Supplementary Material of Language-Guided Global Image Editing via Cross-Modal Cyclic Mechanism

1. Component Analysis

1.1. Data Augmentation

We conduct ablation studies on the proposed two kinds of data augmentation strategy, as shown in Table 1. From top to bottom, *w.o. aug*, *swap* and *swap&random* indicate the model without data augmentation, with only swapping augmentation and with two kinds of augmentation respectively. The quantitative results show that the proposed data augmentation strategy is both necessary and useful for promoting the performance of the generator.

1.2. Cyclic Mechanism

Augmented EDNet. Given the input image x, target image y, and language embedding h, the augmented EDNet produces editing embeddings to supervise the generate G. Theoretically, the augmented EDNet is learned better than the generator through the cyclic mechanism and data augmentation. To validate the superiority of the augmented EDNet, we generate images condition on editing embeddings and perform a comparison with the original model. The process can be illustrated using the following equations:

$$\tilde{x}_h = G(x, h), \tag{1}$$

$$\begin{aligned} e_{x \to y} &= ED(x, y), \\ \tilde{x}_e &= G(x, e_{x \to y}), \end{aligned} \tag{2}$$

where $\tilde{x}_h$ and $\tilde{x}_e$ are images generated using language embeddings h and editing embeddings e. Note that the generator uses language embedding h and editing embedding e to achieve image editing by merely scaling and shifting the visual feature maps for only once. This design prevents the generator from memorizing the target image. We visualize some examples of $\tilde{x}_h$ and $\tilde{x}_e$ in Figure 1. $\tilde{x}_e$ can adjust the hue, contrast, and brightness following the linguistic requests accurately, which is better than $\tilde{x}_h$. Quantitative comparisons between $\tilde{x}_h$ and $\tilde{x}_e$ are also conducted, as shown in Table 1. The results also demonstrate the superiority of $\tilde{x}_e$. Thus, the augmented EDNet can be used to supervise the generator.

Language-Sensitivity. To further examine the language-sensitivity of our model, we propose the image variance σ^2 to measure the diversity of the generated image conditioned on different requests. We apply 8 predefined different language requests to the same input image and output 8 different images. These 8 predefined requests are: "brighten the image", "darken the image", "increase the contrast", "decrease the contrast", "increase the saturation", "decrease the saturation", "enhance the color" and "decrease the color". Then we compute the variance over the 8 images of all pixels and take the average overall spatial locations and color

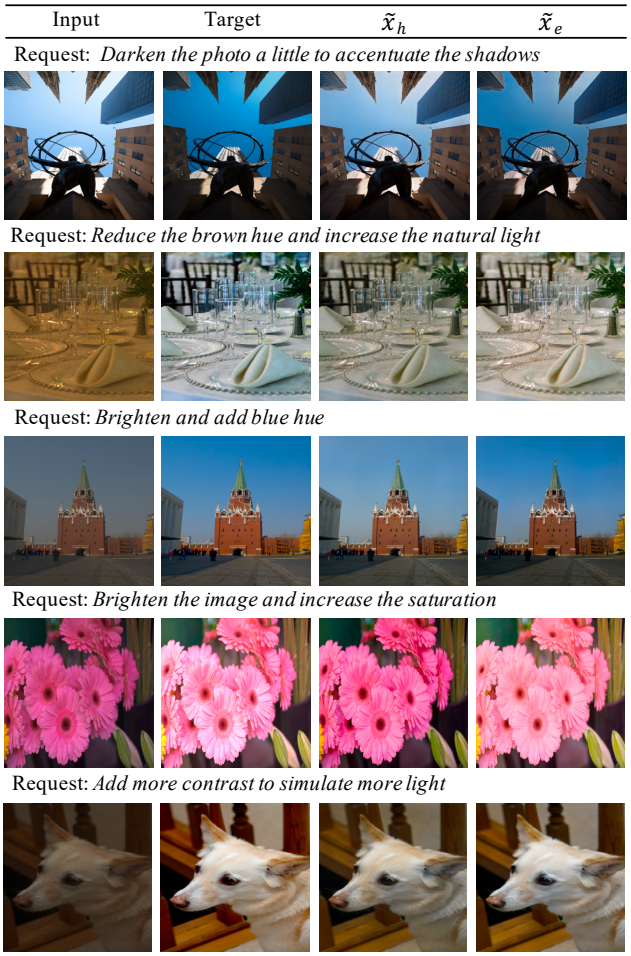

Figure 1. Visualization of $\tilde{x}_h$ and $\tilde{x}_e$. $\tilde{x}_e$ match the requests better.

Method	MA5k-Req						GIER					
	IS↑	FID↓	RSS↑				IS↑	FID↓	RSS↑			
			BLEU-4	CIDEr	METEOR	ROUGE-L			BLEU-4	CIDEr	METEOR	ROUGE-L
w.o. Aug	16.29	10.09	8.39	65.89	11.01	22.20	9.91	44.70	3.81	35.81	9.16	23.23
swap	16.93	10.03	8.57	65.84	11.03	22.56	10.21	43.28	3.94	36.32	9.27	23.35
swap & random ($\tilde{x}_h$)	17.16	9.95	8.66	66.18	11.13	22.83	10.35	42.01	4.09	37.03	9.45	23.60
augmented EDNet ($\tilde{x}_e$)	**18.09**	**9.52**	**9.05**	**68.38**	**11.63**	**23.91**	**10.99**	**40.07**	**4.18**	**38.82**	**9.89**	**24.55**

Table 1. Ablation studies on data augmentation and cyclic mechanism.

	w.o. EDNet	Ours
σ^2	3278.02	**3564.10**

Table 2. Quantitative results of image variance.

channels. Finally, we take the average of the average variance over the entire test set, as shown in Table 2. We can see that our model with the cyclic mechanism has a higher variance, which indicates that the generated images are more diverse in different language conditions. Similar observations can be found in Figure 8 of our original paper. Due to the insufficient and imbalance of data, the vanilla generator is biased and thus not sensitive to different requests. By leveraging the proposed cyclic mechanism and data augmentation, our model is more robust and sensitive, which can edit images through requests actually.

