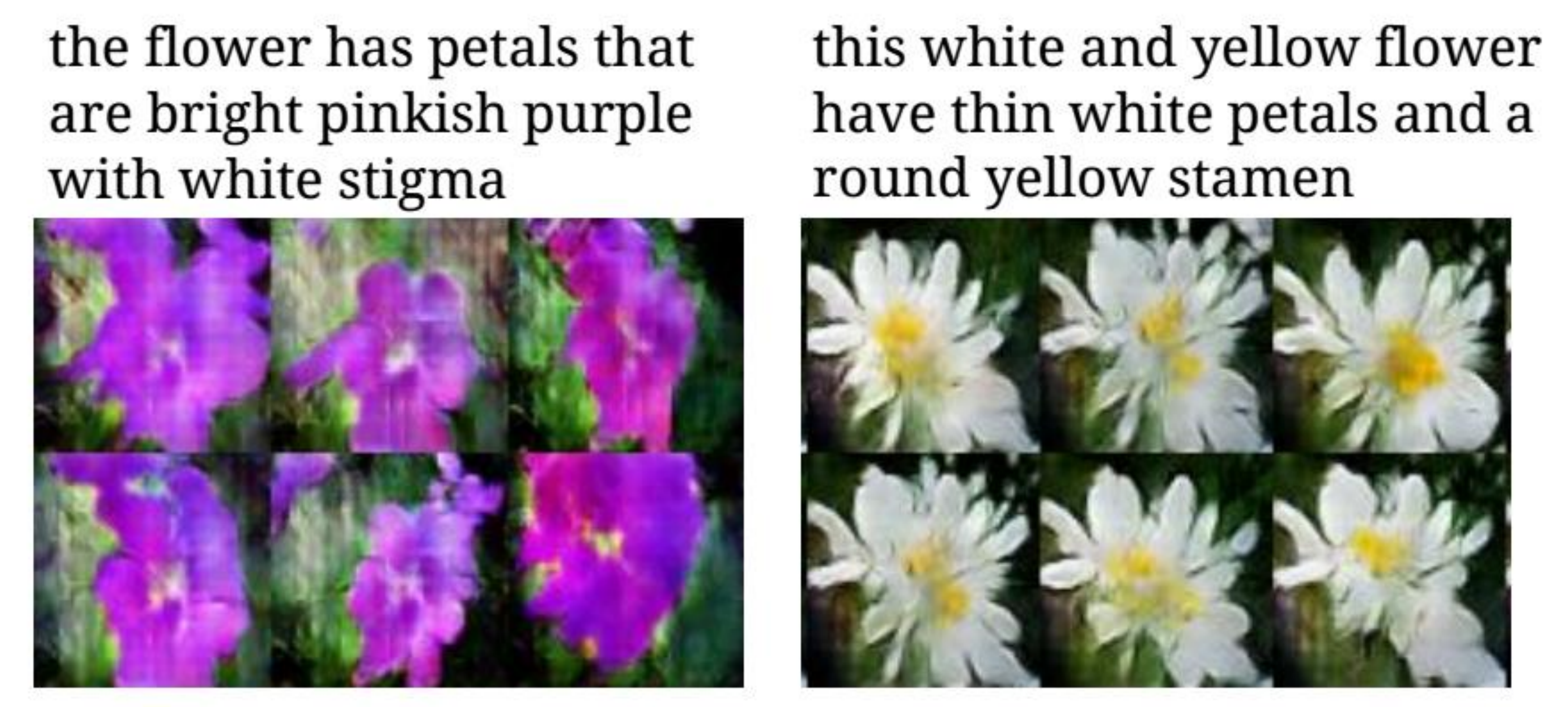
Generative Adversarial Text to Image Synthesis

Reed S, Akata Z, Yan X, et al. ICML 2016

Text2Image



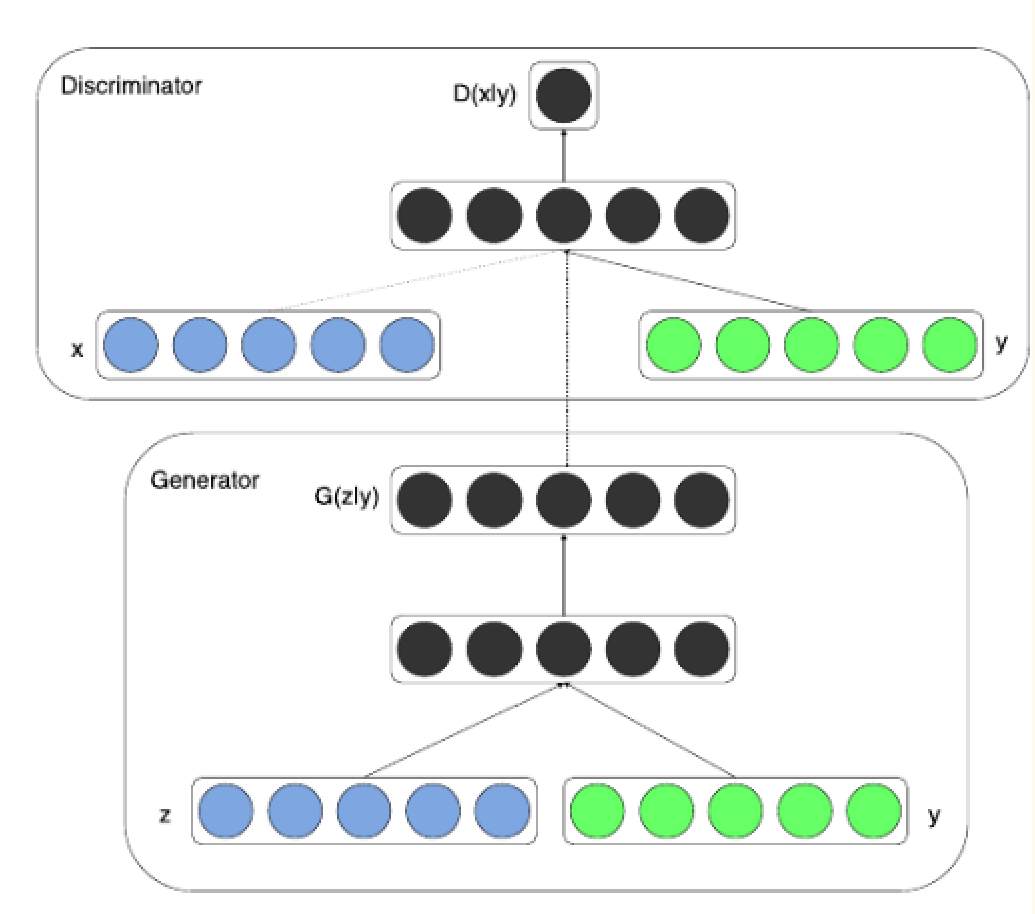
the flower has petals that are bright pinkish purple with white stigma

this white and yellow flower have thin white petals and a round yellow stamen

Text2Image

DCGAN + CGAN，文字描述由一个现成的词嵌入方法生成向量*

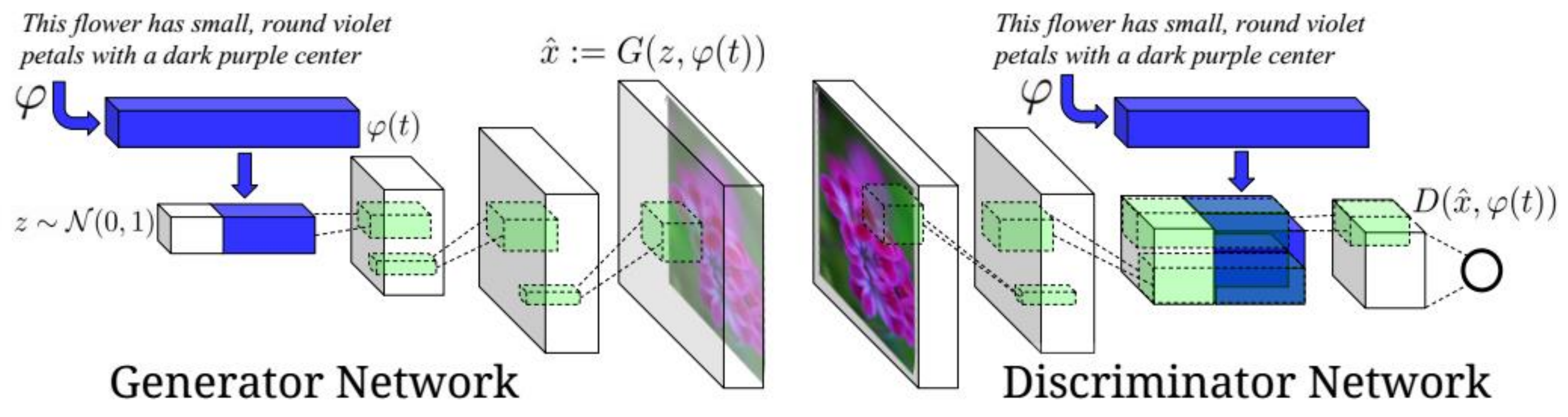
$$\min_{G}\max_{D} V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{x\sim p_{z}(z)}[\log(1 - D(G(z)))]$$

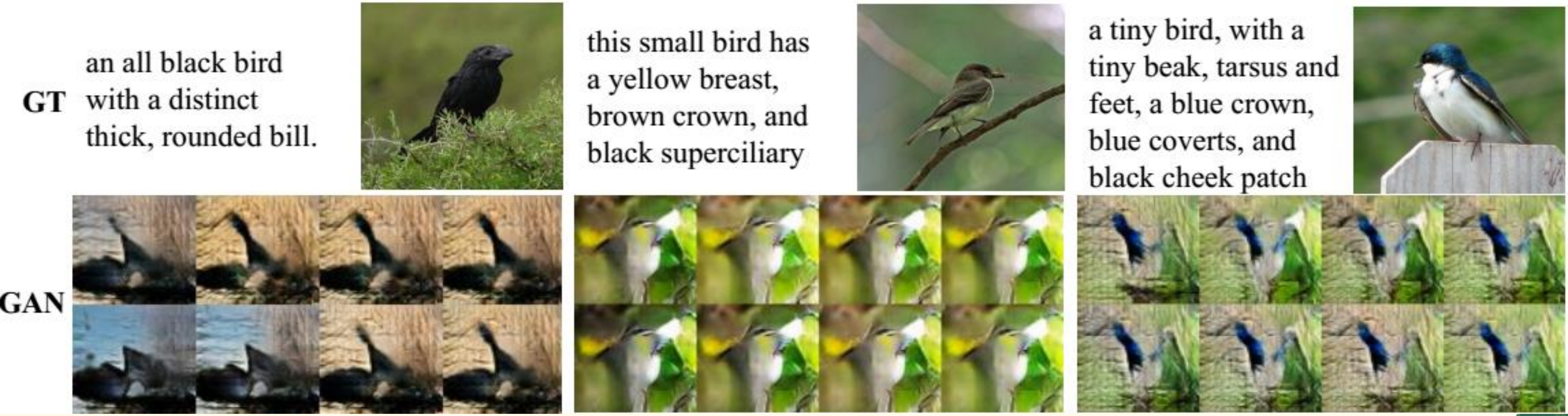


*** Reed, S., Akata, Z., Lee, H., and Schiele, B. Learning deep representations for fine-grained visual descriptions. In CVPR, 2016.**

Text2Image

DCGAN + CGAN，文字描述t由一个现成的词嵌入方法生成向量$\boldsymbol{\varphi(t)}$


This flower has small, round violet petals with a dark purple center
φ
$\varphi(t)$
$z \sim \mathcal{N}(0,1)$
$\hat{x} := G(z, \varphi(t))$
This flower has small, round violet petals with a dark purple center
φ
$D(\hat{x}, \varphi(t))$
Generator Network
Discriminator Network



