Brush Your Text: Synthesize Any Scene Text on Images via Diffusion Model

Lingjun Zhang[1,2*†], **Xinyuan Chen**[2*], **Yaohui Wang**[2], **Yue Lu**[1‡], **Yu Qiao**[2]

[1]East China Normal University, Shanghai, China
[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China
51215904033@stu.ecnu.edu.cn, chenxinyuan@pjlab.org.cn, wangyaohui@pjlab.org.cn, ylu@cs.ecnu.edu.cn, yu.qiao@siat.ac.cn

Abstract

Recently, diffusion-based image generation methods are credited for their remarkable text-to-image generation capabilities, while still facing challenges in accurately generating multilingual scene text images. To tackle this problem, we propose **Diff-Text**, which is a *training-free* scene text generation framework for any language. Our model outputs a photo-realistic image given a text of any language along with a textual description of a scene. The model leverages rendered sketch images as priors, thus arousing the potential multilingual-generation ability of the pre-trained Stable Diffusion. Based on the observation from the influence of the cross-attention map on object placement in generated images, we propose a localized attention constraint into the cross-attention layer to address the unreasonable positioning problem of scene text. Additionally, we introduce contrastive image-level prompts to further refine the position of the textual region and achieve more accurate scene text generation. Experiments demonstrate that our method outperforms the existing method in both the accuracy of text recognition and the naturalness of foreground-background blending. Code: https://github.com/ecnuljzhang/brush-your-text.

Introduction

Minority languages, such as Arabic, Thai, and Kazakh, not only have a significant number (reaching 5000 to 7000), but their low-resource nature also impedes the progress of computer vision, particularly in the domain of image generation. In recent years, with the advancement of diffusion models (Ho, Jain, and Abbeel 2020), significant progress has been made in generating realistic and prompt-aligned images (Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022). However, achieving accurate scene text generation remains challenging due to the fine-grained structure within the scene text.

Recent efforts utilize diffusion models to overcome the limitations of traditional methods and enhance text rendering quality. For instance, Imagen (Saharia et al. 2022) and DeepFloyd (DeepFloydLab 2023) use the T5 series to generate text better. While these methods are capable of generating structurally accurate scene text, they demand a large amount of training data which is not suitable for minority languages and still lack control over the generated scene text. Some researchers (Wu et al. 2019; Yang, Huang, and Lin 2020; Lee et al. 2021; Krishnan et al. 2023) exploit GAN (Goodfellow et al. 2014) based scene text editing methods to generate scene text, which is more controllable. However, these methods are confined to generating scene text at the string level and do not possess the capability to generate complete scene compositions.

To tackle these challenges, we propose a training-free framework, Diff-Text, and a simple yet highly effective approach for multilingual scene text image generation. Our proposed framework inherits the off-the-shelf diffusion model while specializing in text generation by localized attention constraint method along with positive and negative image-level prompts. Specifically, given a text to be rendered, we first render it to a sketch image and then detect the edge map which is used as the control input of our model. Our model generates a realistic scene image according to the control input and the prompt input which contains a description of a scene. However, the control inputs are easily treated as grotesque patterns instead of texts on signs or billboards. Recent research (Hertz et al. 2022) suggests that the input prompts exert their influence on the object placement within the generated images via the cross-attention mechanism. Inspired by this observation, we first identify the keywords in the prompt that correspond to the textual region, such as "sign", "notice", and "billboard", and then constrain the cross-attention maps for these keywords to the textual region. Furthermore, we introduce a positive image-level prompt that further refines the placement of the textual region and a negative image-level prompt that enhances the alignment between the generated scene text and edge image, thereby ensuring greater accuracy in the generated scene text. Experiments demonstrate the effectiveness and robustness of our method.

Related Works

Scene Text Generation automates the creation of scene text images from provided textual content. Notably, SynthText (Gupta, Vedaldi, and Zisserman 2016) is widely used to train scene text recognition models. It employs existing models

*Equal Contribution.
†Work done as an intern at Shanghai AI Laboratory.
‡Corresponding author.

Figure 1: Diff-Text has the ability to generate accurate and realistic scene text images from a given scene text of any language along with a textual description of any scene.

to analyze images, identifies compatible text regions in semantically coherent areas, and places processed text using a designated font. Furthermore, SynthText3D (Liao et al. 2020) and UnrealText (Long and Yao 2020) generate scene text images from a virtual realm using a 3D graphics engine. However, these methods directly overlay text onto the background, resulting in artifacts in text appearing, which leads to a significant disparity between the synthesized and real image distributions. Some methods introduce GANs for realistic image generation. SF-GAN (Zhan, Zhu, and Lu 2019) introduces geometry and appearance synthesizers for realistic scene text generation, but struggles with accurate text placement. Scene text editing methods (Wu et al. 2019; Yang, Huang, and Lin 2020; Roy et al. 2020; Zhang et al. 2021; Lee et al. 2021; Xie et al. 2021; Krishnan et al. 2023; He et al. 2022) attempt tackle this problem. However, these methods concentrate only on generating the text region rather than the entire image.

