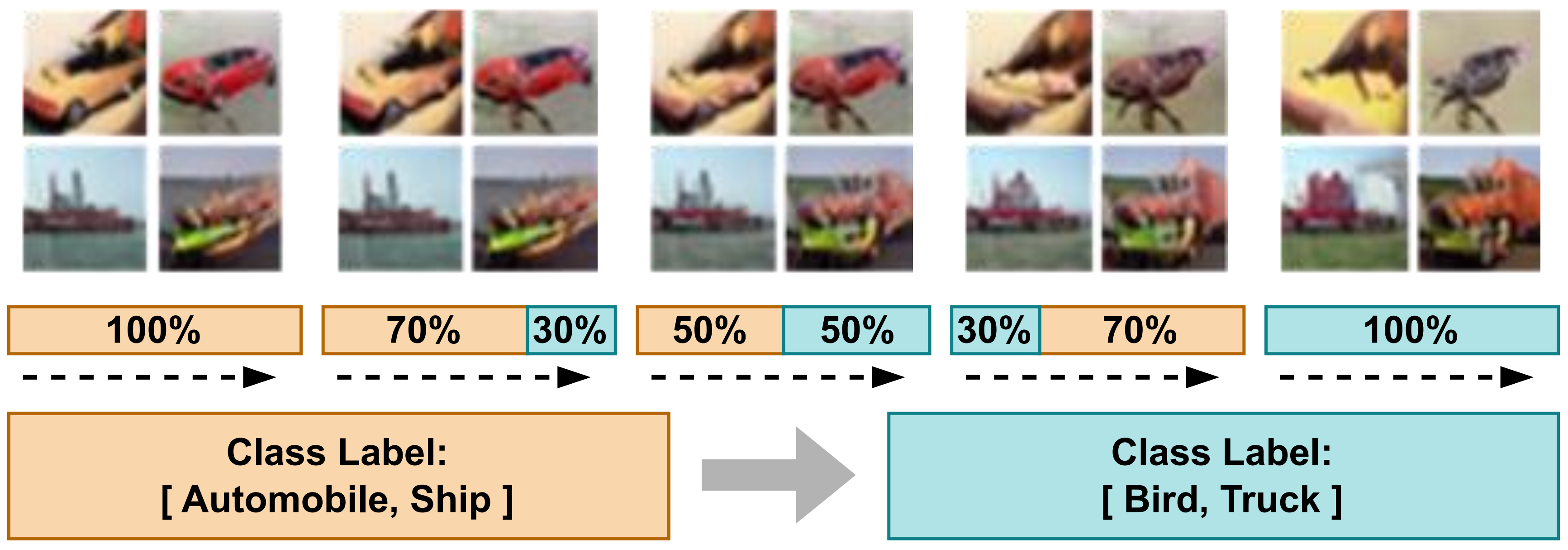
Gemini Diffusion Model Unveils Underlying Division of Denoising Process