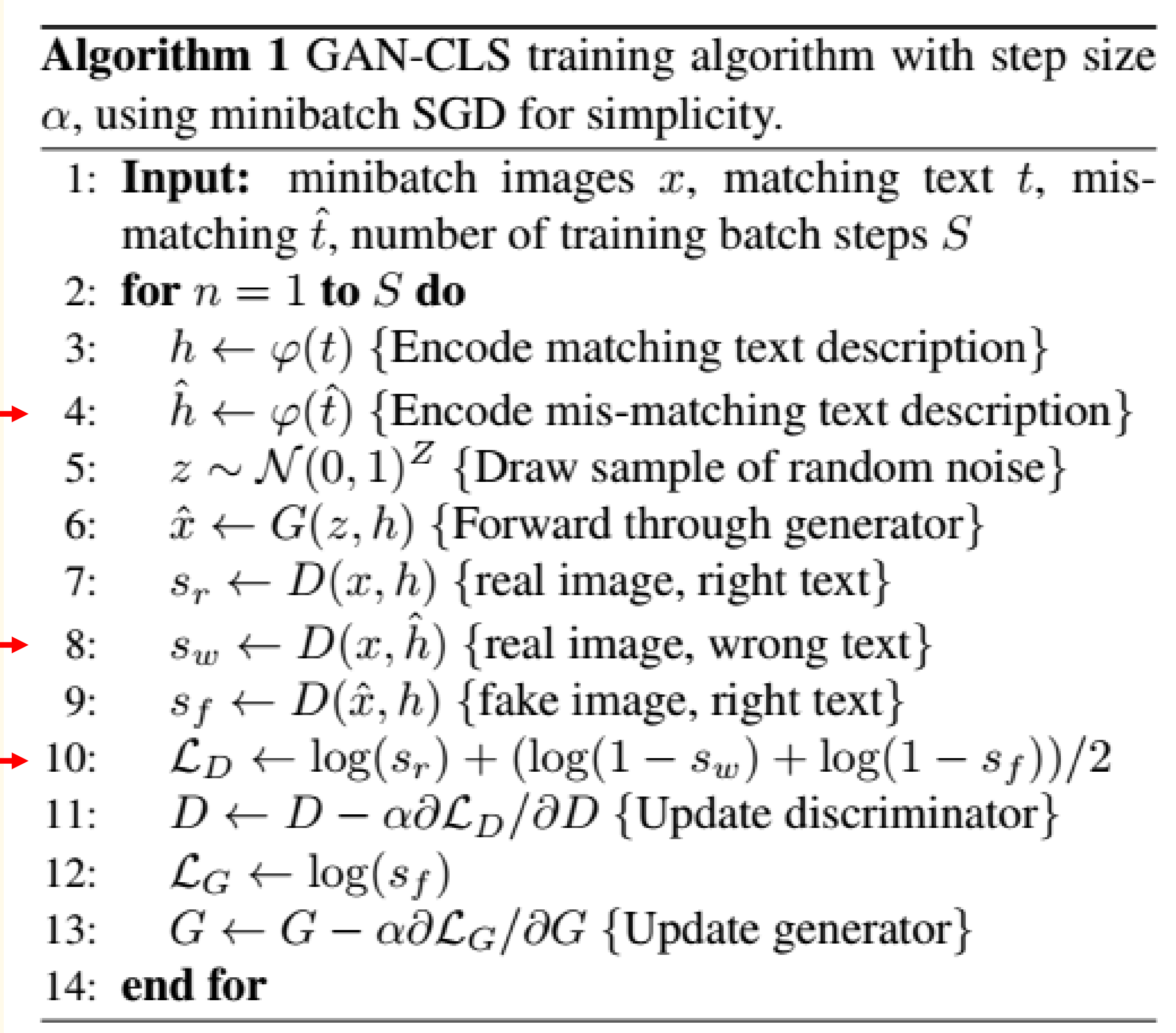
an all black bird with a distinct thick, rounded bill.

this small bird has a yellow breast, brown crown, and black superciliary

a tiny bird, with a tiny beak, tarsus and feet, a blue crown, blue coverts, and black cheek patch

GAN-CLS 通过加入反例来训练D(与G无关)

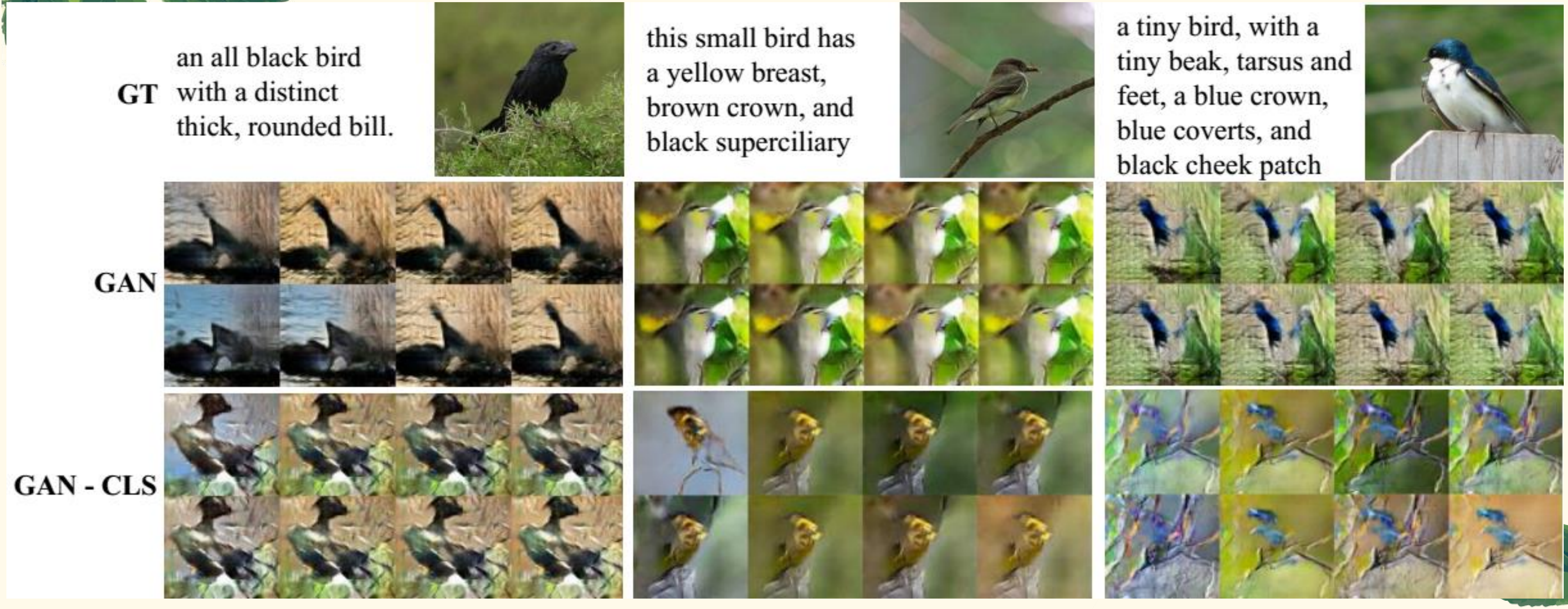
Intuition: the discriminator has no explicit notion of whether real training images match the text embedding context.



Algorithm 1 GAN-CLS training algorithm with step size α, using minibatch SGD for simplicity.

```
Input: minibatch images $x$, matching text $t$, mis-matching $\hat{t}$, number of training batch steps $S$
for $n = 1$ to $S$ do
    $h \leftarrow \varphi(t)$ {Encode matching text description}
    $\hat{h} \leftarrow \varphi(\hat{t})$ {Encode mis-matching text description}
    $z \sim \mathcal{N}(0,1)^Z$ {Draw sample of random noise}
    $\hat{x} \leftarrow G(z, h)$ {Forward through generator}
    $s_r \leftarrow D(x, h)$ {real image, right text}
    $s_w \leftarrow D(x, \hat{h})$ {real image, wrong text}
    $s_f \leftarrow D(\hat{x}, h)$ {fake image, right text}
    $\mathcal{L}_D \leftarrow \log(s_r) + (\log(1 - s_w) + \log(1 - s_f))/2$
    $D \leftarrow D - \alpha \partial \mathcal{L}_D / \partial D$ {Update discriminator}
    $\mathcal{L}_G \leftarrow \log(s_f)$
    $G \leftarrow G - \alpha \partial \mathcal{L}_G / \partial G$ {Update generator}
end for
```

一张图片和他的一个文字描述是一个样本对



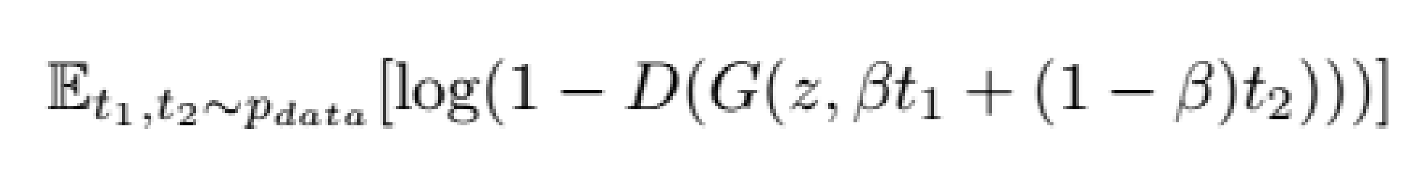
an all black bird with a distinct thick, rounded bill.

this small bird has a yellow breast, brown crown, and black superciliary

a tiny bird, with a tiny beak, tarsus and feet, a blue crown, blue coverts, and black cheek patch

GAN-INT 通过插值法增加文本样本数量(针对G)

Deep networks have been shown to learn representations in which interpolations between embedding pairs tend to be near the data manifold



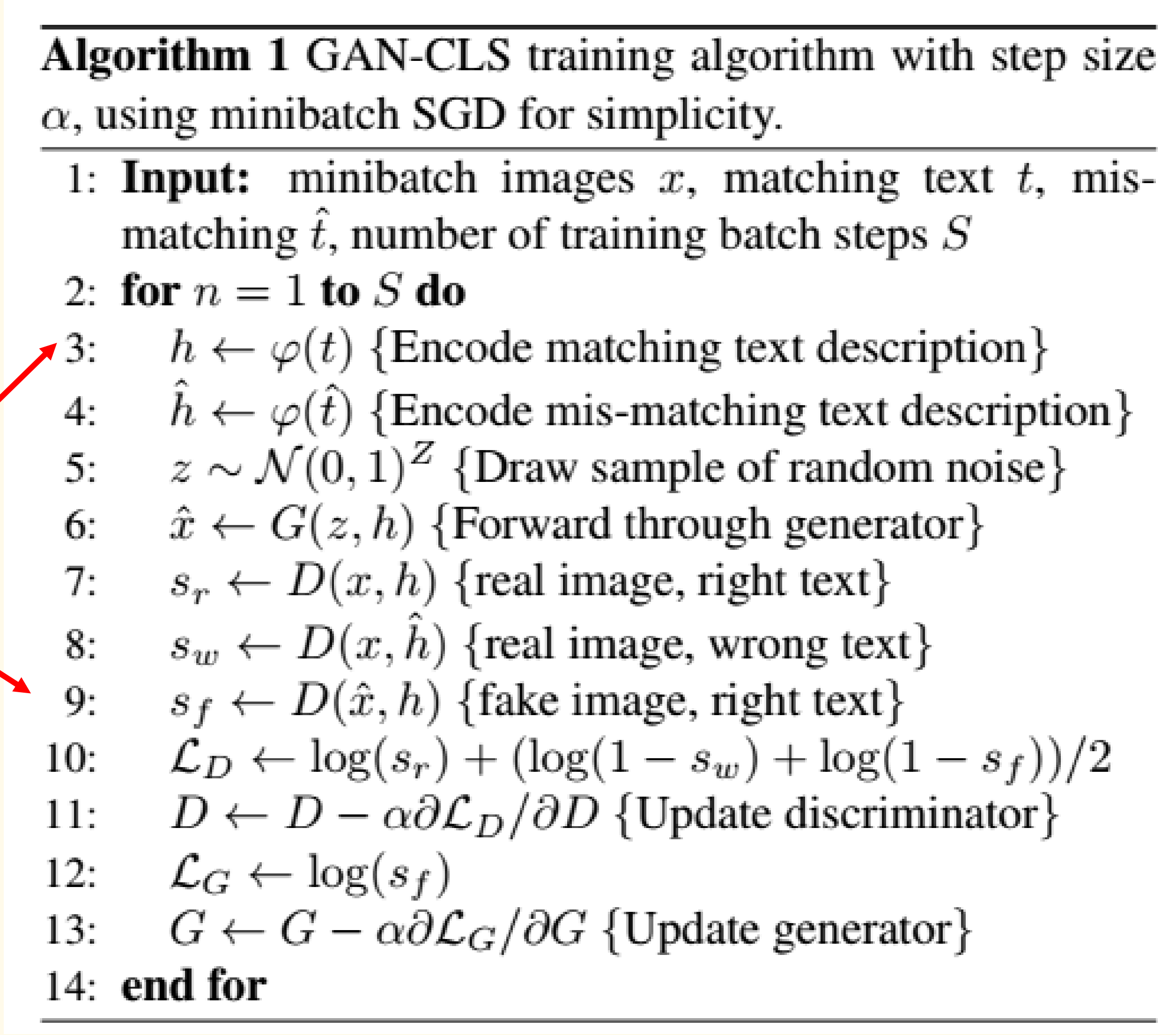
$$\mathbb{E}_{t_1,t_2\sim p_{data}}[\log(1-D(G(z,\beta t_1+(1-\beta)t_2)))]$$

Because the interpolated embeddings are synthetic, the discriminator D does not have “real” corresponding image and text pairs to train on. However, D learns to predict whether image and text pairs match or not. Thus, if D does a good job at this, then by satisfying D on interpolated text embeddings G can learn to fill in gaps on the data manifold in between training points.

