Generating Multiple Objects at Spatially Distinct Locations

Tobias Hinz, Stefan Heinrich, Stefan Wermter

INTRO

- one challenge with Generative Adversarial Nets is that it is hard to control the layout of the generated scenes
- a popular approach for increased control is to use scene layouts as additional input, however, this requires a lot of labeling
- our approach only needs a bounding box (thereby specifying location and shape) and a class label for each of the foreground objects we want to be in the image

METHODS

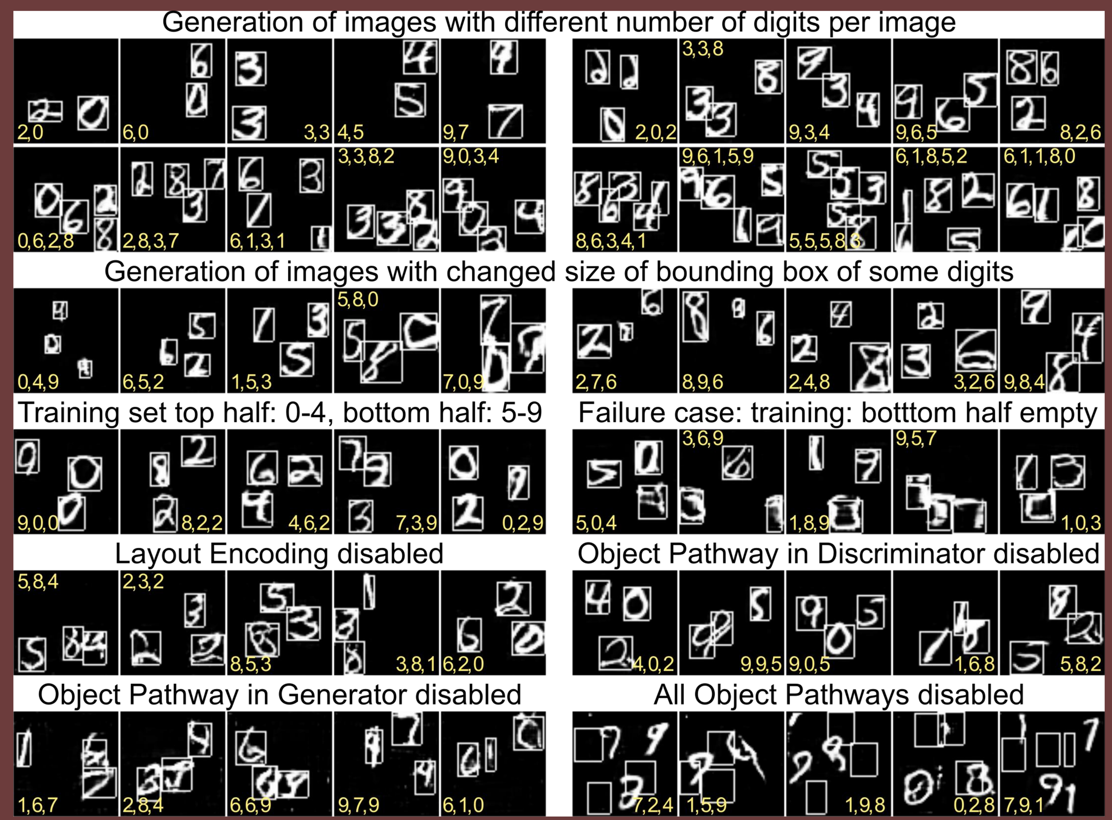
- we add an object pathway to the generator and the discriminator of a normal GAN
- the generator applies the object pathway iteratively at each bounding box location, conditioned by the label
- the discriminator's object pathway checks each bounding box location iteratively and evaluates whether the specified object is actually recognizable