Group 24 310552069 王盈皓、311552004 李明倫、311553009 許瑋芸

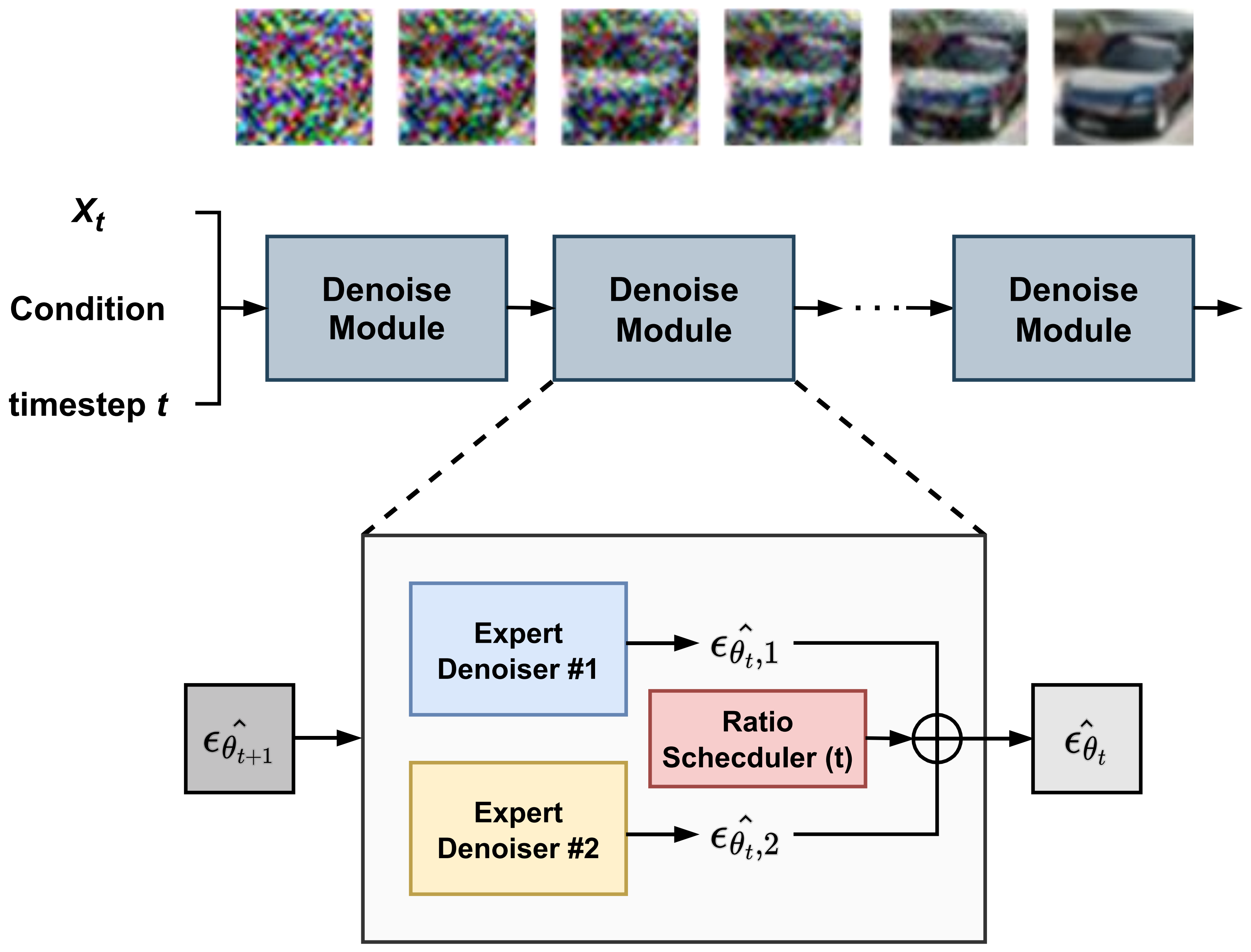
Introduction

Diffusion models have made significant progress in the field of image synthesis, by generating high-quality images with diverse content and realistic details. Motivated by *eDiff-I*, a recent work from NVIDIA, indicated that the synthesis behavior qualitatively changes throughout the denoising process. At the early stage of sampling, the diffusion model mainly relies on the condition to guide the sampling process. As the denoising process continues, the model gradually ignores the class conditioning and focuses on generating high-quality visual features. Based on this observation, we propose to decouple a diffusion model by two expert denoisers, each was responsible for optimizing different parts of the images at different denoising stage. We aim to improve the quality and details of the synthesized images.



Methodology

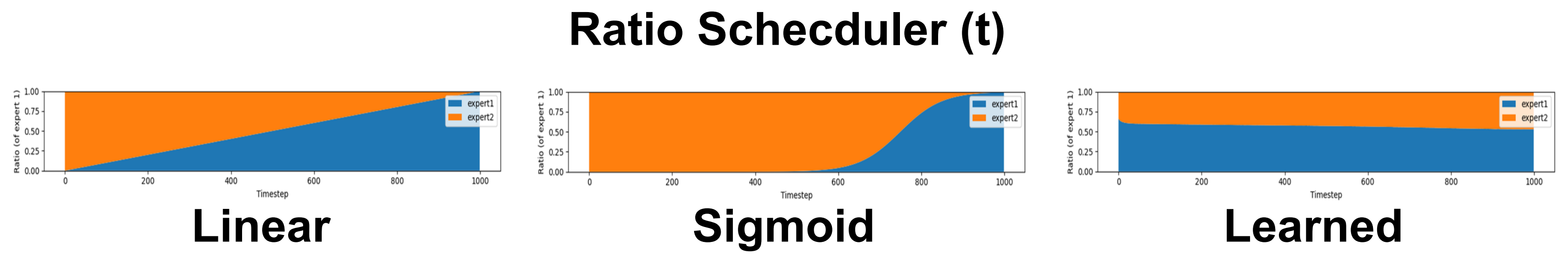
In order to train specialized experts to capture different visual features, we designed a **Denoise Module**, which consist of two expert denoisers and a ratio scheduler. Each expert denoisers will focus on one or more specific features e.g. line, contour, color, etc. For example, one expert denoiser may specialize in generating fine lines and contours in the image, another may be responsible for adding realistic colors and textures. These experts will collaborate effectively throughout the denoising process.



$$Loss = (\epsilon_t - \hat{\epsilon_{\theta_t}})^2$$

Figure 1. Illustration of our proposed denoise module which we train 2 expert denoisers to focus on different features generation.