原来的样本数量不足以使得判别器**D**判断图文的关系，通过插值法增加样本数使得**D**能学得他们之间的关系。

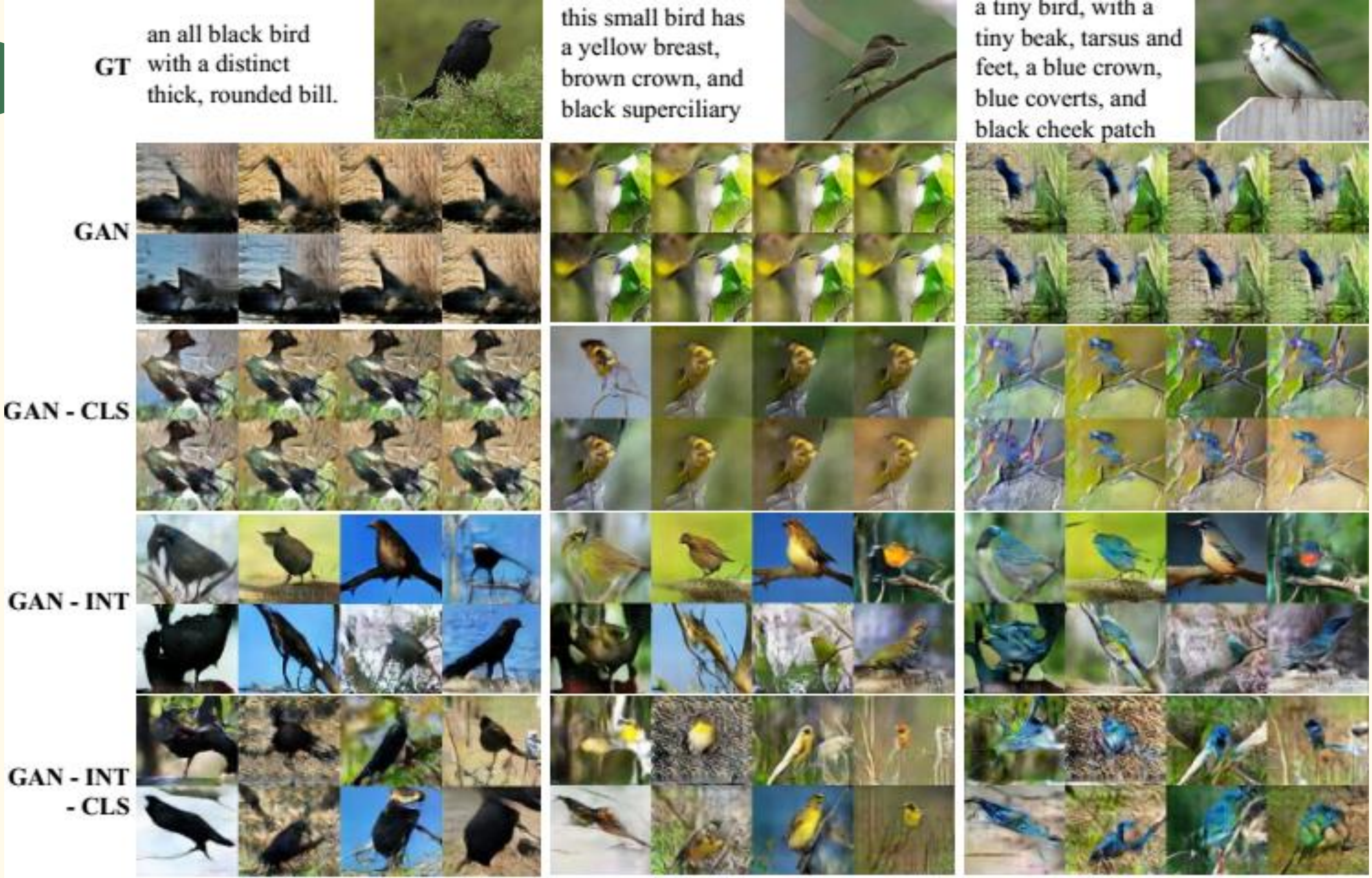
GAN-INT-CLS 两种改进方法结合

差值得到的文字描述当作真实的对应文本



Algorithm 1 GAN-CLS training algorithm with step size α, using minibatch SGD for simplicity.

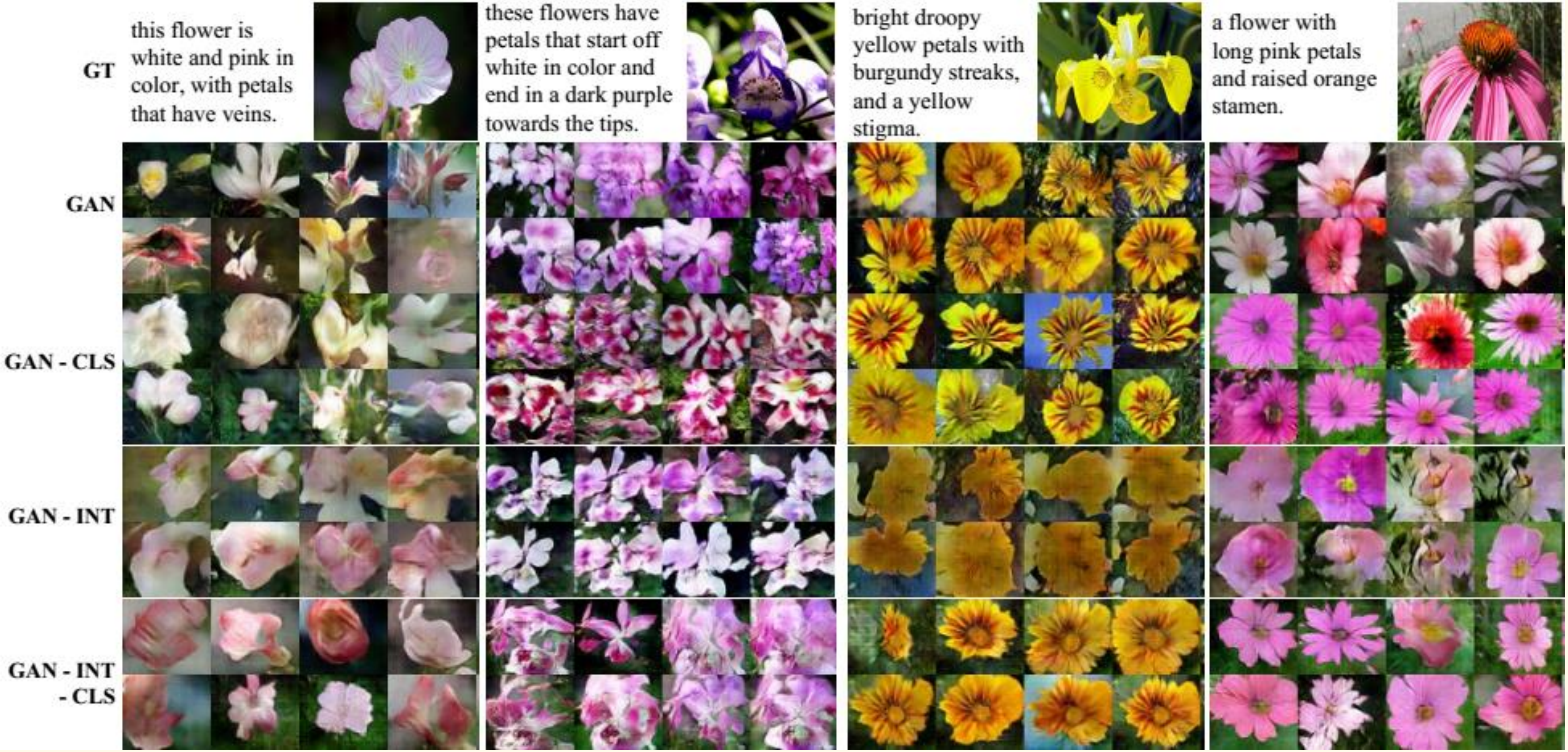
```
Input: minibatch images $x$, matching text $t$, mis-matching $\hat{t}$, number of training batch steps $S$
for $n = 1$ to $S$ do
    $h \leftarrow \varphi(t)$ {Encode matching text description}
    $\hat{h} \leftarrow \varphi(\hat{t})$ {Encode mis-matching text description}
    $z \sim \mathcal{N}(0,1)^Z$ {Draw sample of random noise}
    $\hat{x} \leftarrow G(z, h)$ {Forward through generator}
    $s_r \leftarrow D(x, h)$ {real image, right text}
    $s_w \leftarrow D(x, \hat{h})$ {real image, wrong text}
    $s_f \leftarrow D(\hat{x}, h)$ {fake image, right text}
    $\mathcal{L}_D \leftarrow \log(s_r) + (\log(1 - s_w) + \log(1 - s_f))/2$
    $D \leftarrow D - \alpha \partial \mathcal{L}_D / \partial D$ {Update discriminator}
    $\mathcal{L}_G \leftarrow \log(s_f)$
    $G \leftarrow G - \alpha \partial \mathcal{L}_G / \partial G$ {Update generator}
end for
```



an all black bird with a distinct thick, rounded bill.

this small bird has a yellow breast, brown crown, and black superciliary

a tiny bird, with a tiny beak, tarsus and feet, a blue crown, blue coverts, and black cheek patch



this flower is white and pink in color, with petals that have veins.

these flowers have petals that start off white in color and end in a dark purple towards the tips.

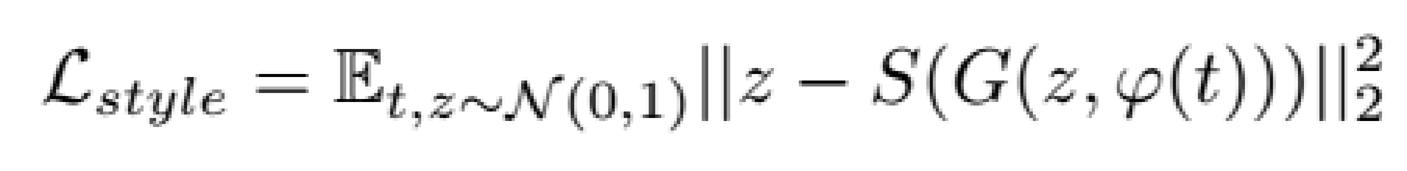
bright droopy yellow petals with burgundy streaks, and a yellow stigma.

a flower with long pink petals and raised orange stamen.

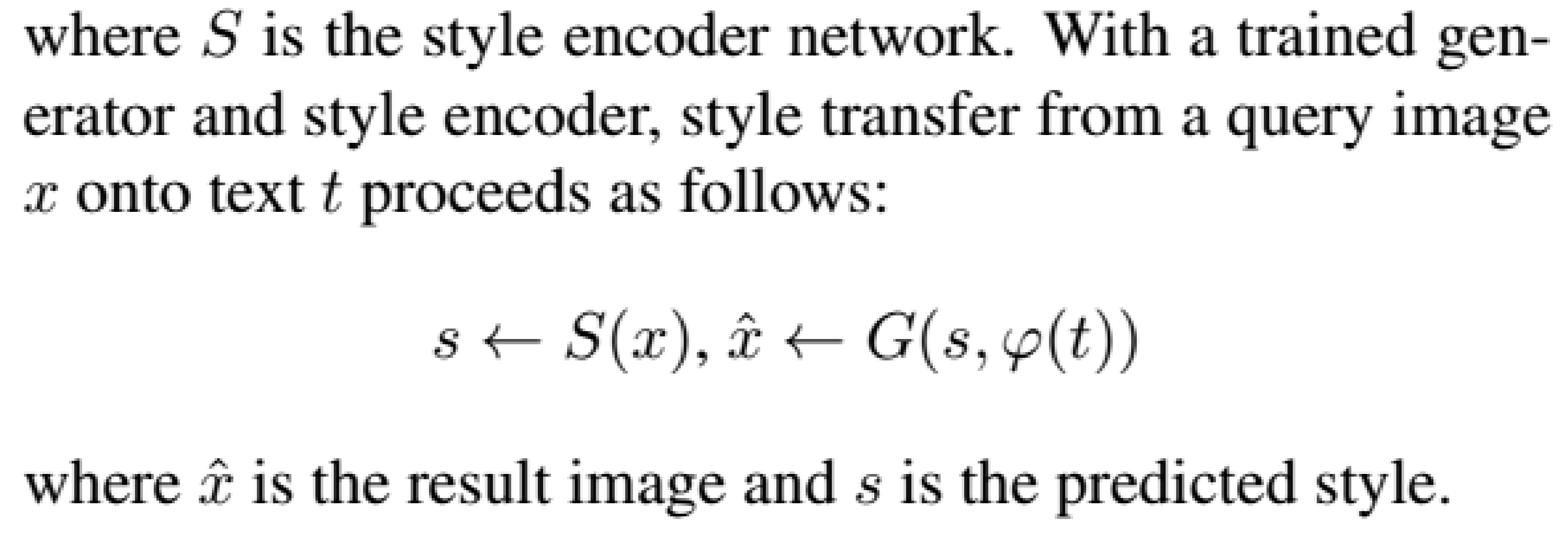
CUB contains 200 bird species with 11,788 images.
Oxford-102 contains 8,189 images of flowers from 102 different categories.

