SUPPLEMENTARY MATERIAL

A Introduction

We compare our layered model to single resolution diffusion models in Fig. A.1. All models in this study are trained for 500k steps when evaluated.

Figure A.1: 512×512 outputs from prompt *A black apple and a green backpack.* Finer detail textures can be seen in the layered model when compared with the single resolution model.

B Architecture

B.1 Noise Scaling

Figure A.2: Outputs of layered 256×256 model from prompt *A black apple and a green backpack.*

Noise Type	Target Resolution	FID	IS
Independent	256×256	17.59	$29.32 \pm .46$
Sinc Interpolation	256×256	**13.38**	**29.62 ± .57**
Independent	512×512	42.46	$28.89 \pm .22$
Sinc Interpolation	512×512	**40.05**	**28.74 ± .47**

Table A.1: Model performance for 256×256 and 512×512 layered models using independently sampled noise and scaled noise via sinc interpolation.

B.2 Cosine Schedule Shifting

Figure A.3: Outputs of layered 256×256 model from prompt *landscape photo of beach with proof watermark*. We note that for a more aggressively shifted noise delay (bottom), higher resolution features can be seen in the waves. We also note the advent of a water-mark looking feature in the first image.

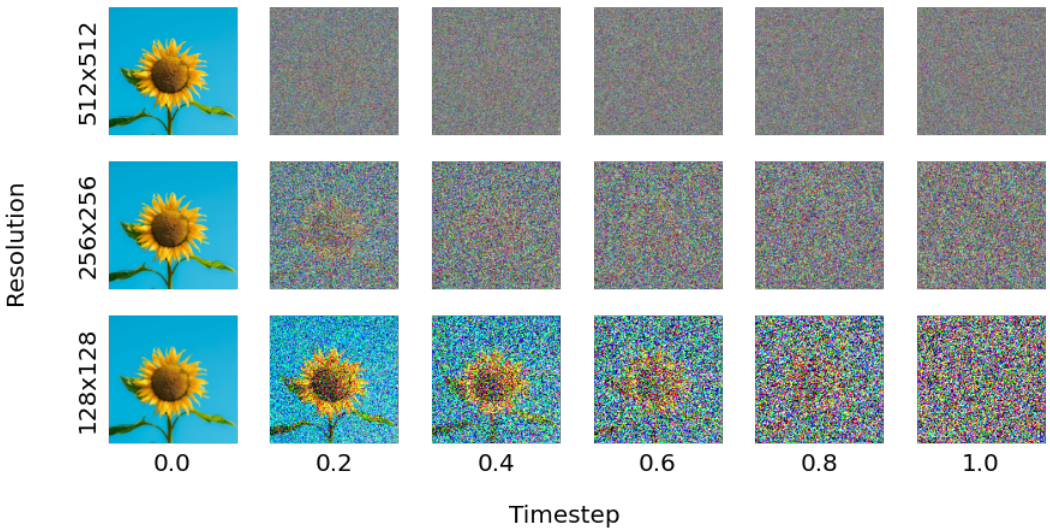

Figure A.4: Demonstration of the increasingly shifted noise schedules for higher resolution on a reference image [1]. No shifting is applied to the cosine noise schedule at 128×128.

