Trace Controlled Text to Image Generation: Supplementary

Kun Yan[1], Lei Ji[2], Chenfei Wu[2], Jianmin Bao[2], Ming Zhou[3], Nan Duan[2], and Shuai Ma[1]

[1] SKLSDE Lab, Beihang University, Beijing, China
{kunyan,mashuai}@buaa.edu.cn
[2] Microsoft Research Asia, Beijing, China
{leiji,chewu,jianbao,nanduan}@microsoft.com
[3] Langboat Technology, Beijing, China
zhouming@chuangxin.com

1 Diffusion Decoder

As illustrated in Fig. 1, we train a diffusion decoder by feed the image latents from the VQGAN encoder into a conditional DDPM. After training the transformer model of TCTIG, we can feed the image latent code into this diffusion decoder and get much better high-quality image.

The diffusion model architecture is following the designs of guided diffusion[1], except where they concatenate the low res image channel-wise with the noised input, we simply skip the encoder layers entirely and inject the VQ latents into the middle block. This allows us to re-use the encoder and decoder weights from the pre-trained model, only the middle blocks needs to be re-trained.

2 SOA Filter Statistics

See Table 1

3 Hyperparameter Details

For the transformer encoder, we use the pretrained text encoder of ViT-B/32 version of CLIP, as well as a 3 layer additional transformer with 12 heads and a hidden size of 768. The number of decoder layers is 12, also with 12 heads and a hidden size of 768.

When training the transformer, we apply DeepSpeed ZeRO[2] Stage 3 to offload gradients and parameters into memory. We update the parameters using CPUAdamW with β_1=0.9, β_2=0.96, ϵ = 1e-8, and weight decay multiplier 4.5e-2. We clip the decompressed gradients by norm using a threshold of 4, prior to applying the Adam update. Gradient clipping is only triggered during the warm-up phase at the start of training. We trained the model using 16 NVIDIA V100 GPUs each with 32GB memory and a total batch size of 768, for a total

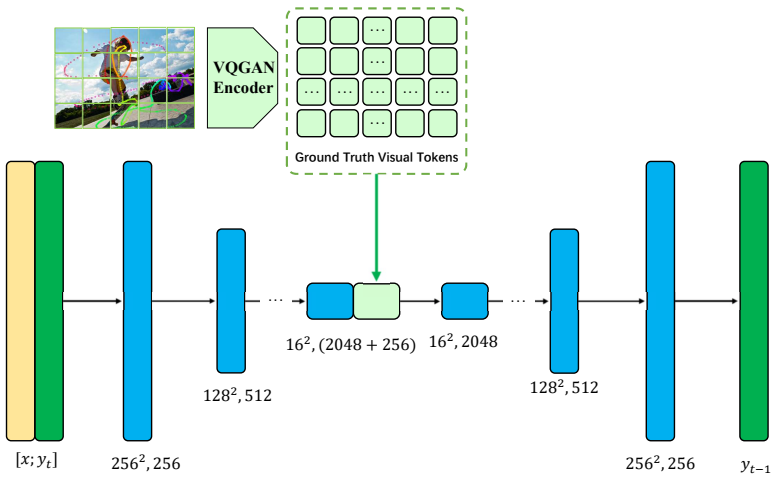

Fig. 1: The data flow of diffusion decoder, where x is the diffusion noise, y_t is the output from last diffusion step

of 80,000 updates. At the start of training, we use a linear schedule to ramp up the learning rate to 5e-4 over 5000 updates and reduce the learning rate to 0 in a cosine schedule.

4 Qualitative study with XMC-GAN

Note that, the generated samples on validation set of XMC-GAN and pretrained model are currently unavailable yet, so we cannot directly compare our results with theirs. Figure 2 compared the presented examples in their paper with our results.

