# Supplementary Material TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering

Jingye Chen<sup>\*13</sup>, Yupan Huang<sup>\* 23</sup>, Tengchao Lv<sup>3</sup>, Lei Cui<sup>3</sup>, Qifeng Chen<sup>1</sup>, Furu Wei<sup>3</sup> <sup>1</sup>HKUST <sup>2</sup>Sun Yat-sen University <sup>3</sup>Microsoft Research qwerty.chen@connect.ust.hk, {yupanhuang,tengchaolv,lecu,fuwei}@microsoft.com, cqf@ust.hk

# A Samples for training the layout planner

As shown in Figure 1, we demonstrate more data for training the layout planner, including samples with keyword specifications and samples without keyword specifications.

## B Choice of the maximum length L

During the training process, the composed sequence (*i.e.*, the prompt combined with text content and position) has a maximum length limit. As shown in Figure 2, by analyzing the MARIO-10M dataset [1], we notice that the composed sequence for 94.0% of the samples is less than 128 in length, and all samples are below this threshold during the evaluation. Obviously, we can increase the maximum length to a larger value, such as 256, accommodating 99.2% of training samples. However, enlarging the length limit would also result in increased computational costs, such as raising the single sample inference time from 6 to 7 seconds. Therefore, the choice of length limit should be made based on practical considerations, balancing between the model's capability and efficiency.

# C More visualizations of layout prediction

As depicted in Figure 3, we showcase more layout prediction results. We specify the keywords to be rendered in the first two rows. The language model has the capacity to organize the specified keywords, placing related keywords in the same line and generating aesthetically pleasing layouts. Notably, the final row of predictions includes words not present within the prompt. For instance, the model substitutes "200g" for "200gram". It is a logical substitution, given that both terms convey the same meaning. Additionally, the model replaces the misspelled term "RRAINBOW" in the prompt with the correct term "RAINBOW". This further showcases the flexibility of the layout planner  $M_1$ .

<sup>\*</sup> Work done during internship at Microsoft Research.

#### D Generation guided by single-point condition

We retrain TextDiffuser-2 and implement a single-point supervision strategy during the training process, such as using the center and top-left points. As illustrated in Figure 4, despite the diversity in text size and angle generated by the single-point conditions, we observe a significant portion of the text to be inaccurate. Given the observation of a decline in accuracy over 20% (as shown in Table 2 in the main paper), we ultimately employ the top-left and bottom-right points as the condition.

#### E Generation with additional angle condition

As shown in Figure 5, we demonstrate samples generated with different angle conditions. Specifically, we retrain TextDiffuser-2 and add 181 angle tokens, ranging from  $-90^{\circ}$  to  $90^{\circ}$ . When constructing the language-format layout, the angle token is placed after the four coordinate tokens. The results show that the generated results align well with the angle instructions.

#### F Details of compared methods and evaluation

We introduce all the baselines and their experimental settings as follows.

**SD 1.5** [14] is a latent diffusion model. It consists of a VAE to compress images to latent features and conducts the diffusion process in the latent space. The resolution of the output image is 512x512.

**SD 2.1** [14] is a refined version of SD 1.5. It uses a deeper backbone and enhances the generation resolution to  $768 \times 768$ .

**Deepfloyd** [4] uses a progressive process by iteratively enhancing the generation resolution to  $1024 \times 1024$ . Besides, it utilizes a more powerful text encoder, T5 [13], rather than the CLIP text encoder used in SD 1.5 and SD 2.1.

**StableCascade** [11] is a fast and efficient framework that progressively enhances the generation resolution to  $1024 \times 1024$ .

**SD-XL** [12] is an improved version of the latent diffusion model [14] with stronger backbone and powerful text embedding. The resolution of the output images is  $1024 \times 1024$ .

**PixArt-** $\alpha$  [2] is a powerful Transformer-based text-to-image diffusion model and is training-efficient. The output resolution is of size 1024×1024.

**Ideogram** [6] is an online website that can produce attractive logos, posters, and other natural images based on prompts. We use the typography mode and manually quote keywords to be rendered. The resolution is  $1024 \times 1024$ .

**DALLE-3** [3] exhibits robust text-to-image capabilities, producing images that precisely conform to the given prompt. It generates high-resolution outputs with  $1024 \times 1024$  resolution. We leverage the official API for the generation process.