Text-to-Image Generation represents a promising result that has seen significant progress in generating realistic and prompt-aligned images (Rombach et al. 2022; Ramesh et al. 2022; Saharia et al. 2022), as well as videos (Singer et al. 2023; Ho et al. 2022; Blattmann et al. 2023; Ge et al. 2023; Wang et al. 2023a; Chen et al. 2023b; Wang et al. 2023b), through the application of diffusion models (Ho, Jain, and Abbeel 2020). GLIDE (Nichol et al. 2022) introduces text conditions into the diffusion process using classifier-free guidance. DALL-E 2 (Ramesh et al. 2022) adopts a diffusion prior module on CLIP text latent and cascaded diffusion decoder to generate high-resolution images. Imagen (Saharia et al. 2022) emphasizes language understanding and proposes to use a large T5 language model for better semantics representation. Stable Diffusion (Rombach et al. 2022) is an open-sourced model that projects the image into latent space with VAE and applies the diffusion process to generate feature maps in the latent level.

In addition to text conditions, a realm of research explores controlling diffusion models through image-level conditions. Certain image editing methods (Meng et al. 2021; Kawar et al. 2023; Mokady et al. 2023; Brooks, Holynski, and Efros 2023) introduce images to be edited as conditions in the denoising process. Image inpainting (Balaji et al. 2022; Avrahami, Lischinski, and Fried 2022; Lugmayr et al. 2022; Bau et al. 2021) constitutes another type of editing method, aiming to generate coherent missing portions of an image based on a specified region while preserving the remaining areas. Additionally, SDG (Liu et al. 2023) represents an alternative approach involving extra conditions, which injects semantic input using a guidance function to direct the sampling process of unconditional DDPM. Some methods (Chen et al. 2023a; Ma et al. 2023) utilize textual layouts or masks as conditions for scene text generation. However, these approaches need extensive labeled datasets of scene text for training, which poses a challenge for low-resource languages.

Moreover, ControlNet (Zhang, Rao, and Agrawala 2023) and T2I-adapter (Mou et al. 2023) are dedicated to offering a comprehensive solution for controlling the generation process by leveraging auxiliary information like edge maps, color maps, segmentation maps, *etc*. These methods exhibit remarkable control and yield impressive results in terms of image quality. In this work, we perceive scene text generation as a text-to-image task with supplementary control (scene text) and incorporate the rendered scene text as an

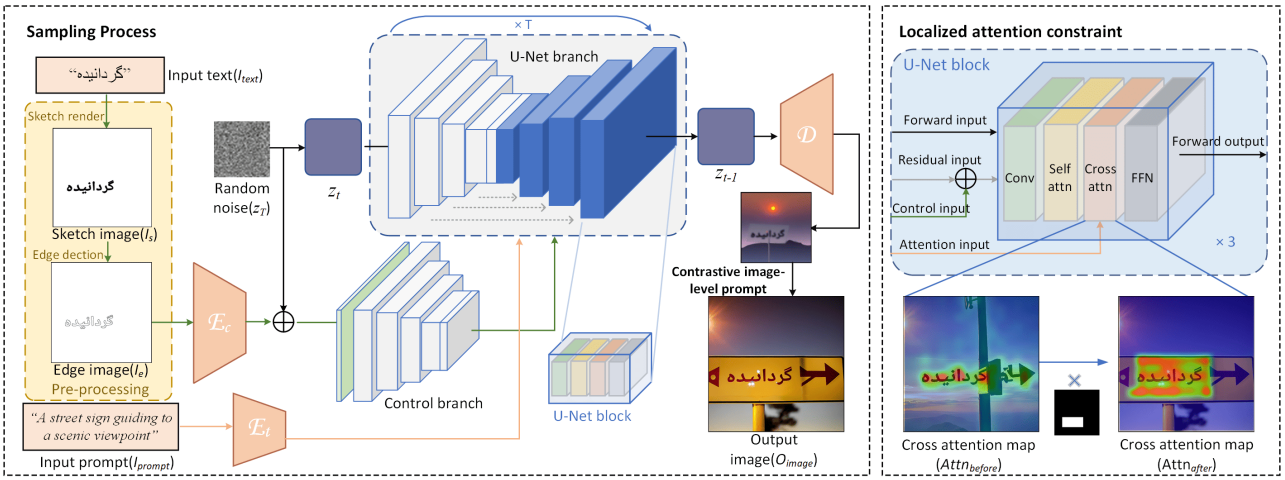

Figure 2: Our model employs input text (I_{text}) of any language to serve as the foreground element. The text is subsequently rendered into a sketch image, and its edges are detected to derive an edge image, which acts as an input of the control branch. Concurrently, our model takes in an input prompt (I_{prompt}) as the description of the background scene. After T denoising iterations, the model generates the final output image (O_{image}). Localized attention constraint and contrastive image-level prompts are employed in the U-Net block's cross-attention layer to enhance textual region positioning for precise scene text generation.