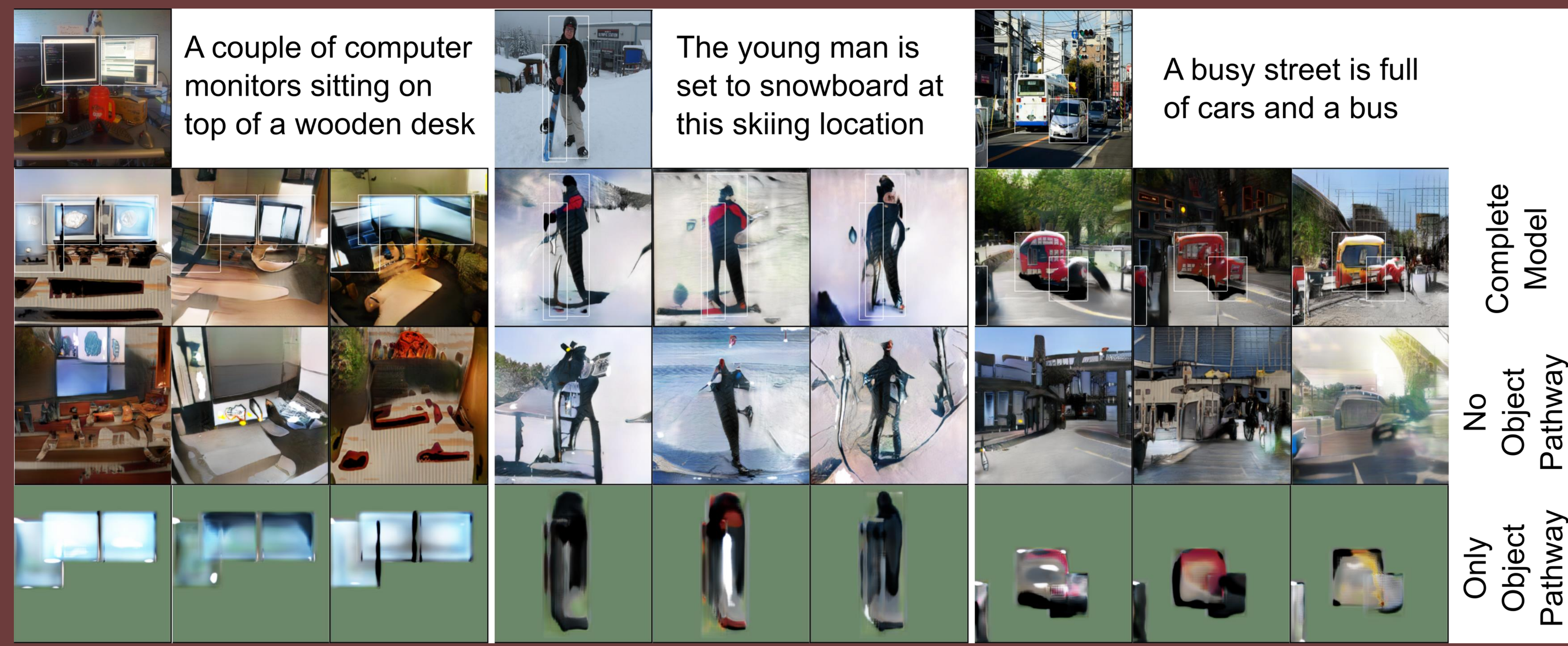
RESULTS

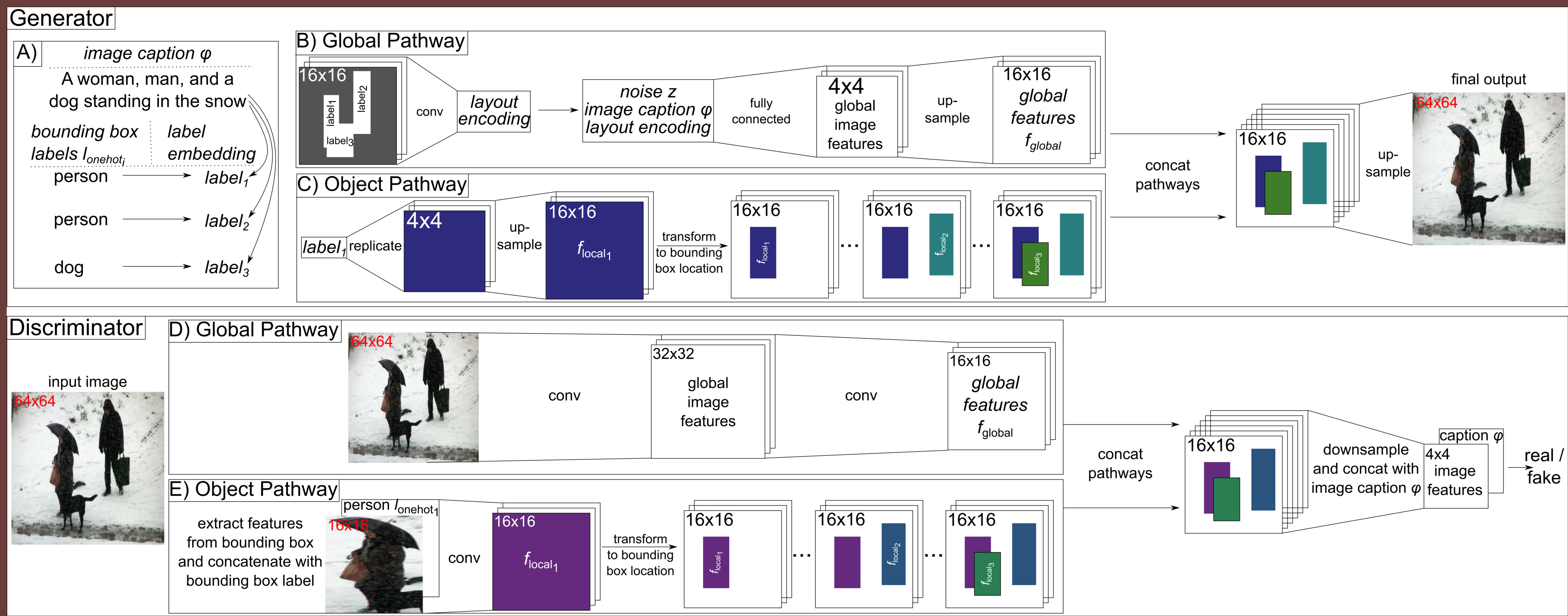
- experiments show that the architecture does lead to increased control over the image generation process
- experiments on the Multi-MNIST and CLEVR datasets show that the architecture can generalize to novel object characteristics and locations, as well as to different numbers of generated objects per scene
- experiments on the MS-COCO dataset show that the object pathway learns features for the individual foreground objects and can lead to an overall higher quality of the generated images (based on IS and FID)

