Appendix: Cross-Modal Contrastive Learning for Text-to-Image Generation

In this appendix, we share implementation details (Sec. A), architecture details (Sec. B), details about our human evaluation procedure (Sec. C), and further qualitative results (Sec. E).

A. Implementation Details

All models are implemented in TensorFlow 2.0. Spectral normalization is used for all convolutional and fully-connected layers in the discriminator. For training all models, we use the Adam optimizer with parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rates for the generator and discriminator are set to $1e^{-4}$ and $4e^{-4}$ respectively. We use two discriminator training steps for each generator training step. During validation, we report results from the generator with exponential moving averaged weights, with a decay rate of 0.999.

Models are trained with a batch size of 256. For reporting results in our paper, models are trained for 1000 epochs, and we report the scores corresponding to the checkpoint with the best FID score on the validation set. For reporting our main results, we train a model with base channel dimensions $ch = 96$ (see Table 2). For ablation experiments in the main paper, we train models with base channel dimensions $ch = 64$.

B. Architecture Details

Detailed generator and discriminator architectures can be found in Tables 2a and 2b respectively. The details of the up-sampling block and down-sampling block are shown in Fig. 1.

C. Human Evaluations

The user interface shown to human evaluators is shown in Fig. 2. Users are requested to rank 4 images from best to worst on (1) image realism and (2) alignment to a given caption. The images are displayed in a random order.

D. Similarities and differences between DAMSM and the proposed contrastive losses

Our proposed contrastive losses bear several similarities to the DAMSM losses of AttnGAN. However, there are several key differences which are crucial to our strong performance:

Loss	IS ↑	FID ↓	R-prec ↑	SOA-C ↑	SOA-I ↑
G	23.69	34.70	40.44	21.61	38.13
D	25.81	26.63	56.62	28.58	49.36
G + D (XMC-GAN)	**31.33**	**11.34**	**73.11**	**42.29**	**61.39**

Table 1: Contrastive losses applied on the generator/discriminator.

- DAMSM losses are only used to train the *generator* (G), while contrastive losses in XMC-GAN are designed to train the *discriminator* (D) also. Features for contrastive losses are calculated from the different heads of the D backbone. This allows D to learn more robust and discriminative features, so XMC-GAN is less prone to mode collapse. This is a key reason that our model does not require multi-stage training. For training G, our contrastive losses are similar to DAMSM, which enforce consistency between generated images and conditional text descriptions. Table 1 compares adding contrastive losses on D and G separately, which highlights the benefits of our proposed method of training the discriminator.

- Second, the motivation behind contrastive losses and DAMSM also differs. As described in Sec. 4.1, we propose maximizing the mutual information between intra-modality and inter-modality pairs. We do this by maximizing the lower bound through optimizing contrastive (InfoICE) losses, consistently using cosine distance as the similarity metric. In contrast, the DAMSM loss in AttnGAN is motivated by information retrieval. Their DAMSM module uses dot product in certain instances (Eq. 7 in AttnGAN), and requires an additional normalization step (Eq. 8 in AttnGAN).

- Last, our training procedure is completely end-to-end, while AttnGAN needs a separate pretraining step. For AttnGAN, their DAMSM module undergoes a separate pretraining step before training the main generator / discriminator models.

E. Qualitative Results

E.1. Effect of random noise on generated images

In Sec. 6.1 of the main paper, we show that XMC-GAN generated images are largely preferred by human raters. XMC-GAN also significantly improves state-of-the-art FID scores. However, we also observe that the IS and SOA scores for CP-GAN are better than XMC-GAN. We conjecture that the issue was with IS and SOA not penalizing intra-class mode dropping (*i.e*. low diversity within a class or caption).

To verify this hypothesis, we conduct experiments to generate images from CP-GAN and XMC-GAN conditioned on the same caption, but with varying noise vectors z. The comparison results are shown in Fig. 3. Both the captions and noise vectors used are selected at random. As shown in the figure, XMC-GAN is able to generate diverse images (*e.g*., different view angles or compositions of the scene) for a fixed caption when different noise vectors are used. In contrast, CP-GAN generated images do not show much diversity despite conditioning on different noise vectors. This verifies our hypothesis that CP-GAN may have less diversity for the same class or caption. XMC-GAN is able to generate high quality and diverse scenes even when conditioned on a single caption.