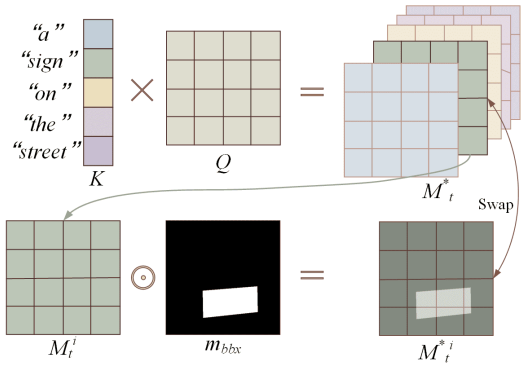

Figure 3: Details of the proposed localized attention constraint method. The "×" signifies matrix multiplication, while "⊙" denotes element-wise multiplication.

image-level condition within the diffusion model.

Methods

Overall Framework

We introduce a training-free scene text generation framework named Diff-Text, applicable to any language. Given an input text I_{text} and a prompt I_{prompt}, our proposed framework can generate scene text images that encompass: (1) precise textual content of I_{text}; (2) scenes that align with the provided prompt I_{prompt}; and (3) seamless integration of textual content with the depicted scenes. The architecture of our framework is presented in Fig. 2 and contains a pre-processing module, a U-Net branch, and a control branch.

Initially, the provided input text I_{text} undergoes pre-processing and is rendered into a sketch image denoted as I_s, depicting black text against a whiteboard backdrop with a randomly chosen font. Subsequently, the Canny edge detection algorithm is applied to derive an edge image denoted as I_e. This image, serving as an image-level condition, is then utilized as input for the control branch. Simultaneously, the provided input prompt I_{prompt} is processed by the text encoder, serving as a text-level condition. Under the guidance of both image-level and text-level conditions, the U-Net branch predicts the noise z_t at time t and utilizes z_t to reconstruct the output image from Gaussian noise.

Due to the independence of control input and prompt input for the U-Net network, there is a risk of incorrect fusion between image-level and text-level controls. For instance, the network might mistake the edges of the character "O" as part of a circular pattern. This issue is particularly prominent in the generation of scene text images for minor languages. To address this concern, we introduce a localized attention constraint method tailored for scene text generation. Simultaneously, to ensure a more rational fusion and enhance the precision of image-level control, we have proposed a contrastive image-level prompt. The localized attention constraint is utilized to confine the cross-attention maps associated with text region descriptors from the prompt input, such as "sign" or "billboard". These maps are limited to areas near the text through a pre-processing module that generates random bounding boxes. Regarding the contrastive image-level prompt, it comprises a positive image-level prompt and a negative image-level prompt.

Localized Attention Constraint

Our goal is to place scene text sensibly within scenes, such as on billboards or street signs. To achieve this, we introduce the localized attention constraint method. As shown in Fig. 3, during one forward pass at each timestep, we