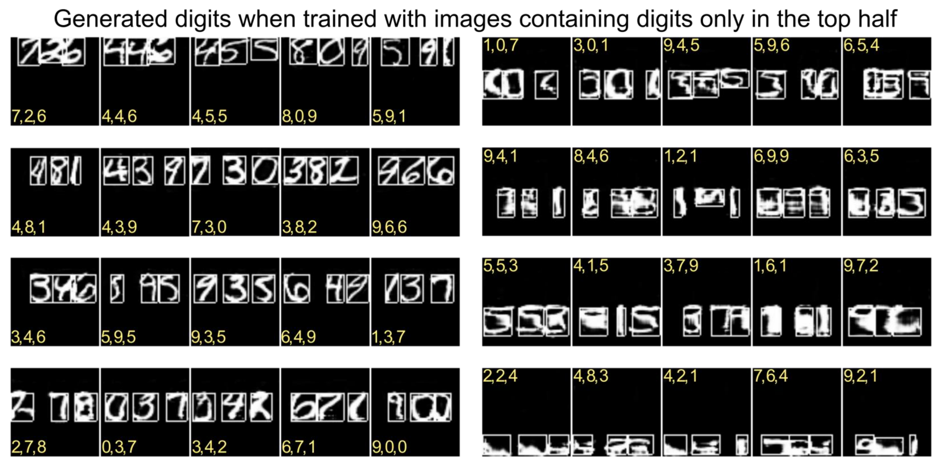
Take a picture to get to our Github page

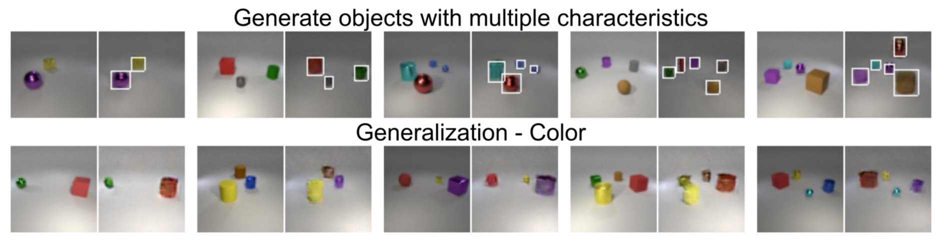
We **extend the GAN architecture** in order to **control object locations** and **identities** in generated images.