E.2. Effect of captions on generated images

In Fig. 4, we present several examples of XMC-GAN generated images given different captions corresponding to the same original image.

Different MS-COCO captions. We observe that the generated images vary widely depending on the given caption, even if they are semantically similar. For example, we observe that in the first row, XMC-GAN generated images for caption #2 and caption #3 produce very different images. For caption #3, "A bus driving in a city area with traffic signs.", we observe that XMC-GAN is able to generate features of a city, with high-rise buildings in the background, and a traffic light to the left of the image. In contrast, in caption #2, which does not mention the city XMC-GAN generates an image that shows the bus next to a curb, in agreement with the caption.

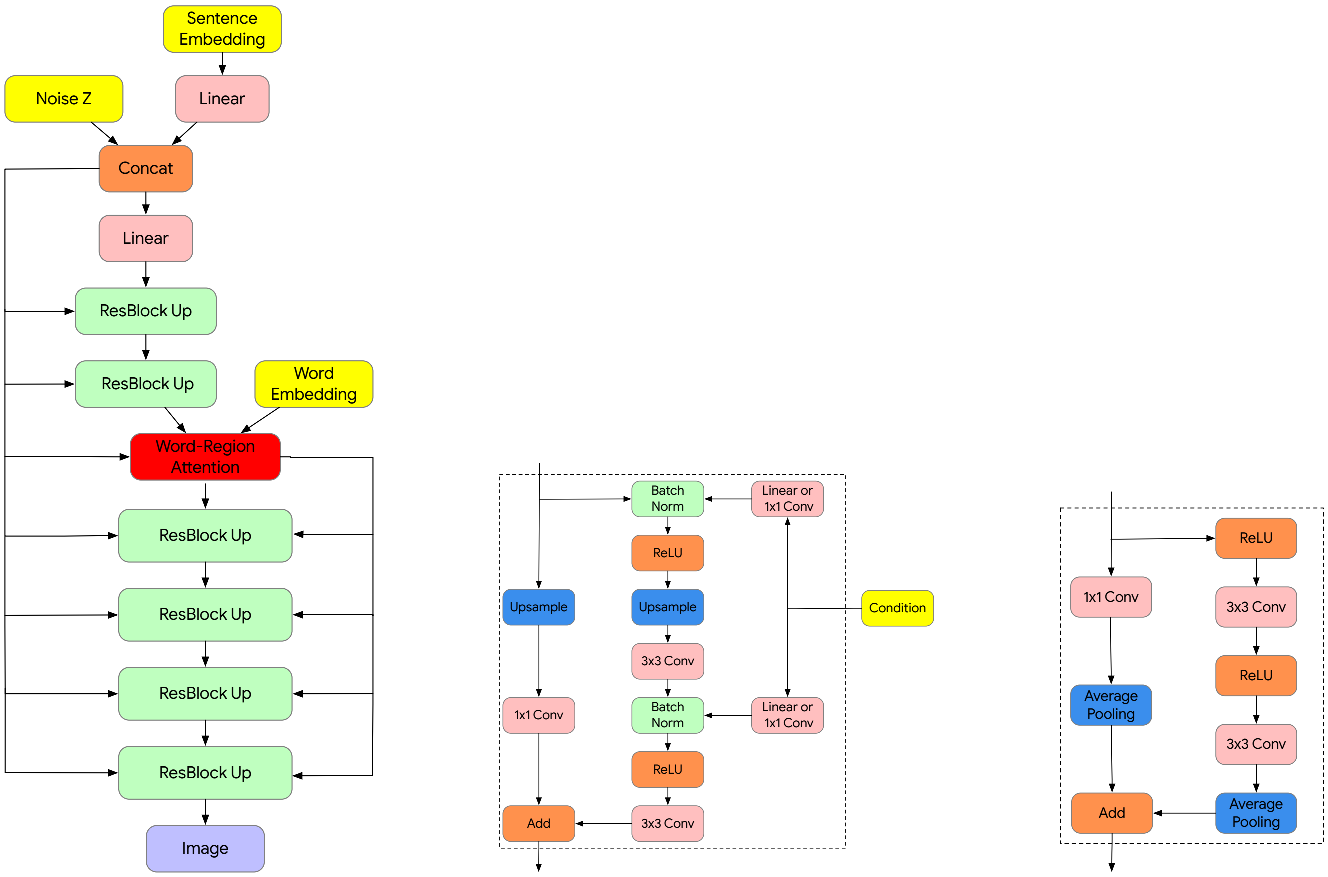
MS-COCO compared to LN-COCO captions. We also observe distinct differences in generated images when conditioned on MS-COCO as compared to LN-COCO captions. LN-COCO captions are much more detailed, which increases image generation difficulty. The increase in difficulty of LN-COCO captions appears to lead to less coherent scenes in general as compared to the MS-COCO model (*e.g*. the third row of Fig. 4).

