

ManiTrans: Entity-Level Text-Guided Image Manipulation via Token-wise Semantic Alignment and Generation: Supplementary Material

Jianan Wang¹ Guansong Lu² Hang Xu² Zhenguo Li² Chunjing Xu² Yanwei Fu¹

¹School of Data Science, Fudan University ²Huawei Noah’s Ark Lab

{jawang19, yanweifu}@fudan.edu.cn {luguansong, xu.hang, li.zhenguo, xuchunjing}@huawei.com

In this supplementary material, we provide the statistics of the datasets, quantitative results of ablation study, and qualitative results to further demonstrate the ability of our ManiTrans, and to discuss failed cases.

1. Experiment Datasets

As we mention in the main paper, we benchmark ManiTrans on three public datasets, including the CUB [5], Oxford [2] and COCO [1] datasets. The number of images and the number of captions per image of each dataset are listed in Table 1. The CUB and Oxford are two datasets about birds and flowers respectively. While there are at least 80 categories of objects with different shape structures and appearances on COCO images. Thus, COCO is a more complicated dataset than CUB and Oxford, not only in model understanding the correspondence between the image and text, but also in image manipulation on the entity level.

Dataset		#images	#captions/image
CUB [5]	train	8855	10
	test	2933	
Oxford [2]	train	7034	10
	test	1155	
COCO [1]	train	80k	5
	test	40k	

Table 1. Statistics of datasets.

2. Quantitative Results of Ablation Study

Effects of Vision Guidance. Vision guidance aims to provide prior structures of entities and to make our model better generate the appearance of entities. Without such prior from vision guidance, our model generates entirely different entities, such as shown in Fig. 2 - Fig. 6. In Table 2, we present the quantitative results of our ManiTrans with and without vision guidance on CUB. Two semantic metrics, CLIP-score and R@10, are higher without vision guidance. The other two image quality metrics, IS and L2-error,

Vision Guidance	IS	CLIP-score	R@10	L2-error
✓	5.02 ± 0.11	23.56	34.82	0.01
–	4.98 ± 0.06	24.02	42.61	0.02

Table 2. Quantitative results on CUB w/ or w/o vision guidance.

SL	SAM	IS	CLIP-score	R@10	L2-error
✓	✓	5.02 ± 0.11	23.56	34.82	0.01
✓	–	4.93 ± 0.08	23.53	32.89	0.01
–	✓	5.01 ± 0.09	22.19	16.06	0.01

Table 3. Quantitative results on CUB. “SL” and “SAM” are for semantic loss and semantic alignment module, respectively.

are still competitive without the prior information by vision guidance.

Effects of Semantic Loss. Table 3 shows the quantitative results of the ablation study on semantic loss (SL). With SL, our ManiTrans achieves a higher CLIP-score and R@10, which demonstrates that SL helps the model improve the relevance between the manipulated images and text guides.

Effects of Semantic Alignment Module. Table 3 compares the quantitative metrics when we apply the semantic alignment module (SAM) in the inference phase or not (*i.e.* word-path alignment). All the semantic metrics (CLIP-score, R@10) and image quality metrics (IS, L2-error) are better with SAM. This suggests the manipulated images by SAM are more realistic and more consistent with the text semantics.

3. Comparison with StyleCLIP

StyleCLIP [3] is one pioneer work on style transfer. However, StyleCLIP is more focused on faces, rather than nature images in our paper. In Fig. 1, we provide a qualitative comparison with the global direction method of StyleCLIP, whose results are generated with official codes & models from StyleCLIP and StyleGAN2. StyleCLIP edits

the latent of image inverted from e4e [4], whose content is different from the original image as shown in Fig. 1. Furthermore, we achieve better manipulation on the tower.

4. Additional Qualitative Results

We provide more qualitative results on CUB (Fig. 2), Oxford (Fig. 3), COCO (Fig. 4, Fig. 5), and a bi-directional entity transformation between bird and flower (Fig. 6). The prompt words of CUB and Oxford are “bird” and “flower”. Almost all the prompt words of COCO are the nouns of their text and we will state the prompt words for special examples.

