Supplementary Material

In the following, we provide additional experiments and visualizations for the GANsformer model. To complement the numerical evaluations with qualitative results, we present in figures 11 and 8 a comparison of sample images produced by the GANsformer and a set of baseline models, over the course the training and after convergence respectively, while section A specifies the implementation details, optimization scheme and training configuration of the model. Finally, in section B and figure 7, we measure the degree of spatial compositionality of the GANsformer attention mechanism and sheds light upon the roles of the different latent variables.

A. Implementation and Training Details

To evaluate all models under comparable conditions of training scheme, model size, and optimization details, we implement them all within the TensorFlow codebase introduced by the StyleGAN authors (Karras et al., 2019). See tables 4 for particular settings of the GANsformer and table 5 for comparison of models' sizes. In terms of the loss function, optimization and training configuration, we adopt the settings and techniques used in the StyleGAN2 model (Karras et al., 2020), including in particular style mixing, Xavier Initialization, stochastic variation, exponential moving average for weights, and a non-saturating logistic loss with lazy R1 regularization. We use Adam optimizer with batch size of 32 (4 times 8 using gradient accumulation), equalized learning rate of 0.001, $\beta_1 = 0.9$ and $\beta_1 = 0.999$ as well as leaky ReLU activations with $\alpha = 0.2$, bilinear filtering in all up/downsampling layers and minibatch standard deviation layer at the end of the discriminator. The mapping layer of the generator consists of 8 layers, and ResNet connections are used throughout the model, for the mapping network synthesis network and discriminator. We train all models on images of 256×256 resolution, padded as necessary. The CLEVR dataset consists of 100k images, the FFHQ has 70k images, Cityscapes has overall about 25k images and the LSUN-Bedroom has 3M images. The images in the Cityscapes and FFHQ datasets are mirror-augmented to increase the effective training set size. All models have been trained for the same number of training steps, roughly spanning a week on 2 NVIDIA V100 GPUs per model.

B. Spatial Compositionality

To quantify the compositionality level exhibited by the model, we employ a pre-trained segmentor to produce semantic segmentations for the synthesized scenes, and use them to measure the correlation between the attention cast by the latent variables and the various semantic classes. We derive the correlation by computing the maximum intersection-over-union between a class segment and the attention segments produced by the model in the different layers. The mean of these scores is then taken over a set of 1k images. Results presented in figure 7 for the Bedroom and Cityscapes datasets, showing semantic classes which have high correlation with the model attention, indicating it decomposes the image into semantically-meaningful segments of objects and entities.

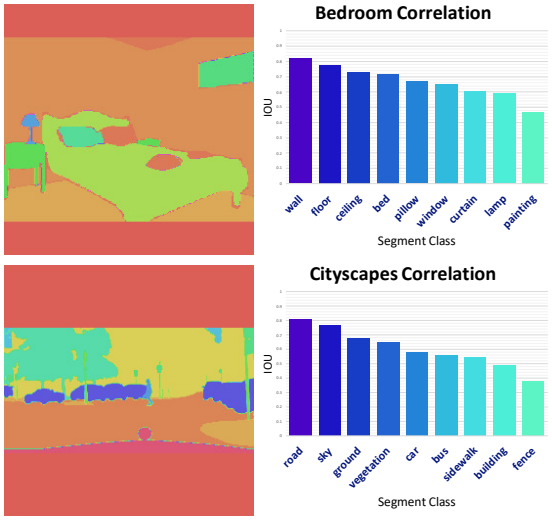

Figure 7. **Spatial Compositionality.** Correlation between attention heads and semantic segments, computed over 1k sample images. Results presented for the Bedroom and Cityscapes datasets.

Table 4. **Hyperparameter choices**. The latents number (each can be multidimensional) is chosen based on performance among $\{8, 16, 32, 64\}$. The overall latent dimension is chosen among $\{128, 256, 512\}$ and is then used both for the GANsformer and the baseline models. The R1 regularization factor γ is chosen among $\{1, 10, 20, 40, 80, 100\}$.

