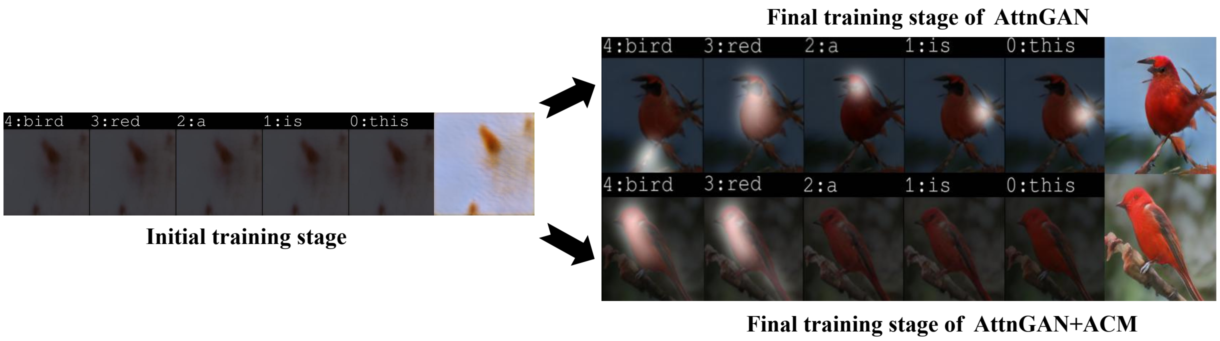
Supplementary Material for SEGAN

Section A: Analysis of Attention Regularization Term


Final training stage of AttnGAN

Initial training stage

Final training stage of AttnGAN+ACM

Figure 1: Word-level attention weight maps generated of initial training stage (image on the left), final training stage of AttnGAN (image on the upper right corner) and final training stage of AttnGAN+ACM (image on the lower right corner).

The ACM (**page 3, 4**) contains the DAMSM in AttnGAN and *proposed attention regularization term* ($\mathcal{L}_c$), namely $\mathcal{L}_{ACM} = \mathcal{L}_{DAMSM} + \lambda_1 \mathcal{L}_c$ (**Eq.2, page 4**). This section gives a more thorough discussion on the attention regularization term $\mathcal{L}_c = \sum_{i,j} (\min(r_{i,j}, \alpha))^2$ (**page 3**).

In Figure 1, left, we show the word-level attention weight maps generated in the initial training stage. Using AttnGAN, after the training stage, the attention weight maps are shown in the upper right. Using our AttnGAN + ACM, after the training, the attention weight maps are shown in the botton right. In this visualization, for the subregions whose semantic meanings have corresponding expression in the description text, these regions are highlighted in white and listed together with their most relevant words.

In the initialization phase of the training, the attention weights (r_{ij}) for all words are zero, and all subregions in images are dark. Attention weights (r_{ij}) of all words are lower than threshold α. During the training process, attention weights from key visual words play roles on cross-modal similarity matching in $\mathcal{L}_{DAMSM}$. The $\mathcal{L}_{DAMSM}$ in AttnGAN would then tend to push the attention weights of *visually important words* to exceed the threshold α. And their attention weights will be preserved.

Visually unimportant words produce redundant information for cross-modal similarity matching in $\mathcal{L}_{DAMSM}$, and are not conducive to better synthesized images. Hence, $\mathcal{L}_{DAMSM}$ becomes irrelavent to *visually unimportant words* and does not push their attention weights to exceed the threshold α. In contrast, to minimize $\mathcal{L}_c$, their attention weights should move towards 0.

As shown in Figure 1, if we introduce the ACM into AttnGAN, in final training stage, the AttnGAN+ACM could better focus on visually important words and synthesize higher-quality images (image on the lower right corner). Without using our ACM, the AttnGAN pays attention on every word including unimportant ones. But such kind of attention may result in strange synthesized subregions (see the result image on the upper right corner).

Section B: More Visual Comparison Results

In this section, we show more visual comparison results between our SEGAN and AttnGAN (baseline) on the Bird and COCO dataset in Figure 2, Figure 3, Figure 4, and Figure 5. These visual comparison results further demonstrate the generalization ability of the SEGAN.