5. Additional Failed Cases

We additionally present and discuss failed cases on the CUB (Fig. 7), Oxford (Fig. 8) and COCO (Fig. 9), where the black patches in the **Masked Patch** are the selected entity tokens by our semantic alignment module, to be manipulated, roughly reflected on the original image.



Figure 1. Comparison with StyleCLIP.

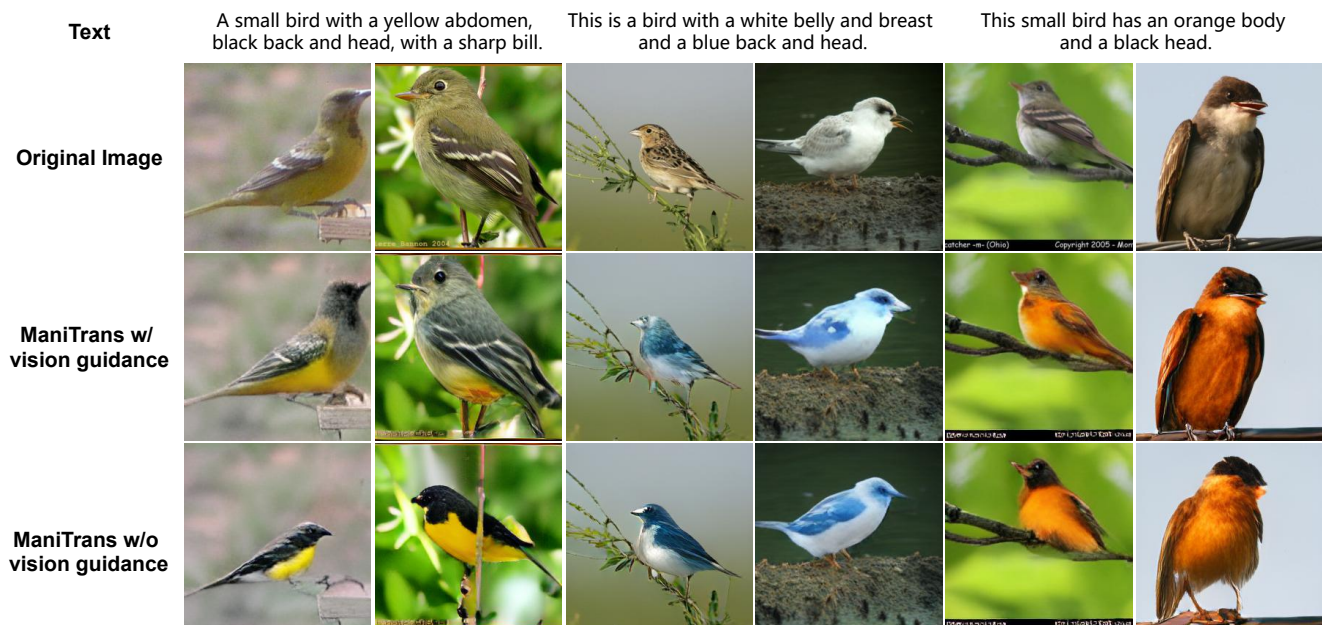


Figure 2. Our manipulation results w/ and w/o vision guidance on CUB.