	FFHQ	CLEVR	Cityscapes	Bedroom
# Latent vars	8	16	16	16
Latent var dim	16	32	32	32
Latent overall dim	128	512	512	512
R1 reg weight (γ)	10	40	20	100

Table 5. **Model size** for the GANsformer and competing approaches, computed given 16 latent variables and an overall latent dimension of 512. All models have comparable size.

	# G Params	# D Params
GAN	34M	29M
StyleGAN2	35M	29M
k-GAN	34M	29M
SAGAN	38M	29M
GANsformer$_s$	36M	29M
GANsformer$_d$	36M	29M

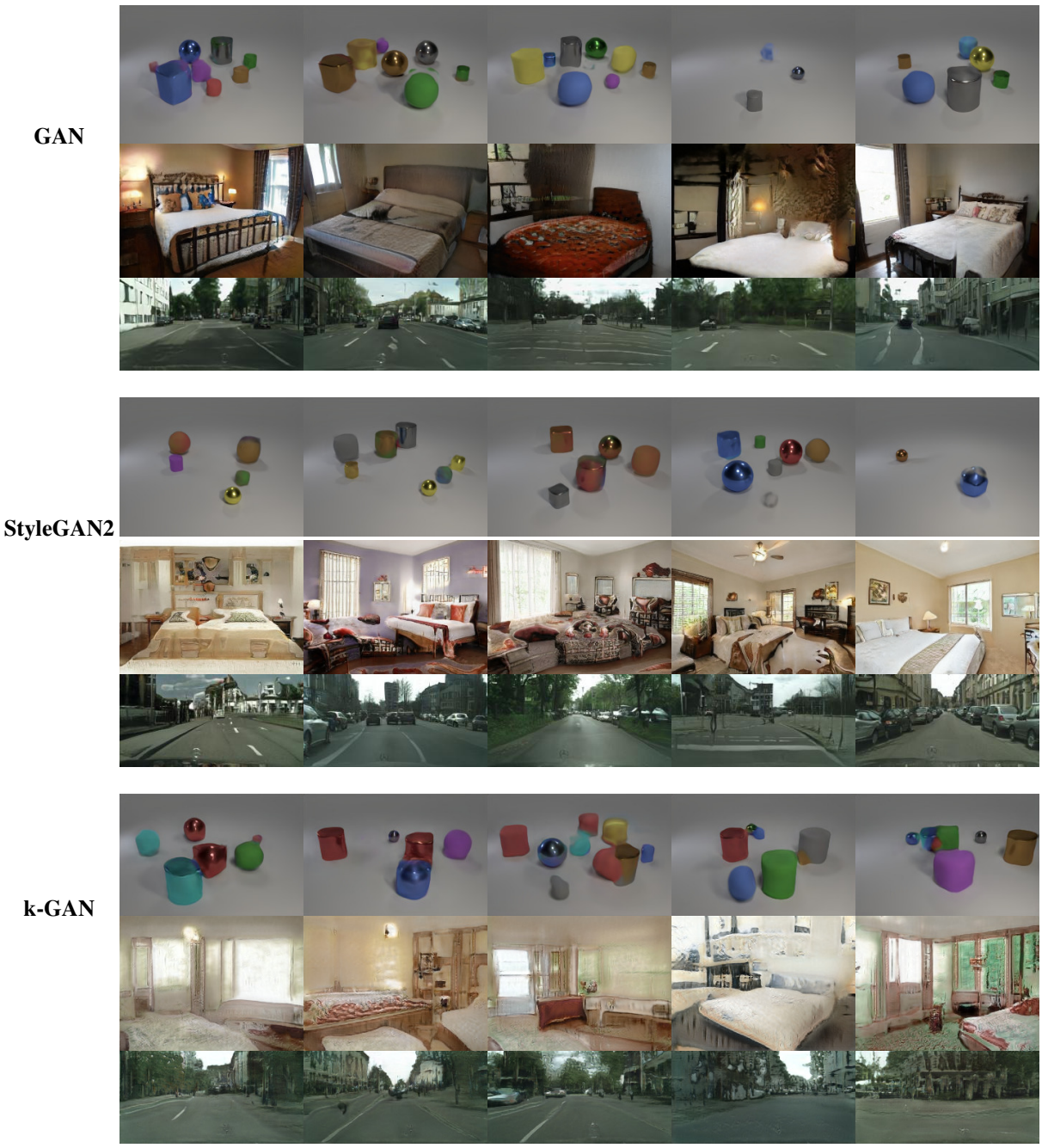

Figure 8. **State-of-the-art Comparison.** A comparison of models' sampled images for the CLEVR, LSUN-Bedroom and Cityscapes datasets. All models have been trained for the same number of steps, which ranges between 5k to 15k samples. Note that the original StyleGAN2 model has been trained by its authors for up to generate 70k samples, which is expected to take over 90 GPU-days for a single model. See next page for image samples by further models. These images show that given the same training length the GANsformer model's sampled images enjoy high quality and diversity compared to the prior works, demonstrating the efficacy of our approach.