Figure 6, Figure 7, and Figure 8 visualize the attention weight maps on synthesized images. For sub-regions whose semantic meaning are expressed in the description text, the attentions are allocated to their most relevant words (bright regions in Figure 6, Figure 7 and Figure 8). Compared with AttnGAN, SEGAN could better focus on visually important words and synthesize higher-quality images.

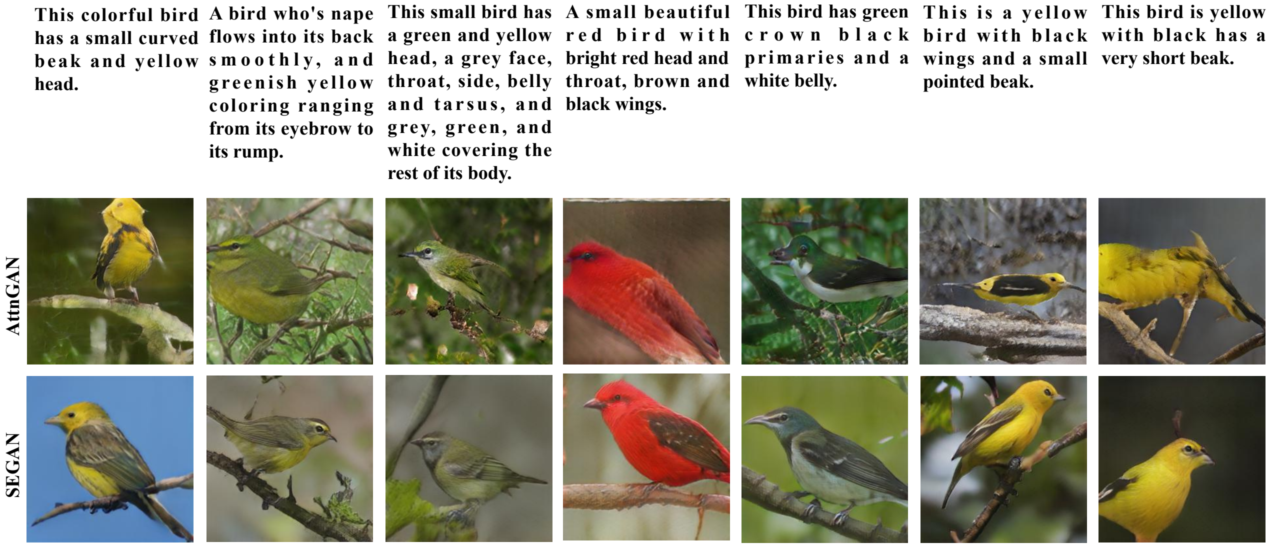

This colorful bird has a small curved beak and yellow head.
A bird who's nape flows into its back smoothly, and greenish yellow coloring ranging from its eyebrow to its rump.
This small bird has a green and yellow head, a grey face, throat, side, belly and tarsus, and grey, green, and white covering the rest of its body.
A small beautiful red bird with bright red head and throat, brown and black wings.
This bird has green crown black primaries and a white belly.
This is a yellow bird with black wings and a small pointed beak.
This bird is yellow with black has a very short beak.
AttnGAN
SEGAN

Figure 2: Images of 256 × 256 resolution are generated by our SEGAN and AttnGAN conditioned on text descriptions from CUB test datasets.

