A. Model specifications and hyperparameters

The hyperparameters used across our experiments are the same as the compared baselines, this is in order to perform a concise evaluation. Still, for clarity and completeness of this work, we indicate the hyperparemeters used for each model version.

A.1. CIFAR10

In CIFAR10, we used the baselines DDPM [9]. The model is a UNet architecture, with the following hyperparameters. The UNet had a depth of 4 downsampling (and upsampling) blocks, with a base number channel size of 128 and channel multiplier of [1,2,2,2]. Each block contained a residual block with 2 residual layers, and an attention block at the 16x16 resolution. The model was subject to dropout of 0.1. Following the baseline, the linear noise schedule was from 1e-4 to 2e-2 in 1000 steps. Training was done on 4 GPUs, with a batch size of 128x4, for 1M iterations. The Adam optimizer was used with learning rate of 2e-4, and EMA decay of 0.9999.

For the cosine noise schedule, we used IDDPM [23]. The improved UNet model included a scale-shift GroupNorm instead of the standard GroupNorm, three residual layers in each block, attention on both the 16x16 and 8x8 resolutions, 4 attention heads instead of 1, and a cosine noise schedule.

A.2. CelebA

DDPM was also selected for CelebA evaluation of 64x64 image resolution. The difference from its CIFAR10 counterpart, is the addition of a fifth downsampling layer, with the same base channel size of 64x4, and a channel multiplier of 4. Model was trained for 500K iterations, using batch size of 32, and Adam optimizer with learning rate 1e-5.

A.3. ImageNet

The evaluated model on ImageNet is based on ADM [6]. This architecture has some major differences from the previous ones. The model had classifier condition, which were being added to the time condition. Instead of pooled downsampling and interpolated upsampling, a learned up/downsampling was applied through the residual block. In addition, each block has a base channel size of 256, with channel multipliers of [1,1,2,3,4]. Attention of 4 heads was applied on the 32x32, 16x16, and 8x8 resolutions. A dropout of 0.1 and scle-shift GroupNorm. Noise schedule was the default linear schedule. Finally, we trained only the decoder weights, while using the pretrained weights of the baseline for the rest of the model. Model was trained for 80K iterations, with batch size of 32x4 (with 8 mini-batches of 4). Adam optimizer with learning rate of 1e-5.

B. Generated images

In addition to the images in the paper, we provide additional generated images for the various datasets, for further inspection.