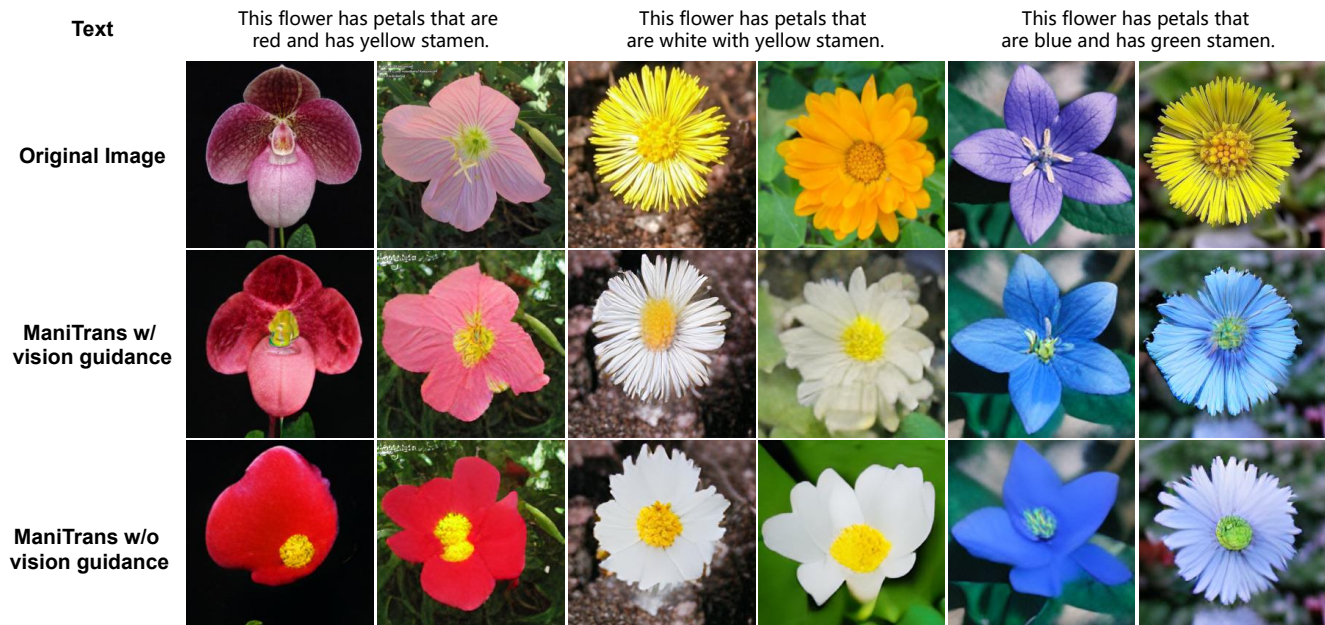


Figure 3. Our manipulation results w/ and w/o vision guidance on Oxford.

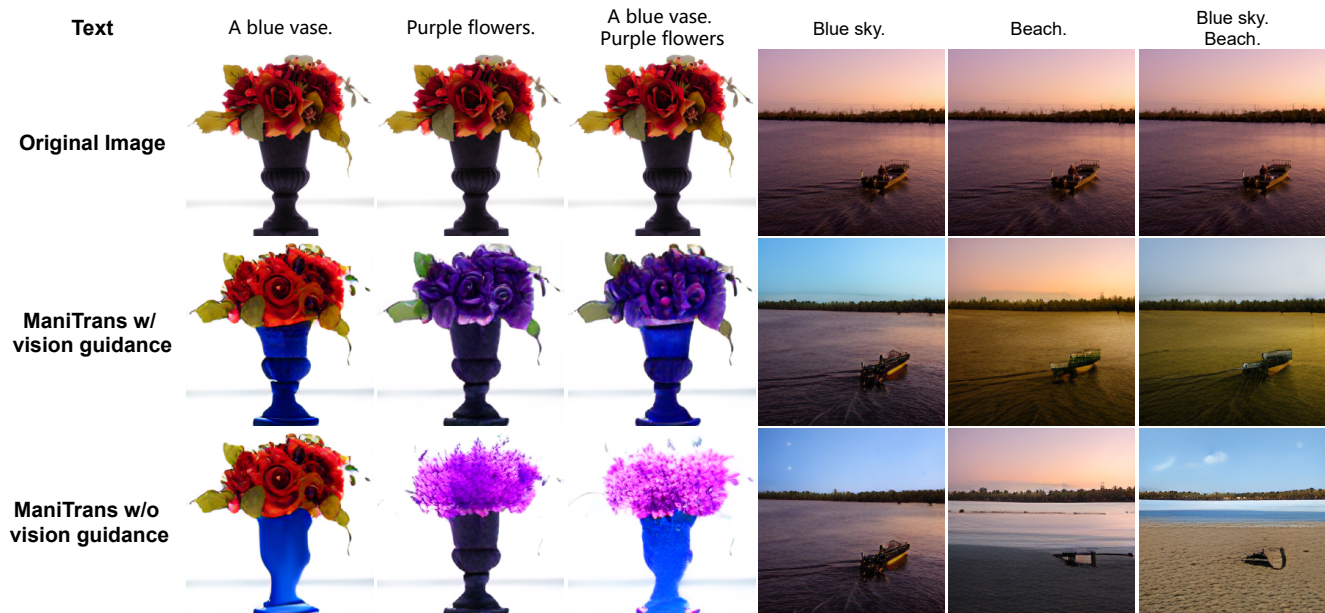


Figure 4. Our manipulation results w/ and w/o vision guidance on COCO. Our ManiTrans can manipulate different entities separately and together on one image. The prompt word for the text “Beach.” is “river”.

