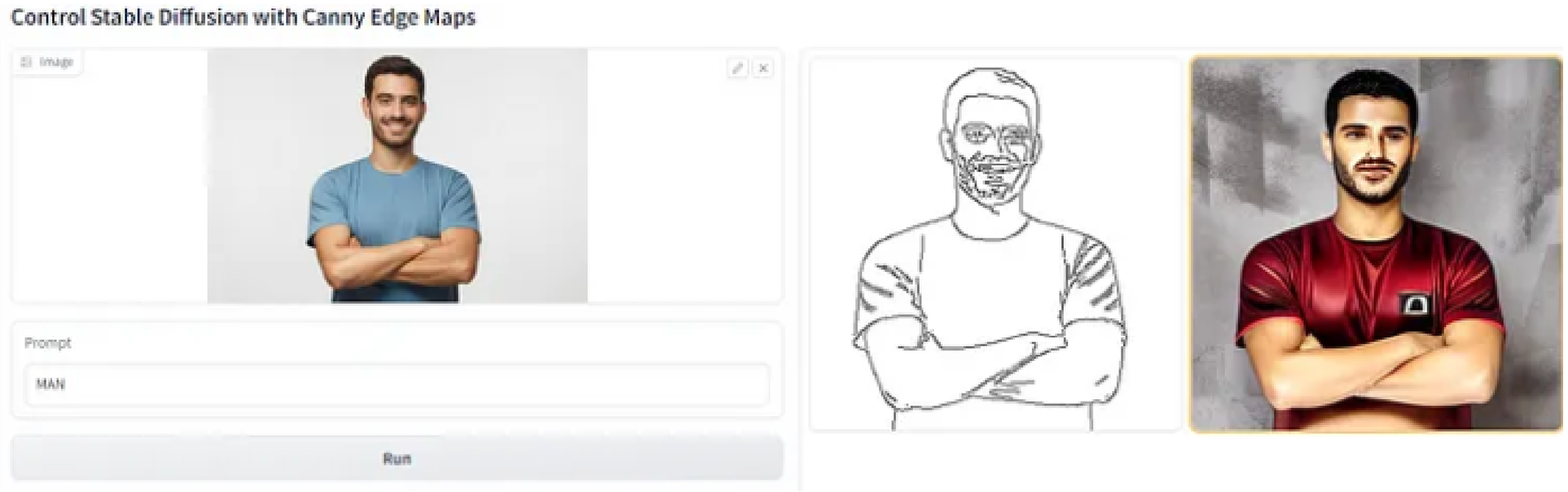
Final Presentation

Stable Diffusion

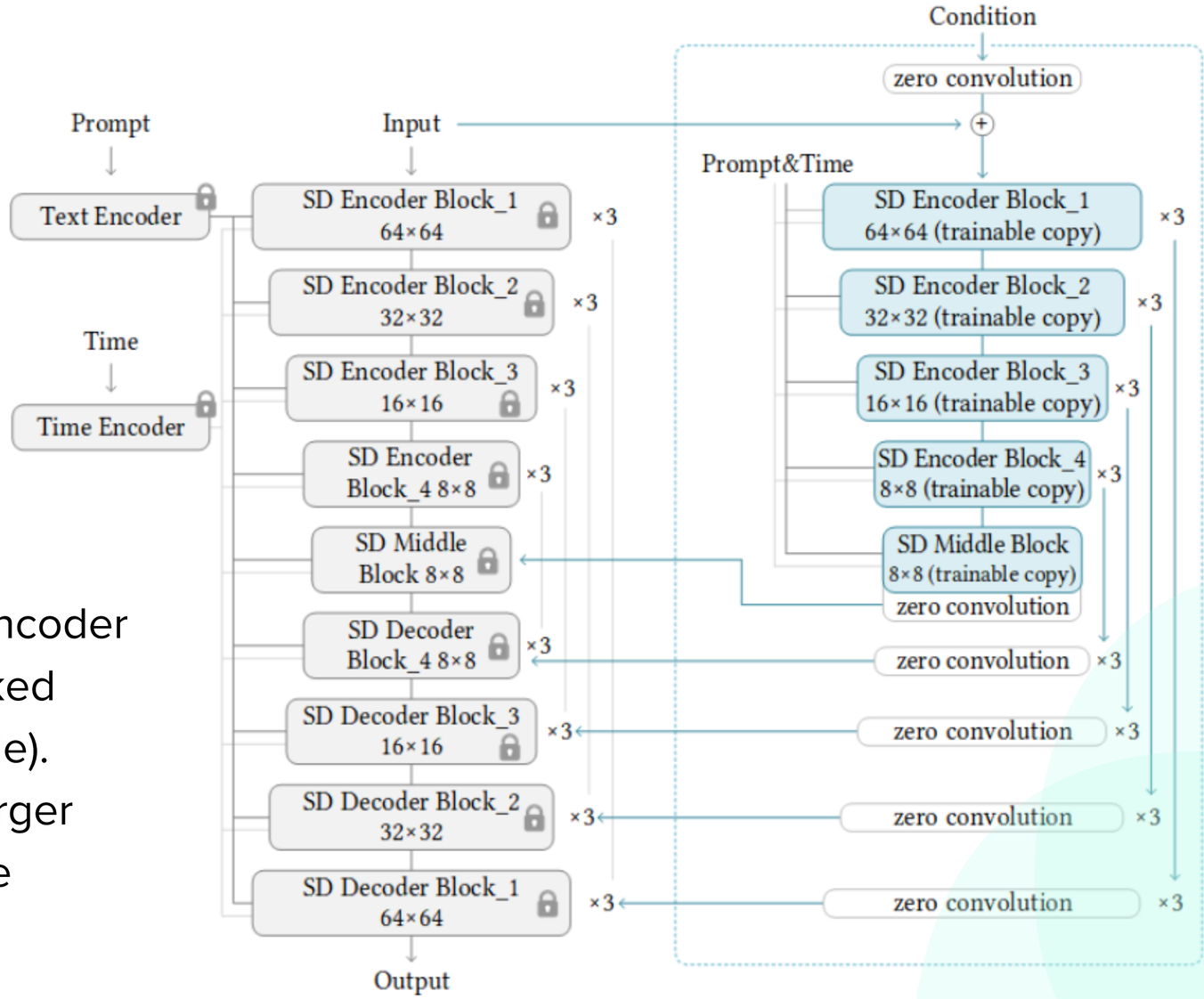
Stable Diffusion is deep learning, text-to-image AI/machine learning model released in 2022. It is primarily used to generate detailed images conditioned on text descriptions, though it can also be applied to other tasks such as inpainting, outpainting, and generating image-to-image translations guided by a text prompt.