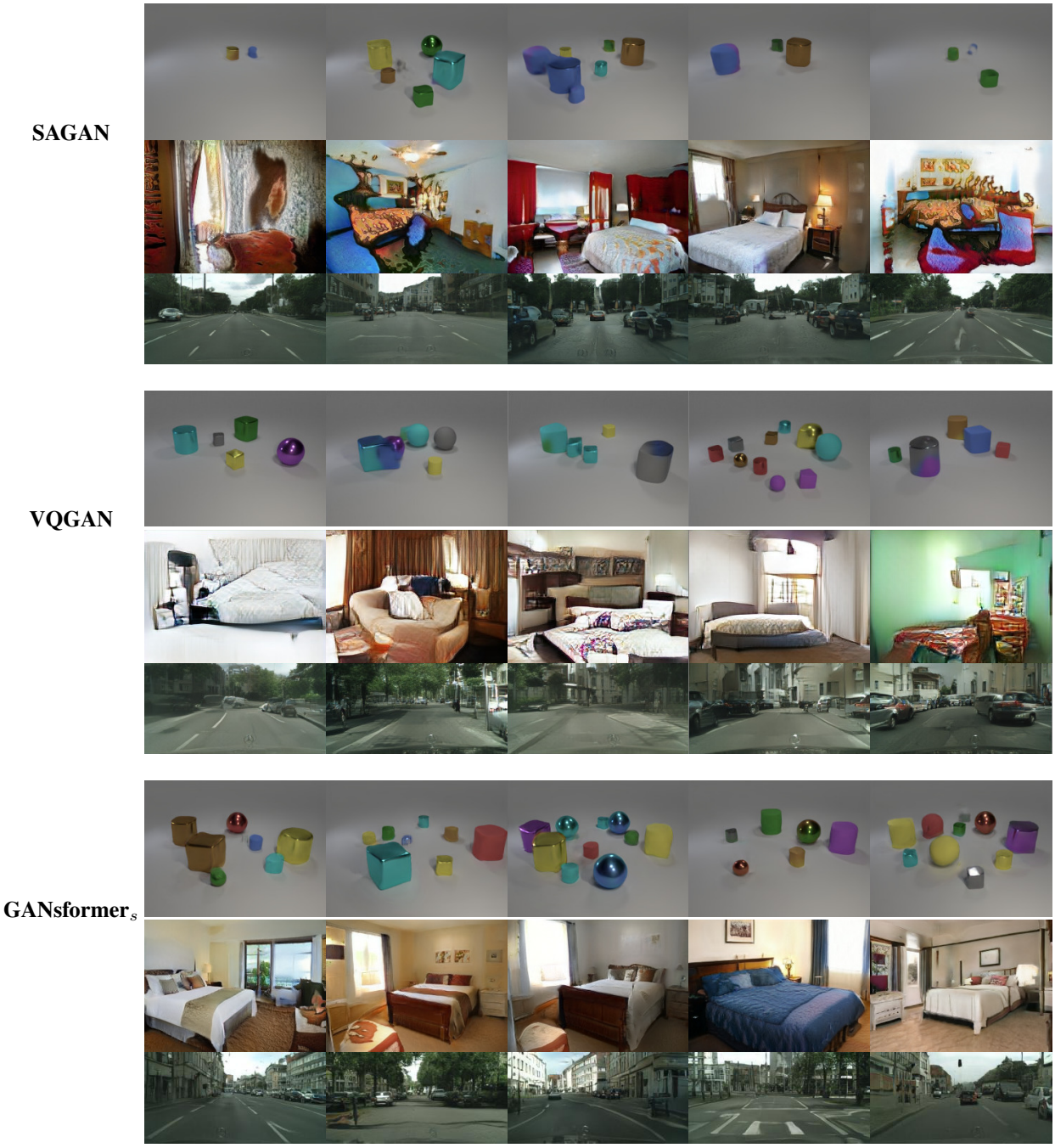

Figure 9. A comparison of models' sampled images for the CLEVR, LSUN-Bedroom and Cityscapes datasets. See figure 8 for further description.