E.3. Random samples

COCO-14 Random qualitative samples from COCO-14 are presented in Fig. 5. We observe that even over randomly selected captions, XMC-GAN appears to generate images that are significantly clearer and more coherent. Scenes often depict clear objects, as compared to previous methods.

LN-COCO Random qualitative samples from LN-COCO are presented in Fig. 6. The longer captions increase the challenge of realistic text-to-image synthesis, but we observe clear improvements from previous methods in most images. In particular, XMC-GAN appears to generate objects and people that are more clear and distinct.

LN-OpenImages Random qualitative samples from LN-OpenImages are presented in Fig. 7. As this dataset was previously untested on, we simply display the original images against XMC-GAN generated images. Despite the increase in complexity and diversity of images, XMC-GAN generates very strong results, with especially convincing scene generation capability (*e.g*. first column, second and third last rows). We hope that our results will inspire future work to advance on tackling this very challenging dataset.



(a) (b) (c)

Figure 1: (a) The generator achitecture for XMC-GAN. (b) The residual block (ResBlock Up) of XMC-GAN's generator. For the self-modulation ResBlock Up, the condition are noise z and global sentence embedding. For attentional self-modulation ResBlock Up, the condition are noise z, global sentence embedding and attentional work context. (c) The Residual Block (ResBlock Down) of XMC-GAN's discriminator.

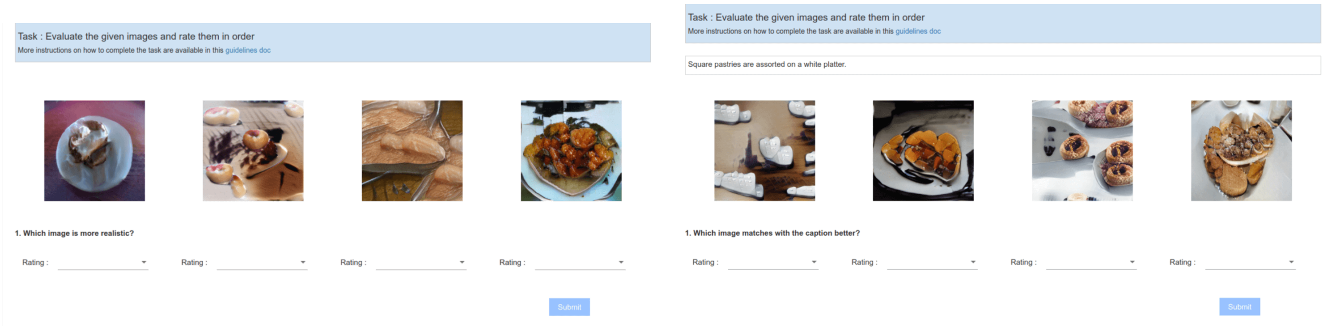
$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$, $e_s \in \mathbb{R}^{768}$, $e_w \in \mathbb{R}^{T\times 768}$
Linear (768) $\to$ 128 # projection for e_s
Linear $(128 + 128) \to 4 \times 4 \times 16ch$
Self-modulation ResBlock up $\to 8 \times 8 \times 16ch$
Self-modulation ResBlock up $\to 16 \times 16 \times 8ch$
Linear Layer $(8ch) \to 768$ # projection for attention
Attentional Self-modulation ResBlock up $\to 32 \times 32 \times 8ch$
Attentional Self-modulation ResBlock up $\to 64 \times 64 \times 4ch$
Attentional Self-modulation ResBlock up $\to 128 \times 128 \times 2ch$
Attentional Self-modulation ResBlock up $\to 256 \times 256 \times ch$
Attentional Self-modulation, 3×3 Conv $\to 256 \times 256 \times 3$