**GlyphControl** [17] utilizes the framework of ControlNet [19] and the pretrained model of SD 2.1 [14], producing the output image of size  $768 \times 768$ . It takes glyph images with multiple text lines as the condition. Specifically, we re-train GlyphControl using the MARIO-10M dataset for six epochs for a fair comparison.

**TextDiffuser** [1] is a two-stage framework that can convert user prompts into images. It relies on users to specify keywords for rendering. TextDiffuser is pre-trained based on SD 1.5 [14], and the resolution of the generated images is  $512 \times 512$ . We use the tool provided by the authors to convert the generated text layouts into character-level segmentation masks.

**ControlNet** [19] appends an additional branch to the original stable diffusion architecture. ControlNet offers different kinds of control signals, including the canny map, segmentation map, depth map, etc. To employ ControlNet for the text rendering task, we utilize the checkpoint of canny map ControlNet, and use the glyph map produced by GlyphControl as the condition. We observe that ControlNet heavily relies on the contour of glyph images and generates text with limited diversity. We employ ControlNet to generate images with size  $512 \times 512$ . **Brush Your Text** [18] takes advantage of the ControlNet architecture and employs a training-free strategy guided by text region attention maps to render text images. Similar to ControlNet, Brush Your Text mainly depends on the given contour and lacks diversity regarding the text style. The produced image size is  $512 \times 512$ .

**AnyText** [16] employs two conditions, including the glyph images and text region masks for the text rendering process. AnyText generates an image with resolution  $512 \times 512$ .

For evaluation, we utilize the metrics employed in TextDiffuser [1] and also use Microsoft Read API to evaluate the OCR performance. We utilize the code provided by authors of TextDiffuser [1] to calculate the metrics. All compared methods are sampled with 50 steps for the generation process.

# G Incorporation of the proposed techniques to existing works

The proposed TextDiffuser-2 introduces two components, including the layout planning and layout encoding to improve the text rendering quality.

The layout planner can be easily incorporated into existing works. For example, TextDiffuser [1] can directly replace the original Layout Transformer with the proposed layout planner and obtain the character-level segmentation masks using the original strategy for the generation. Since GlyphControl [17] can not generate images from prompts, we use the layouts produced by TextDiffuser-2 to produce the glyph image using the code provided by the authors. As for AnyText [16], we first obtain the glyph image following GlyphControl guided by layouts generated by the proposed layout planner, and use the mask of rectangle box as the text region masks. For training-free ControlNet [19] and Brush Your Text [18], we convert the produced layouts into the glyph image generated by GlyphControl [17] as the condition.

The layout encoding technique can be naturally integrated into the training process for existing works. One only needs to add new position tokens and

character tokens to the original tokenizer, and appends the language-format layout to the original prompts. We conduct experiments with TextDiffuser [1], and GlyphControl [17] with available training scripts.

#### H Comparisons with samples in other papers

Since the source code, pre-trained weight, or demo is not available for Character-Aware Model [8] and GlyphDraw [9], we conduct comparisons with samples in their corresponding papers. As demonstrated in Figure 6, we visualize four samples for each compared method. Notably, TextDiffuser-2 shows better rendering accuracy compared with the Character-Aware Model, which contains several typos, including the missing "r" in "from" and the incorrect spelling of "Chimpanzees". Besides, the Character-Aware Model enhances visual text rendering by utilizing language models with a larger parameter size (*e.g.*, T5-XXL [13] with 11B parameters). We have demonstrated that even with a smaller-scaled CLIP text encoder with 63M parameters, superior text rendering performance can be achieved by virtue of explicit positional and content supervision. Besides, TextDiffuser-2 outperforms GlyphDraw as TextDiffuser-2 can render images with multiple text lines.

### I More details about text inpainting

As for the training process, similar to TextDiffuser [1], by appending another five-dimension feature, including the one-dimension mask and four-dimension non-inpainted area features, to the input of U-Net, TextDiffuser-2 can be trained for the text inpainting task. Specifically, 14,400 parameters will be added, which accounts for a small proportion of the whole architecture containing 922M parameters. We set the classifier-free guidance to 7.5 and used 50 sampling steps, which cost 6 seconds for generation using one A100 GPU card.

For the quantitative evaluation, we randomly sample 10,000 cases from the test set of MARIO-10M dataset [1]. For each sample, we randomly mask a text region and use the original text label as the inpainting guidance. To evaluate whether the inpainted result matches the given text label, we employ the OCR tool to determine if there is a detected text region where the intersection over union (IoU) with the mask area exceeds 0.7 and the text content corresponds accurately with the text label.