C. Progressive generation

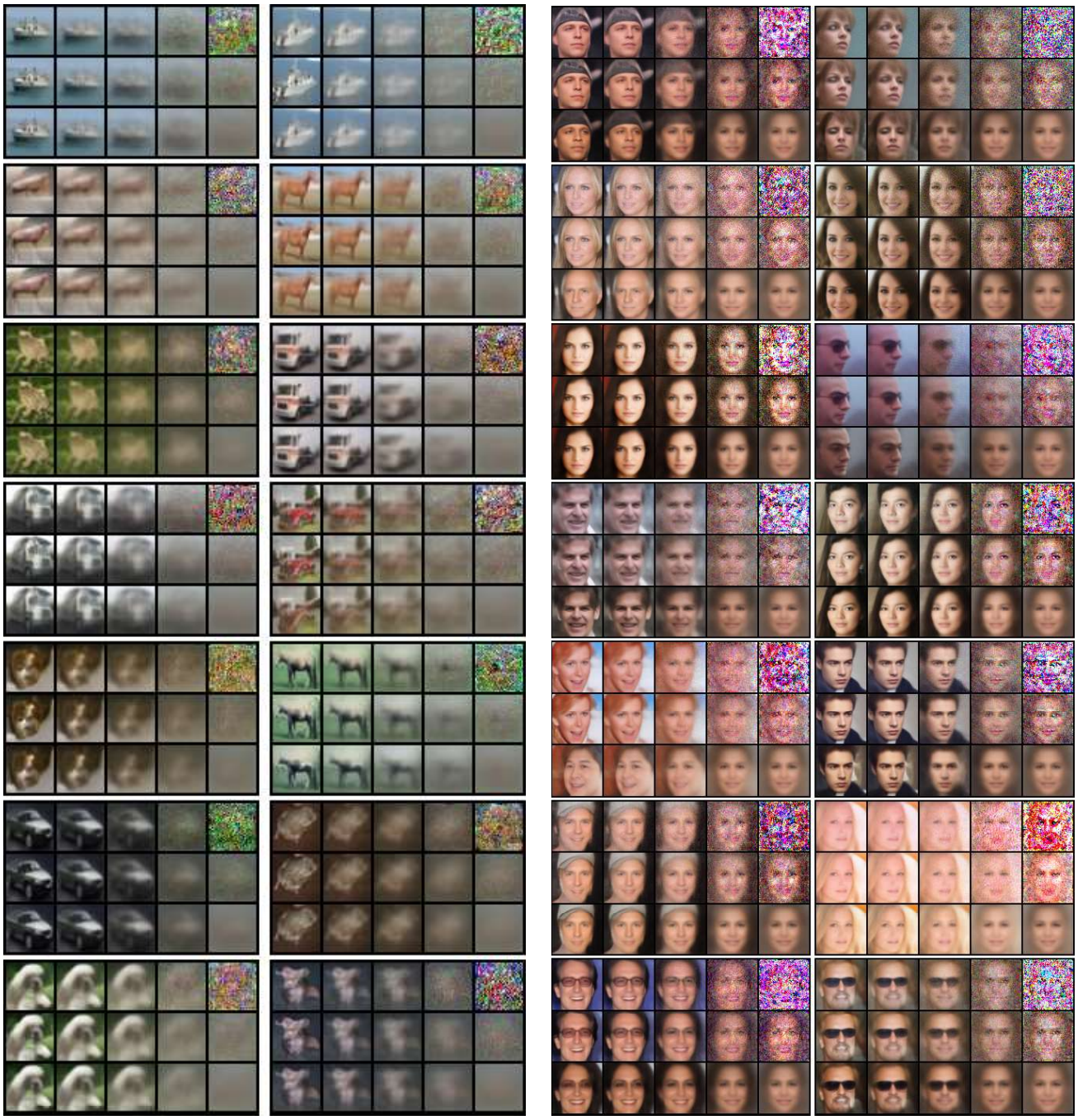
Fig. 9 shows progressive generation results for CIFAR10 and CelebA. All grids show the intermediate results of the three paths ϵ_θ, *dual*, and x_θ, from top to bottom. It can be seen how in all cases, ϵ_θ starts very noisy, while x_θ is blurry. All paths end with a similar image, but the dual method provides a sharper and less noisy result. In CelebA we noted more difference between the images. The additive path often produced darker images, and the noise in the final result of ϵ_θ is very noticeable.

D. Effect of iteration count

Fig. 10 shows generation for both CIFAR10 and CelebA with a different number of denoising iterations. Iteration are monotonically increasing from left to right (5, 10, 20, 50, 100). The effect of the number of iterations is very clear as the image becomes more detailed and sharp when more denoising iterations are applied. Sometimes there is also a change in appearance, but an improvement in quality is always present. However, it can be seen that the change in quality is relatively low, and an already good image is achieved with few iterations.