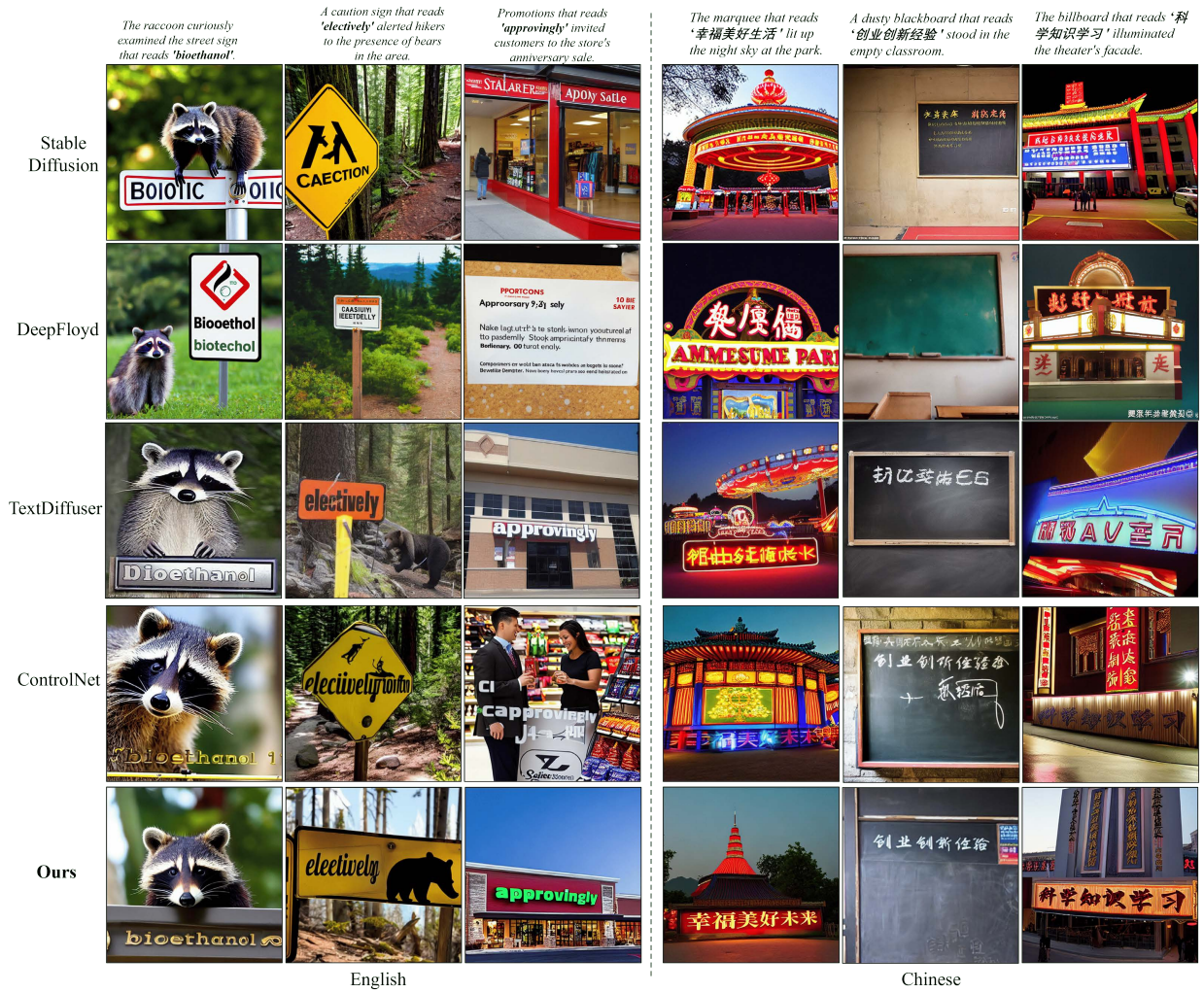

Figure 4: Visualizations of scene text generation in English and Chinese, compared with existing methods. The first three columns represent the generated results of English scene text, while the last three columns depict the generated results of Chinese scene text.

traverse through all layers of the diffusion model and manipulate the cross-attention map. The cross-attention map is denoted as $M_t \in R^{HW \times d_t}$, where HW refers to the width and height of z_t at different scales, and d_t represents the maximum length of tokens. In the framework, the positions of text within I_s are either user-specified or randomly placed, which means, obtaining the corresponding text bounding box is straightforward. We use this bounding box to derive a mask image of the text region, which we define as $m_{bbx} \in R^{H \times W}$. Then, assuming that the indices of tokens corresponding to words that may contain text in the prompt are represented by the set I, we resize m_{bbx} to HW and compute the new cross-attention map $M_t^* = \{\lambda \times M_t^i \odot m_{bbx} | \forall i \in I\}$. Finally, M_t^* is involved in the calculation of the z_{t-1}^*. After applying the localized attention constraint, we find a sensible and appropriate position to place the scene text. This approach also enhances the natural integration of foreground text with the background, resulting in a more realistic scene text generation.

Contrastive Image-level Prompts

The limited availability of images for minority languages within the training dataset of Stable Diffusion frequently results in the misinterpretation of edge images as object outlines. This misinterpretation often leads to the introduction of additional strokes, ultimately resulting in unrecognizable scene text generation. Indeed, the effectiveness of the localized attention constraint method depends on the presence of objects in the generated image that can accommodate the placement of text. In other words, if $M_t^i, i \in I$ approaches 0 and M_t^* remains the same as M_t, the localized attention constraint will not yield the desired output.

To tackle this issue, we introduce the definition of contrastive image-level prompt. In this regard, we consider the edge image I_e as the foundation of the image-level prompt, which we extend into a positive image-level prompt (PIP)