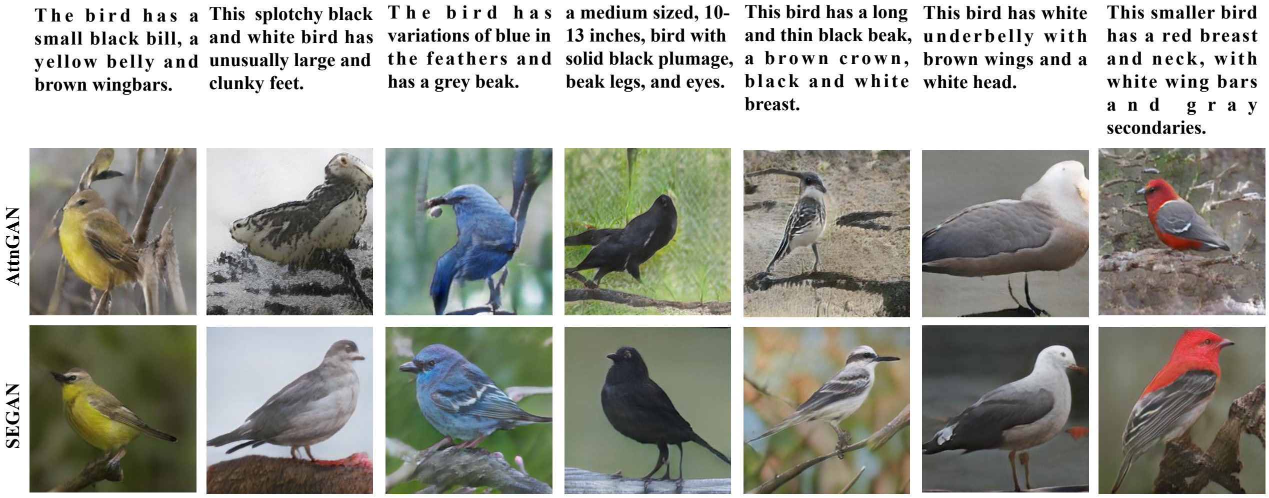

The bird has a small black bill, a yellow belly and brown wingbars.
This splotchy black and white bird has unusually large and clunky feet.
The bird has variations of blue in the feathers and has a grey beak.
a medium sized, 10-13 inches, bird with solid black plumage, beak legs, and eyes.
This bird has a long and thin black beak, a brown crown, black and white breast.
This bird has white underbelly with brown wings and a white head.
This smaller bird has a red breast and neck, with white wing bars and gray secondaries.
AttnGAN
SEGAN

Figure 3: Images of 256 × 256 resolution are generated by our SEGAN and AttnGAN conditioned on text descriptions from CUB test datasets.

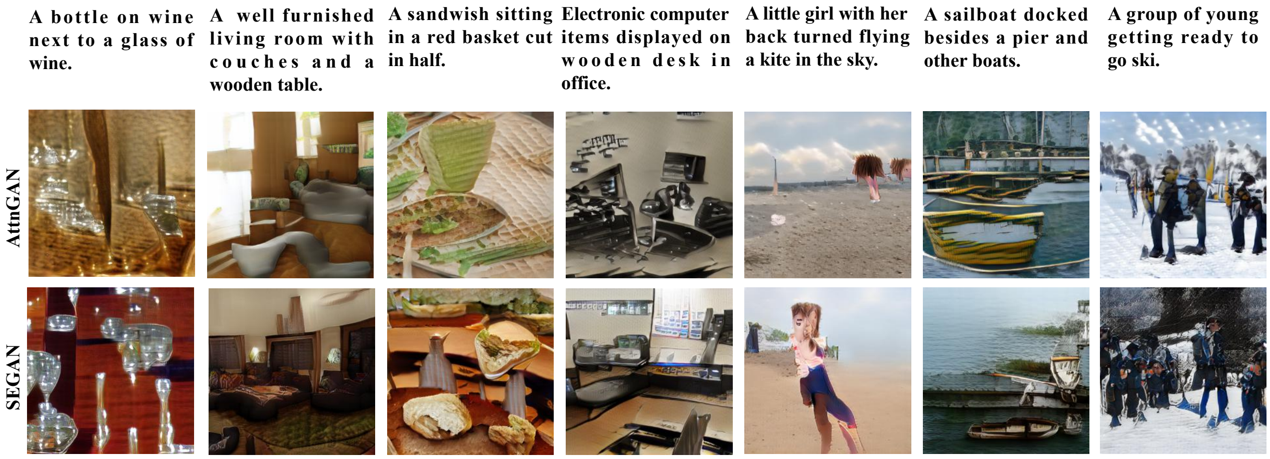

A bottle on wine next to a glass of wine.
A well furnished living room with couches and a wooden table.
A sandwish sitting in a red basket cut in half.
Electronic computer items displayed on wooden desk in office.
A little girl with her back turned flying a kite in the sky.
A sailboat docked besides a pier and other boats.
A group of young getting ready to go ski.
AttnGAN
SEGAN

Figure 4: Images of 256 × 256 resolution are generated by our SEGAN and AttnGAN conditioned on text descriptions from COCO test datasets.