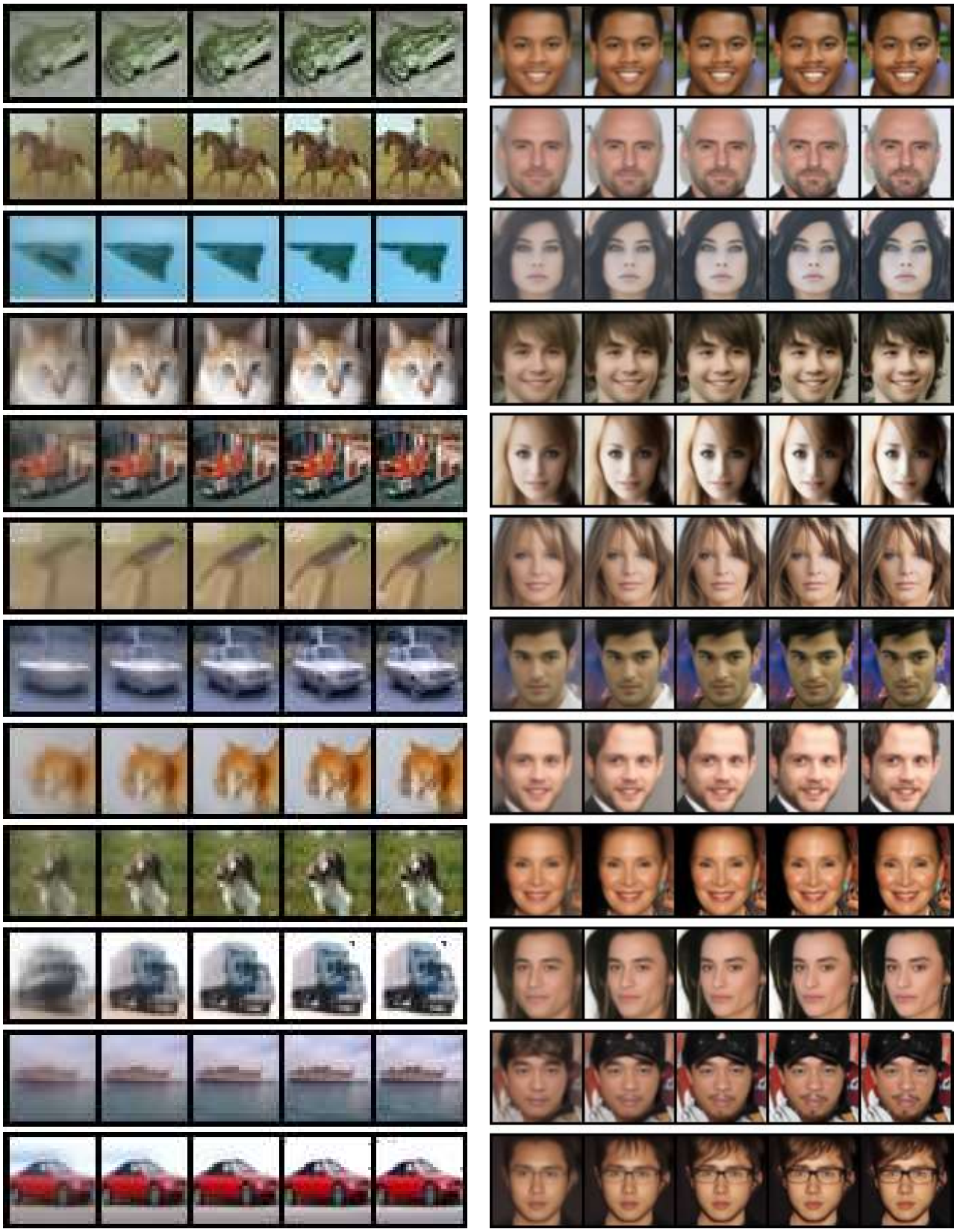
E. High quality ImageNet result (128×128)

Fig. 11 shows generated images from ImageNet, using 50 denoising iterations. There is a high variety in the images, and the class condition successfully represents the chosen category. Considering that the model was only fine-tuned for 80K steps, and the denoising is done with only 50 iterations, the image quality is quite good.



(a) (b)

Figure 9. Progressive generation (a) CIFAR10 and (b) CelebA. From top to bottom: ϵ_θ, *dual*, and x_θ.



(a) (b)

Figure 10. Generation on (a) CIFAR10 and (b) CelebA. From left to right: 5, 10, 20, 50, 100 iterations.