Given a time step *t*, the ratio scheduler produces a weighting factor *r* to blend the predicted epsilons from two experts. We introduce three types of ratio schedules: linear, sigmoid and learned schedules:



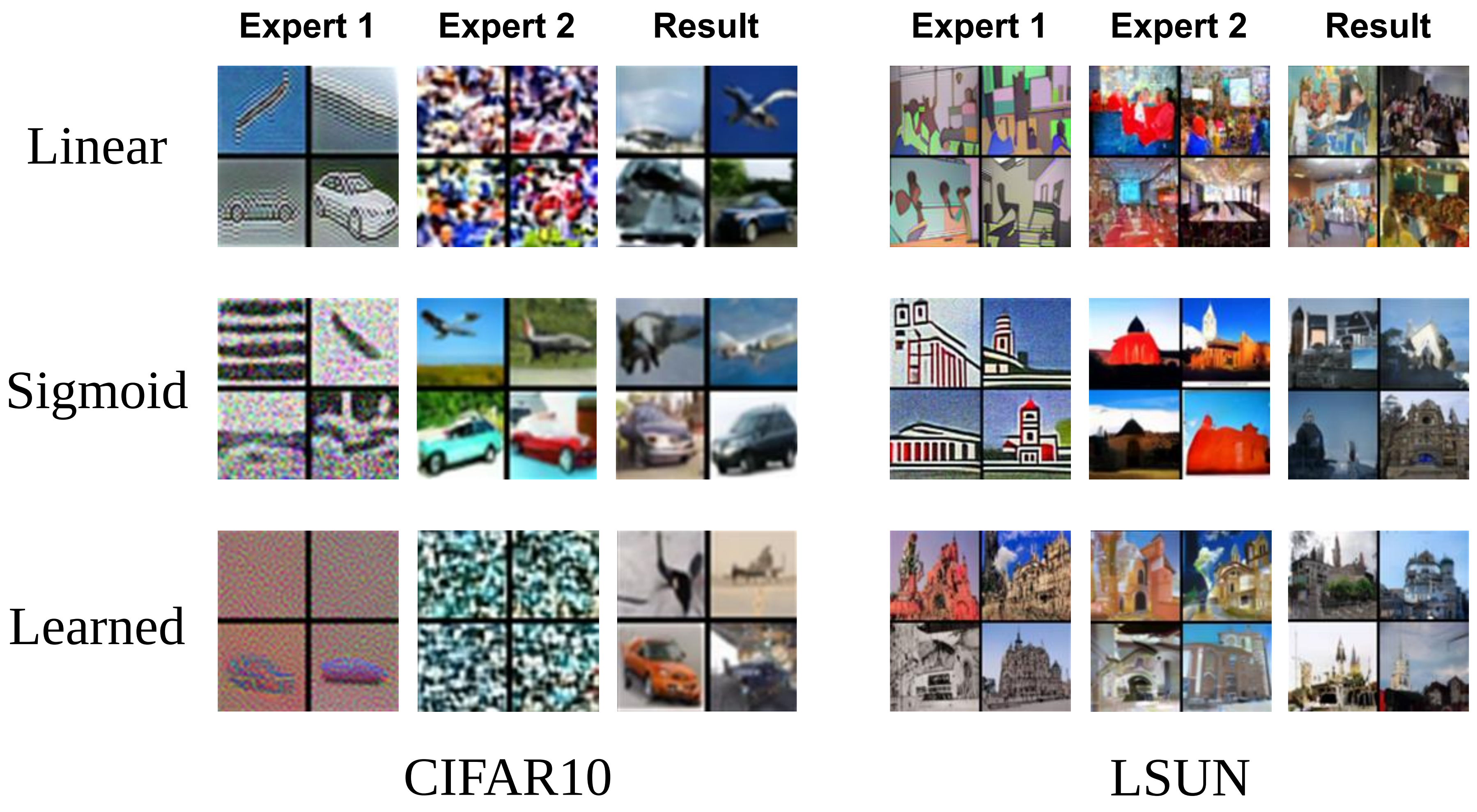
The learned scheduler is the only one that incorporates learnable parameters. These schedule types enable denoise modules to learn different aspects during the diffusion process. We trained three models using each ratio schedule and compare with a baseline model.

Experiment

We use CIFAR10 and LSUN as our datasets. We only select four classes in LSUN due to our limited computing resource. Models trained on CIFAR10 are trained with 200 epochs from scratch. For models trained on LSUN, we use a base model that has been trained for 735000 iterations, and train another 55000 iterations based on this model.

Moreover, to gain a deeper insight into the aspects of diffusion that each expert denoiser learns, we observe the images that they generate. For performance evaluation, we calculate FID between 10000 sampled images and the validation set of each dataset.

	FID	
Models	CIFAR10	LSUN
Baseline	17.02	36.33
Our (Linear)	16.48	43.70
Our (Sigmoid)	16.56	**36.16**
Our (Learned)	**15.84**	61.02



Conclusion

Our main contribution is to improve the quality of the synthesized images by replacing a diffusion model with two expert denoisers and different ratio schedules. As we can see, different expert denoisers allow us to specialize the model for different behaviors during different intervals of the iterative synthesis process. Moreover, we showed our method has a chance to outperform baseline.