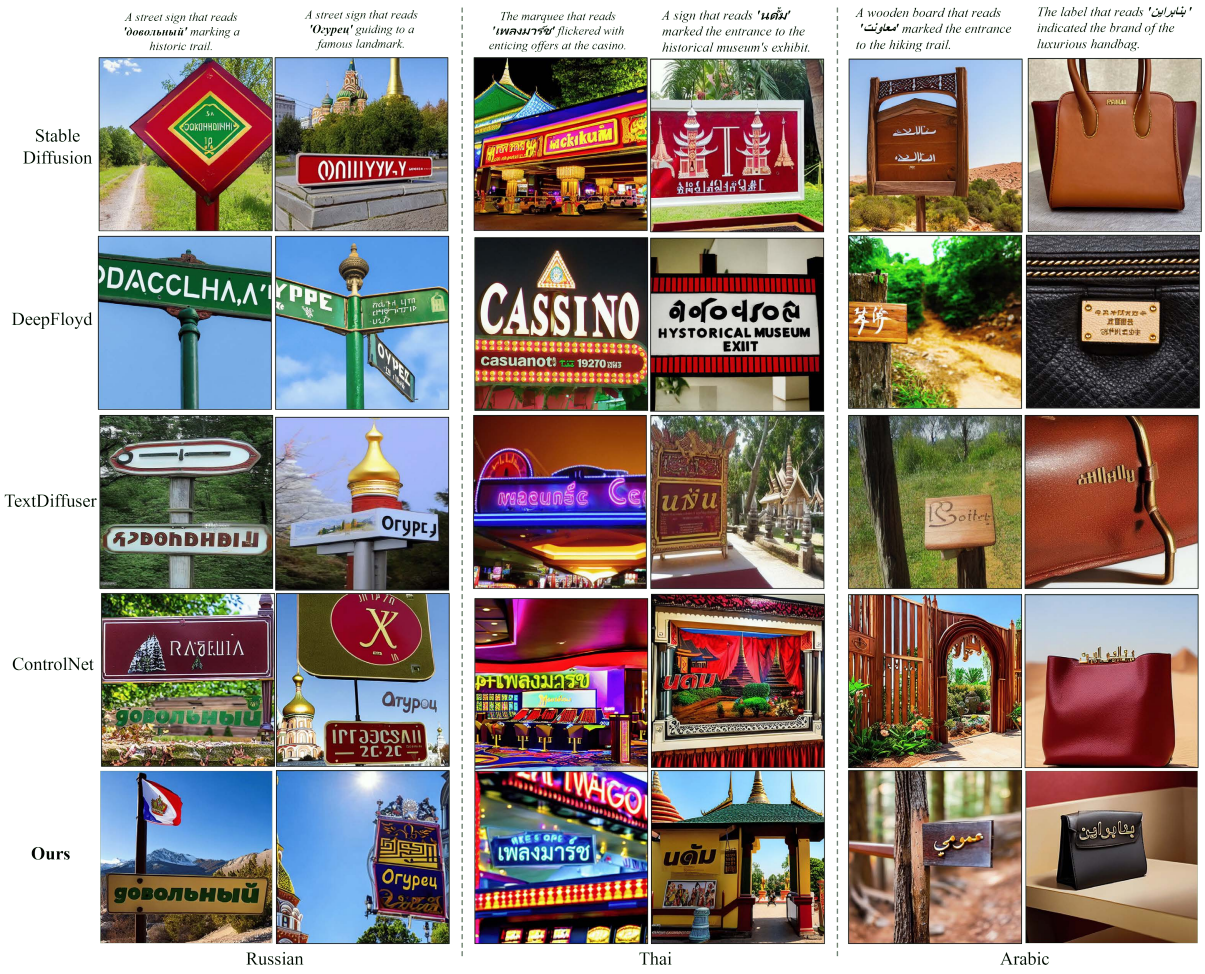

Figure 5: Visualizations of scene text generation in Russian, Thai, and Arabic, compared with existing methods. The first and second columns present the results of Russian scene text, the third and fourth columns depict Thai scene text, and the final two columns illustrate Arabic scene text.

and a negative image-level prompt (NIP). The edge image for PIP is the original edge image incorporating the depiction of a bounding box, while the sketch image for NIP is purely white. These two conditional inputs, denoted as $I_e^{'}$ and $\varnothing$, respectively, serve as the basis for the contrastive image-level prompt. They are then incorporated into the denoising process through the following equation:

$$
\begin{aligned}
&z_{t-1} = \widetilde{\epsilon}(z_t, I_e, I_{prompt}) \\
&= \epsilon(z_t, \varnothing, \varnothing) + s_{cfg}(\epsilon(z_t, \varnothing, I_{prompt}) - \epsilon(z_t, \varnothing, \varnothing)) \\
&+ s_{neg}(\epsilon^{'}(z_t, I_e, I_{prompt}) - \epsilon(z_t, \varnothing, I_{prompt})), \\
&\epsilon^{'}(z_t, I_e, I_{prompt}) = \epsilon(z_t, I_e, I_{prompt}) \\
&+ s_{pos}(\epsilon(z_t, I_e^{'}, I_{prompt}) - \epsilon(z_t, I_e, I_{prompt})),
\end{aligned} \tag{1}
$$

where s_{cfg} and s_{neg} are used to finely adjust the respective effects of the PIP item and NIP item on the predictions, which will be discussed in our ablation study (see Fig. 6). PIP provides a subtle hint to the network, compelling it to include objects suitable for placing scene text in the generated image. On the other hand, NIP is used to control the clarity and visibility of the scene text. Through this contrastive image-level prompt, we provide the model with both a negative direction and a positive direction which enables the model to generate clear and precise scene text while maintaining a rational background.