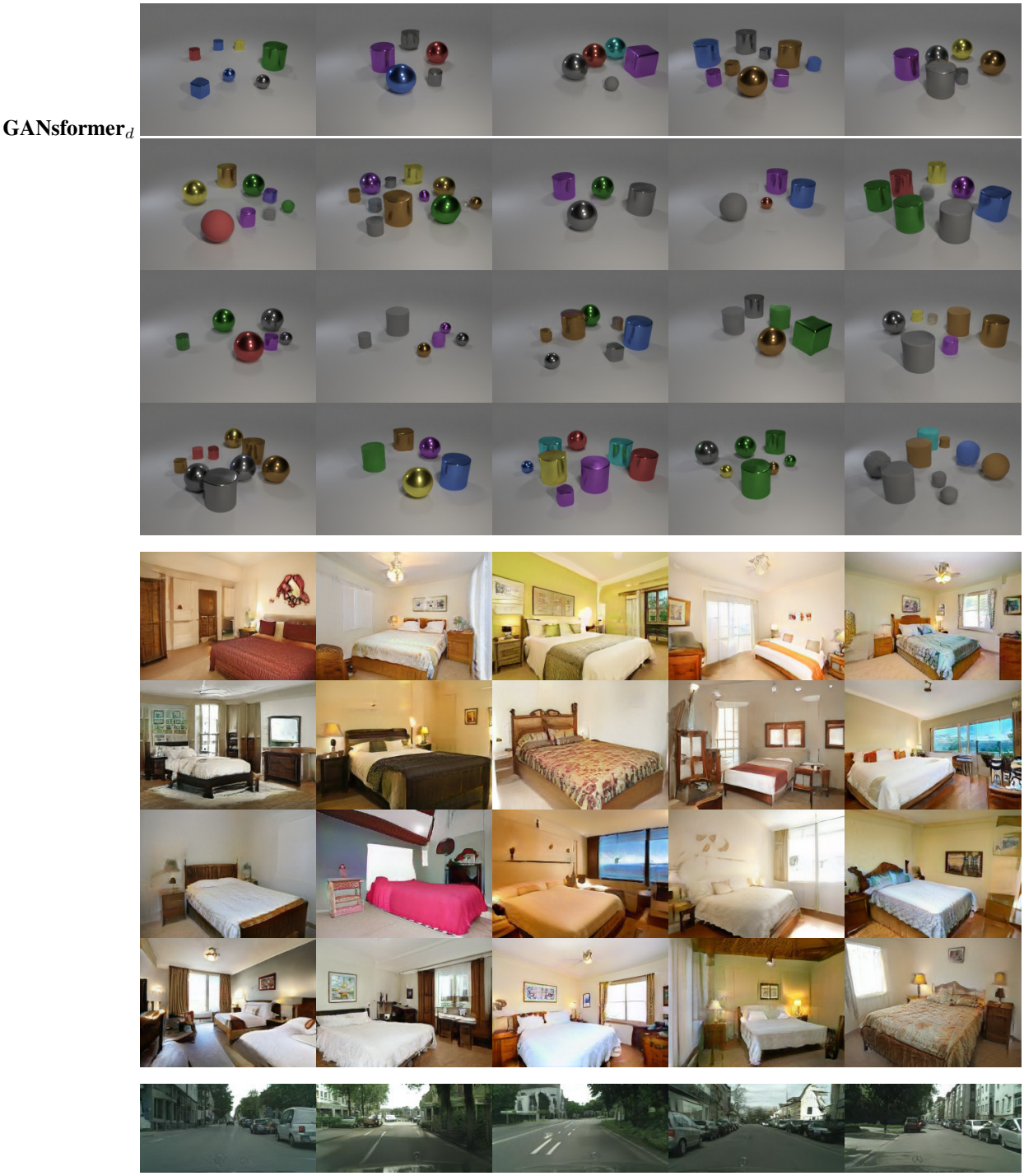

Figure 10. A comparison of models' sampled images for the CLEVR, LSUN-Bedroom and Cityscapes datasets. See figure 8 for further description.

