Table 1: Quantitative comparison: number of parameters in generator (NoP-G) and discriminator (NoP-D) (million), inference time for generating 100 new modified images (IT), accuracy (Acc), realism (Real), and FID. The numbers for Acc and Real indicate the average percentage of images favoured by users for the method. For Acc and Real, higher is better, for others, lower is better.

	CUB						COCO					
Method	NoP-G	NoP-D	IT (s)	Acc	Real	FID	NoP-G	NoP-D	IT (s)	Acc	Real	FID
ManiGA	41.1M	169.4M	4.71	34.06	42.18	9.75	53.3M	377.6M	10.03	22.03	32.47	25.08
Ours	5.4M	1.6M	1.08	65.94	57.82	8.02	7.4M	3.5M	1.12	77.97	67.53	14.79

Figure 1: Top: qualitative comparison between our method, SISGAN [4] and TAGNA [18] on CUB (left) and COCO (right). Bottom: various colours on a same image by our method.

R3: Comparison. Please see Table 1 for inference time and number of parameters in generator and discriminator.

R3: Blurry results on COCO. Because our model has much fewer parameters, and generating realistic images on COCO with text is more difficult. However, our method still outperforms ManiGAN both qualitatively and quantitatively.

R3: User study. Thanks for the suggestion. For each dataset, we randomly select 30 images with 1 randomly chosen description (for COCO, each text-image pair is from the same category). Thus, there are 60 samples in total from two datasets for each method. Then, we ask workers to compare two results after looking at the input image, given text and outputs based on two criteria: (1) accuracy: whether the visual attributes of the manipulated image match the text, and the text-irrelevant contents are preserved, and (2) realism: whether the manipulated image looks realistic. Finally, we collected 1380 results from 23 workers, shown in Table 1. For both the accuracy and the realism, our results are most preferred by workers. Also, both IS (Table 1 of the paper) and FID further verify the better performance of our method.

R3, R4: Technical novelty. The architecture of our method is fundamentally different from ManiGAN, as ManiGAN has two modules: the main module and DCM. Once the main module is optimised, ManiGAN sets it to eval mode and trains the DCM subsequently. However, our method can be trained end-to-end with much fewer parameters (see Table 1). As for the proposed discriminator, it is not a simple extension of the auxiliary classifier, as we need to consider how to combine cross-domain features together, and how to provide appropriate word-level target labels to calculate word-level feedback. It is also much more accurate than other word-level discriminators [13, 18] discussed in Sec. 4.2.

R4: Limited context of general descriptions. Both generator and discriminator objectives (Eqs 5 and 6) have included the full sentence (S) as conditional adversarial loss to convey rich context. Our word-level discriminator does not have trainable weights, and just works as a new approach to calculate an auxiliary loss, which encourages the generator to better disentangle different attributes, and actually may not affect the alignment of rich text context to the image.

R4: More baselines. Thanks for your suggestion; we will include more baselines in the paper. We only compare our method with ManiGAN, because both SISGAN and TAGAN fail to achieve an effective manipulation on both datasets, shown in Fig. 1 (top). As for the user study, please see above "R3: User study".

R4: Clarity. Following ManiGAN, there are no ground-truth modified images in our training. We just use paired data (I, S)$\rightarrow I$ to train the model, and our model is required to jointly solve text-to-image generation ($S \rightarrow I$) and text-irrelevant contents reconstruction ($I \rightarrow I$). $\mathcal{L}_{\text{DAMSM}}$ is referred to [28], and manipulative precision is to [14].

R5: Lightweight. Thanks for your suggestion. Saying "lightweight" means that our model has a much simpler structure with fewer parameters, but it still achieves a competitive performance. One reason to achieve this is that our word-level discriminator provides much accurate word-level training feedback related to each specific attributes, helping the generator to recognise them easily and thus enabling the possibility to simplify the architecture. The better disentanglement can be verified by the visualisation of attention in Fig. 6. We will include more details in our paper.

R5: Notation meaning. L represents the number of words in the word features w. Single optimisation epoch means the running time per epoch. We will clarify this in the paper and add corresponding notation definition in Fig. 2.

R5: Model size comparison. Please see Table 1: our model has fewer parameters for both generator and discriminator.

R5: No SISGAN and TAGAN results. In Fig. 1 (top), SISGAN and TAGAN fail to achieve an effective manipulation on CUB, and cannot produce realistic images on COCO. Thus, we only compare our method with ManiGAN.

R5: Other adjective types. For CUB, most descriptions are colours of different parts of a bird. For COCO, captions are mainly about "type of objects + type of locations". So, our method mainly focuses on modifying the colours of a bird for CUB, and the background or global style for COCO. See Fig. 1 (bottom) for various colours on the same image.

R5: CUB result of Fig. 6. Thanks for pointing this out. It is a mistake that a different random seed is chosen for "Ours w/o Dis."; we will correct it in the paper. There is no specific meaning for the order; we just present some words with corresponding highlighted visual regions to compare the performance of different word-level discriminators.

R5: Results on other parts of speech. Our proposed discriminator is used to provide additional word-level training feedback related to visual attributes, enabling a better disentanglement. We think there is no need to include other parts of speech in the proposed discriminator, because (1) it may not improve the manipulation ability, and (2) objective functions have included the full sentence (S) as conditional adversarial losses to convey rich context of text.

R6: Other evaluation matrix. Thanks for your suggestion. Please see Table 1, our method achieves better FID scores.

R6: Further improvement. Thanks for your suggestion. First, how to produce higher-quality results involving cross-domain features on the complex COCO dataset is still a challenge. Also, how to achieve an effective geometric translation (e.g., horse <-> car) by using language is the other direction of our future work.