Supplementary Material for DiffUTE: Universal Text Editing Diffusion Model

Anonymous Author(s)
Affiliation
Address
email

In this supplementary, we will first provide the detailed structure and training strategies of DiffUTE in Section 1 to ensure better understanding and reproducibility. Then, we will provide more visual results of text editing tasks in Section 2. Moreover, we provide the source code of DiffUTE in **CODE** folder, making it easy to reproduce our method.

1 Details of DiffUTE

Model Architecture. Our DiffUTE is composed of VAE, glyph encoder and UNet. (i) The VAE uses the same structure as in *stable-diffusion-2-inpainting* [1], with a downsampling factor of 8. (ii) The glyph encoder employs the pre-trained TrOCR model Li et al. [2023], specifically the *trocr-large-printed* [2] version. The TrOCR model is an encoder-decoder model, consisting of an image Transformer as encoder, and a text Transformer as decoder. The image encoder was initialized from the weights of BEiT Bao et al. [2021], while the text decoder was initialized from the weights of RoBERTa Liu et al. [2019]. And the TrOCR model is fine-tuned on the SROIE dataset Huang et al. [2019]. Note that we only use image encoder of TrOCR. Given a character image, the glyph encoder will return a latent feature of size 577×1024. This output is just the right size to be fed directly into the conditioned Unet as a condition. (iii) The UNet uses the same structure as in *stable-diffusion-2-inpainting*.

Training Details. We adopt the Stable Diffusion Rombach et al. [2022] as our baseline model and choose their publicly released v2 model for image inpainting as initialization for VAE and UNet. For the glyph encoder, we use its pre-trained checkpoints for initialization and freeze its weights during training. To improve the reconstruction ability of VAE, we use progressive training strategy. The experimental setting of VAE and UNet is shown in Table S1. Upon completion of VAE training, we proceed to train UNet while keeping the weights of VAE frozen.

Table S1: Training setting for VAE.

Module	Batch size	Crop Image Size	Iterations	Learning Rate
VAE	48	64 128 256 512	0−8w 8w−16w 16w−24w 24w−32w	5e-6
UNet	256	256	0−10w	1e-5

[1] https://huggingface.co/stabilityai/stable-diffusion-2-inpainting
[2] https://huggingface.co/microsoft/trocr-large-printed

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

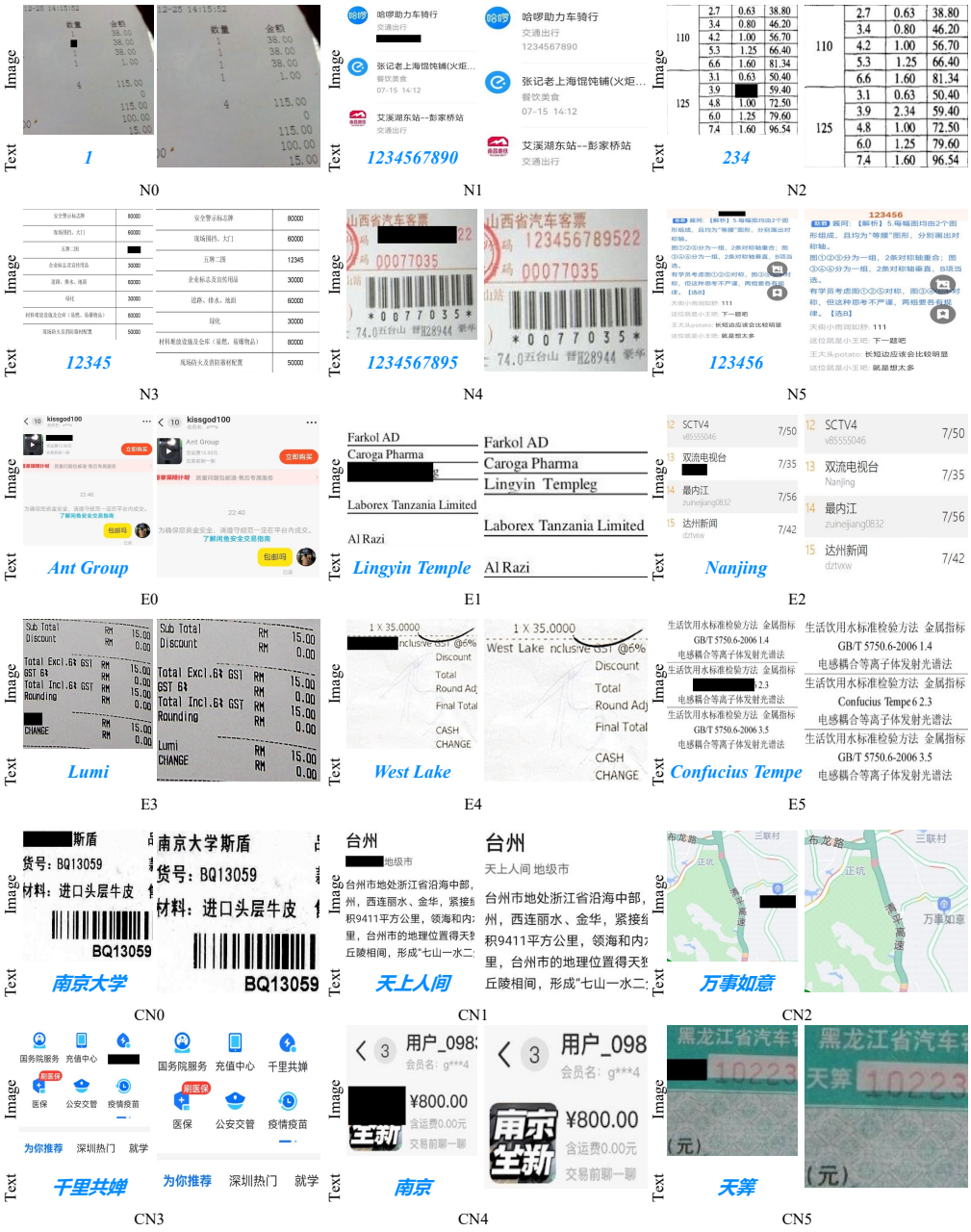

Figure S1: More visualization results of text editing.

2 Visualization Results

We provide additional generated images for editing text in image by our method DiffUTE in Figure S1. DiffUTE consistently generates correct visual text, and the texts naturally follow the same text style, i.e. font, and color, with other surrounding texts. We can see from the experiment that DiffUTE has a strong generative power. (i) In sample N1, DiffUTE can automatically generate slanted text based on the surrounding text. (ii) As shown in sample N2, the input is 234, and DiffUTE can automatically add the decimal point according to the context, which shows that DiffUTE has some document context understanding ability. (iii) In the sample CN4, DiffUTE can generate even artistic characters very well.

References

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. In *AAAI*, 2023.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *ICDAR*, pages 1516–1520, 2019.

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.