Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 7

Supplemental Material of Continuous Layout Editing of Single Images with Diffusion Models

Zhiyuan Zhang [1†], Zhitong Huang[1†] and Jing Liao[1‡]

[1]City University of Hong Kong, Hong Kong SAR, China

1. Robustness to coarse masks

We tested the robustness of our method to object masks by resizing the ground truth mask to different scales. Fig. 2 illustrates the results of switching the positions of two dogs given masks in different coarseness. As shown in Tab. 1(f), our method is robust to mask noises within scales 0.9, 1.1, and 1.2. However, when masks are too large(1.5) and overlapped, our method may fail to rearrange the positions of the objects. When masks are too small (0.5 and 0.8) to cover the whole object, our method can still rearrange positions but may cause changes in the appearance of the objects, resulting in lower visual similarity.

2. Quantitative results of the ablation study

We run 5 edits for each ablation study. Besides the examples shown in the paper, we also show the results of additional examples in supplementary. The quantitative results are as below:

1. Textual inversion: Our method achieves the highest visual similarity as shown in Tab. 1(a), outperforming Dreambooth, textual inversion with finetune, and masked textual inversion without finetune. The additional examples used are shown in Fig. 3.
2. Layout control: Our method excels in visual similarity and layout alignment as shown in Tab. 1(b). The additional examples used are shown in Fig. 5.
3. Optimization loss: Our method achieves the best visual similarity and layout alignment as shown in Tab. 1(c). The additional examples used are shown in Fig. 4.
4. Iterative optimization: Our method outperforms other iteration steps, aligning better with the task mask as shown in Tab. 1(d). The additional examples used are shown in Fig. 7.
5. Blending steps: As shown in Tab. 1(e), the visual similarity of our method is 29% higher than no blending and 2% higher than using fewer blending steps($t > 0.8$). Blending for all steps gives a 2% higher visual similarity, but it reduces the background's flexibility to adapt to objects' new layout. The additional examples used are shown in Fig. 6.

3. Multi-run results

We conducted 5 runs for each of the 15 examples in the user study. The average visual similarity is 0.57 (±0.0048), while the average layout alignment is 0.0096 (±0.0001). Our method exhibited stability with small variances in both metrics and visual quality. We also illustrate some images generated in Fig. 8.

4. Background elements editing

Our method has the capability to edit the layout of background elements, such as sky and lake, by treating them as single objects. The challenge is that background elements usually have larger sizes and occlusions. We provide two examples in row 1&2 of Fig. 9, demonstrating our method over different background elements.

5. Additional results

We provide additional examples, demonstrating our method of editing different elements in the image, as shown in Fig. 9, Fig. 10, Fig. 11 and Fig. 12.

References

[BTYLD23] Bar-Tal O., Yariv L., Lipman Y., Dekel T.: Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113 2* (2023). 3

[GAA*22] Gal R., Alaluf Y., Atzmon Y., Patashnik O., Bermano A. H., Chechik G., Cohen-Or D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022). 3

[KZZ*23] Kumari N., Zhang B., Zhang R., Shechtman E., Zhu J.-Y.: Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 1931–1941. 3

[LLW*23] Li Y., Liu H., Wu Q., Mu F., Yang J., Gao J., Li C., Lee Y. J.: Gligen: Open-set grounded text-to-image generation. *CVPR* (2023). 3

[RLJ*23] Ruiz N., Li Y., Jampani V., Pritch Y., Rubinstein M., Aberman K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 22500–22510. 3

[ZA23] Zhang L., Agrawala M.: Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023). 3

Table 1: *Quantitative results of ablation study on: (a) Textual inversion methods; (b) Layout control methods; (c) Optimization loss of layout control; (d) Iterative optimization; (e) Blending steps; (f) Mask coarseness.*

(a)

	Visual similarity ↑	Layout alignment ↑
Dreambooth	0.56	0.0183
Textual inversion with finetune	0.44	0.0085
Masked textual inversion without finetune	0.46	0.0237
Our	**0.64**	**0.0187**

