Supplementary Material of DreamText: High Fidelity Scene Text Synthesis

Yibin Wang[1,2], **Weizhong Zhang**[1,4], **Honghui Xu**[5], **Cheng Jin**[1,3†]
[1]Fudan University [2]Shanghai Innovation Institute
[3]Innovation Center of Calligraphy and Painting Creation Technology, MCT
[4]Shanghai Key Laboratory of Intelligent Information Processing [5]Zhejiang University of Technology
yibinwang1121@163.com, weizhongzhang@fudan.edu.cn, xhh@zjut.edu.cn, jc@fudan.edu.cn

A. Experimental Settings for Baselines

In this study, we conduct a comprehensive comparison of our method against several state-of-the-art baselines, encompassing both GAN-based and diffusion-based approaches. Specifically, we evaluate our method against MOSTEL [3], Stable Diffusion-inpainting (v2.0) [4], DiffSTE [2], TextDiffuser [1], AnyText [5] and UDiffText [6]. For MOSTEL, we utilize it to generate text within the masked region and then integrate the output back into the original image. Regarding Stable Diffusion, we employ the publicly available pre-trained model "stable-diffusion-2-inpainting" from Hugging Face, setting its prompt as "[word to be rendered]" for a fair comparison. For TextDiffuser, we utilize their inpainting variant, where the desired text is rendered in a standard font (Arial) within the masked region, serving as input for their proposed segmentor. Finally, for UDiffText, AnyText, and DiffSTE, we follow the settings outlined in their respective original papers.

B. Ablation Study on Balanced Supervision

We analyze the efficacy of our balanced supervision for character attention by comparing it with unsupervised and self-supervised learning approaches. We additionally incorporate the mean Intersection over Union (mIoU) metric to evaluate the alignment between their latent character masks and ground truth character segmentation masks on the LAION-OCR dataset. As illustrated in Tab. 1, when conducting unsupervised learning, we observed a significant deterioration in the model's performance across all metrics. This outcome is reasonable since the model lacks prior knowledge about determining the ideal character locations at the outset. Consequently, it fails to concentrate attention effectively around character regions, leading to significant deviations between the characters' positions in generated latent character masks and their actual positions. As a result, the autonomous alternate optimization process is adversely affected.

Furthermore, we explore self-supervised learning by guiding attention calibration using the cross-entropy objective between latent character masks and character segmentation masks. Although self-supervised learning demonstrates a notable improvement compared to unsupervised learning, there remains a gap when compared to our balanced supervision. This discrepancy arises from the overly strong constraint of the characters' position, limiting the model's flexibility in estimating optimal positions, which hinders its ability to adapt to varied and complex scenarios.

Table 1. Ablation study results on balanced supervision.

Setting	*Average SeqAcc*		FID	mIoU
	Recon	Editing		
unsupervised	0.212	0.157	62.36	0.203
supervised	0.862	0.813	14.92	0.617
balanced supervision (**Ours**)	**0.940**	**0.887**	**12.13**	**0.722**

C. Human Study

We conduct a human study to compare our method with UdiffText. The results are visualized in the accompanying figure. A total of 50 cases were prepared, with each case generating four images using both methods to evaluate diversity. Additionally, one image per case was randomly selected for quality assessment. We report the percentage of queries receiving positive votes, with a black box highlighting the cases where the majority consensus was achieved.

D. Additional Visualized Attention Results

Additional visualized attention cases of all characters across several steps during training are exhibited in Fig. 1.

E. Additional Qualitative Comparative Results

Additional qualitative comparisons against the baselines are exhibited in Fig. 3 and 4.

Figure 1. Additional visualized attention cases of all characters across several steps during training.

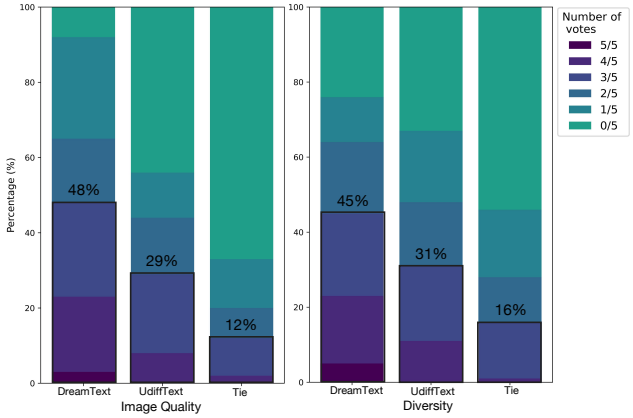

Figure 2. Visualized results of human study.

F. Additional Visual Results

Additional visual results generated by our are exhibited in Fig. 5 and Fig. 6.

G. Limitations

While our method demonstrates promising capabilities in synthesizing scene text, it is limited by its inability to simultaneously modify multiple regions within an image, which restricts its applicability. Future research will explore techniques to address this limitation, aiming to develop more efficient and versatile text synthesis methods capable of simultaneously generating multiple texts within an image. Besides, its application prompts considerations regarding privacy. The generation of realistic text, including personal signatures or identifiable information, may pose risks if misused, potentially compromising individuals' privacy and security. These concerns underscore the importance of implementing robust safeguards and ethical guidelines to address potential privacy risks and ensure the responsible use of this technology.

H. Societal Impact

The advancement of scene text synthesis technology in our work holds significant societal implications. Firstly, it contributes to cultural preservation by enabling the generation of text in diverse styles and languages, aiding in the digitization and conservation of historical scripts and manuscripts. Additionally, our method has applications in art, design, and advertising, empowering creators to produce visually captivating compositions and typography designs. However, ethical considerations surrounding the potential misuse of synthesized text for fraudulent purposes must be carefully addressed through the development of robust safeguards and guidelines. Therefore, while our work offers promising possibilities for text synthesis, it necessitates thoughtful consideration of its societal impacts and ethical implications.

I. Ethical Statement

In this work, we affirm our commitment to ethical research practices and responsible innovation. To the best of our

Figure 3. Qualitative comparative results against AnyText [5].

knowledge, this study does not involve any data, methodologies, or applications that raise ethical concerns. All experiments and analyses were conducted in compliance with established ethical guidelines, ensuring the integrity and transparency of our research process.

References

[1] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. In *NeurIPS*, 2023. 1

[2] Jiabao Ji, Guanhua Zhang, Zhaowen Wang, Bairu Hou, Zhifei Zhang, Brian Price, and Shiyu Chang. Improving diffusion models for scene text editing with dual encoders. *arXiv preprint arXiv:2304.05568*, 2023. 1

[3] Yadong Qu, Qingfeng Tan, Hongtao Xie, Jianjun Xu, Yuxin Wang, and Yongdong Zhang. Exploring stroke-level modifications for scene text editing. In *AAAI*, pages 2119–2127, 2023. 1

[4] 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. 1

[5] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *ICLR*, 2024. 1, 3

[6] Yiming Zhao and Zhouhui Lian. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In *ECCV*, 2024. 1

Figure 4. Additional qualitative comparative results against state-of-the-art methods.

Figure 5. Additional visual results generated by our .

Figure 6. Additional visual results generated by our .