Experiments

Implementation Details

Experimental Settings Our model is built with Diffusers. The pre-trained models are "runwayml/stable-diffusion-v1-5" and "lllyasviel/sd-controlnet-canny". While predicting, the size of the output images is 512×512. We use one A100 GPU for inference. The localized attention constraint is applied in both the U-Net branch and the control branch. The λ in the localized attention constraint is 6.0. The s_{cfg}, s_{neg} and s_{pos} are respectively 7.5, 2.0 and 0.1. The wordlist for localized attention constraint includes "sign", "billboard",

Language	Metrics	Stable Diffusion	DeepFloyd	TextDiffuser	ControNet	Ours
Arabic	CLIPScore	0.7961	0.7335	0.8084	0.8067	**0.8138**
	Accuracy	0.000	0.000	0.000	3.291	**33.13**
	Edit_accuracy	16.65	13.17	11.58	34.80	**72.93**
Thai	CLIPScore	0.7733	0.7926	0.7873	0.8059	**0.8164**
	Accuracy	0.000	0.000	0.000	7.160	**38.41**
	Edit_accuracy	10.70	14.51	11.64	36.34	**82.97**
Russian	CLIPScore	0.7948	0.8201	0.8335	0.8306	**0.8632**
	Accuracy	0.000	0.000	1.375	9.790	**39.29**
	Edit_accuracy	14.60	26.05	37.72	39.21	**80.58**
English	CLIPScore	0.7879	0.8658	**0.8666**	0.7334	0.8649
	Accuracy	0.083	16.67	43.91	12.88	**61.03**
	Edit_accuracy	32.75	66.20	84.84	40.04	**89.52**
Chinese	CLIPScore	0.8265	0.8347	0.8201	0.8312	**0.8351**
	Accuracy	0.000	0.000	0.000	5.875	**32.40**
	Edit_accuracy	3.890	6.830	9.598	26.81	**68.75**

Table 1: Quantitative comparison with existing methods across five languages. The bold numbers represent the best results among all compared methods.

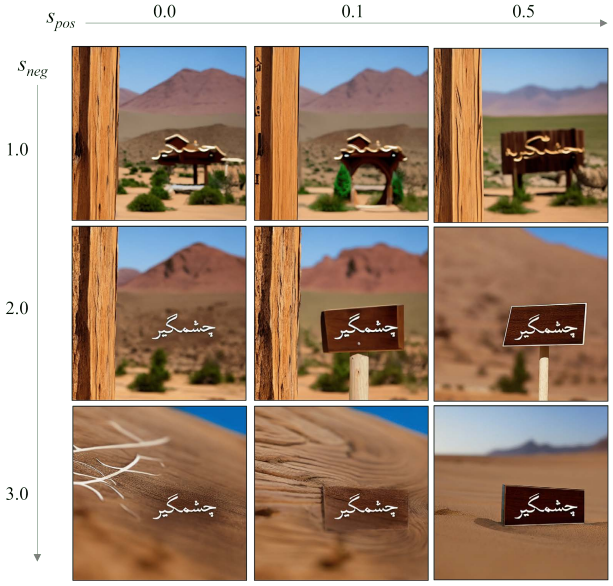

Figure 6: The image-level prompt comprises both positive and negative components, denoted as s_{pos} and s_{neg}, respectively. s_{pos} controls the intensity of "sign" occurrences in the background, while s_{neg} controls the clarity of the scene text.

and so on. More details are shown in Appx. in https://arxiv.org/abs/2312.12232.

Evaluation Due to the lack of publicly available multilingual benchmarks, we use multilingual vocabularies in the work of Zhang et al. (Zhang et al. 2021) and Xie et al. (Xie et al. 2023) as the input texts and generate corresponding input prompts using chatGPT (Ouyang et al. 2022). We select five languages and filter out words with fewer than five characters. From the remaining set, we randomly choose 3000 words for each language. Ultimately, we generate 15,000 multilingual images for evaluation for each comparative method. We conduct both quantitative and qualitative comparative experiments. In the quantitative comparison, we utilize three metrics: CLIP Score (Hessel et al. 2021; Huang et al. 2021; Radford et al. 2021), accuracy, and normalized edit distance (Shi et al. 2017). To ensure equitable capabilities across all languages for OCR tools, we use a multilingual OCR, namely easy-OCR (JadedAI 2020).

Comparison with Existing Methods

In this subsection, we compare our method with existing open-source methods capable of scene text generation, *i.e.,* Stable Diffusion (Rombach et al. 2022), DeepFloyd (DeepFloydLab 2023), TextDiffuser (Chen et al. 2023a) and ControlNet (Zhang, Rao, and Agrawala 2023). DeepFloyd uses two super-resolution modules to generate higher resolution 1024×1024 images compared with 512×512 images generated by other methods. We employ the template-to-image mode of the TextDiffuser method and utilize our sketch image as the template image.

Quantitative Comparison In the quantitative comparison, we selected the following three metrics: (1) **CLIPScore** is used to measure the similarity between the generated images and the input prompts. (2) **Accuracy evaluation** employs OCR tools to detect and calculate the recognition accuracy to assess whether the scene text in the generated images matches the input text. (3) **Normalized edit distance** is used to compare the similarity between the scene text in the generated images and the input text. We demonstrate the quantitative results compared with existing methods in Table 1. As shown in Table 1, Although training-free, our method still achieves a competitive CLIP score and significantly enhances the recognition accuracy of generated images. For each specific language, our method demonstrates an average improvement in accuracy of 25% compared to the existing method.