(b)

	Visual similarity ↑	Layout alignment ↑
ControlNet	0.43	0.0089
GLIGEN	0.47	0.0068
MultiDiffusion	0.48	0.0048
Our	**0.57**	**0.0226**

(c)

	Visual similarity ↑	Layout alignment ↑
Mean loss	0.58	0.0003
Max loss	0.44	0.0044
Mean+max (ours)	**0.62**	**0.0117**

(d)

Optimization timesteps	Visual similarity ↑	Layout alignment ↑
No iterative opt.	0.60	0.0120
1.0	**0.61**	0.0124
0.8	0.57	0.0157
0.6	0.58	0.0096
0.4	0.59	0.0116
0.2	0.60	0.0087
0.02	0.59	0.0107
1.0+0.8	0.58	0.0214
1.0+0.8+0.6 (ours)	0.55	**0.0249**
1.0+0.8+0.6+0.4	0.53	0.0204
1.0+0.8+0.6+0.4+0.2	0.56	0.0023

(e)

Blending timesteps	Visual similarity ↑	Layout alignment ↑
No blending.	0.51	0.0065
> 0.8	0.64	0.0116
> 0.6	0.64	0.0122
> 0.4	0.66	0.0117
> 0.2	**0.67**	0.0130
> 0 (all)	0.66	**0.0147**
> 0.7 (ours)	0.65	0.0126

(f)

Mask coarseness	Visual similarity ↑
0.5	0.57
0.8	0.53
0.9	0.68
1.0	**0.74**
1.1	0.72
1.2	0.71
1.5	0.66

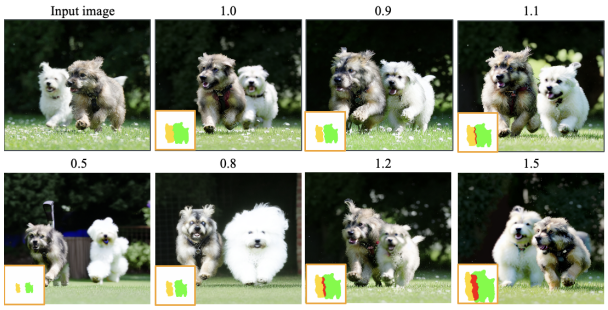

Figure 2: *Results of switching the positions of the two dogs given masks in different coarseness. The number above the image indicates the scaling levels for the masks.*

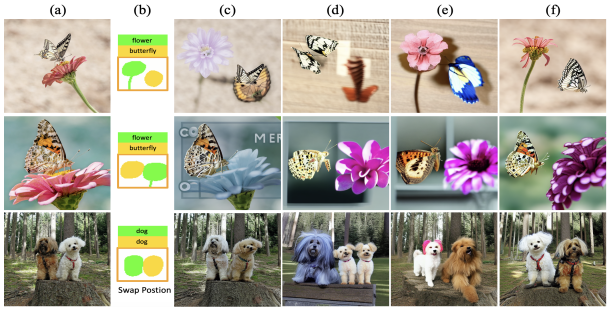

Figure 3: *Additional ablation study on different textual inversion methods. From left to right: (a) Input images; (b) Target layouts; (c) Inversion with Dreambooth [RLJ*23]; (d) Textual inversion [GAA*22] + finetune [KZZ*23]; (e) Masked textual inversion w/o finetune; (f) Our full inversion method with masked textual inversion and finetune.*

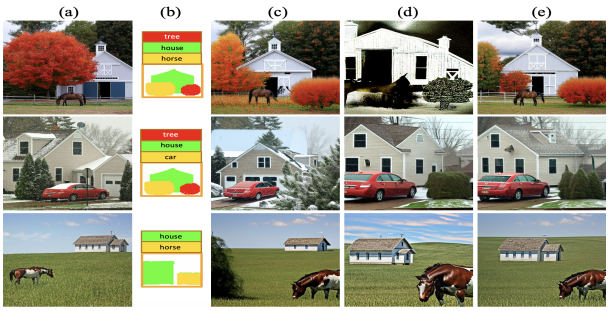

Figure 4: *Additional ablation study on optimization loss of layout control. From left to right: (a) Input images; (b) Target layouts; (c) mean as loss; (d) max as loss; (e) mean + max as loss.*