风格转换 – 文字只描述了内容，风格一定在噪声z中蕴含

Solution: train a convolutional network S to invert G to regress from samples x ˆ back onto z.



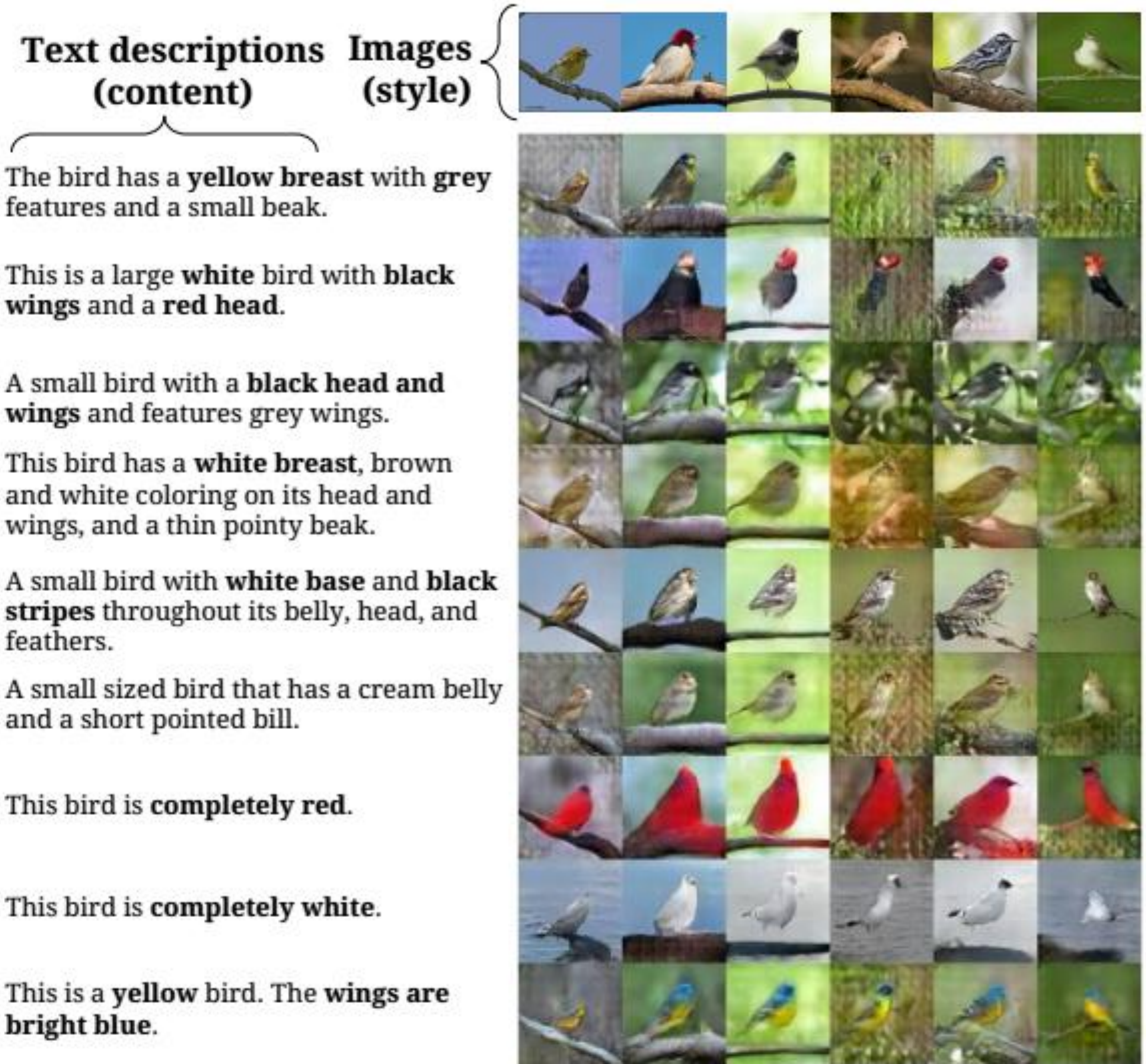
$$\mathcal{L}_{style} = \mathbb{E}_{t,z\sim\mathcal{N}(0,1)}||z - S(G(z,\varphi(t)))||_2^2$$



where S is the style encoder network. With a trained generator and style encoder, style transfer from a query image x onto text t proceeds as follows:

$$s \leftarrow S(x), \hat{x} \leftarrow G(s,\varphi(t))$$

where $\hat{x}$ is the result image and s is the predicted style.



The bird has a **yellow breast** with **grey** features and a small beak.

This is a large **white** bird with **black wings** and a **red head**.

A small bird with a **black head and wings** and features grey wings.

This bird has a **white breast**, brown and white coloring on its head and wings, and a thin pointy beak.

A small bird with **white base** and **black stripes** throughout its belly, head, and feathers.

A small sized bird that has a cream belly and a short pointed bill.

This bird is **completely red**.

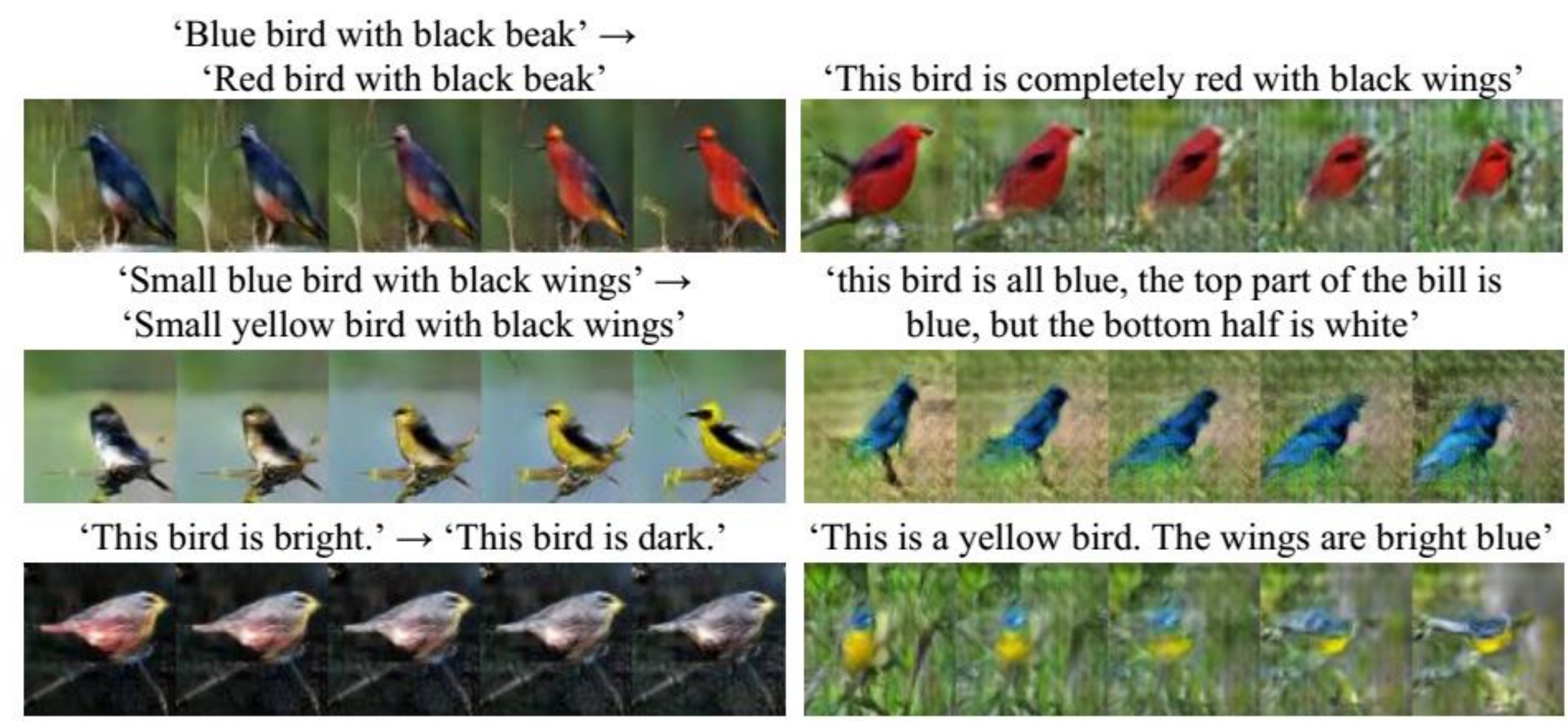
This bird is **completely white**.

This is a **yellow** bird. The **wings are bright blue**.

$$s \leftarrow S(x)$$

$$\hat{x} \leftarrow G(s, \varphi(t))$$

插值法实验 - 左边噪声z固定，对文字描述插值，右边相反



‘Blue bird with black beak’ → ‘Red bird with black beak’

‘This bird is completely red with black wings’

‘Small blue bird with black wings’ → ‘Small yellow bird with black wings’

‘this bird is all blue, the top part of the bill is blue, but the bottom half is white’

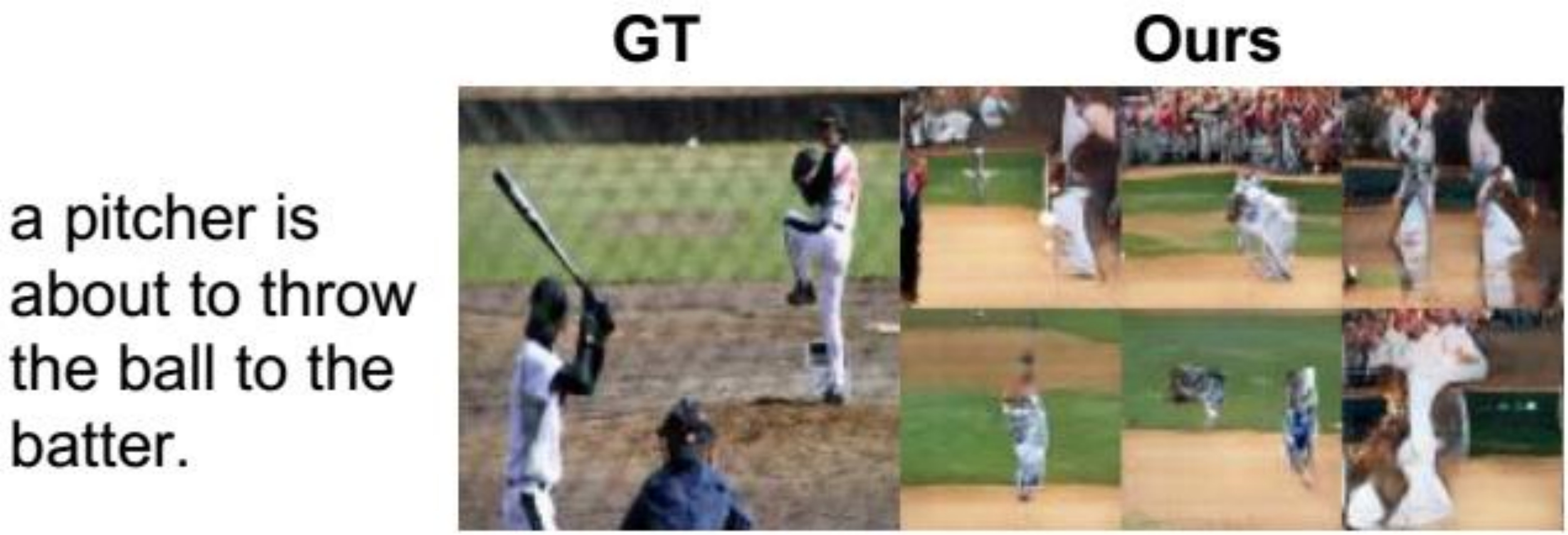
‘This bird is bright.’ → ‘This bird is dark.’

‘This is a yellow bird. The wings are bright blue’

Problem

the generated scenes are not usually coherent; for example the human-like blobs in the baseball scenes lack clearly articulated parts.

Low resolution (64x64像素)



a pitcher is about to throw the ball to the batter.

StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks

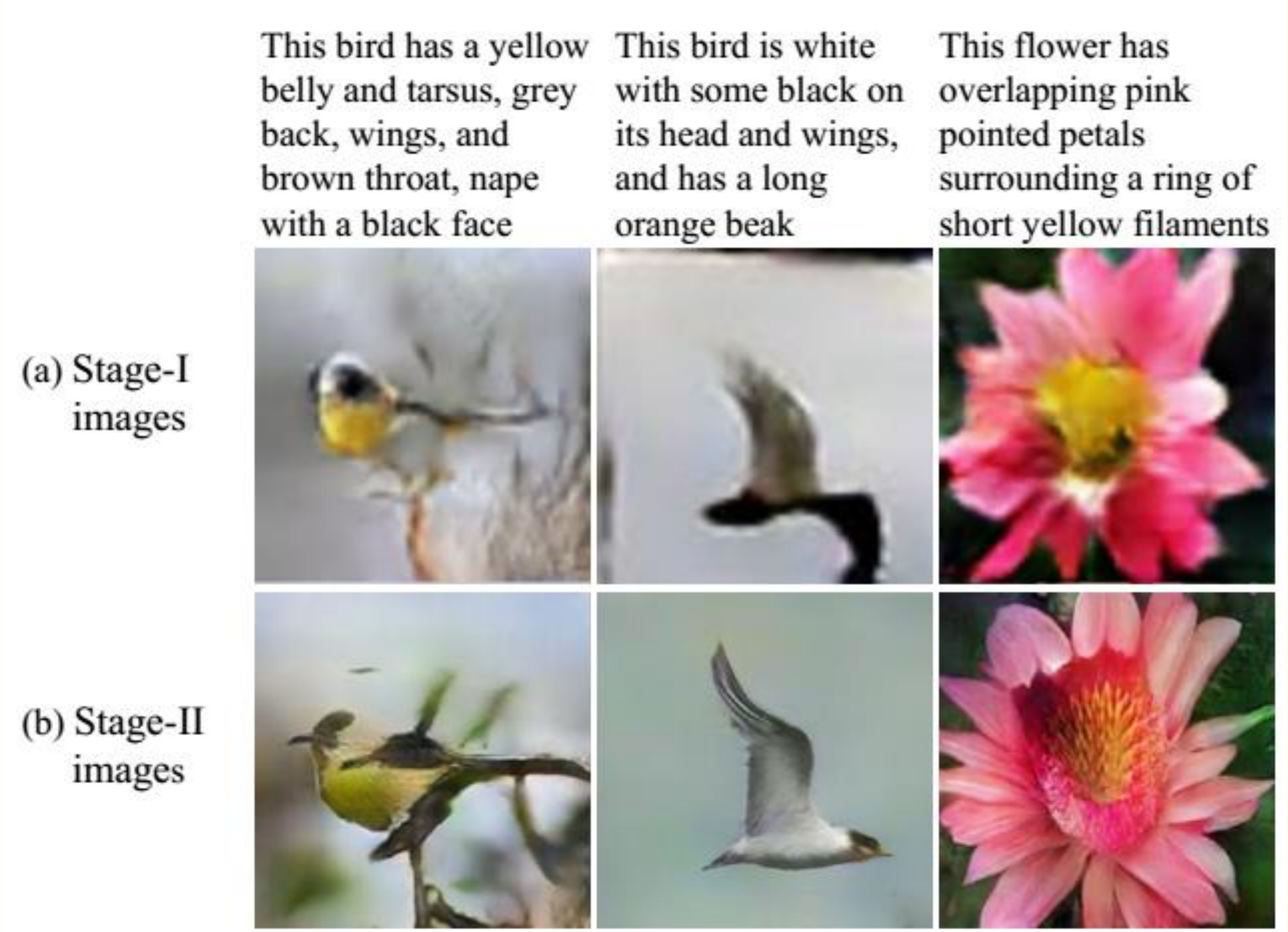
Zhang H, Xu T, Li H, et al.
arXiv preprint. 2016.

Text2Image

the space of plausible images given text descriptions is multimodal. There are a large number of images that correctly fit the given text description

给定文字描述作为条件，输入是噪声z拼上文字向量。一阶段只生成大概的轮廓和颜色的低分辨率图像

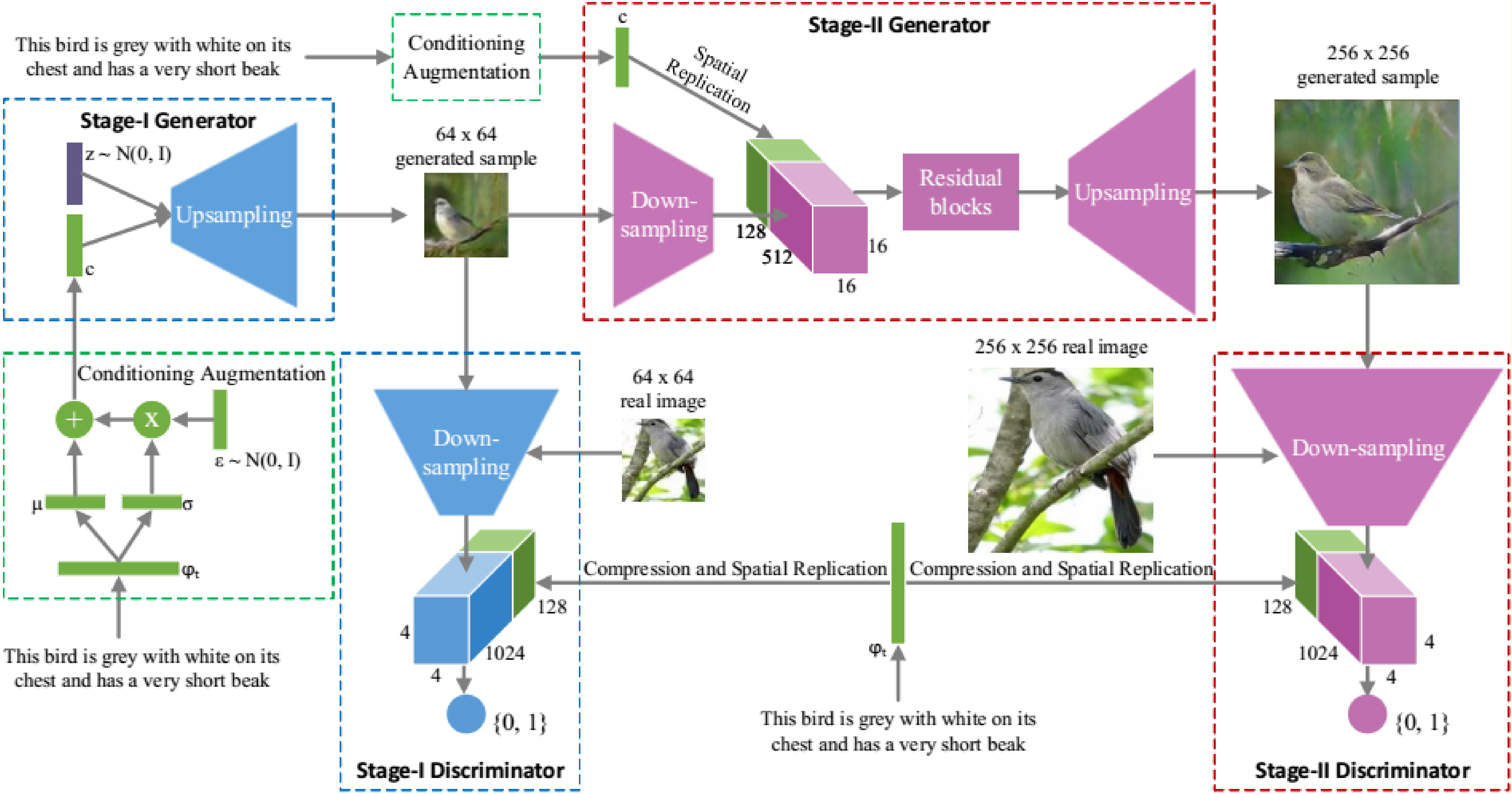
给定文字描述和一阶段的图像作为输入，二阶段生成细节和对一阶段改进



This bird has a yellow belly and tarsus, grey back, wings, and brown throat, nape with a black face

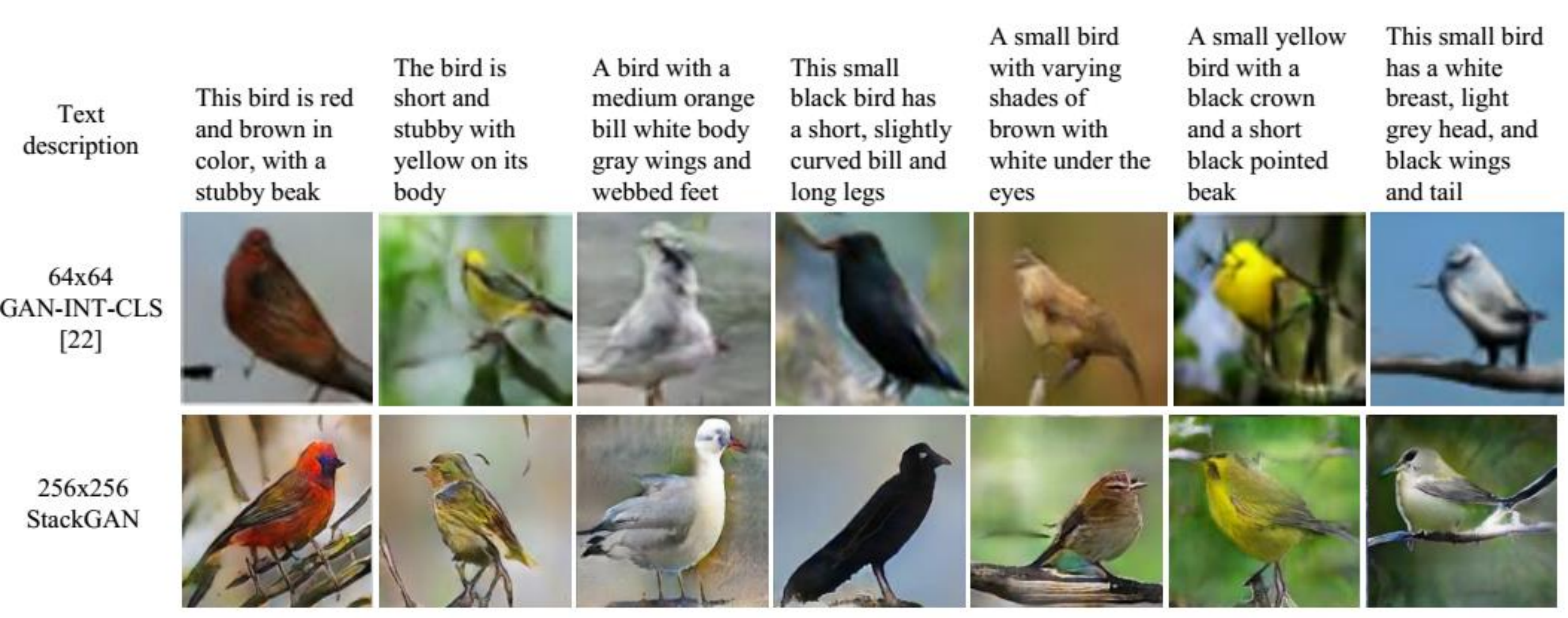
This bird is white with some black on its head and wings, and has a long orange beak

This flower has overlapping pink pointed petals surrounding a ring of short yellow filaments


This bird is grey with white on its chest and has a very short beak
Conditioning Augmentation
c
Stage-II Generator
Spatial Replication
256 x 256 generated sample
Stage-I Generator
z ~ N(0, I)
c
Upsampling
64 x 64 generated sample
Down-sampling
128
512
16
16
Residual blocks
Upsampling
256 x 256 real image
Conditioning Augmentation
ε ~ N(0, I)
μ
σ
φ_t
64 x 64 real image
Down-sampling
Down-sampling
Compression and Spatial Replication
Compression and Spatial Replication
128
4
1024
4
{0, 1}
128
1024
4
4
{0, 1}
φ_t
This bird is grey with white on its chest and has a very short beak
This bird is grey with white on its chest and has a very short beak
Stage-I Discriminator
Stage-II Discriminator

we adopt the GAN-CLS for both stages.

