which are crucial for improving training and rendering reliability.

4.3. Human evaluation

To complement the automated evaluation using OCR and edit distance, which provided a quantitative measure of the models' ability to generate text within images, we recognized the need for human insight to capture nuances that machines might miss. Therefore, we conducted a comprehensive human evaluation to assess how well the images generated by T2I models followed the prompts and produced accurate, clear text. We evaluated 1,000 randomly selected images, and each image was assessed by at least 2 human evaluators who answered two questions: whether the image follows the prompt (Prompt Following) and whether the text in the image is accurate and clear (Text Accuracy).

The correlation between edit distance and text accuracy is -0.20, indicating that as the edit distance increases, the accuracy and clarity of the text decrease. Figure 8 illustrates that with an increase in edit distance, the percentage of "good" counts for text accuracy drastically decreases. In contrast, the correlation between edit distance and prompt following is -0.06, suggesting a fairly weak negative trend. This finding corroborates our CLIP scores, implying that while the models can generate images that align with the prompts, they struggle to produce precise and legible text within those images. Further details including inter-rate agreement analysis are provided in Appendix C.

One factor behind these observations is the diversity of models used in our study. We see "Bad" text accuracy at low edit distances (<5) and "Good" accuracy at higher distances (>50) largely due to the range of models – from TextDiffuser to SD 3.5 and Flux. Some models fail even on single words, while others handle texts up to 70 words fairly well.

4.4. Impact of VAE performance

Building upon our analysis of how texts and prompts influence the performance of T2I models, it becomes evident that the interplay between textual input and model responsiveness is a critical determinant of successful image generation. Meanwhile, the generation process does not rely solely on prompts and model understanding of texts and prompts. The VAE plays a fundamental role in translating textual embeddings into coherent visual representations, particularly when embedding text within images. Understanding its functionality and limitations is essential to comprehensively evaluate the overall effectiveness of T2I models.

Thus, we hypothesized that the VAE component serves as a critical bottleneck affecting text fidelity in diffusion models. To validate this hypothesis, we conducted a comprehensive investigation into the VAE's role in text reconstruction. This section delineates our methodology, findings, and the profound implications of the VAE's performance on text rendering within generated images.

To thoroughly assess the effectiveness of VAE compo-

Ground Truth Image | Flux VAE Reconstructed | SD 2.1 VAE Reconstructed | SD 1.5 VAE Reconstructed

Figure 9. Decreases in quality of VAE reconstructed images leading to non-readable letters. Photo by Jeroen den Otter on Unsplash

nents in accurately reconstructing textual information, we selected VAEs from the SD 1.5, SD 2.1, and Flux models for evaluation. Our dataset comprised 1,000 images containing textual content sourced from diverse media such as advertisements, web pages, book covers, posters, and billboards. These images were obtained from internet sources ([17, 28] and featured a wide range of text characteristics, including varying font sizes, styles, colors, and alignments, all with a minimum resolution of 640 pixels in width.

Each image was encoded and decoded using its respective VAE to produce a reconstructed version. We then employed OCR models (see Appendix B for details on the OCR models) to extract textual content from both the original (ground truth) and the VAE-reconstructed images. By comparing the OCR-extracted texts, we focused on two critical metrics: word retention, which measures the proportion of exactly retained words, and partial accuracy, which assesses the similarity of partially correct words using letter-level edit distance. (See Appendix D for detailed word retention and partial accuracy calculations).

VAE is a Bottleneck: Our analysis revealed significant discrepancies in text retention and reconstruction accuracy among the VAEs. The SD 1.5 model exhibited a mean word retention rate of only 39%, while the SD 2.1 model demonstrated a modest improvement with a retention rate of 42%. In contrast, the Flux model achieved a retention rate of 51%, indicating better performance in maintaining whole words during reconstruction. These results suggest that all models struggle to preserve textual content effectively.

In terms of partial accuracy, the SD 1.5 model had an average letter edit distance of 127, and the SD 2.1 model slightly outperformed it with an average of 122. The Flux model, however, achieved 107. A smaller edit distance indicates a closer approximation to the original text at the character level; nonetheless, all models exhibited substantial deviations from the ground truth (see Figure 9).

These findings indicate that the VAE component acts as a significant bottleneck in diffusion models concerning text fidelity. The low word retention rates and high edit distances highlight the inadequacy of current VAEs in encoding and reconstructing textual information accurately. This inability to preserve text leads to considerable degradation in text quality within the reconstructed images.(See Appendix D for more examples).

