**Improved Diffusion Models for Text-to-Image Synthesis**

**Introduction:** In recent years, generative models have improved their capacity to produce natural language that is human-like, infinitely varied human speech and music, along with high-quality, synthetic pictures. These models can be applied in a variety of contexts, such as learning relevant feature representations or generating images from text prompts. The adversarial learning process used by GANs does not scale well to modeling complex, multi-modal distributions; it has been found that their findings are primarily limited to data with fairly low variability. The state-of-the-art in class-conditional image synthesis and super-resolution is defined by diffusion models, which are constructed from a hierarchy of denoising autoencoders. Diffusion models have recently demonstrated their ability to achieve outstanding outcomes in image synthesis and beyond. An updated diffusion model named latent diffusion model also introduced to train massive models with efficiency. But there is still some scope of improvement in image generation in terms of making images more fine-tuned as well as improving generating more accurately described images from the text.

**Background**: On a high level, diffusion models sample from a distribution by reversing a gradual noising process. In particular, sampling starts with noise xT and produces gradually less-noisy samples xT -1, xT -2,…until reaching a final sample x0. Each timestep t corresponds to a certain noise level, and xt can be thought of as a mixture of a signal x0 with some noise where the signal to noise ratio is determined by the timestep t. For the remainder of this paper, we assume that the noise is drawn from a diagonal Gaussian distribution, which works well for natural images and simplifies various derivations.

A diffusion model learns to produce a slightly more “denoised” xt-1 from xt. Ho et al. parameterize this model as a function θ (xt; t) which predicts the noise component of a noisy sample xt. To train these models, each sample in a minibatch is produced by randomly drawing a data sample x0, a timestep t, and noise, which together give rise to a noised sample xt. The training objective is then ||(xt; t) - ||2, i.e., a simple mean-squared error loss between the true noise and the predicted noise.

It is not immediately obvious how to sample from a noise predictor θ (xt, t). Recall that diffusion sampling proceeds by repeatedly predicting xt-1 from xt, starting from xT. Ho et al.[1] show that, under reasonable assumptions, we can model the distribution pθ(xt-1|xt) of xt-1 given xt as a diagonal Gaussian N (xt-1; µθ (xt; t); Σθ (xt; t)), where the mean µθ (xt; t) can be calculated as a function of θ (xt, t). The variance Σθ (xt, t) of this Gaussian distribution can be fixed to a known constant [25] or learned with a separate neural network head, and both approaches yield high-quality samples when the total number of diffusion steps T is large enough.

According to Ho et al.[1], the simple mean-squared error objective, Lsimple, performs better in real-world applications than the variational lower limit Lvlb that results from viewing the denoising diffusion model as a VAE. They also point out that training for this objective and applying their related sampling method is identical to the denoising score matching model from Song and Ermon [2], who sample from a denoising model trained with several noise levels to obtain high-quality image samples. We often use “diffusion models'' as shorthand to refer to both classes of models. [3]

**Related Work:** Diffusion models are a class of powerful deep generative models[3] which have received a lot of attention recently. In image generation, diffusion models have been highly successful [1] [4] [5]without training instability and mode collapse problems, exceeding GANs in fidelity and diversity [6] [4]. In terms of text-to-image conversion, Autoregressive models[7] , GANs [8] [9], VQ-VAE Transformer-based methods [10] [11]**,** DALL-E2 [10] and diffusion models [12] [4] have made huge progress. Such an example is the concurrent IMAGEN, which consists of a text encoder that converts text to a sequence of embeddings and a cascade of conditional diffusion models that map these embeddings to images of increasing resolution[13]**.** GLIDE [4]also uses cascaded diffusion models for text-to-image conversion, but imagen uses large pretrained frozen language models, which have been found to be essential to both image fidelity and image-text alignment[13]. These models do not provide fine-grained control over a generated image and use text guidance only. In particular, given text descriptions, it is difficult or impossible to preserve a subject's identity consistently across several images, as modifying the context in the prompt also modifies the subject's appearance [14]**.**Their methodology looks for the optimal embedding that can represent the concept; as a result, it is limited by the expressiveness of the textual modality and constrained to the original output domain of the model [14].Personalization has recently gained prominence in a number of areas of machine learning, including recommendation systems[15]**.** There aren't many works that address the issue of novel synthesis of subjects using GANs within the vision and graphics sector. Casanova[16] introduced a method to condition GANs on instances such that variants of an instance can be generated. Generated subjects share features with the conditioning instance, but are not identical and thus cannot solve the problem tackled in our work. Nitzan[17] proposed MyStyle to fine-tune a face synthesis GAN on a specific identity, in order to build a personalized prior. But MyStyle requires around 100 images to learn an adequate prior and it is constrained to the face domain.

# Reference

|  |  |
| --- | --- |
| [1] | A. J. a. P. A. Jonathan Ho, "Denoising diffusion probabilistic models," 2022. |
| [2] | Y. S. a. S. Ermon, "Improved techniques for training score-based generative models," 2020. |
| [3] | E. A. W. N. M. a. S. G. Jascha Sohl-Dickstein, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," 2015. |
| [4] | A. N. a. P. Dhariwal, "Improved denoising diffusion probabilistic models," 2021. |
| [5] | J. W. K. C. H. A. R. G. G. S. A. G. S. A. A. P. M. J. C. G. K. a. I. S. Alec Radford, "Learning Transferable Visual Models From Natural Language Supervision," 2021. |
| [6] | J. D. a. K. S. Andrew Brock, " Large scale gan training for high fidelity natural image synthesis," 2018. |
| [7] | E. P. J. L. B. a. R. S. Elman Mansimov, "Generating Images from Captions with Attention," 2016. |
| [8] | P. Z. Q. H. H. Z. Z. G. X. H. a. X. H. Tao Xu, "AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Network," 2018. |
| [9] | J. Y. K. J. B. H. L. a. Y. Y. Han Zhang, " Cross-Modal Contrastive Learning for Text-to-Image Generation," 2021. |
| [10] | M. P. G. G. S. G. C. V. A. R. M. C. a. I. S. Aditya Ramesh, "Zero-Shot Text-to-Image Generation," 2021. |
| [11] | A. P. O. A. S. S. D. P. a. Y. T. Oran Gafni, "Make-a-scene: Scene-based text-to-image generation with human priors," 2022. |
| [12] | A. B. D. L. P. E. a. B. O. Robin Rombach, "High-Resolution Image Synthesis with Latent Diffusion Models," 2022. |
| [13] | W. C. S. S. L. L. J. W. E. D. S. K. S. G. B. K. A. S. S. M. R. G. L. T. S. J. H. D. J. F. M. N. Chitwan Saharia, "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding," 2022. |
| [14] | Y. L. V. J. Y. P. M. R. a. K. A. Nataniel Ruiz, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," 2022. |
| [15] | F. A. a. T. J. Justin Basilico Ashok Chandrashekar, "Artwork personalization at netflix," 2022. |
| [16] | M. C. J. V. M. D. a. A. R. S. Arantxa Casanova, "Instance-conditioned gan. Advances in Neural Information Processing Systems," 2021. |
| [17] | K. A. Q. H. O. L. M. Y. Y. G. I. M. Y. P. a. D. C.-O. Yotam Nitzan, " Mystyle: A personalized generative prior.," 2022. |