(a) Generator

RGB images $x \in \mathbb{R}^{256\times 256\times 3}$, $e_s \in \mathbb{R}^{768}$, $e_w \in \mathbb{R}^{T\times 768}$
ResBlock down $\to 128 \times 128 \times ch$
ResBlock down $\to 64 \times 64 \times 2ch$
ResBlock down $\to 32 \times 32 \times 4ch$
ResBlock down $\to 16 \times 16 \times 8ch$
Linear $(4ch) \to 768$ # projection for word-region contrastive
ResBlock down $\to 8 \times 8 \times 8ch$
ResBlock down $\to 4 \times 4 \times 16ch$
ResBlock $\to 4 \times 4 \times 16ch$
Global sum pooling
Linear $(768) \to 16ch$ # projected$(e_s) \cdot h$
Linear $(16ch) \to 1$

(b) Discriminator

Table 2: XMC-GAN generator and discriminator architectures.



(a) UI for ranking image realism.

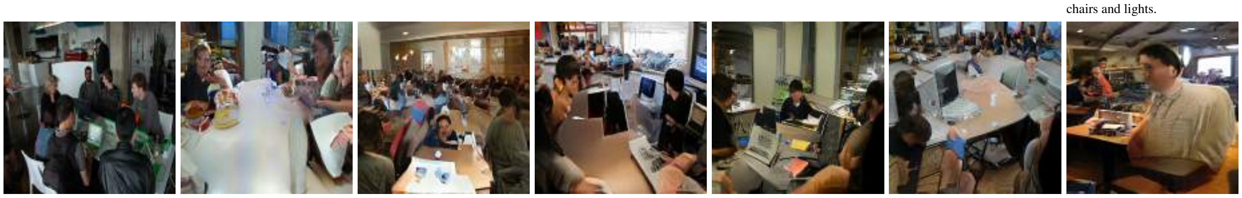
(b) UI for ranking text alignment.

Figure 2: User interface for collecting human evaluations.

Figure 3: Comparison of CP-GAN and XMC-GAN generated images for the same caption with different noise vectors.

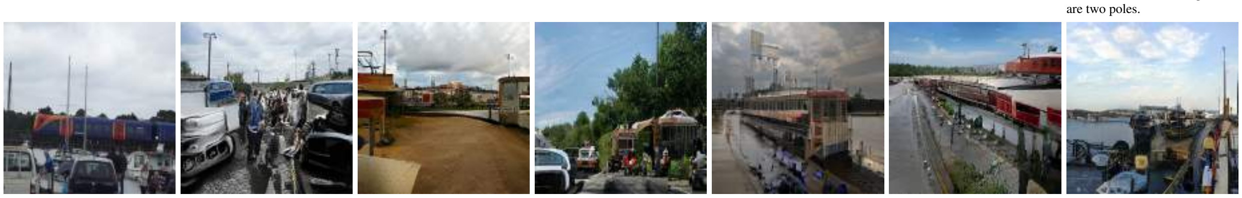
Real Image	Caption #1	Caption #2	Caption #3	Caption #4	Caption #5	LN-COCO
	The bus is pulling off to the side of the road.	A bus pulls over to the curb close to an intersection.	A bus driving in a city area with traffic signs.	a public transit bus on a city street	Bus coming down the street from the intersection	in this image there are some vehicles on the road and behind the vehicles one big building is there on the right side there are some persons are walking on the street and the background is little bit sunny.

Real Image	Caption #1	Caption #2	Caption #3	Caption #4	Caption #5	LN-COCO
	A group of people sitting around a table with laptops and notebooks.	Seven people seated at table talking and working on computer devices.	A group of people at a table working on small laptops.	A group of people sitting at a table using computers.	Several friends are visiting at a table with tablets.	In the center of the image there is a table and there are people sitting around the table. We can see bottles, laptops and wires placed on the table. In the background there is a man standing. We can see a counter table, chairs and lights.



Real Image	Caption #1	Caption #2	Caption #3	Caption #4	Caption #5	LN-COCO
	A group of people are walking and one is holding an umbrella.	these people are walking together down a road	Three young people walking behind a large crowd.	Three men who are walking in the sand.	A group of people walking down a road.	In this image, in the middle there are some people walking, in the right side there is a man standing and he is holding a umbrella, in the background there are some cars, there is a bus, there are some green color trees, in the top there is a sky which is cloudy and in white color.