We also compare TextDiffuser-2 with another concurrent work UDiffText [20]. UDiffText employs the attention map guidance to enhance the text inpainting result. The visualizations are shown in Figure 7. It is observed that the proposed TextDiffuser-2 can produce text that is more harmonized with surrounding texts. Besides, the alphabet of UDiffText is limited since it can only tackle uppercase and lowercase characters. We also conduct the quantitative experiment, and the results reveal that the accuracy of TextDiffuser-2 outperforms UDiffText by a large margin (76.42% v.s. 69.13%).

# J More samples on text diversity

More visualizations about text diversity are shown in Figure 8. It is observed that images generated TextDiffuser-2 exhibit more diversity regarding the text style and text positions.

#### K More details about user studies

We conduct comprehensive user studies on four aspects, including text quality, style diversity, layout aesthetics, and inpainting ability.

- To evaluate the text quality, we compare our TextDiffuser-2 with six other methods that can perform text-to-image without other conditions. The six methods include SD-XL [12], StableCascade [11], Ideogram [6], Midjourneyv6 [10], DALLE-3 [3].
- To evaluate the style diversity, we compare TextDiffuser-2 with other five methods that can stably generate text, including ControlNet [19], Brush Your Text [18], TextDiffuser [1], GlyphControl [17], and AnyText [16].
- For evaluating the layout aesthetics, we conduct comparisons with TextDiffuser [1], which uses the Layout Transformer for the prediction.
- In order to evaluate the inpainting ability, we compare our method with two other text inpainting methods, including DiffSTE [7] and TextDiffuser [1].

The details of the questions are displayed in Figure 9. In addition to human involvement, we incorporate GPT-4V [5] in our user studies. Specifically, we design prompts to encourage GPT-4V to proceed step-by-step, deriving the final answer through logical analysis. The dialogue record is shown in Figure 10. It suggests that GPT-4V exhibits impressive literacy skills, and its logical chain is reasonable. The results are shown in Figure 11, indicating that GPT-4V favors TextDiffuser-2 in most cases.

#### L Natural image generation without text

As is shown in Figure 12, TextDiffuser-2 can indeed generate images without text even after fine-tuning with domain-specific datasets.

#### M Details about the compatibility to ControlNet.

The proposed TextDiffuser-2 can also incorporate ControlNet [19]. As shown in Figure 13, TextDiffuser-2 can take advantage of the pre-trained ControlNets, such as the canny map ControlNet and the depth map ControlNet, for generation. Besides, we can also train a text mask ControlNet to facilitate a more artistic text style.

#### N Generation based on overlapping layouts.

We visualize some samples generated with overlapping layouts in Figure 14. It is observerd that TextDiffuser-2 is more robust under this situation compared with other methods.

# O Conformity of the positional guidance

To verify whether the text generated in the images adheres to positional guidance, we employ a widely-used text detection tool [15] for assessment. The precision, recall and F-measure are 0.9524, 0.9635, and 0.9579, respectively. This indicates that in the vast majority of cases, the generated text is able to comply with the provided positional constraints.

# P Generation guided by quadrilateral boxes

In addition to using horizontal boxes to provide positional information, we remain curious whether TextDiffuser-2 can be guided by quadrilateral boxes, which could more accurately describe slanted text. To investigate this, we make two modifications. First, we train a layout planner **M1** to output each line in the format of "textline x0, y0, x1, y1, x2, y2, x3, y3". Secondly, we adapt the layout encoder **M2** to encode this sequence. We set the maximum length limit L to 256 to accommodate longer input sequences. Visualizations are shown in Figure 15. We notice that the generated results align well with the guidance of quadrilateral boxes. For future work, we plan to use more control points to represent the boxes, allowing for rendering more artistic text.