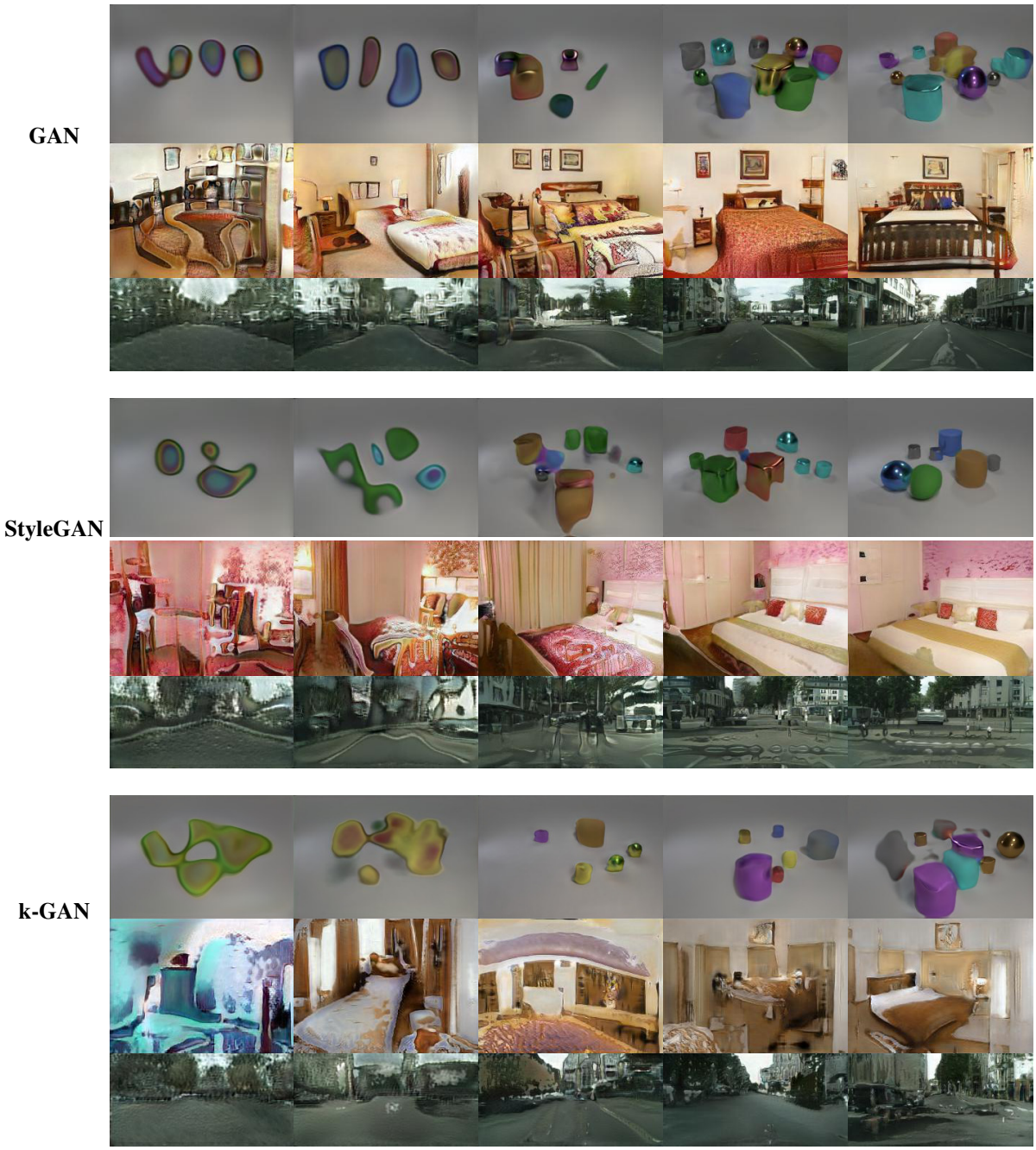

Figure 11. **State-of-the-art Comparison over training.** A comparison of models' sampled images for the CLEVR, LSUN-Bedroom and Cityscapes datasets, generated at different stages throughout the training. Sampled image from different points in training of based on the same sampled latents, thereby showing how the image evolves during the training. For CLEVR and Cityscapes, we present results after training to generate 100k, 200k, 500k, 1m, and 2m samples. For the Bedroom case, we present results after 500k, 1m, 2m, 5m and 10m generated samples while training. These results show how the GANsformer, and especially when using duplex attention, manages learn a lot faster than the competing approaches, generating impressive images very early in the training.

