

GlyphMastero: A Glyph Encoder for High-Fidelity Scene Text Editing

Supplementary Material

1. Effects of Classifier-Free Guidance

Classifier-free guidance (CFG) has demonstrated effectiveness in controlling the strength of prompt-following behavior in text-to-image diffusion models. Recognizing its potential utility in scene text editing, we incorporate CFG by training our model with a probabilistic null glyph condition (we also trained DiffUTE with CFG for fair comparison, though not in their original work).



Prompt: "八百标兵奔北坡"

Figure S1. Effect of classifier-free guidance (CFG). Original image with target area mask shown top-left. Without CFG (i.e. CFG=1), GlyphMastero produces unreadable text. CFG=3 improves readability while maintaining style. CFG=5 generates overly thick text, deviating from the original region.

Our experiments with CFG reveal a crucial trade-off in scene text editing. As demonstrated in Figure S1, we found that in inference, a higher CFG scale results in stronger glyph control, producing clearer and thicker text. This allows for improved readability when editing texts. However, our findings show that this comes at a cost to style preservation. Conversely, lower CFG scales excel at maintaining the original text style, though occasionally at the expense of target text accuracy. This insight offers a new approach to balancing readability and style preservation in the scene

text editing task.

2. Example Failure Cases

As shown in Figure S2, our method encounters limitations when the selected editing region substantially exceeds the target text length. In such scenarios, the model struggles to maintain coherent text generation, resulting in irregularly sized characters and occasional repetition patterns in the output. These artifacts emerge as the model attempts to distribute textual elements across disproportionately large spatial regions.



Figure S2. Example of a failure case. The upper image displays the source text with regions marked by red boxes. The middle and bottom images show two unsuccessful generation attempts. The intended target text appears at the bottom left of each generated result.

3. Additional Results

3.1. Quantitative Comparison

To cross-validate the effectiveness of our method, we also evaluated our method on TextCtrl's ScenePair dataset (1,285 test cases), re-evaluating all other methods using TextCtrl's GitHub-published results and scripts. Table S1 showed our superior generation accuracy and strong performance across style metrics, except for a slightly higher FID.

3.2. Qualitative Comparison

We present additional qualitative comparisons in Figure S3, S4, S5, and S6.

Figure S3 presents additional examples from our curated test set, demonstrating the efficacy of our approach for stylistic scene text editing. Additionally, Figures S4 and S5 showcase random samples from the AnyText-Eval

Methods	Accuracy		Similarity			
	W.Acc \uparrow	NED \uparrow	SSIM \uparrow	PSNR \uparrow	MSE \downarrow	FID \downarrow
SRNet	16.64	0.4790	26.66	14.08	5.61	49.23
MOSTEL	35.16	0.5570	27.46	14.46	5.19	49.20
DiffSTE	29.14	0.5255	26.91	13.49	6.07	118.60
TextDiffuser	51.48	0.7190	27.02	13.99	5.72	57.48
AnyText	47.97	0.7186	31.19	13.58	6.36	52.07
TextCtrl	78.91	0.9199	37.93	14.92	4.58	31.98
Ours	83.52	0.9572	47.58	16.25	3.97	32.03

Table S1. Performance comparison on English *ScenePair* testset. SSIM and MSE scaled by $\times 10^{-2}$. W.Acc: Word Accuracy.

benchmark dataset - Figure S4 illustrates our model’s performance on English text using the LAION dataset, while Figure S5 highlights its capabilities with Chinese text from the Wukong dataset. Figure S6 provides comparative results on TextCtrl’s ScenePair test set as previously discussed. These qualitative results align consistently with our quantitative evaluations, validating the effectiveness of our proposed method.





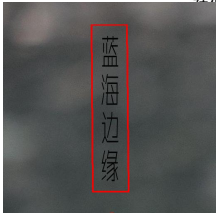
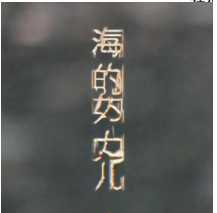
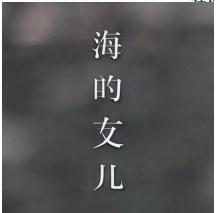
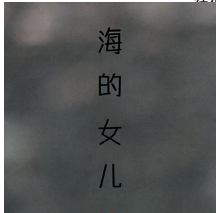











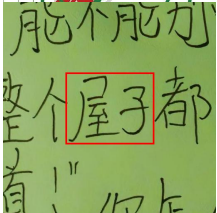
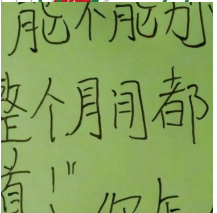
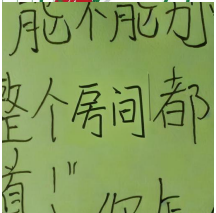
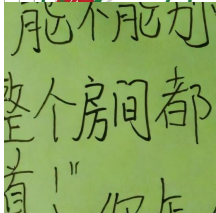
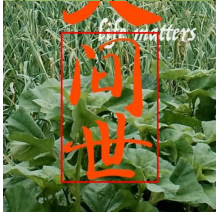



Prompt	Masked Source Image	DiffUTE	AnyText	Ours
2026-03-09				
海的女儿 (the daughter of the sea)				
34				
羊 (sheep)				
喜多多 (Xi'duo'duo)				
房间 (room)				
世间 (the world)				

Figure S3. Comparison results on our test set with stylish scene texts

Prompt	Masked Source Image	DiffUTE	AnyText	Ours
WINNER				
DANCE				
200				
IC				
HITS				
S				
Strategic				

Figure S4. Comparison results on English (LAION) test set

Prompt	Masked Source Image	DiffUTE	AnyText	Ours
注意安全 (watch out)				
小仙女专用座 (dedicated seat for girls)				
帮人难处： (help others)	<div>帮人难处：</div> <div>锦上添花易，雪中送炭难！ 我们应该在别人 困难之时帮扶一把， 不管别人是否记得， 自己要做到无愧于心。</div>	<div>帮人难处：</div> <div>锦上添花易，雪中送炭难！ 我们应该在别人 困难之时帮扶一把， 不管别人是否记得， 自己要做到无愧于心。</div>	<div>帮人难处：</div> <div>锦上添花易，雪中送炭难！ 我们应该在别人 困难之时帮扶一把， 不管别人是否记得， 自己要做到无愧于心。</div>	<div>帮人难处：</div> <div>锦上添花易，雪中送炭难！ 我们应该在别人 困难之时帮扶一把， 不管别人是否记得， 自己要做到无愧于心。</div>
居住证 (residency ID)				
先进集体 (advanced group of people)				
精度高圆板牙 (high precision bolt)				
寓 (apartment)				

Figure S5. Comparison results on Chinese (Wukong) test set

Source Image	Target Text	SRNet	MOSTEL	DiffSTE	TextDiffuser	AnyText	TextCtrl	GlyphMaster
	"Learning"							
	"Buses"							
	"FLASH"							
	"ROYAL"							
	"Bernd"							
	"and"							
	"WEATHER"							
	"EU-funded"							
	"SPRINKLER"							
	"FOSTERS"							
	"HAMBLION"							
	"Centre"							
	"DRIVERS"							
	"WSDL"							
	"Authorised"							
	"National"							
	"Section"							
	"Centre"							
	"order"							
	"LONG"							
	"work"							
	"PLEASE"							
	"Knowledge"							
	"and"							
	"EMMERICH"							

Figure S6. Comparison of different scene text editing methods on the ScenePair dataset