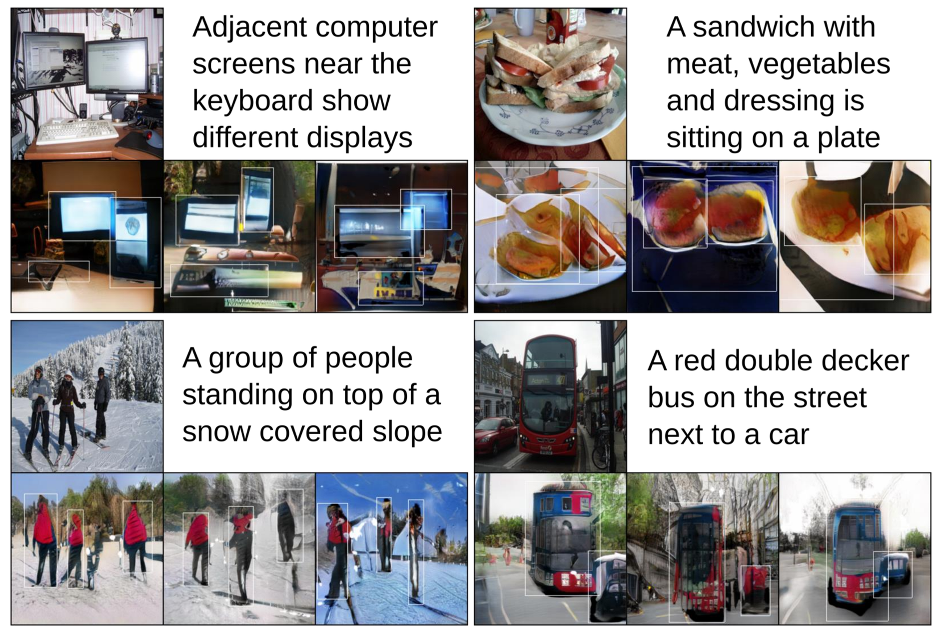








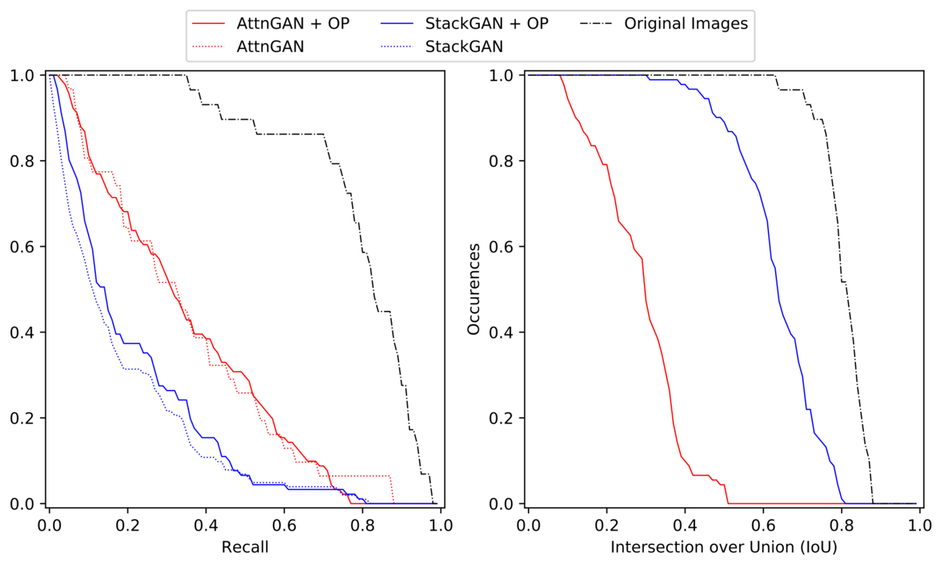




Model	Resolution	IS ↑	FID ↓
GAN-INT-CLS	64×64	7.88 ± 0.07	60.62
StackGAN-V2	256×256	8.30 ± 0.10	81.59
StackGAN	256×256	8.45 ± 0.03[1]	74.05
PPGN	227×227	9.58 ± 0.21	
ChatPainter (StackGAN)	256×256	9.74 ± 0.02	
Semantic Layout	128×128	11.46 ± 0.09[2]	
HDGan	256×256	11.86 ± 0.18	71.27 ± 0.12[3]
AttnGAN	256×256	23.61 ± 0.21[4]	33.10 ± 0.11[3]
StackGAN + Object Pathways (Ours)[5]	256×256	12.12 ± 0.31	55.30 ± 1.78
AttnGAN + Object Pathways (Ours)	256×256	24.76 ± 0.43	33.35 ± 1.15

[1] Recently updated to 10.62 ± 0.19 in its source code.
[2] When using the ground truth bounding boxes at test time (as we do) the IS increases to 11.94 ± 0.09.
[3] FID score was calculated with samples generated with the pretrained model provided by the authors.
[4] The authors report a "best" value of 25.89 ± 0.47, but when calculating the IS with the pretrained model provided by the authors we only obtain an IS of 23.61. Other researchers on the authors' Github website report a similar value for the pretrained model.
[5] We use the updated source code (IS of 10.62) as our baseline model.

Table 1: Comparison of the Inception Score (IS) and Fréchet Inception Distance (FID) on the MS-COCO data set for different models. Note: the IS and FID values of our models are not necessarily directly comparable to the other models, since our model gets, in addition to the image caption, up to three bounding boxes and their respective object labels as input at test time.



Take a picture to get to our Github page