5 More Examples

See Fig. 3

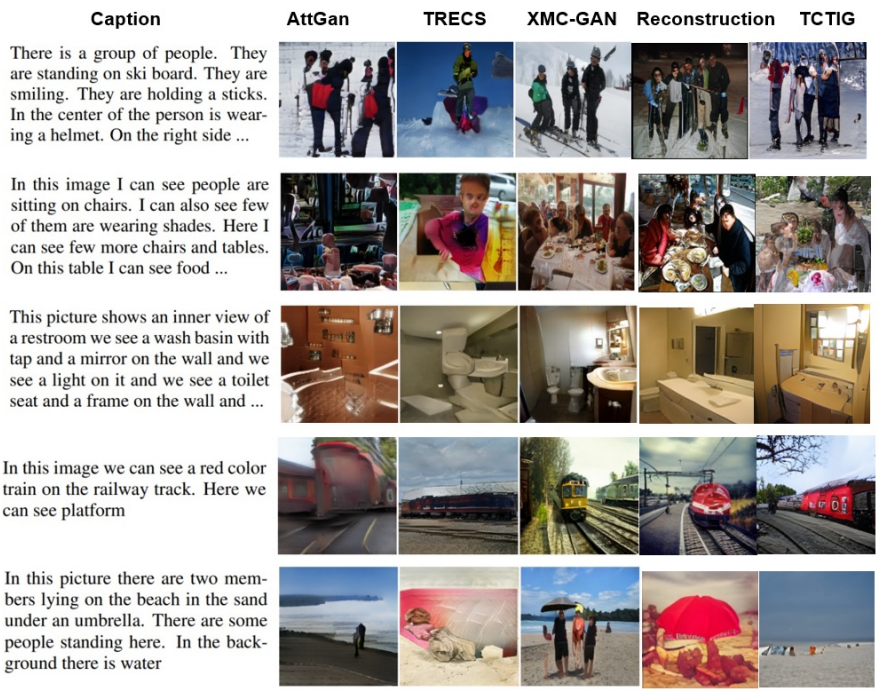

Fig. 2: Qualitative study on results in XMC-GAN. We can find that our generated results is highly spatially aligned with the reconstructed ground truth picture. And our model is sensitive with descriptive words such as color in the fourth row.

