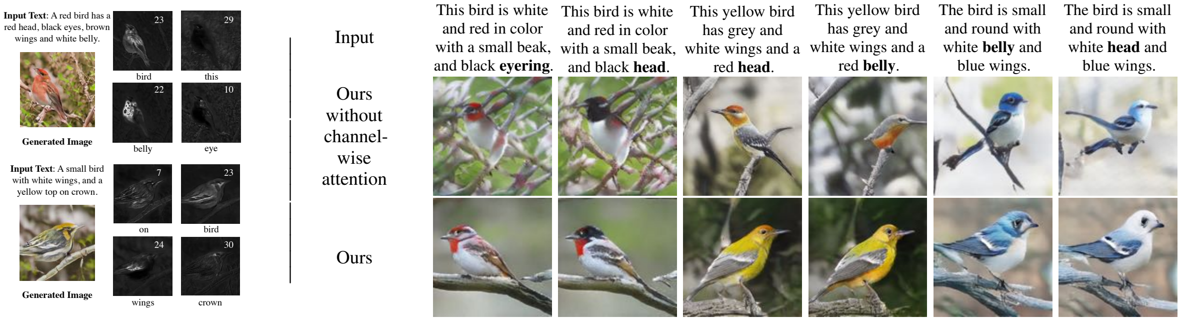
Input Text: A red bird has a red head, black eyes, brown wings and white belly.
Generated Image
23
bird
29
this
22
belly
10
eye
Input Text: A small bird with white wings, and a yellow top on crown.
Generated Image
7
on
23
bird
24
wings
30
crown
Input
Ours without channel-wise attention
Ours
This bird is white and red in color with a small beak, and black eyering.
This bird is white and red in color with a small beak, and black head.
This yellow bird has grey and white wings and a red head.
This yellow bird has grey and white wings and a red belly.
The bird is small and round with white belly and blue wings.
The bird is small and round with white head and blue wings.

Figure 1: Left: visualisation of feature channels. The number on the right top corner is the channel number. The word that has the highest correlation $\alpha_{i,j}$ in Eqn.1 with the channel is shown under the image. Right: ablation study.

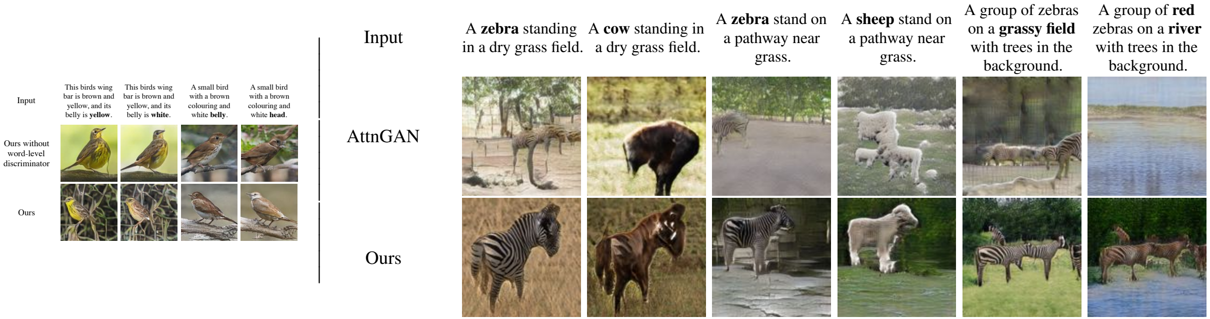
Input
Ours without word-level discriminator
Ours
This birds wing bar is brown and yellow, and its belly is yellow.
This birds wing bar is brown and yellow, and its belly is white.
A small bird with a brown colouring and white belly.
A small bird with a brown colouring and white head.
Input
AttnGAN
Ours
A zebra standing in a dry grass field.
A cow standing in a dry grass field.
A zebra stand on a pathway near grass.
A sheep stand on a pathway near grass.
A group of zebras on a grassy field with trees in the background.
A group of red zebras on a river with trees in the background.

Figure 2: Left: ablation study. Right: Qualitative comparison of two methods on the COCO dataset. Odd-numbered columns show the original text and even-numbered ones the modified text. Please zoom in to see clearer.

R1: The effectiveness of channel-wise attention. We experimentally find that the channel-wise attention highly correlates with semantically meaningful parts while the spatial attention focuses on colour descriptions. Fig.1(L) shows several channels of feature maps transformed with our channel-wise attention. It shows that channels closely correlate to semantic parts, such as belly, eye, wing, crown. This phenomenon is further verified by the results shown in Fig.1(R). Without channel-wise attention, the model fails to generate controllable results when we modify the text related to parts of a bird. In contrast, our full model with channel-wise attention can generate much better controllable results.

R1: The necessity of attention. In the paper Fig. 6 and Tab. 2, *SPM only* means the model does not incorporate the channel-wise attention but still has the spatial attention. We will clarify this in our main paper. Fig. 6 *SPM only* verifies that spatial attention works well if only colour information is modified. Our channel-wise attention is responsible for parts related modification as described above.

R1: Word-level discriminator. The word-level discriminator can help better disentangle different visual attributes and facilitate image manipulation as shown in Fig.2(L). With word-level discriminator, our model achieves better results, e.g., the original shape and colour of the bird are well preserved.

R1: Novelty of SPM. To our best knowledge, we are the first to apply perceptual loss in controllable text-to-image generation and show its effectiveness on reducing randomness involved in generation, which makes manipulation stable.

R2: Channel-wise attention. Please refer to "R1: The effectiveness of channel-wise attention."

R2: The importance of self-attention. We adopt the pre-trained RNN used in AttnGAN. The objective function used to train RNN is to improve text-image matching score based on cosine similarity. Thus, it is reasonable to use self-attention to reduce impact from less important words which has small cosine similarity.

R2: SPM and diversity of the model. The SPM would not affect the diversity of the model, as the diversity comes from the random noise sampled from a normal distribution in stage 0. We will clarify this in our paper.

R2: The controllability and out-of-distribution on COCO. In this rebuttal, we focus on the animal subset of COCO. As shown in Fig.2, our model can effectively manipulate the images compared with AttnGAN and also work well on out of distribution queries, e.g., red zebras on the river.

R3: Shape of the visual features, sensitivity. v in equation 1 indicates the generated image features whose shape is $B \times N \times (H * W)$, where N is 32, $H * W$ is $64 * 64$ in stage 1 and $128 * 128$ in stage 2. The v in equation 2 indicates real image features whose shape is same as generated images. We will clarify this.

R3: CUB and COCO experiments COCO has much more diversified descriptions where each only contains few examples. The low density of corresponding images is a reason for the different behaviours on the two datasets. Also, captions in COCO are more abstract and focus on the category of objects, which makes text-to-image generation be more challenging on COCO. However, we still produce much better results compared with AttnGAN shown in Fig.2(R).