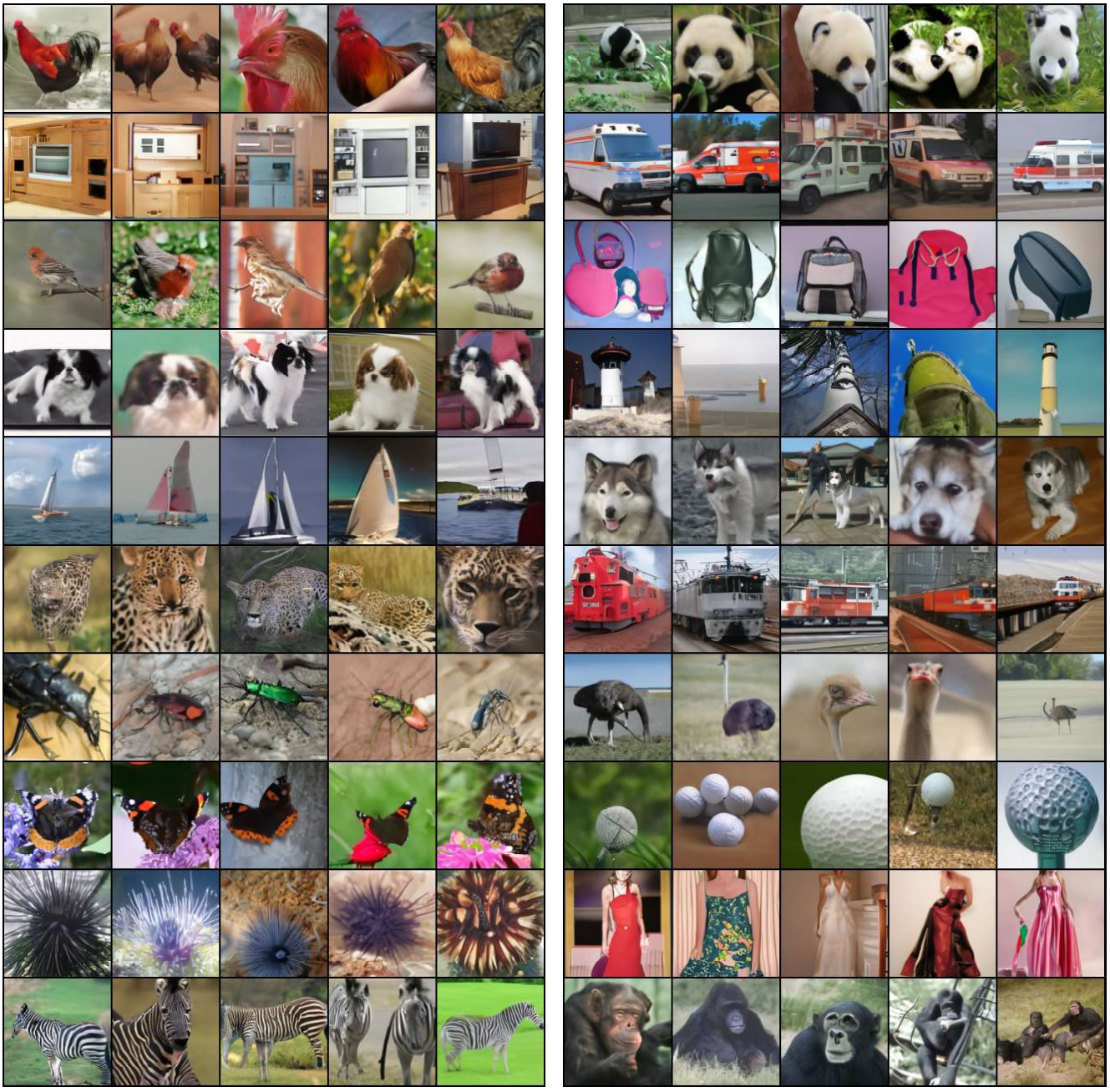


Figure 11. Generation on ImageNet with 50 iterations.