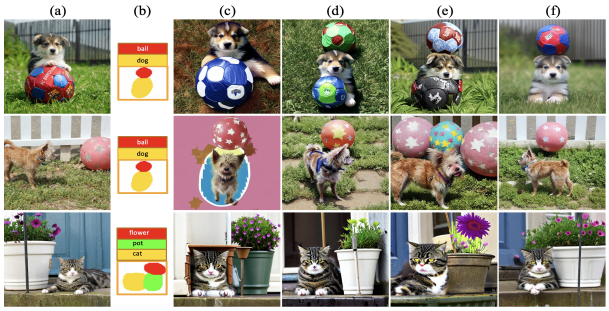

Figure 5: *Additional ablation study on different layout control methods. From left to right: (a) Input images; (b) Target layouts; (c) Our inversion + ControlNet [ZA23]; (d) Our inversion + GLIGEN [LLW*23]; (e) Our inversion + MultiDiffusion [BTYLD23]; (f) Our full methods.*

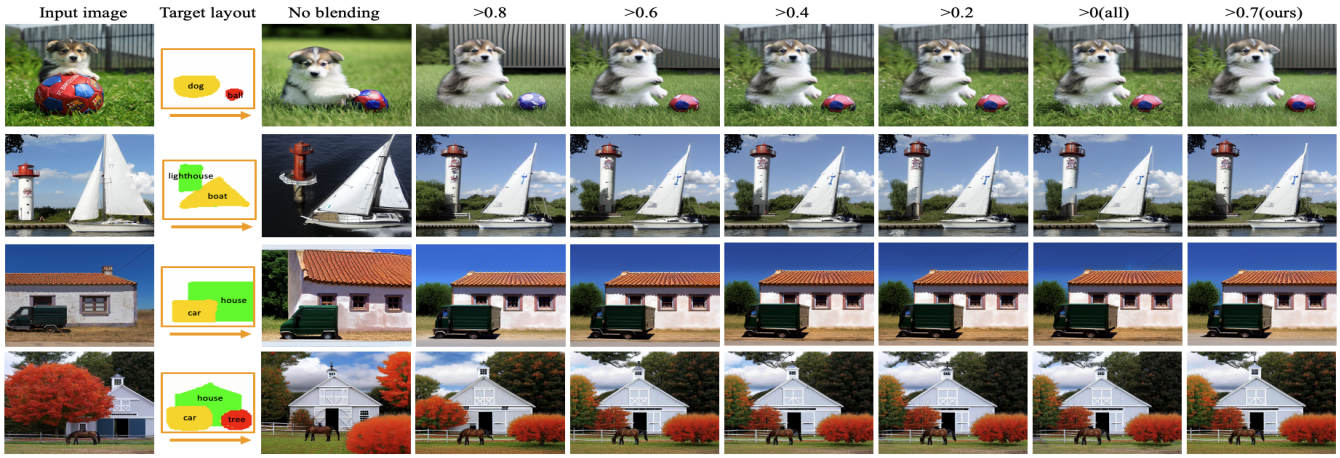

Figure 6: *Additional ablation study on blending steps. The number above the image indicates the number of steps where blending is applied.*