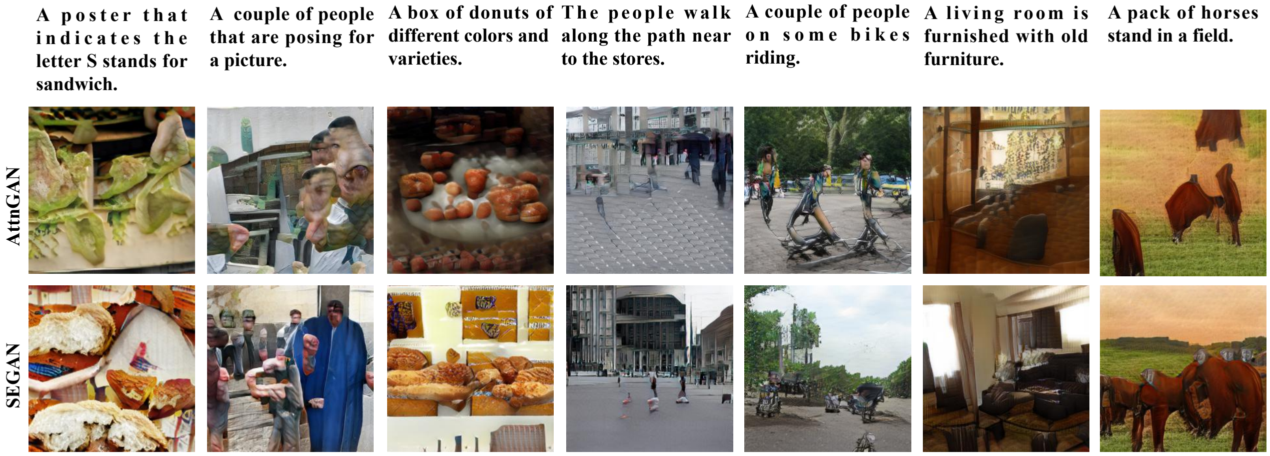

A poster that indicates the letter S stands for sandwich.
A couple of people that are posing for a picture.
A box of donuts of different colors and varieties.
The people walk along the path near to the stores.
A couple of people on some bikes riding.
A living room is furnished with old furniture.
A pack of horses stand in a field.
AttnGAN
SEGAN

Figure 5: Images of 256 × 256 resolution are generated by our SEGAN and AttnGAN conditioned on text descriptions from COCO test datasets.

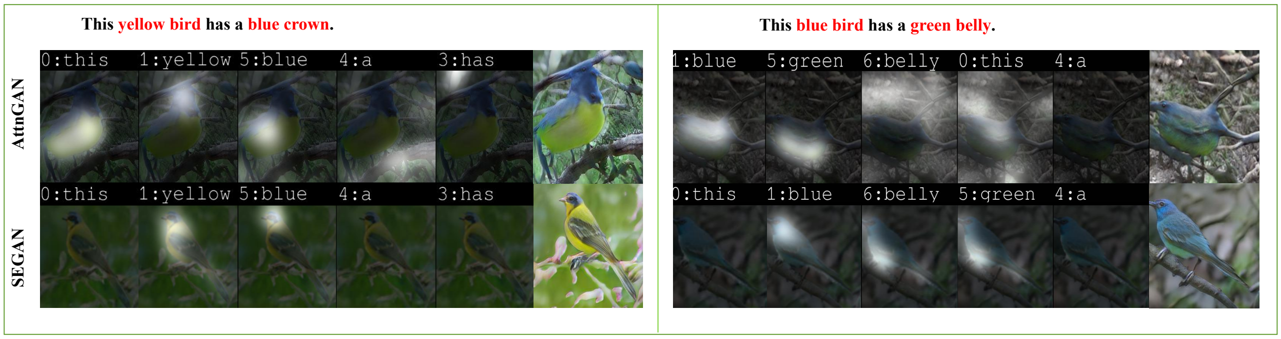

This yellow bird has a blue crown.
This blue bird has a green belly.
AttnGAN
SEGAN
0:this 1:yellow 5:blue 4:a 3:has
0:this 1:yellow 5:blue 4:a 3:has
1:blue 5:green 6:belly 0:this 4:a
0:this 1:blue 6:belly 5:green 4:a

Figure 6: Word-level Attention Weight Maps generated from AttnGAN and our SEGAN. In the text description, red font is the key words, black font is the non-key words.