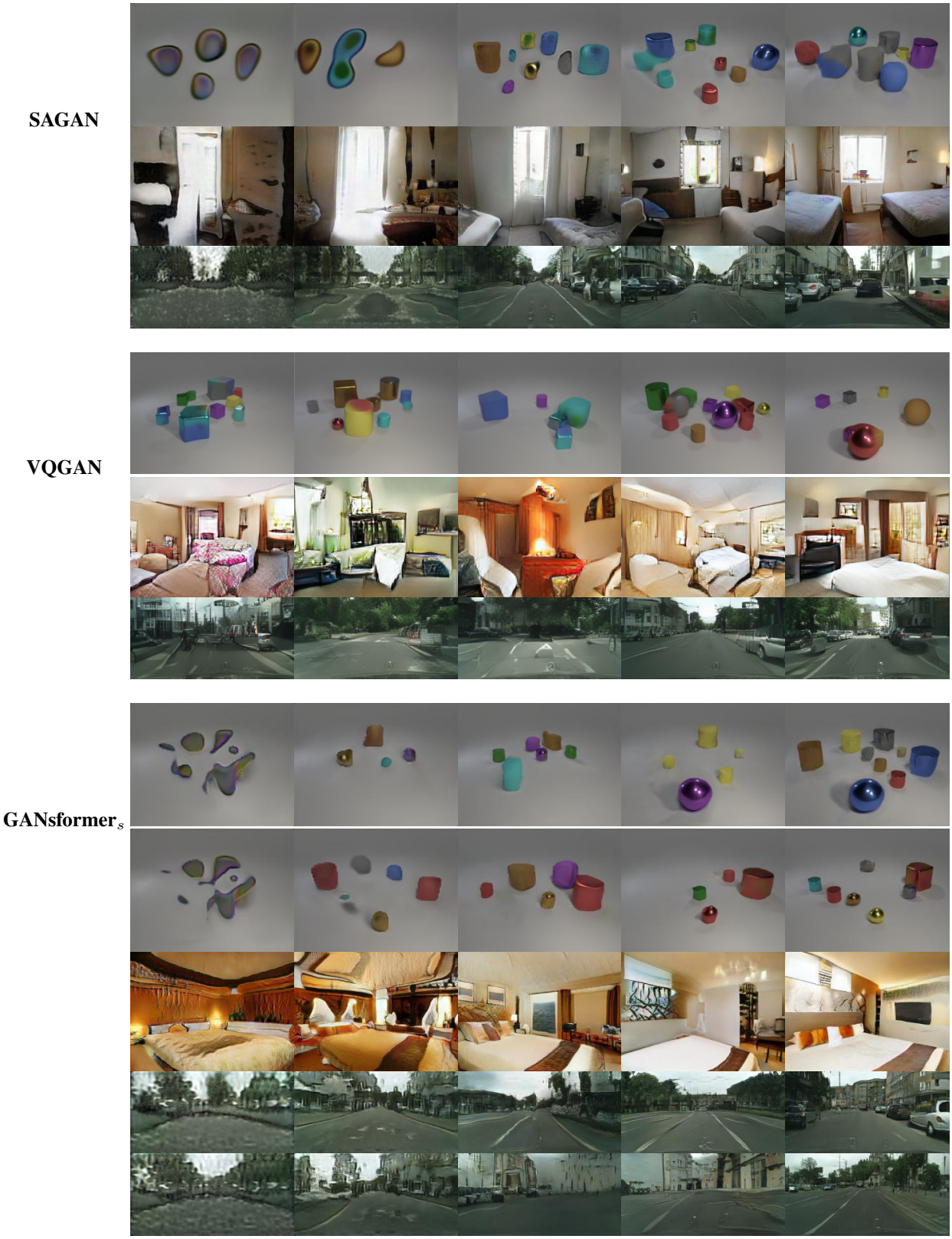

Figure 12. A comparison of models' sampled images for the CLEVR, LSUN-Bedroom and Cityscapes datasets throughout the training. See figure 11 for further description.