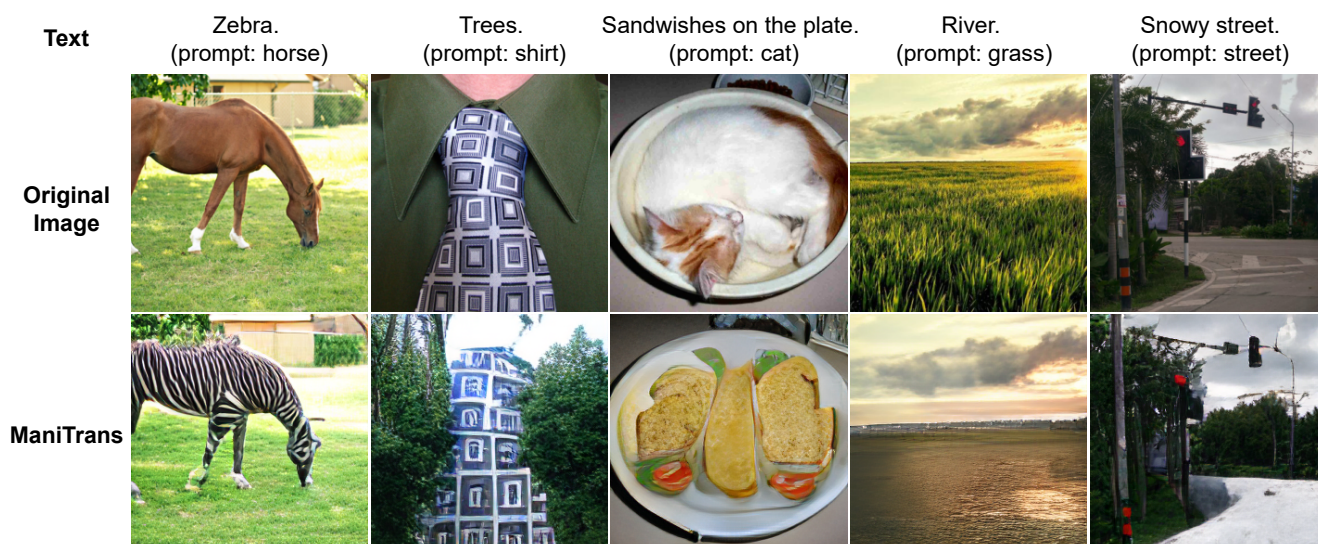


Figure 5. More manipulation results across categories w/o vision guidance on COCO.