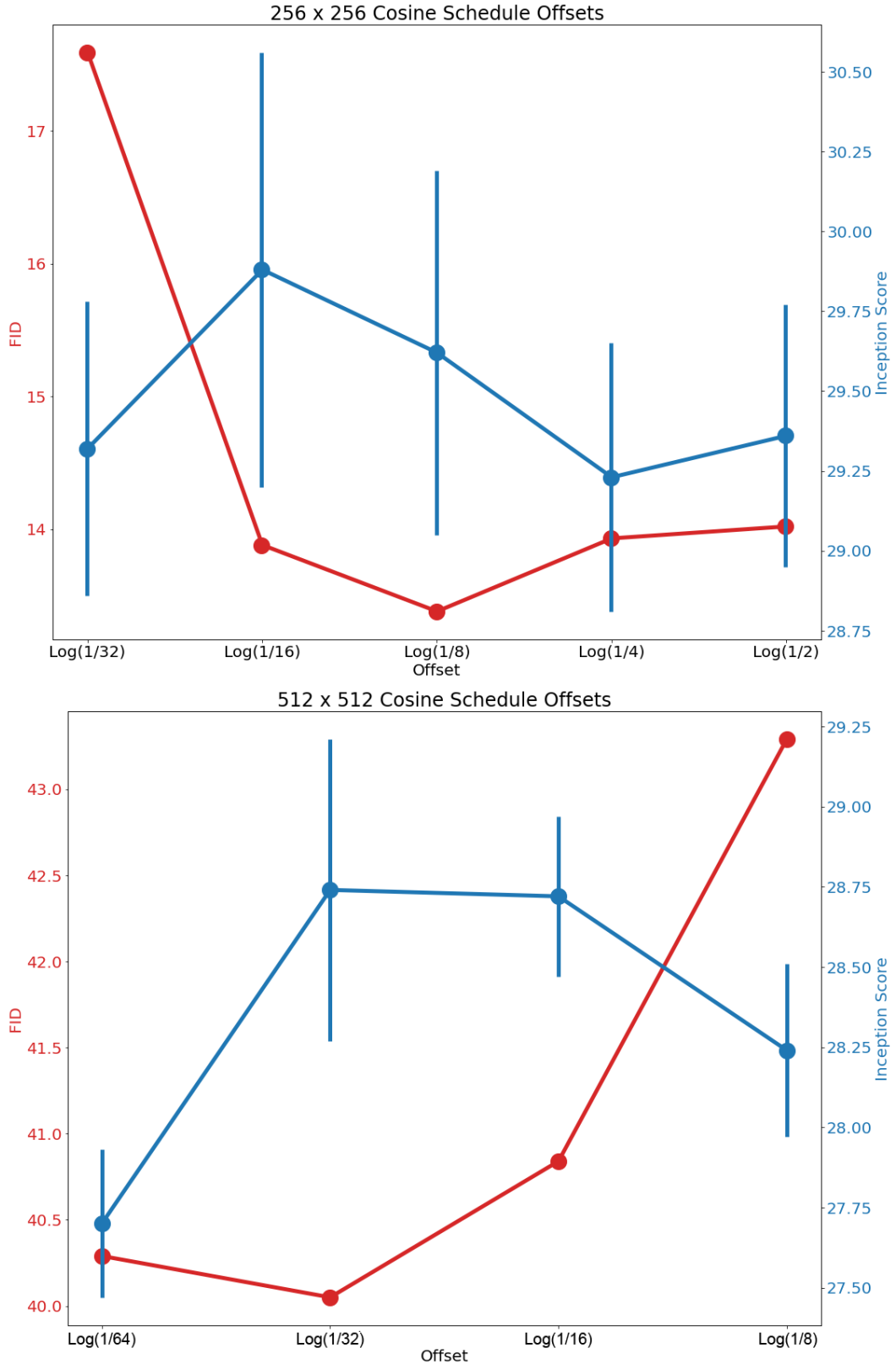

Figure A.5: Sweeps showing IS and FID values on MSCOCO validation set for models trained with varying noise offsets. For the 512×512 model, we use a cosine schedule offset of $log\frac{1}{8}$ for the 256×256 layer.

C Training Optimizations

C.1 Strategic Cropping

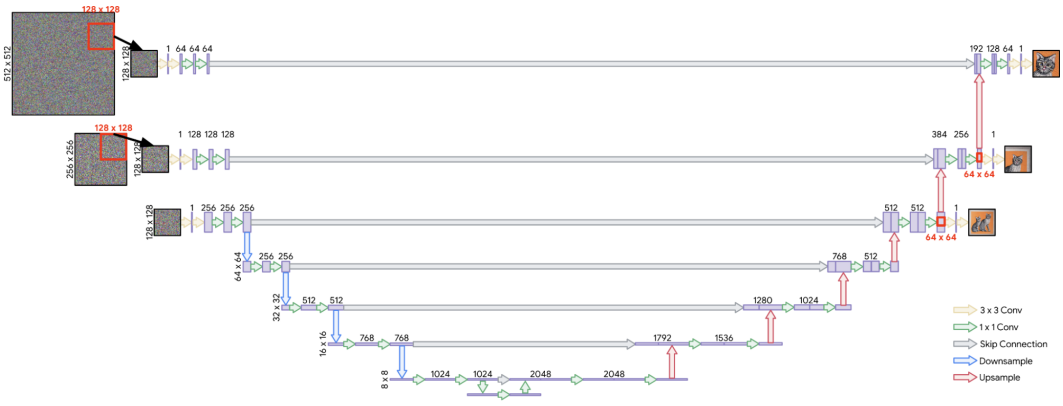

Figure A.6: Model architecture with cropping applied. During training, all input images are 128×128. Observe in the upsampling stack, we take 64×64 crops prior to the upsampling convolution to ensure that the output is a 128×128 crop in the correct area.

C.2 Model Stacking

Model	Resolution	FID	IS
(a) Random Initialization	256×256	**13.38**	**29.62 ± .57**
(b) Load 128×128 Model	256×256	13.91	$28.85 \pm .03$
Random Initialization	512×512	**40.05**	**28.74 ± .47**
Load (a)	512×512	42.21	$27.82 \pm .21$
Load (b)	512×512	43.53	$26.85 \pm .26$

Table A.2: Model performance for 256×256 and 512×512 layered models utilizing different amount of pre-training. Despite parts of the model essentially encountering a higher number of training steps and thereby images, we see degraded FID and IS.

Figure A.7: 512×512 output images for prompt *A pizza on the right of a suitcase*. We note decreased image quality with increased initialization. However, we observe the opposite trend from the perspective of image alignment, where only the most initialized model generates the suitcase.

References

[1] Unsplash. Photo by michelle francisca lee on unsplash.