Consequently, this bottleneck fundamentally limits the effectiveness of T2I models in generating images with clear and accurate text. The current performance of VAEs restricts the potential of diffusion models to produce high-quality images where text accuracy is paramount. Addressing the limitations of the VAE component is imperative for advancing the capabilities of diffusion models in text rendering. Future work should focus on developing more robust mechanisms within the model architecture to faithfully capture and reconstruct textual elements.

5. Conclusion

We present TextInVision benchmark, which establishes clear, specific criteria tailored to assess the unique challenges of integrating text into images. This includes evaluating the text's legibility, accuracy, and contextual relevance, ensuring a thorough performance analysis. The TextInVision Benchmark's prompts are chosen with specific purposes in mind, designed to evaluate the model's performance in realistic and relevant scenarios. Each category and subcategory of prompts is selected to mirror real-world applications. This strategic approach ensures that the evaluation is thorough and meaningful, providing insights into the practical capabilities of the models. While other benchmarks often fail to provide insights into why T2I models struggle with accurate text generation, TextInVision includes analyses highlighting potential bottlenecks in the text rendering process by examining different components, including the VAE. By providing a detailed assessment and pinpointing specific challenges, TextInVision serves as a crucial tool for advancing the field of visual text generation. We hope that this benchmark will guide future research efforts, fostering the development of more accurate and contextually relevant T2I models capable of seamlessly integrating text into images.

Acknowledgments

MP, AC, CB, and YY are supported by US NSF RI grant #2132724. We thank the NSF NAIRR initiative, the Research Computing (RC) at Arizona State University (ASU), and cr8dl.ai for their generous support in providing computing resources. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

[1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. Ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint*, arXiv:2211.01324, 2022.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science*, 2(3):8, 2023.

[3] Black Forest Labs. FLUX.1 [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024. Accessed: october, 2024.

[4] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. *arXiv preprint*, arXiv:2311.16465, 2023.

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[6] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. In *NeurIPS*, 2024.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, and Yam Levi et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint*, arXiv:2403.03206, 2024.

[9] Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. Llm blueprint: Enabling text-to-image generation with complex and detailed prompts. *arXiv preprint*, arXiv:2310.10640, 2023.

[10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020.

[12] DeepFloyd Lab. Deepfloyd if. `https://github.com/deep-floyd/IF`, 2023.

[13] Sanyam Lakhanpal, Shivang Chopra, Vinija Jain, Aman Chadha, and Man Luo. Refining text-to-image generation: Towards accurate training-free glyph-enhanced image generation. *arXiv preprint*, arXiv:2403.16422, 2024.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[15] Rosanne Liu, Daniel H. Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, R. J. Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. In *ACL*, 2022.

[16] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint*, arXiv:2112.10741, 2021.

[17] Pixabay. Pixabay. `https://pixabay.com/service/about/api/`, 2024. Accessed: 2024-04-01.

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

[19] Oxford University Press. The oxford 5000™ by cefr level. `https://www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000`, n.d.

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, arXiv:2204.06125, 2022.

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, arXiv:2204.06125, 2022.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, pages 234–241. Springer, 2015.

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022.

[25] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Aaron Mullis, Mehdi Cherti, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. `https://laion.ai/laion-400/`, 2021. Accessed: 2024-06-05.

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint*, arXiv:2010.02502, 2020.

[27] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint*, arXiv:2311.03054, 2023.

[28] Unsplash. Unsplash. `https://unsplash.com/license`, 2024. Accessed: 2024-04-01.

[29] Alex Jinpeng Wang, Dongxing Mao, Jiawei Zhang, Weiming Han, Zhuobai Dong, Linjie Li, Yiqi Lin, Zhengyuan Yang, Libo Qin, Fuwei Zhang, et al. Textatlas5m: A large-scale dataset for dense text image generation. *arXiv preprint arXiv:2502.07870*, 2025.

[30] Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Chris Knutsen, Cyrus Rashtchian, Jordi Pont-Tuset, et al. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *arXiv preprint*, arXiv:2404.16820, 2024.

[31] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. In *NeurIPS*, 2024.

[32] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.

[33] Lingjun Zhang, Xinyuan Chen, Yaohui Wang, Yue Lu, and Yu Qiao. Brush your text: Synthesize any scene text on images via diffusion model. In *AAAI Conference on Artificial Intelligence*, pages 7215–7223, 2024.

[34] Yiming Zhao and Zhouhui Lian. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. *arXiv preprint*, arXiv:2312.04884, 2023.