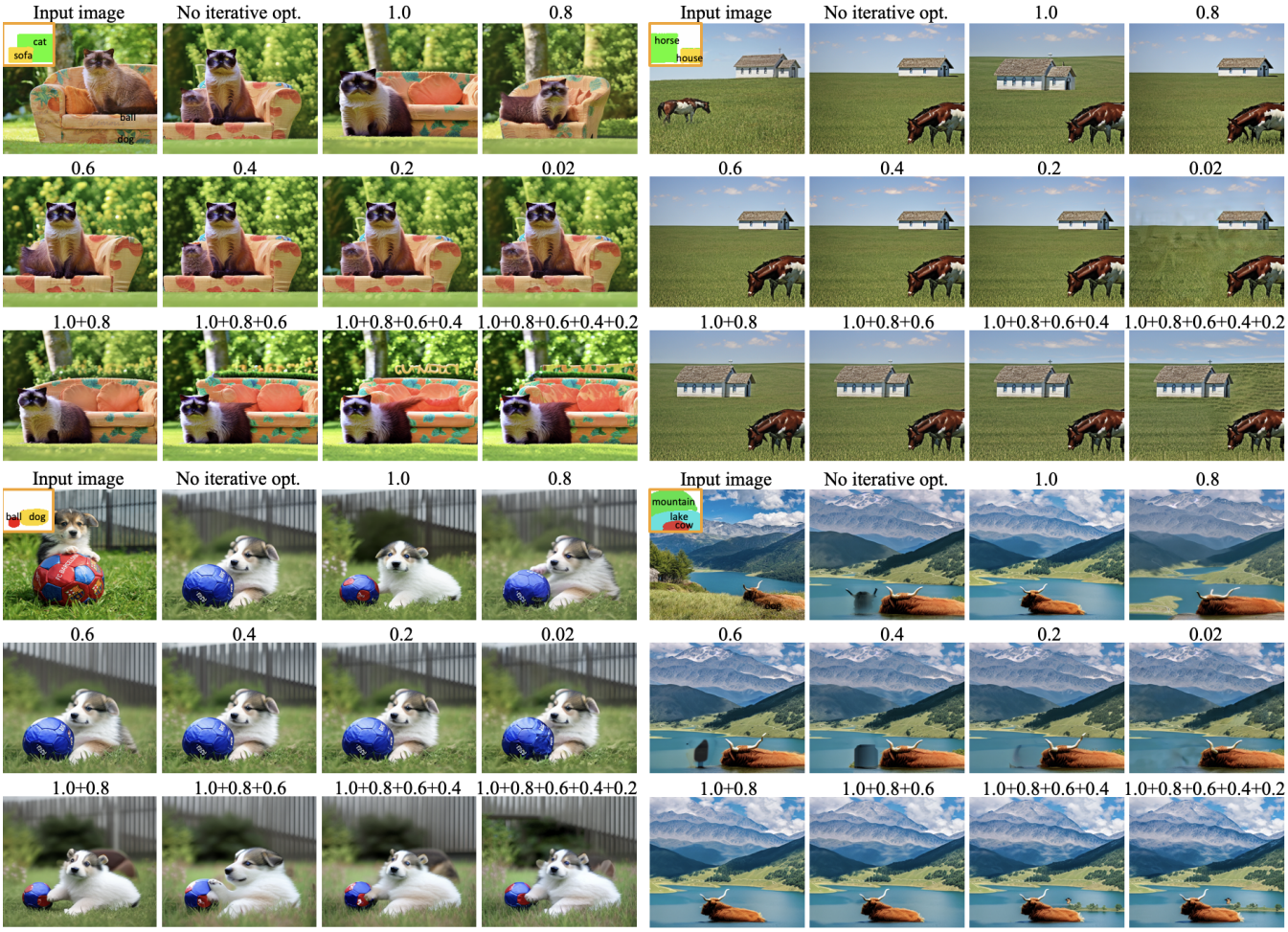

Figure 7: *Additional ablation study on iterative optimization. The number above the image indicates the denoising step on which iterative optimization is applied. If more than one number is labeled, iterative optimization is applied to multiple denoising steps.*