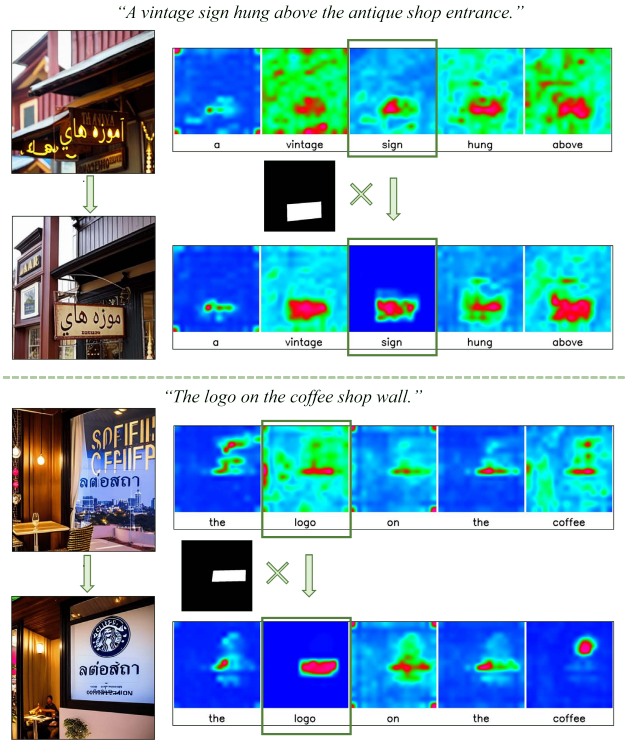

Figure 7: Visualization of ablation experiments on the localized attention constraint method. The heatmaps illustrate the average cross-attention map corresponding to different tokens across all diffusion steps.

Qualitative comparison Fig. 4 and Fig. 5 show the comparison between our method and existing methods in generating scene text images for majority and minority languages, respectively. From Fig. 4, it can be observed that for English, which has a significant presence in the training dataset of existing methods, the generated images possess a certain level of recognizability. However, Stable Diffusion and DeepFloyd may exhibit instances of generating multiple or missing characters. TextDiffuser, with the sketch image as an input template, addresses the issue of multiple and missing characters. Nevertheless, due to insufficient strictness in control, TextDiffuser still encounters problems with erroneous character generation. Despite utilizing edge images for strict control, ControlNet still results in the generation of scene text appearing in unreasonable positions or having additional strokes. In contrast, our method can generate clear, precise, and reasonably positioned scene text. For the languages with a smaller presence in the training dataset (Chinese, Arabic, Thai, Russian), Stable Diffusion, DeepFloyd, and TextDiffuser fail to generate recognizable scene text. TextDiffuser may generate some English letters instead of similar characters from other languages. ControlNet still encounters issues of generating text in unreasonable positions, and when dealing with characters resembling special patterns, such as Arabic characters, ControlNet merges the text with the background, rendering the generated text unidentifiable. Our method, on the other hand, successfully generates scene text images for all languages.

Method	CLIP	Accuracy	Edit accuracy
W/o constraint	0.8065	31.42	74.30
W/o PIP	0.7935	27.68	70.92
W/o NIP	0.7718	10.39	50.95
Full model	**0.8108**	**35.48**	**77.22**

Table 2: Quantitative ablation studies on localized attention constraint and image level prompt. "W/o constraint" denotes the exclusion of the localized attention constraint method, "W/o NIP" denotes the exclusion of the negative image-level prompt, and "W/o PIP" denotes the exclusion of the positive image-level prompt. The results indicate that our full model achieves the best generation results.

Ablation Study

To validate the effectiveness of the proposed localized attention constraint and contrastive image-level prompt, we conduct the ablation study. Table 2 presents the quantitative analysis of the ablation experiments. As demonstrated in Table 2, it is evident that the full model achieves the best performance in both the CLIP score and the accuracy of generated characters. In addition, we also conduct qualitative analysis for the ablation study, and the results are presented in Fig. 6 and 7. The seed is fixed at 2345 to generate visualized results. In Fig. 6, we discuss the impact of different parameters for PIP and NIP (i.e., s_{pos} and s_{neg}) on the generated images. From Fig. 6, it can be observed that as s_{pos} increases, the sign in the background becomes more prominent, but excessively high s_{pos} can cause the sign to appear too pronounced and flat. On the other hand, as s_{neg} increases, the scene text in the foreground becomes clearer, but excessively high s_{neg} can result in scene text floating in unreasonable positions. Fig. 7 showcases the results with and without our localized attention constraint. It can be observed that when we constrain the cross-attention map corresponding to the "sign" and "logo" to the scene text region, the generated images appear more reasonable and realistic.

Discussion and Conclusion

Currently, the bounding box of the text region is obtained either through user specification or random generation, and the tokens in the prompt that require localized attention constraint are determined by manually given wordlists. In future work, it is possible to integrate these two parts with GPT4 API for a more rational selection of bounding boxes and wordlists. However, our model still faces challenges in generating small-scale scene text and achieving precise text color control. Moreover, the generated scene text still occasionally includes unintended textual elements.