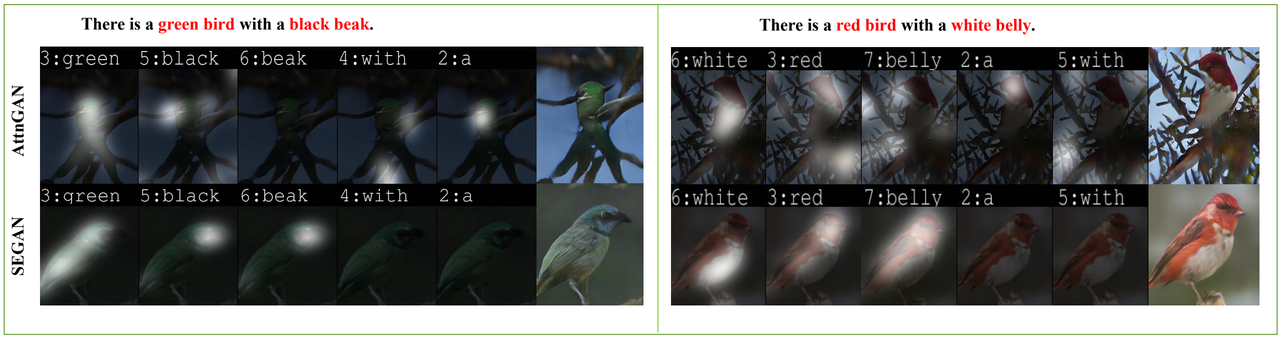

There is a green bird with a black beak.
There is a red bird with a white belly.
AttnGAN
SEGAN
3:green 5:black 6:beak 4:with 2:a
3:green 5:black 6:beak 4:with 2:a
6:white 3:red 7:belly 2:a 5:with
6:white 3:red 7:belly 2:a 5:with

Figure 7: Word-level Attention Weight Maps generated from AttnGAN and our SEGAN. In the text description, red font is the key words, black font is the non-key words.

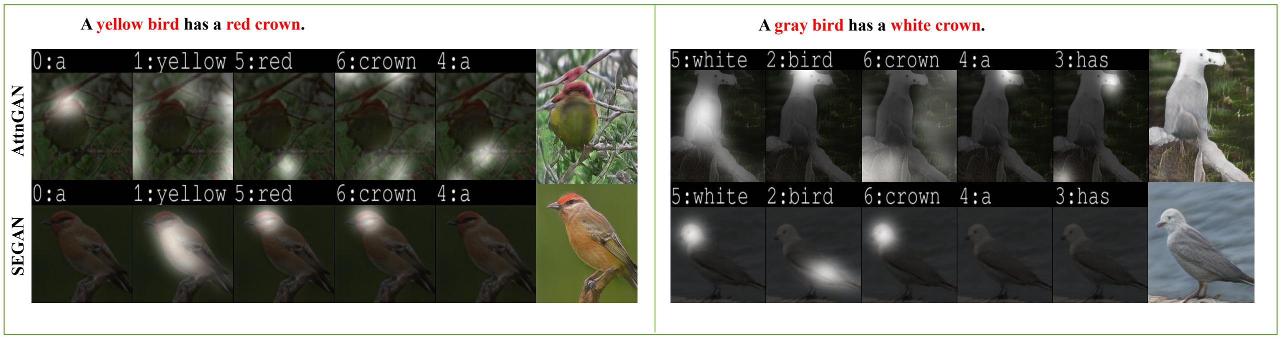

A yellow bird has a red crown.
A gray bird has a white crown.
AttnGAN
SEGAN
0:a 1:yellow 5:red 6:crown 4:a
0:a 1:yellow 5:red 6:crown 4:a
5:white 2:bird 6:crown 4:a 3:has
5:white 2:bird 6:crown 4:a 3:has

Figure 8: Word-level Attention Weight Maps generated from AttnGAN and our SEGAN. In the text description, red font is the key words, black font is the non-key words.