#### References

- Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. In: NeurIPS (2023) 1, 3, 4, 5
- Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023) 2
- DALLE-3: Link: https://openai.com/dall-e-3 (2023), https://openai.com/dalle-3 2, 5
- DeepFloyd: Github link: https://github.com/deep-floyd/if (2023), https://github.com/deep-floyd/IF 2
- 5. GPT-4: Link: https://openai.com/gpt-4 (2023), https://openai.com/gpt-4 5, 14
- 6. ideogram: Link: https://ideogram.ai/ (2023), https://ideogram.ai/ 2, 5
- Ji, J., Zhang, G., Wang, Z., Hou, B., Zhang, Z., Price, B., Chang, S.: Improving diffusion models for scene text editing with dual encoders. Transactions on Machine Learning Research (2024) 5
- Liu, R., Garrette, D., Saharia, C., Chan, W., Roberts, A., Narang, S., Blok, I., Mical, R., Norouzi, M., Constant, N.: Character-aware models improve visual text rendering. In: ACL (2023) 4, 11

- Ma, J., Zhao, M., Chen, C., Wang, R., Niu, D., Lu, H., Lin, X.: Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. arXiv preprint arXiv:2303.17870 (2023) 4, 11
- Midjourney-v6: https://www.midjourney-v6.com/ (2023), https://www.midjourney-v6.com/ 5
- Pernias, P., Rampas, D., Richter, M.L., Pal, C., Aubreville, M.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In: ICLR (2024) 2, 5
- Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 2, 5
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research (2020) 2, 4
- 14. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2, 3
- Sun, Y., Ni, Z., Chng, C.K., Liu, Y., Luo, C., Ng, C.C., Han, J., Ding, E., Liu, J., Karatzas, D., et al.: Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In: ICDAR (2019) 6
- Tuo, Y., Xiang, W., He, J.Y., Geng, Y., Xie, X.: Anytext: Multilingual visual text generation and editing. arXiv preprint arXiv:2311.03054 (2023) 3, 5
- 17. Yang, Y., Gui, D., Yuan, Y., Ding, H., Hu, H., Chen, K.: Glyphcontrol: Glyph conditional control for visual text generation. In: NeurIPS (2023) 2, 3, 4, 5
- Zhang, L., Chen, X., Wang, Y., Lu, Y., Qiao, Y.: Brush your text: Synthesize any scene text on images via diffusion model. arXiv preprint arXiv:2312.12232 (2023) 3, 5
- Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023) 2, 3, 5
- Zhao, Y., Lian, Z.: Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. arXiv preprint arXiv:2312.04884 (2023) 4

Input Prompt	Output Layout	Visualization
Sunshine Blogger Award - Sunflowers.	SUNSHINE 41,41,85,53 BLOGGER 41,55,84,68 AWARD 41,72,85,85	EUNSHINE RLOGGER AVIARD
google adwords.	Google 0,25,100,76 AdWords 6,76,55,97	Google AdWords
Macleod Cricket Club App club mix	MACLEOD 15,6,107,28 CRICKET CLUB 7,92,116,119	DRICKET CLUB
GRET Professionals for Fair Development	GRET 14,21,115,62 Professionals for 0,82,126,98 Fair Development 0,97,126,114	Polestinals for Fair Development
A movie poster with text The Adventures of Milo and Otis on it [Keywords] 'The Adventures Of', 'Milo and Otis'	The Adventures Of 29,1,98,8 Milo and Otis 0,7,127,26	In the second se
Twenty Eight. [Keywords] 'Twenty', 'Eight'	Twenty 22,33,89,74 Eight 41,50,106,95	Twenty Eight
Gems en Vogue Clearance Up to 65% OFF at ShopHQ. [Keywords] 'GEMS EN', 'VOGUE', 'Clearance', 'UP TO', '65% OFF'	GEMS EN 39,32,88,41 VOGUE 45,42,83,51 Clearance 35,55,92,71 UP TO 47,77,80,85 65% OFF 41,86,87,95	DEXIS EN VOQUE Disearance DP TO EXTR OFF
Best Boutique for Designer Gowns. [Keywords] 'Best Boutique', 'for Designer', 'Gowns'	Best Boutique 18,71,112,84 for Designer 23,85,107,98 Gowns 42,99,88,111	Best Boulique for Designer Boens

Fig. 1: Visualization of data for training the layout planner.



Fig. 2: The cumulative distribution function to analyze the length of the composed sequences. When setting the maximum length L at 128, the vast majority of samples (94%) are covered.



Fig. 3: More visualizations of the layout predictions. The specified keywords are marked in blue color in the first two rows.



**Fig. 4:** Visualizations of Generation guided by single-point conditions, including the center point and the top-left point. The prompt is "A text image of hello". The samples highlighted by red boxes in the last row denote the rendered text is incorrect.



Fig. 5: Visualizations of generation with different angle guidance. The prompt is "A text image of hello".



Fig. 6: Comparisons with Character-Aware Model (CA Model) [8] and GlyphDraw [9] using samples in their papers.