Experiment



Text description

This bird is red and brown in color, with a stubby beak

The bird is short and stubby with yellow on its body

A bird with a medium orange bill white body gray wings and webbed feet

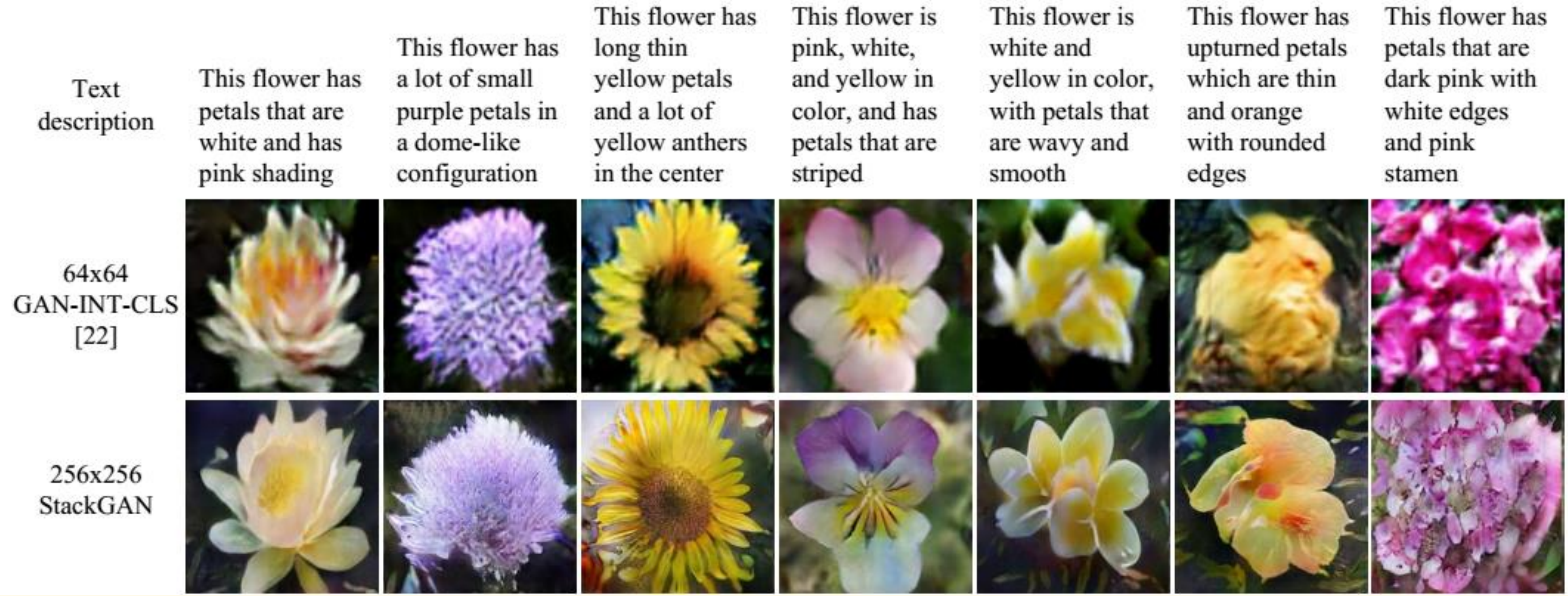
This small black bird has a short, slightly curved bill and long legs

A small bird with varying shades of brown with white under the eyes

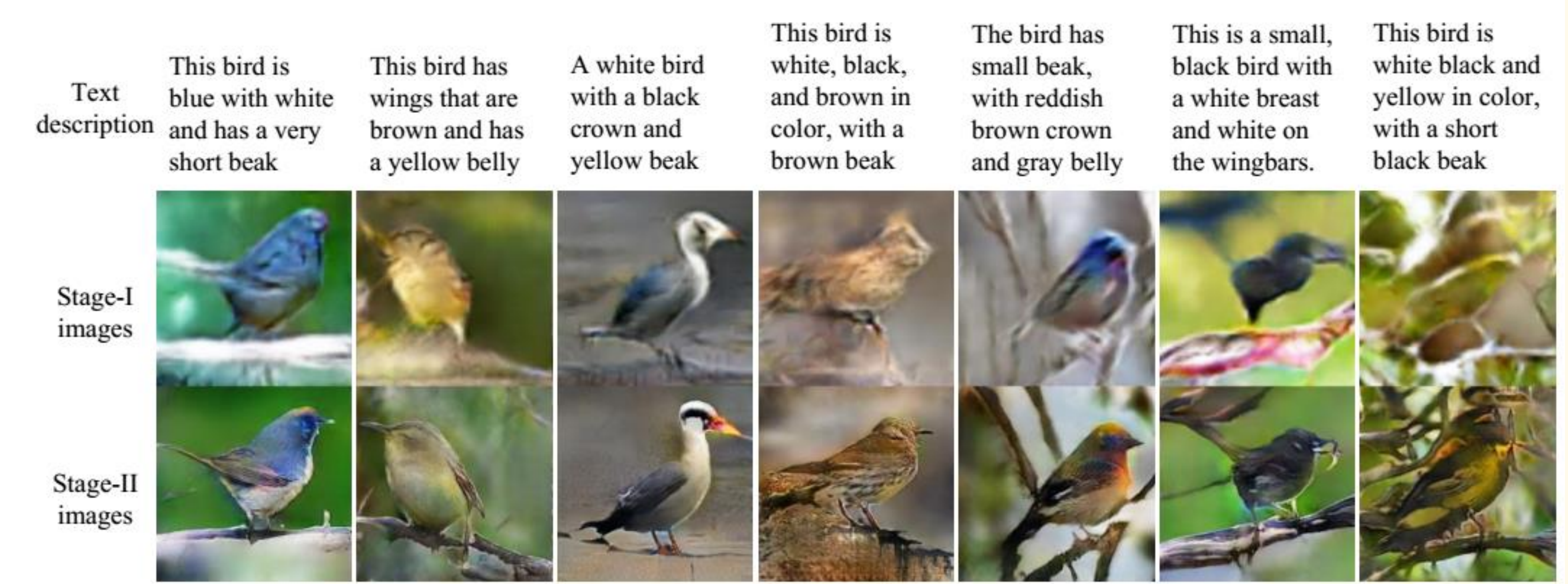
A small yellow bird with a black crown and a short black pointed beak

This small bird has a white breast, light grey head, and black wings and tail

CUB contains 200 bird species with 11,788 images.
Oxford-102 contains 8,189 images of flowers from 102 different categories.

Text description
This flower has petals that are white and has pink shading
This flower has a lot of small purple petals in a dome-like configuration
This flower has long thin yellow petals and a lot of yellow anthers in the center
This flower is pink, white, and yellow in color, and has petals that are striped
This flower is white and yellow in color, with petals that are wavy and smooth
This flower has upturned petals which are thin and orange with rounded edges
This flower has petals that are dark pink with white edges and pink stamen
64x64 GAN-INT-CLS [22]
256x256 StackGAN

一阶段和二阶段生成的图像有显著进步，说明分阶段生成图片的可行性



Text description

This bird is blue with white and has a very short beak

This bird has wings that are brown and has a yellow belly

A white bird with a black crown and yellow beak

This bird is white, black, and brown in color, with a brown beak

The bird has small beak, with reddish brown crown and gray belly

This is a small, black bird with a white breast and white on the wingbars.

This bird is white black and yellow in color, with a short black beak

通过将生成图片从训练集中找出5个最近邻，发现生成的图片和训练集中的图片差别较大，

说明了stackGAN 不是记住了训练样本，而是找到了语言和图像之间的联系（更强的泛化能力）

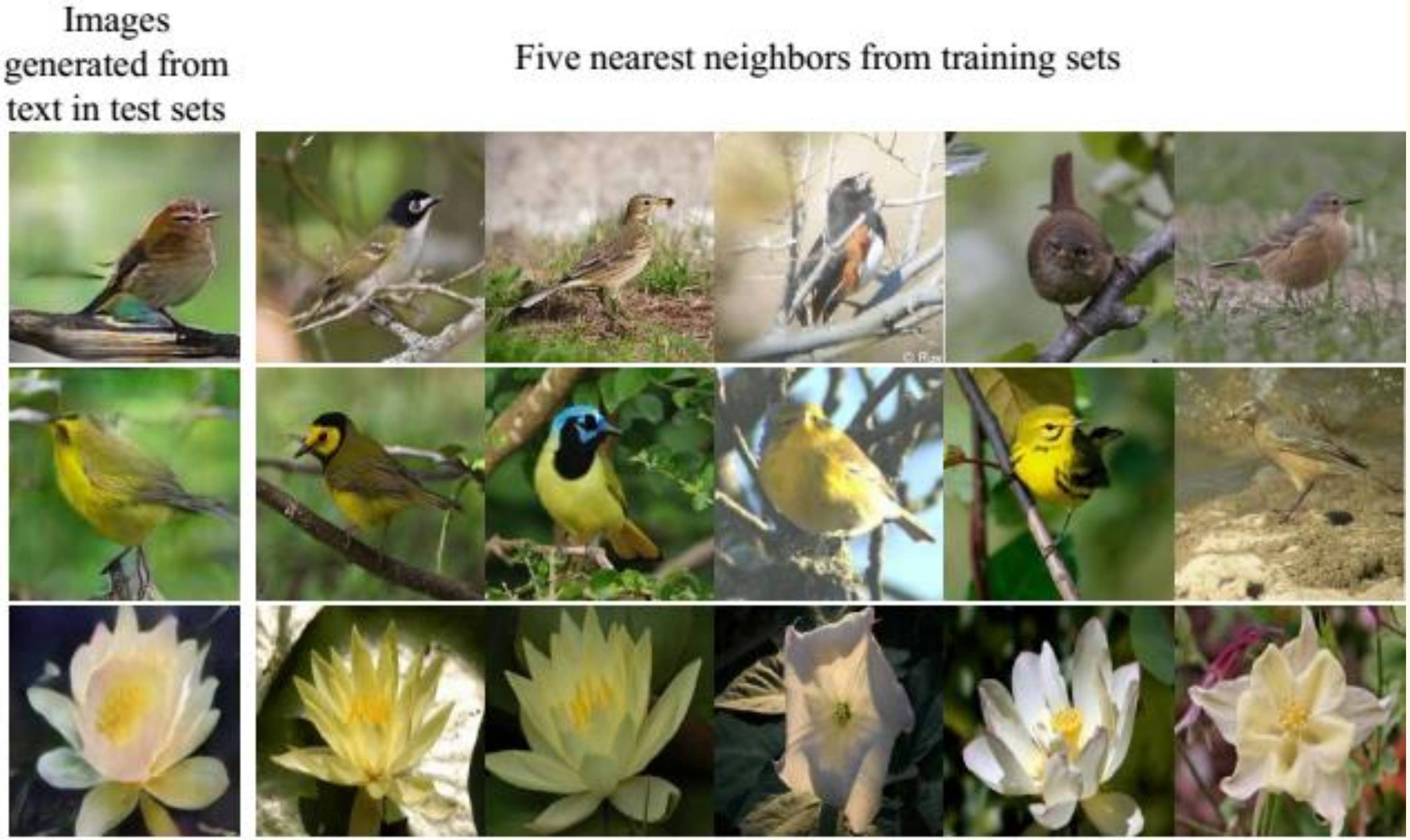

Images generated from text in test sets
Five nearest neighbors from training sets

Figure 6. For generated images (column 1), retrieving their nearest training images (columns 2-6) by utilizing Stage-II discriminator D to extract visual features. The $L2$ distances between features are calculated for nearest-neighbor retrieval.