2. Qualitative Comparison

We conduct additional qualitative comparisons with baselines on the GIER dataset [1] and MA5k-Req dataset [2], as shown in Figure 2 and 3. Extensive experiments have demonstrated the superiority of our method.

References

[1] Jing Shi, Ning Xu, Trung Bui, Franck Dernoncourt, Zheng Wen, and Chenliang Xu. A benchmark and baseline for language-driven image editing. *arXiv preprint arXiv:2010.02330*, 2020. 2

[2] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13590–13599, 2021. 2

Figure 2. Qualitative comparison with baseline models on GIER dataset. Best viewed in color.

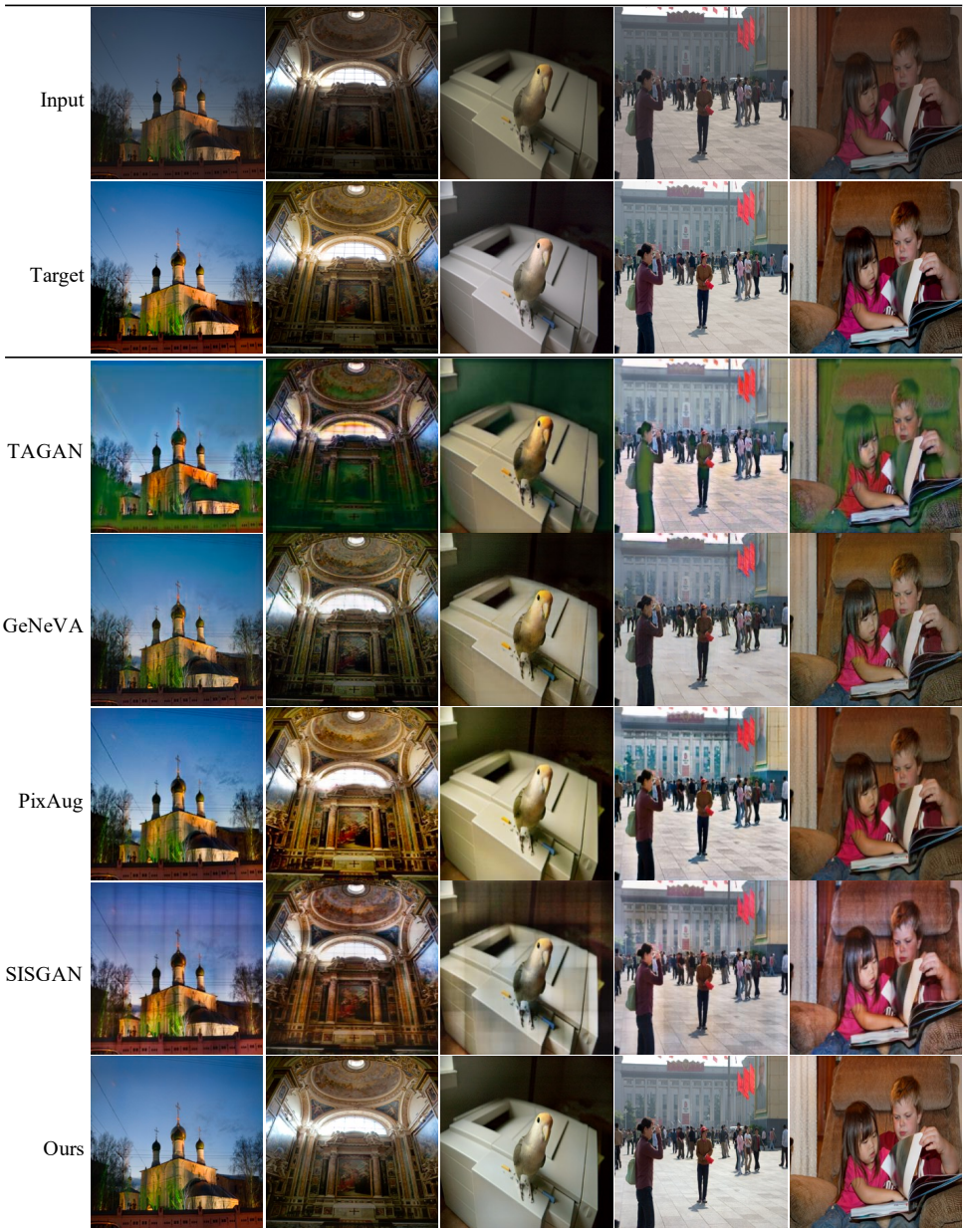

Figure 3. Qualitative comparison with baseline models on MA5k-Req dataset. Best viewed in color.