Real Image	Caption #1	Caption #2	Caption #3	Caption #4	Caption #5	LN-COCO
	People are in a parking lot beside the water, while a train is in the background.	Colorful commuter train goes through a marina area on a cloudy day	A parking lot next to a marina next to a railroad	Group of people standing beside their cars on a pier.	A train crosses as a bunch of gathered vehicles watch.	Bottom left side of the image there are two vehicles behind the vehicles there are few ships on the water and there are few people are standing. In the middle of the image there is a train on the bridge. Behind the train there are some trees and clouds. In the middle of the image there are two poles.



Real Image	Caption #1	Caption #2	Caption #3	Caption #4	Caption #5	LN-COCO
	A calculator and cell phone lay on a desk in front of a keyboard	A cell phone on top of a calculator near a computer keyboard.	a table with a calculator and phone siting on it	A picture of a cell phone Calculator and a computer.	There is a phone on top of a calculator	In the picture we can see a calculator which is black in color and on it there is a mobile phone and it is also black in color, in the background we can see a keyboard which is white in color placed on white paper on the wooden table.

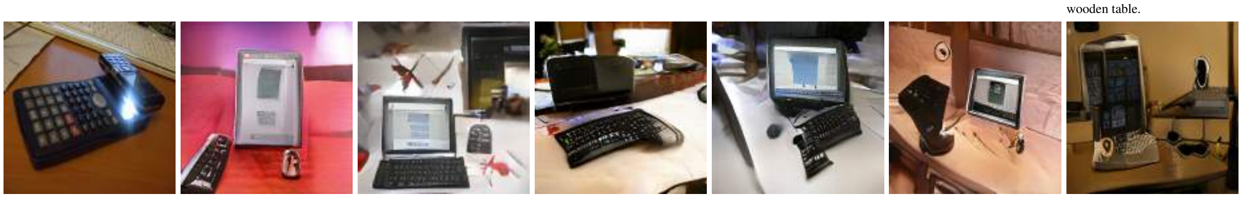


Figure 4: Generated images for varying captions from COCO-14 and LN-COCO corresponding to the same original image.

Figure 5: Generated images for random examples from COCO-14.

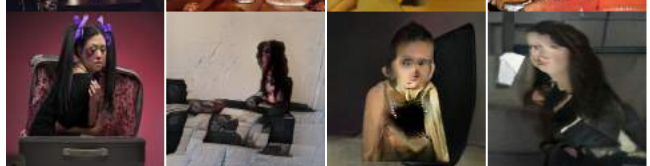
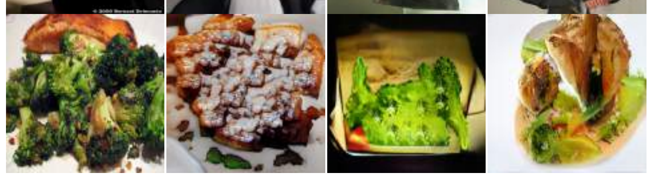
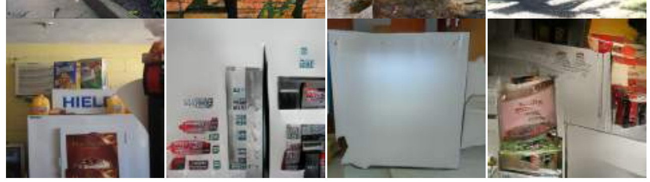
Caption	Original	AttnGAN	TRECS	XMC-GAN
In this picture we can see a pole in front, in the bottom there are some leaves, in the background we can see a white color and black color cars, on the right side of the image we can see a tree, in the background there is a building and a hill.				
In this image we can see zebra and giraffe standing in grass, And there are so many plants, lake with water, mountain with trees.				
In this image we can see both of the children are standing, and smiling and cooking, in front here is the stove and pan on it, here is the spoon, at side here is the vessel, and at back here is the table, here is the wall, and here is the glass door.				
In this image there are group of persons who are sitting around the table in a restaurant and having some food and there are water glasses on the table,at the background of the image there is a door,mirror and some paintings attached to the wall.				
Here we can see a woman sitting in the briefcase. And this is wall.				
There is a man in white color shirt, wearing a black color tie, standing. In the background, there is a yellow wall near the white ceiling.				
The picture consists of food items on a white color plate like object.				
In this image i can see person holding a bat and a wearing a white helmet. He is wearing blue shirt and white pant. At the back side I can see three person sitting. There is a net. The person is holding a umbrella which is in green and white color. Back Side i can see vehicle.				
Here we can see a bench and this is road. There are plants and this is grass. In the background there is a wall.				
This image consists of refrigerator. On that there are cans and boxes. There is light on the top. There is magnum sticker on refrigerator.				