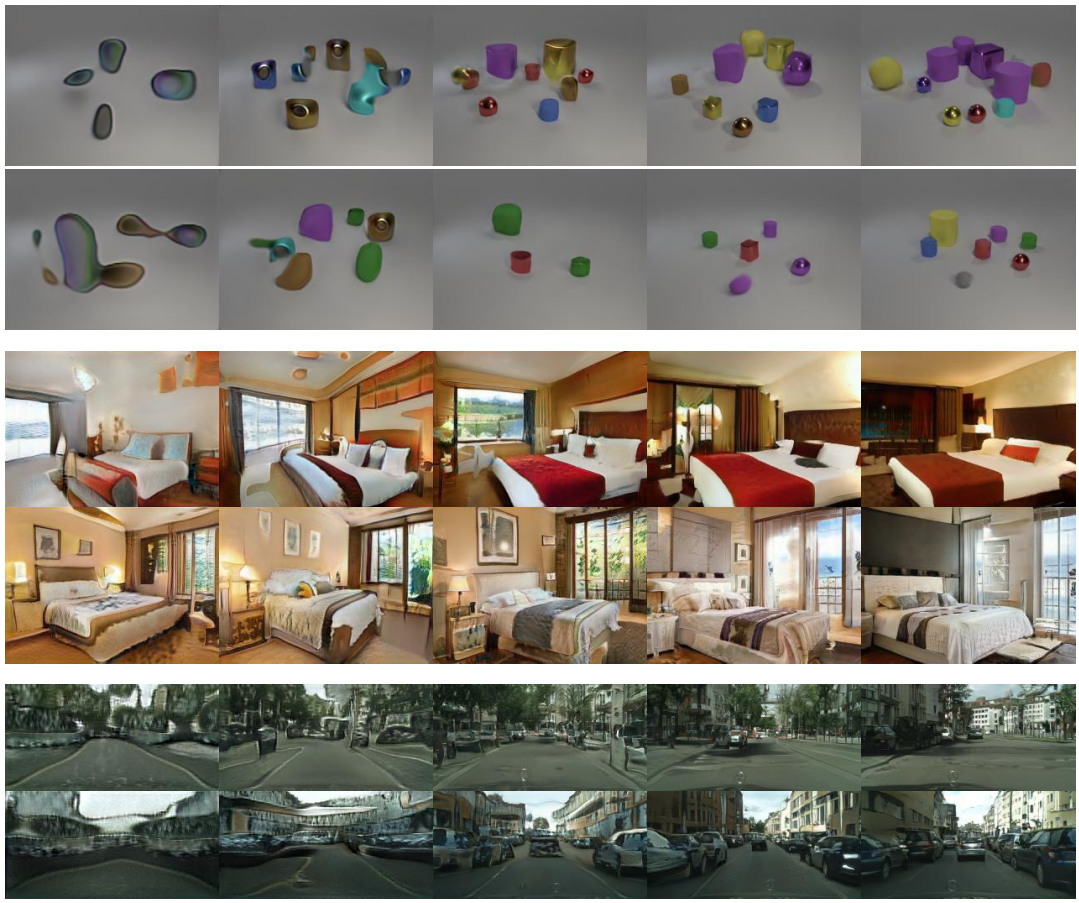

Figure 13. A comparison of models' sampled images for the CLEVR, LSUN-Bedroom and Cityscapes datasets throughout the training. See figure 11 for further description.