由于类似VAE组建的存在，对输入的文字描述加入了一定的扰动，使得即使固定噪声和文字输入，也能输出多样化的图片

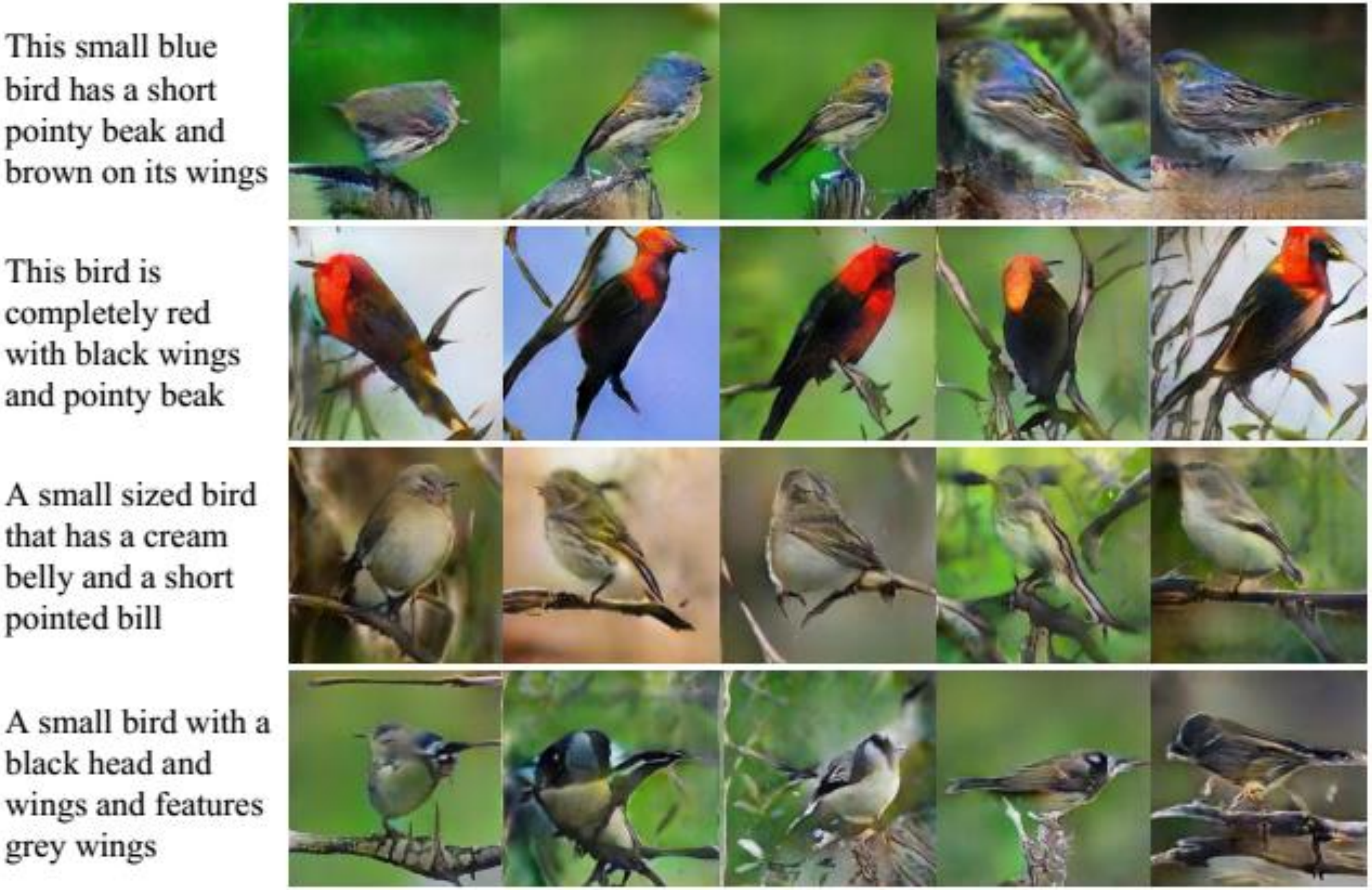
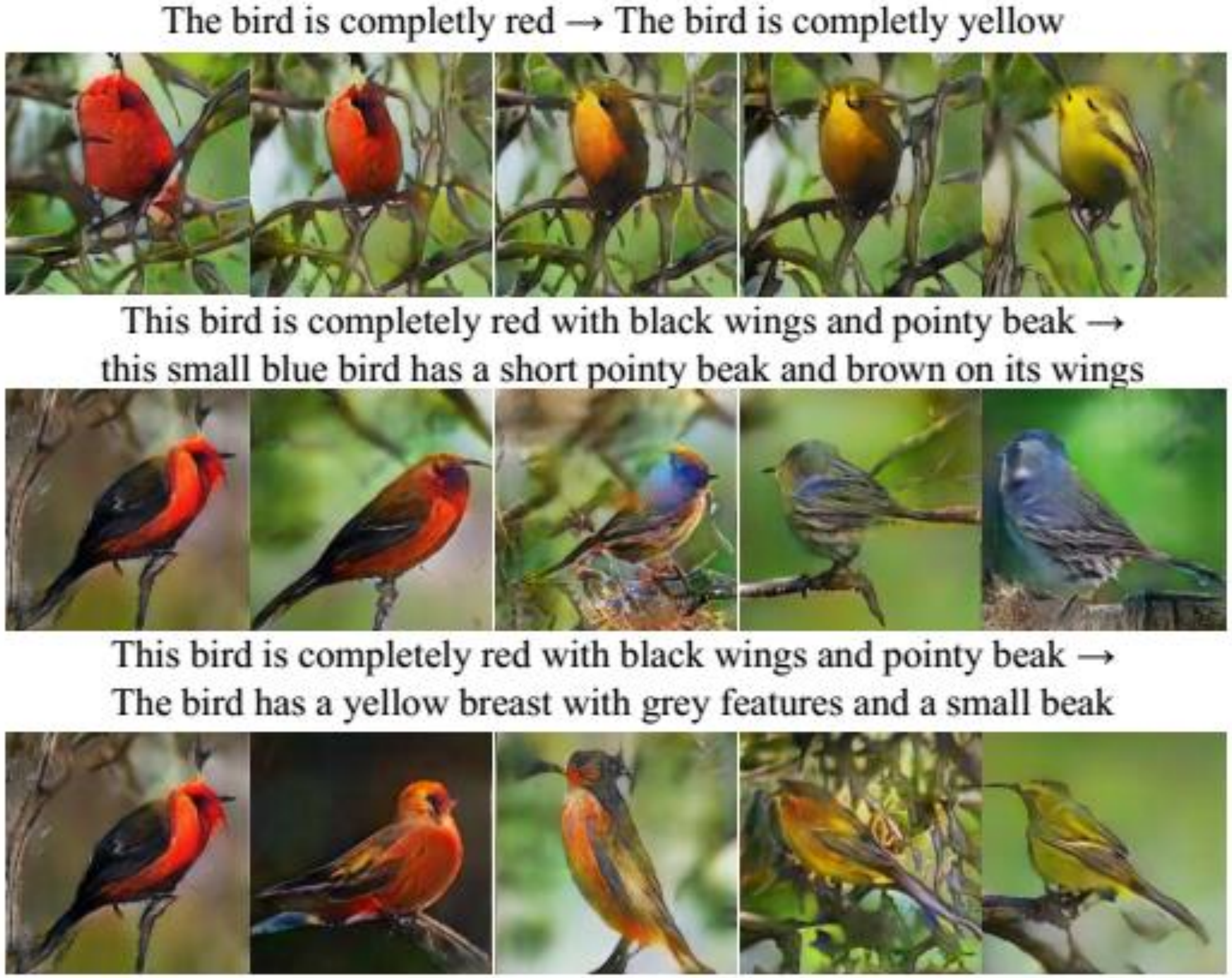

This small blue bird has a short pointy beak and brown on its wings
This bird is completely red with black wings and pointy beak
A small sized bird that has a cream belly and a short pointed bill
A small bird with a black head and wings and features grey wings

Figure 7. Birds with different poses and viewpoints generated with the same input text embedding by our StackGAN. The noise vector z and text embedding are fixed for each row.

To demonstrate that our StackGAN learns a smooth latent data manifold, we use it to generate images from linearly interpolated sentence embeddings



The bird is completly red → The bird is completly yellow

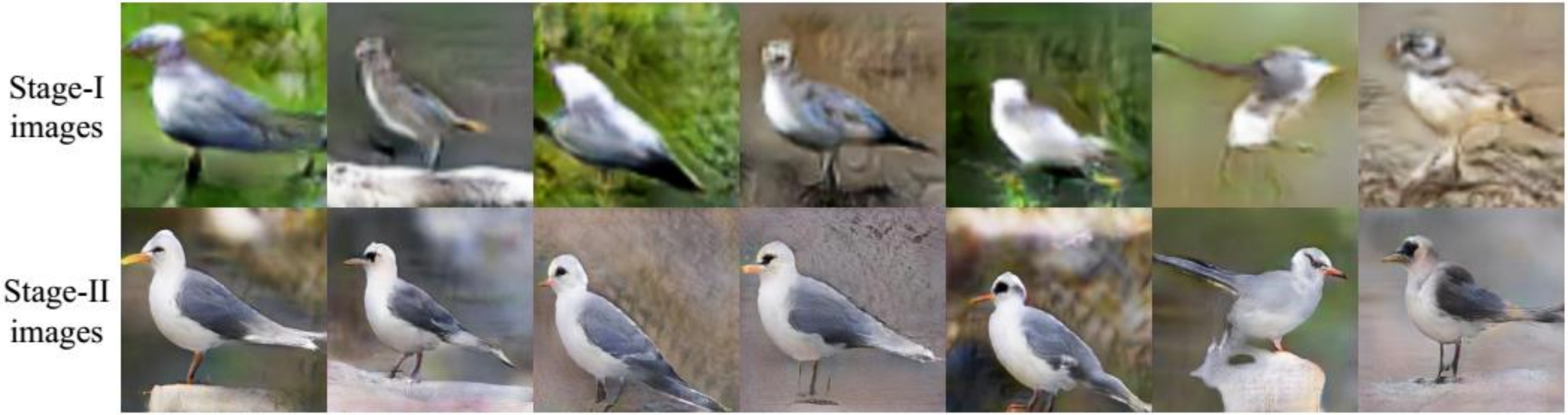
This bird is completely red with black wings and pointy beak → this small blue bird has a short pointy beak and brown on its wings

This bird is completely red with black wings and pointy beak → The bird has a yellow breast with grey features and a small beak

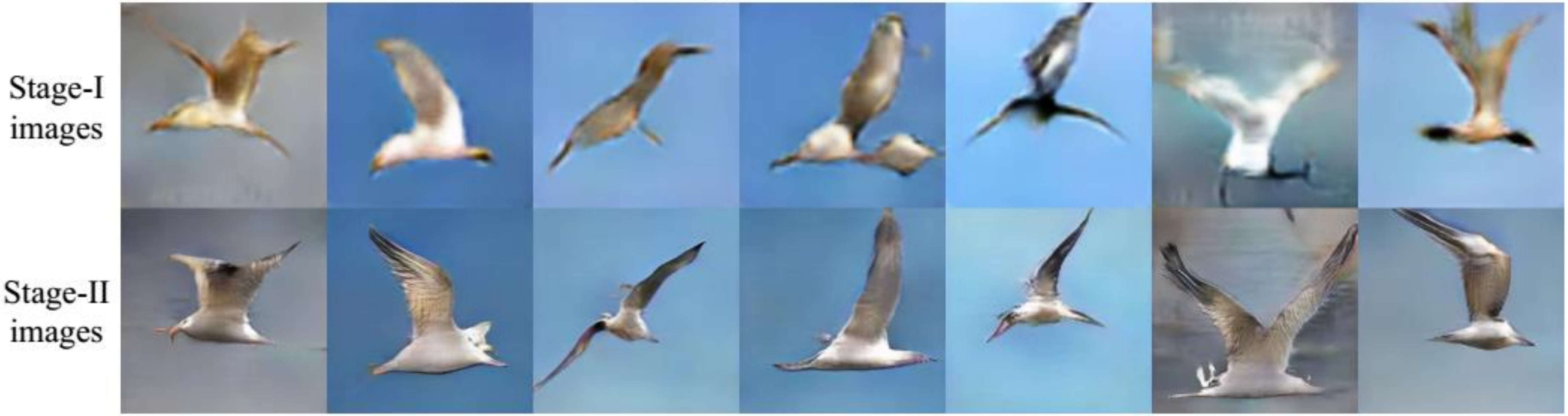
Figure 8. (Left to right) Images generated by interpolating two sentence embeddings. Gradual appearance changes from the first sentence's meaning to that of the second sentence can be observed. The noise vector z is fixed to be zeros for each row.

More example

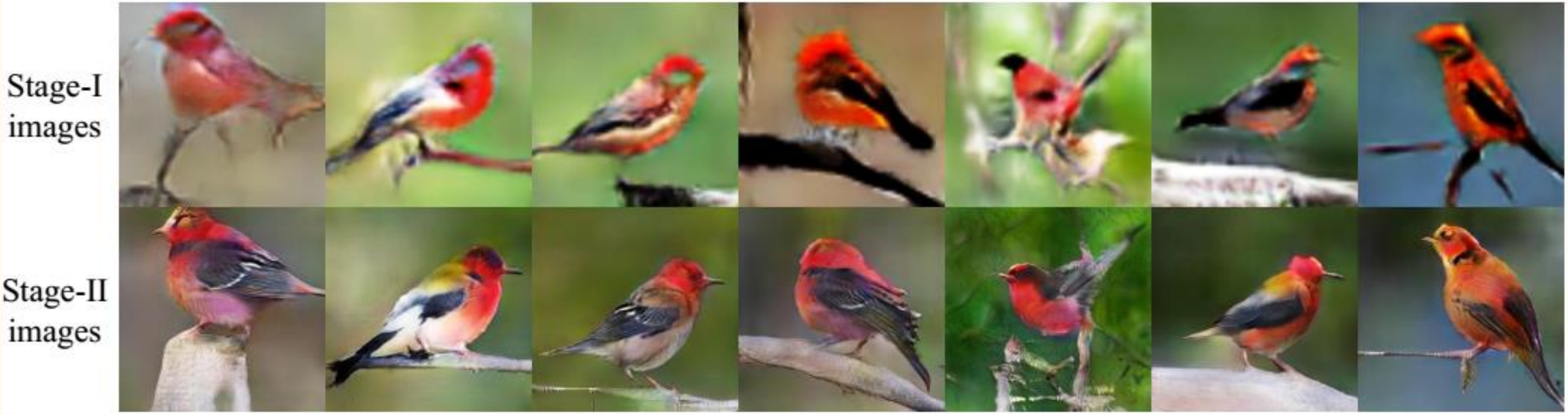
This bird sits close to the ground with his short yellow tarsus and feet; his bill is long and is also yellow and his color is mostly white with a black crown and primary feathers



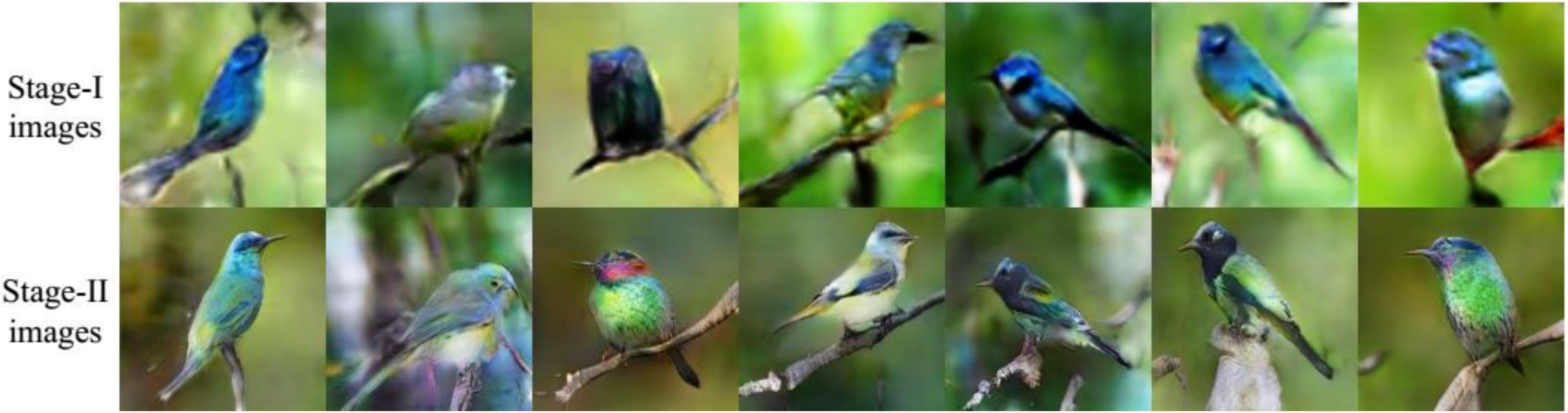
A large bird has large thighs and large wings that have white wingbars