Figure 6: Original and generated images for random examples from LN-COCO.

Caption	Original	XMC-GAN	Caption	Original	XMC-GAN
In this picture I can see the cars on the grass in the top right hand side there is a vehicle. In the background there may be the buildings.			In this image I can see a mirror with some text written on it. In the background I can see a car the trees and the buildings with some text written on it.		
In this image I can see a cat on a sidewalk and I can see a dark color.			In this image we can see people sitting on chairs. Also we can see packets on chairs. There are two people standing. Also we can see cupboards with books. And there is a pillar. And there is a table ...		
In this image we can see vehicles a fence and a pole. At the top there is sky. At the bottom there are plants and we can see grass.			In this picture we can see a grill meat piece in black plate which is placed on the wooden table top.		
In front of the image there is a person running on the track. Beside the track there is a sponsor board. At the bottom of the image there is grass on the surface.			In this image in the foreground we can see a sculpture and in the background we can see many branches of a tree.		
This is an aerial view and here we can see buildings and trees. At the top there is sky.			In front of the image there is an army personnel holding some objects in his hand. Behind the person there are a few army personnel. In the background of the image there are photo frames and doors on ...		
In this image I can see cake on the table. There is hand of a person holding the knife also there are hands of another person holding food item in one hand. And there are some other objects.			In this picture we see a plastic glass containing the ice cream is placed on the white table. We see the tissue papers and a paper glass are placed on the table. In the background we see a grey color object is placed ...		
In this image we can see a bunch of flowers to the plants. We can also see the wooden surface.			In this picture I can see few plants with leaves and I can see the flowers.		
In the foreground I can see grass a fence a net light poles and wires. In the background I can see water house plants some objects the trees and the sky.			It is an edited image with different shaped designs.		
In this image there is dried grass on the ground. In the top left side of the image I can see a tree. In the background there is sky.			In this image there are birds on a pathway and I can see a duck in the water.		
In this image I can see a pen which is black in color on the white colored surface.			In this image I can see the cat on the mat and I can see few objects.		

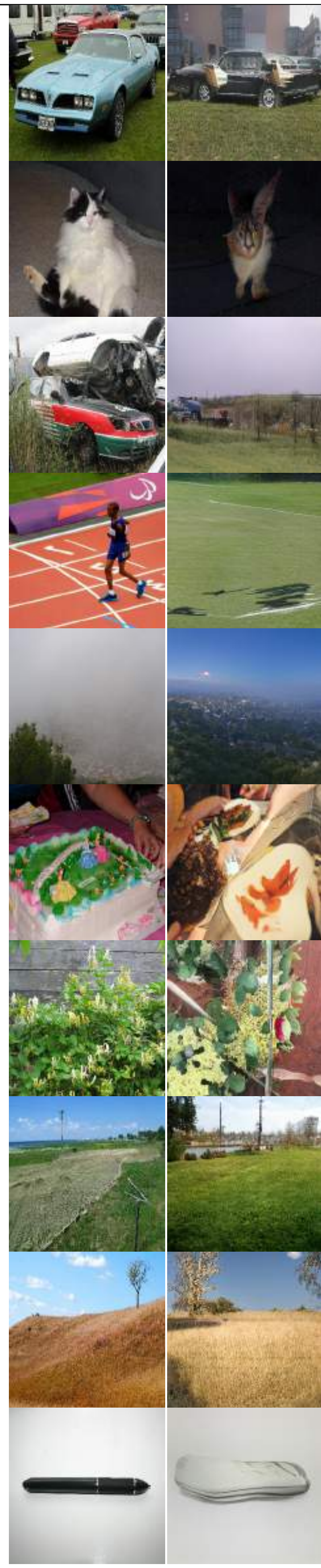


Figure 7: Original and generated images for random examples from LN-OpenImages.