In this paper, we introduce a training-free framework, named Diff-Text. This framework is designed to apply to scene text generation in any language. Localized attention constraint method and contrastive image-level prompt are proposed to enhance the precision, clarity, and coherence of generated scene text images.

Acknowledgements

This work was jointly supported by the National Natural Science Foundation of China under Grant No. 62102150, No. 62176091, the National Key Research and Development Program of China under Grant No. 2020AAA0107903.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18208–18218.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Bau, D.; Andonian, A.; Cui, A.; Park, Y.; Jahanian, A.; Oliva, A.; and Torralba, A. 2021. Paint by word. *arXiv preprint arXiv:2103.10951*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2023a. TextDiffuser: Diffusion Models as Text Painters. *arXiv preprint arXiv:2305.10855*.

Chen, X.; Wang, Y.; Zhang, L.; Zhuang, S.; Ma, X.; Yu, J.; Wang, Y.; Lin, D.; Qiao, Y.; and Liu, Z. 2023b. SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction. *arXiv preprint arXiv:2310.20700*.

DeepFloydLab. 2023. DeepFloyd IF. https://github.com/deep-floyd/IF. Accessed: 2023-1-20.

Ge, S.; Nah, S.; Liu, G.; Poon, T.; Tao, A.; Catanzaro, B.; Jacobs, D.; Huang, J.-B.; Liu, M.-Y.; and Balaji, Y. 2023. Preserve your own correlation: A noise prior for video diffusion models. *arXiv preprint arXiv:2305.10474*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2315–2324.

He, H.; Chen, X.; Wang, C.; Liu, J.; Du, B.; Tao, D.; and Qiao, Y. 2022. Diff-Font: Diffusion Model for Robust One-Shot Font Generation. *arXiv preprint arXiv:2212.05895*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, Y.; Xue, H.; Liu, B.; and Lu, Y. 2021. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1138–1147.

JadedAI. 2020. EasyOCR. https://github.com/JaidedAI/EasyOCR. Accessed: 2020-3-14.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Krishnan, P.; Kovvuri, R.; Pang, G.; Vassilev, B.; and Hassner, T. 2023. Textstylebrush: Transfer of text aesthetics from a single example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lee, J.; Kim, Y.; Kim, S.; Yim, M.; Shin, S.; Lee, G.; and Park, S. 2021. Rewritenet: Realistic scene text image generation via editing text in real-world image. *arXiv preprint arXiv:2107.11041*, 1.

Liao, M.; Song, B.; Long, S.; He, M.; Yao, C.; and Bai, X. 2020. SynthText3D: synthesizing scene text images from 3D virtual worlds. *Science China Information Sciences*, 63: 1–14.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2023. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 289–299.

Long, S.; and Yao, C. 2020. Unrealtext: Synthesizing realistic scene text images from the unreal world. *arXiv preprint arXiv:2003.10608*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Ma, J.; Zhao, M.; Chen, C.; Wang, R.; Niu, D.; Lu, H.; and Lin, X. 2023. GlyphDraw: Learning to Draw Chinese Characters in Image Synthesis Models Coherently. *arXiv preprint arXiv:2303.17870*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images

using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Roy, P.; Bhattacharya, S.; Ghosh, S.; and Pal, U. 2020. STEFANN: scene text editor using font adaptive neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13228–13237.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Shi, B.; Yao, C.; Liao, M.; Yang, M.; Xu, P.; Cui, L.; Belongie, S.; Lu, S.; and Bai, X. 2017. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th iapr international conference on document analysis and recognition (ICDAR)*, volume 1, 1429–1434. IEEE.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; Parikh, D.; Gupta, S.; and Taigman, Y. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *ICLR*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023a. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. *arXiv preprint arXiv:2309.15103*.

Wang, Y.; Ma, X.; Chen, X.; Dantcheva, A.; Dai, B.; and Qiao, Y. 2023b. LEO: Generative Latent Image Animator for Human Video Synthesis. *arXiv preprint arXiv:2305.03989*.

Wu, L.; Zhang, C.; Liu, J.; Han, J.; Liu, J.; Ding, E.; and Bai, X. 2019. Editing text in the wild. In *Proceedings of the 27th ACM international conference on multimedia*, 1500–1508.

Xie, Y.; Chen, X.; Sun, L.; and Lu, Y. 2021. DG-Font: Deformable Generative Networks for Unsupervised Font Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5130–5140.

Xie, Y.; Chen, X.; Zhan, H.; and Shivakum, P. 2023. Weakly Supervised Scene Text Generation for Low-resource Languages. *arXiv preprint arXiv:2306.14269*.

Yang, Q.; Huang, J.; and Lin, W. 2020. Swaptext: Image based texts transfer in scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14700–14709.

Zhan, F.; Zhu, H.; and Lu, S. 2019. Spatial fusion gan for image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3653–3662.

Zhang, L.; Chen, X.; Xie, Y.; and Lu, Y. 2021. Scene Text Transfer for Cross-Language. In *International Conference on Image and Graphics*, 552–564. Springer.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.