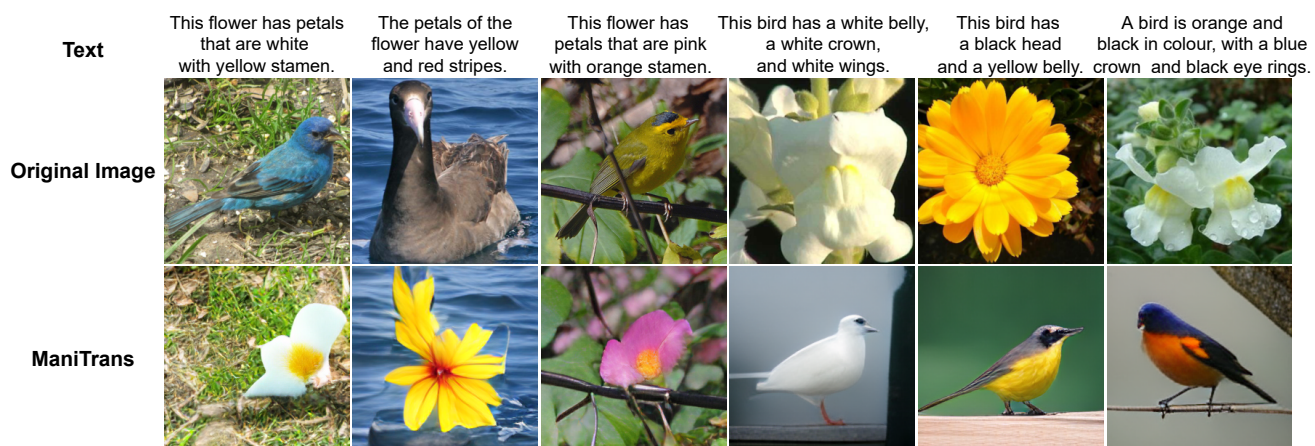


Figure 6. Our manipulation results from bird to flower and flower to bird.








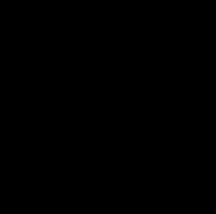


Text	Original Image	Entity Segmentation	Masked Patch	ManiTrans w/ vision guidance	ManiTrans w/o vision guidance
This bird has an orange crown as well as a black bill.					
					

Figure 7. Our failed cases on CUB. The first row shows a failed case when without vision guidance. With the entity shape information of vision guidance, our ManiTrans can manipulate the whole bird according to the text. While, without the entity shape information across the two-part segmentation, ManiTrans only generates a bird head within the limited right part region. The second row shows a failed case in preserving the background when without vision guidance. The bird with a white belly of the second row is difficult to be discriminated with the bright leaves, and the entity segmentation recognizes the whole image as an entity. Thus, without the vision guidance, our ManiTrans did a generation task and lost the original background. Most failed cases on CUB happen in these two situations.

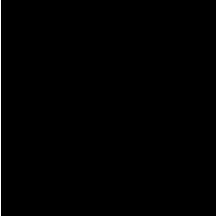

Text	Original Image	Entity Segmentation	Masked Patch	ManiTrans w/ vision guidance	ManiTrans w/o vision guidance
This flower has petals that are purple with white stamen.					
					

Figure 8. Our failed cases on Oxford. Most failed cases for flower manipulation fail in background preservation, for the flowers in Oxford are too large and difficult to discriminate with the background, especially the background of green leaves. Here two examples change the background and the flower at the same time, either with vision guidance or not.

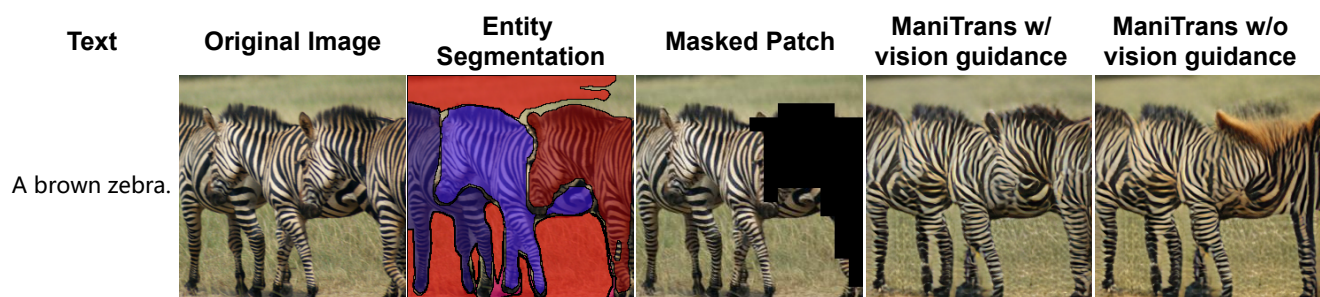


Figure 9. Our failed cases on COCO. Here we aim to manipulate one zebra on the image. However, confused by another preserved zebra next to the tokens to be manipulated, ManiTrans generates a brown body for the nearest zebra instead of generating a whole brown zebra. To manipulate an entity that is overlapped by other entities in the same category is to be improved.

References

- [1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [1](#)
- [2] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. [1](#)
- [3] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. [1](#)
- [4] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. [2](#)
- [5] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. [1](#)