Figure 8: *Multi-run results for our method.*

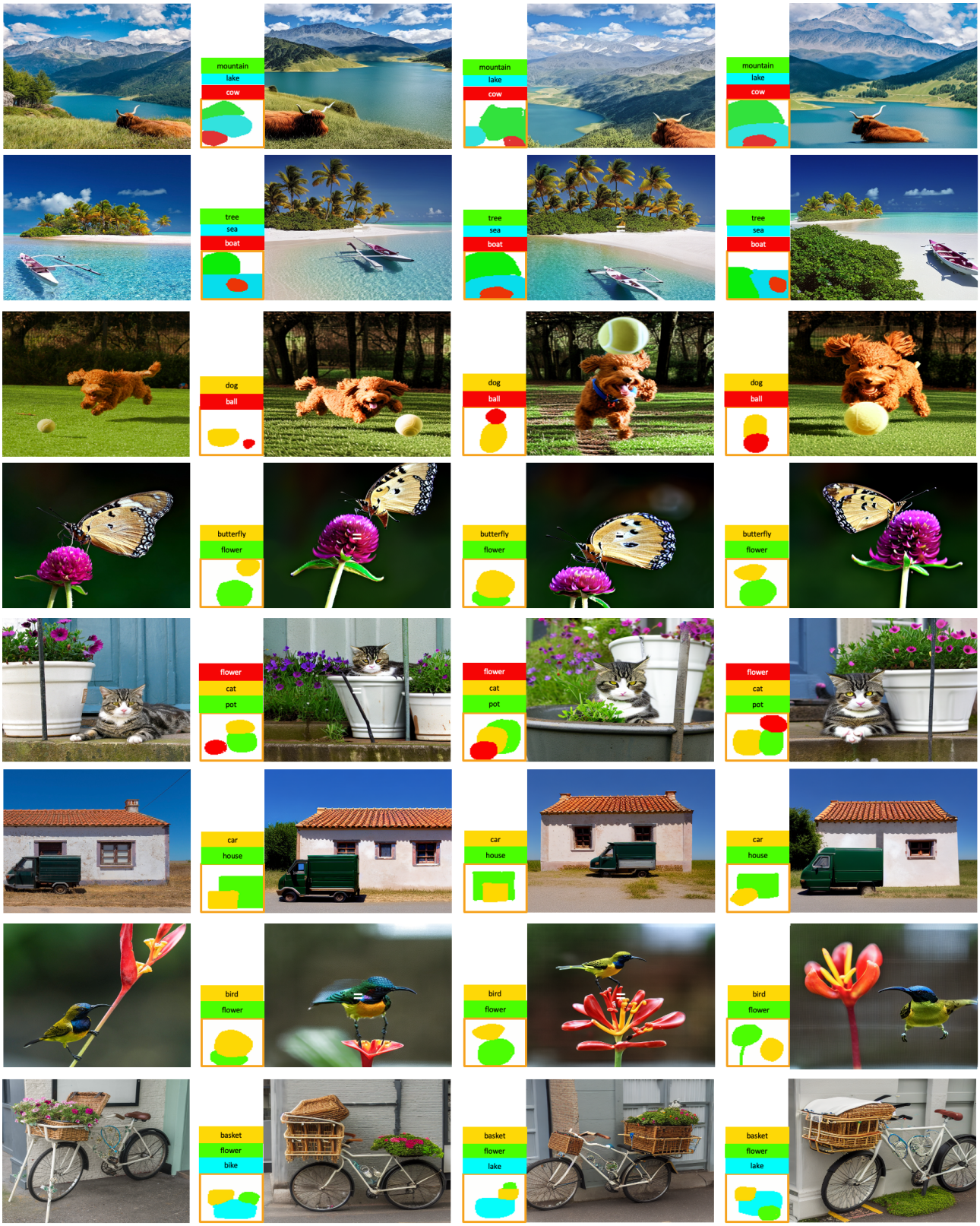

Figure 9: *Additional results on continuous layout editing with different target layout*

Figure 10: *Additional results on continuous layout editing with different target layout*

Figure 11: *Additional results on continuous layout editing with different target layout*

Figure 12: *Additional results on continuous layout editing with different target layout*