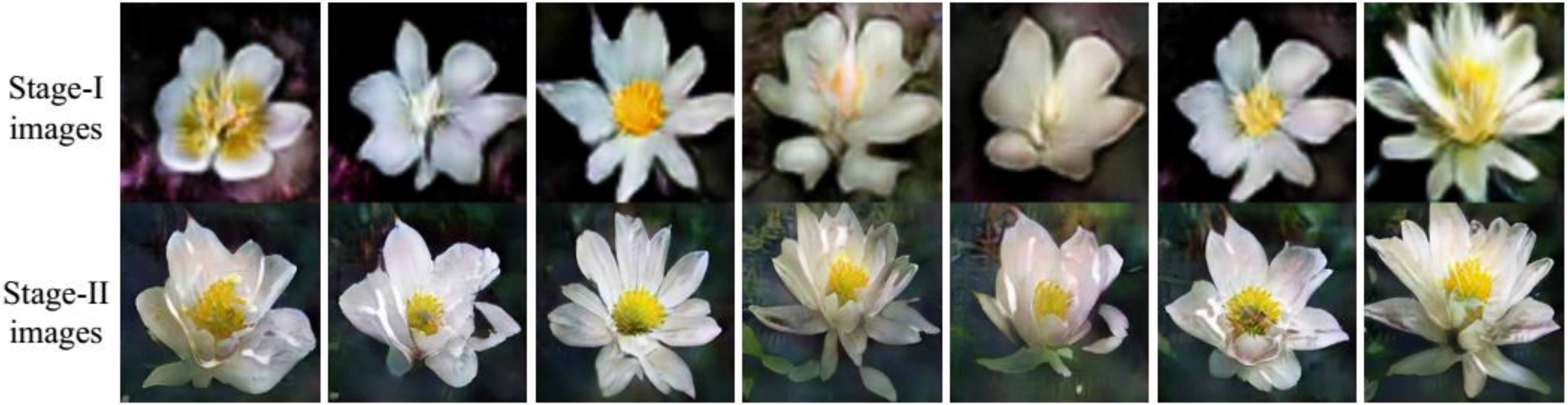
The small bird has a red head with feathers that fade from red to gray from head to tail



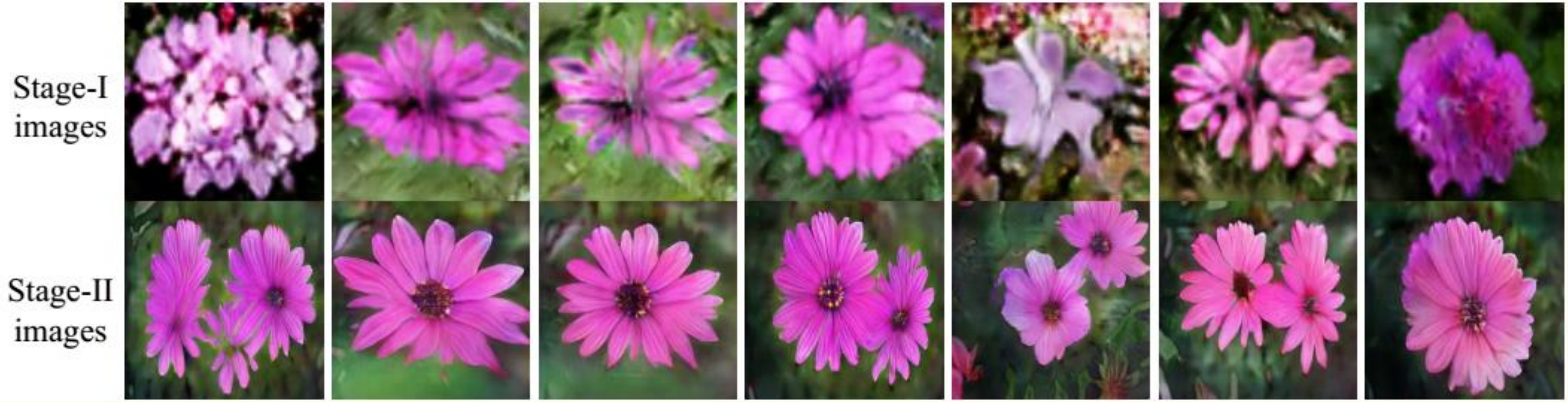
This bird is black with green and has a very short beak



This flower has white petals with a yellow tip and a yellow pistil

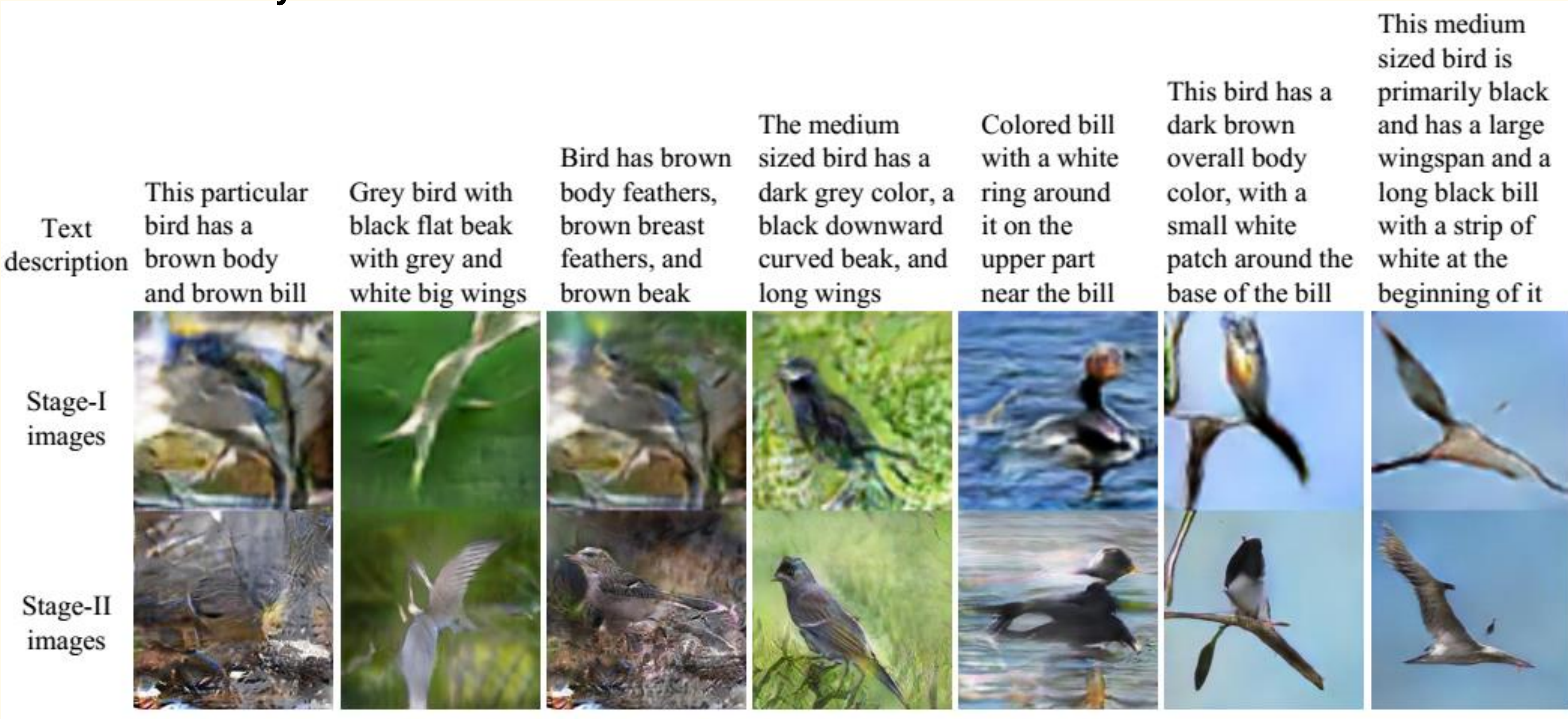


A flower with small pink petals and a massive central orange and black stamen cluster



Failure cases

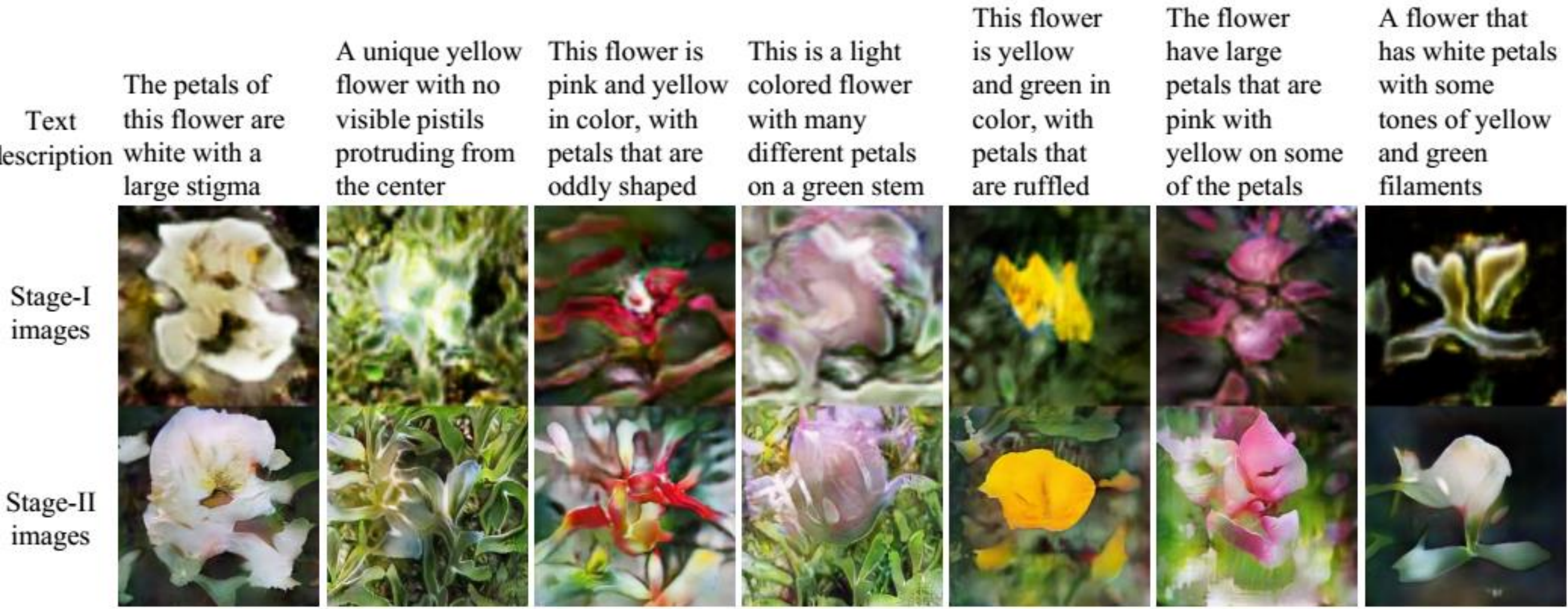
The main reason for failure cases is that Stage-I GAN fails to generate plausible rough shapes or colors of the objects



Text description	This particular bird has a brown body and brown bill	Grey bird with black flat beak with grey and white big wings	Bird has brown body feathers, brown breast feathers, and brown beak	The medium sized bird has a dark grey color, a black downward curved beak, and long wings	Colored bill with a white ring around it on the upper part near the bill	This bird has a dark brown overall body color, with a small white patch around the base of the bill	This medium sized bird is primarily black and has a large wingspan and a long black bill with a strip of white at the beginning of it

Failure cases

The main reason for failure cases is that Stage-I GAN fails to generate plausible rough shapes or colors of the objects



The petals of this flower are white with a large stigma

A unique yellow flower with no visible pistils protruding from the center

This flower is pink and yellow in color, with petals that are oddly shaped

This is a light colored flower with many different petals on a green stem

This flower is yellow and green in color, with petals that are ruffled

The flower have large petals that are pink with yellow on some of the petals

A flower that has white petals with some tones of yellow and green filaments

Thank you