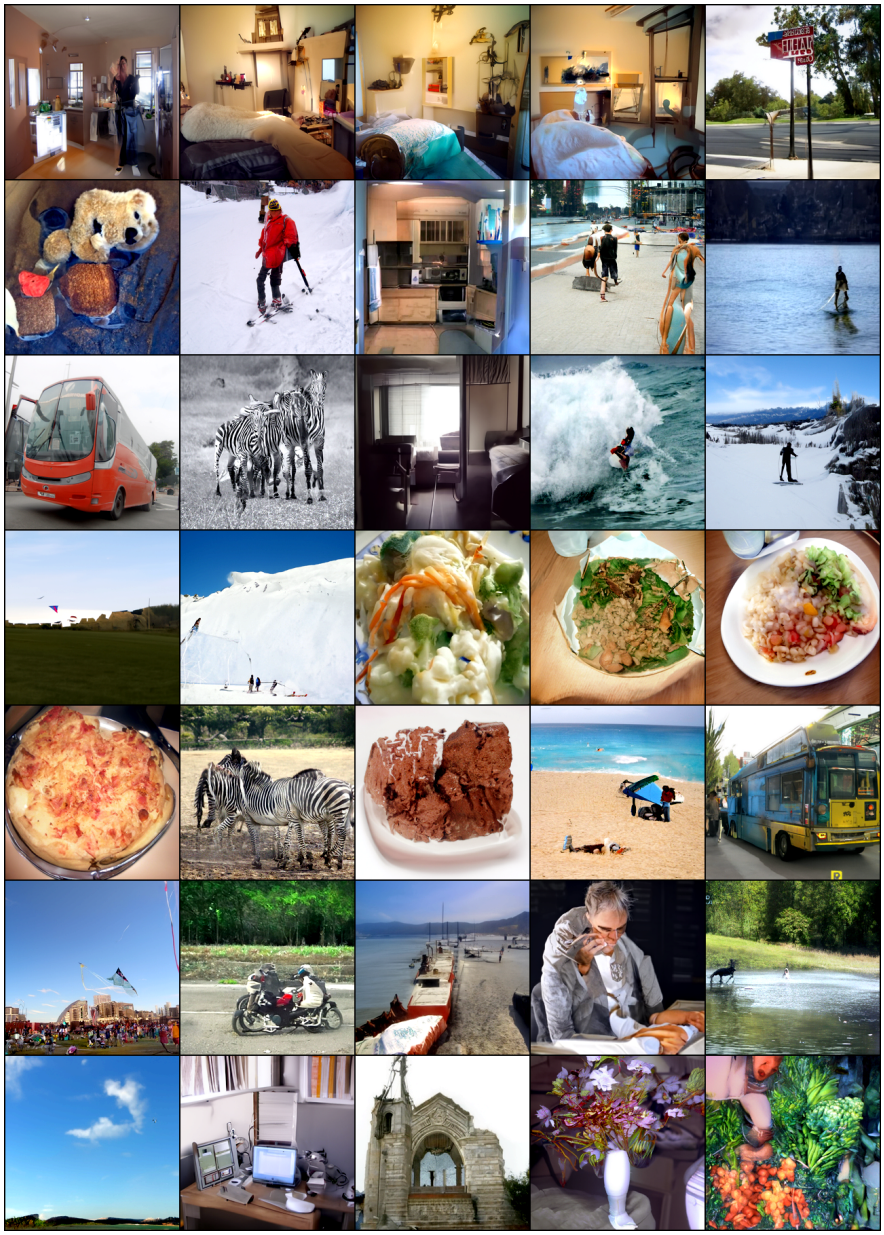

Fig. 3: Random generated samples from validation captions

Label	Words in Captions	Excluded Strings	# of Sent
person	children,women,person,man,human,men,woman,people,child		4361
bicycle	bicycle,bike	motorbike,motor bike,motorcycle,dirt bike	269
car	auto,car	train car,car window,side car,passenger car,subway car,car tire,rail car,tram car,street car,trolly car	858
motorbike	scooter,motorbike,dirt bike,motorcycle		60
aeroplane	plane,aeroplane,aircraft,jet		149
bus	bus		256
train	train		235
truck	truck		78
boat	ship,boat		212
traffic light	traffic light		66
fire hydrant	hydrant,fire hydrant		51
stop sign	stop sign		14
parking meter	parking meter		3
bench	bench		268
bird	bird		138
cat	kitten,cat		320
dog	pup,dog	hot dog,hotdog,hot-dog,cheese dog,chili dog,corn dog	291
horse	horse		187
sheep	sheep		76
cow	cow		73
elephant	elephant	toy elephant,stuffed elephant	147
bear	bear	teddy bear,stuffed bear,care bear,toy bear	45
zebra	zebra		142
giraffe	giraffe		183
backpack	rucksack,backpack		35
umbrella	umbrella		232
handbag	handbag,purse		31
tie	tie	to tie	130
suitcase	suitcase		41
frisbee	frisbee		13
skis	skis		9
snowboard	snowboard		13
sports ball	ball,sports ball		286
kite	kite	kite board,kiteboard	114
baseball bat	baseball bat		37
baseball glove	baseball glove		3
skateboard	skateboard		168
surfboard	surfboard		102
tennis racket	tennis racket,racket		158
bottle	bottle		551
wine glass	wine glass		26
cup	cup		417
fork	fork		179
knife	knife		170
spoon	spoon		239
bowl	bowl	toilet bowl	353
banana	banana		126
apple	apple	pineapple	76
sandwich	sandwich		14
orange	orange,oranges		191
broccoli	broccoli		48
carrot	carrot		72
hot dog	chili dog,cheese dog,hot dog,corn dog		16
pizza	pizza		176
donut	donut		19
cake	cake	cupcake	158
chair	chair		729
sofa	couch,sofa		292
pottedplant	pottedplant,potted plant		0
bed	bed		245
diningtable	desk,table,diningtable		1760
toilet	toilet		178
tvmonitor	monitor,tvmonitor,screen,tv		173
laptop	laptop		256
mouse	mouse,computer mouse		98
remote	remote		98
keyboard	keyboard		117
cell phone	cell phone,mobile phone		90
microwave	microwave		32
oven	oven	microwave oven	60
toaster	toaster		4
sink	sink		154
refrigerator	refrigerator,fridge		116
book	book		335
clock	clock		251
vase	vase		96
scissors	scissors,scissor		40
teddy bear	teddybear,teddy bear		86
hair drier	hair drier		0
toothbrush	toothbrush		17

Table 1: SOA Filter Statistics.

References

1. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34** (2021)
2. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: Zero: Memory optimizations toward training trillion parameter models. ArXiv (October 2019), https://www.microsoft.com/en-us/research/publication/zero-memory-optimizations-toward-training-trillion-parameter-models/