**Fig. 7:** Comparisons with UDiffText. It is observed that the proposed TextDiffuser-2 can render text that is more harmonized with surrounding texts.

12 Authors Suppressed Due to Excessive Length



A raccoon stands in front of the blackboard with the words "Deep Learning" written on it

Fig. 8: Visualizations on text diversity compared with other methods.



Fig. 9: The user studies comprise four questions. Specifically, users should make a single selection for questions pertaining to layout aesthetics, style diversity, and inpainting ability. For questions regarding text quality, users have the freedom to select multiple options.

8 Let's conduct a user study. I will show you an image and a question each time. Please start your response with 'Let's think step by step.' and end with 'Therefore, the answer is'. Please treat each case individually and do not be influenced by previous cases. If you have understo the requirements. Please answer "Yes" and we will start. Yes 8 hello Hello (a) (b) (d) (e) Which of the following images has the best text rendering quality? (Multiple Choice) Tips: Focus more on the accuracy of the text, and whether the text is coherent with backgrounds \$ The images that have the best text rendering quality, focusing on the accuracy of the text and its coherence with the backgrounds, are: (b) The text "Hello World" is clear, sharp, and the style matches the aesthetic of the image, creating a coherent look. (c) The text is artistically integrated with the image, matching the illustration style and providing a clear, coherent message. (g) The text is also clear and sharp, with a font style that fits well with the overall image, making it look coherent. Images (a), (d), (e), and (f) have text rendering that is either less clear, less coherent with the background, or both. For instance, the text in (f) is blurry and does not integrate well with the image. image Ø 🧛 純 🚊 . . Which set of images exhibits a richer style and more diverse text? (Single Choice) \$ The set of images that exhibits a richer style and position diverse text is set (f). In these images, the placement, orientation, and styling of the text "Snowman" vary more significantly compared to the other sets. The text is integrated into the scenes with different angles, curvatures, and Interactions with the elements of the images, such as being placed on the some sown being placed on the sow, behind the snowman, and even looking like it's hanging from the snowman's stick arm. This variety in the text styling and positioning makes set (f) stand out in terms of style and position diversity. 8 Prompt: A cat is reading a book with the title 'How to catch mice The above two figures represent layout planning based on prompt. Which figure best accomplishes this task? (Single Choice) \$ Let's think step by step. The prompt given is "A cat is reading a book with the title 'How to catch mice." We have two figures (a) and (b), that are meant to represent layout planning based on this prompt. Figure (a) shows a more consolidated layout where the title is in a single block, which could represent the cover of a book. Figure (b), however, breaks the title into separate blocks which seems more disjointed and less representative of a book title. For a book cover layout, which is what the prompt suggests, having a cohesive and clear title is important. Therefore, the answer is figure (a) best accomplishes this task. 8 (a) (b) (c) 06 nal Im sting Mack Text FREE FREE FREE FREE npt: Make Curve Tex Which image exhibits better result of text inpainting? (Single Choice) \$ Let's think step by step. We are comparing the results of text inpainting in images (a), (b), and Let's think step by step. We are comparing the results of text inpainting in images (a), (b), and (c) with the original image and the inpainting mask. Image (a) shows some misalignment in the text's curvature and spacing. The word "Curve" is also not following the circular path pecisely. Image (b) has the text "Curve" more accurately following the circular path, but the letters are not evenly spaced, and there is a noticeable break in the text flow. Image (c) has the text "Curve" accurately following the circular path, and the spacing between the letters is even and consistent with the rest of the text. Considering the consistency of the text curvature and the letter spacing, image (c) exhibits the best result of text inpainting because it maintains a consistent curvature and spacing that closely matches the original image's style. Therefore, the answer is image (c). answer is image (c).

Fig. 10: The illustration of using GPT-4V [5] to conduct user studies.



Fig. 11: The result of user studies conducted by GPT-4V.



Fig. 12: Visualizations of TextDiffuser-2 on generation withou text.



(b) Generation with trained text mask ControlNet.

Fig. 13: Visualizations of TextDiffuser-2 generated with pre-trained and trained ControlNets.



**Fig. 14:** Comparative visualizations of generation results using overlapping layouts. TextDiffuser-2 demonstrates enhanced robustness compared with other methods.



Fig. 15: Demonstration of TextDiffuser-2's generation guided by quadrilateral bounding boxes, showcasing the model's ability to align text accurately within the specified geometrical constraints.