ControlNet

ControlNet is a neural network structure to control diffusion models by adding extra conditions.It can be used in combination with Stable Diffusion.

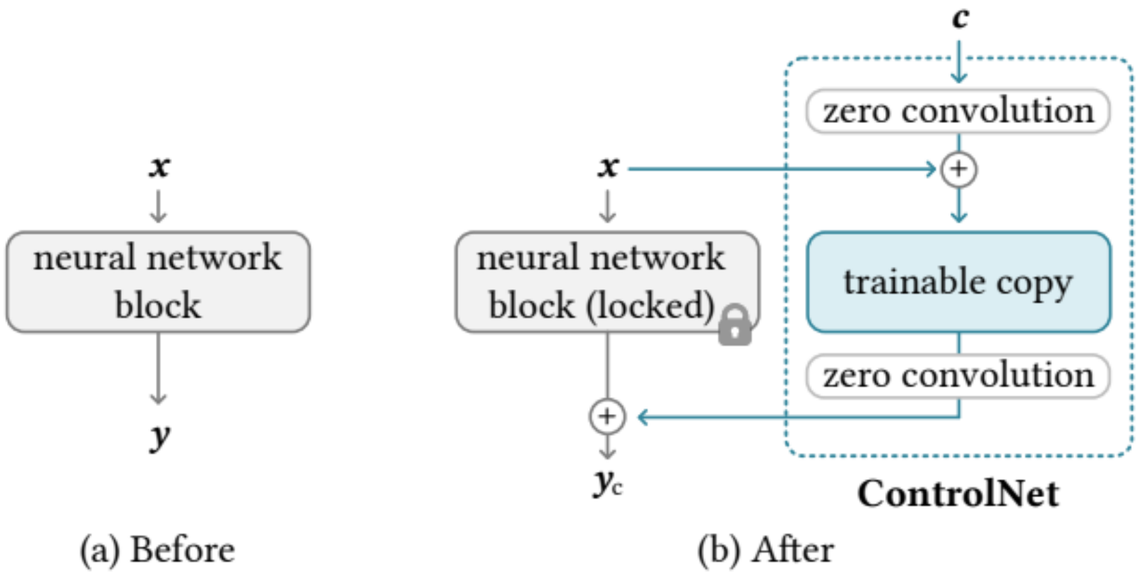
Note that the way we connect layers is computational efficient. The original SD encoder does not need to store gradients (the locked original SD Encoder Block 1234 and Middle). The required GPU memory is not much larger than original SD, although many layers are added.



By repeating the following structure 14 times, we can control Stable Diffusion

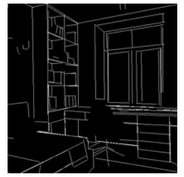
What it does?

It copys the weights of neural network blocks into a "locked" copy and a "trainable" copy.The "trainable" one learns your condition. The "locked" one preserves your model.Training with small dataset of image pairs will not destroy the production-ready diffusion models. The "zero convolution" is 1×1 convolution with both weight and bias initialized as zeros. Before training, all zero convolutions output zeros, and ControlNet will not cause any distortion. No layer is trained from scratch. You are still fine-tuning. Your original model is safe.

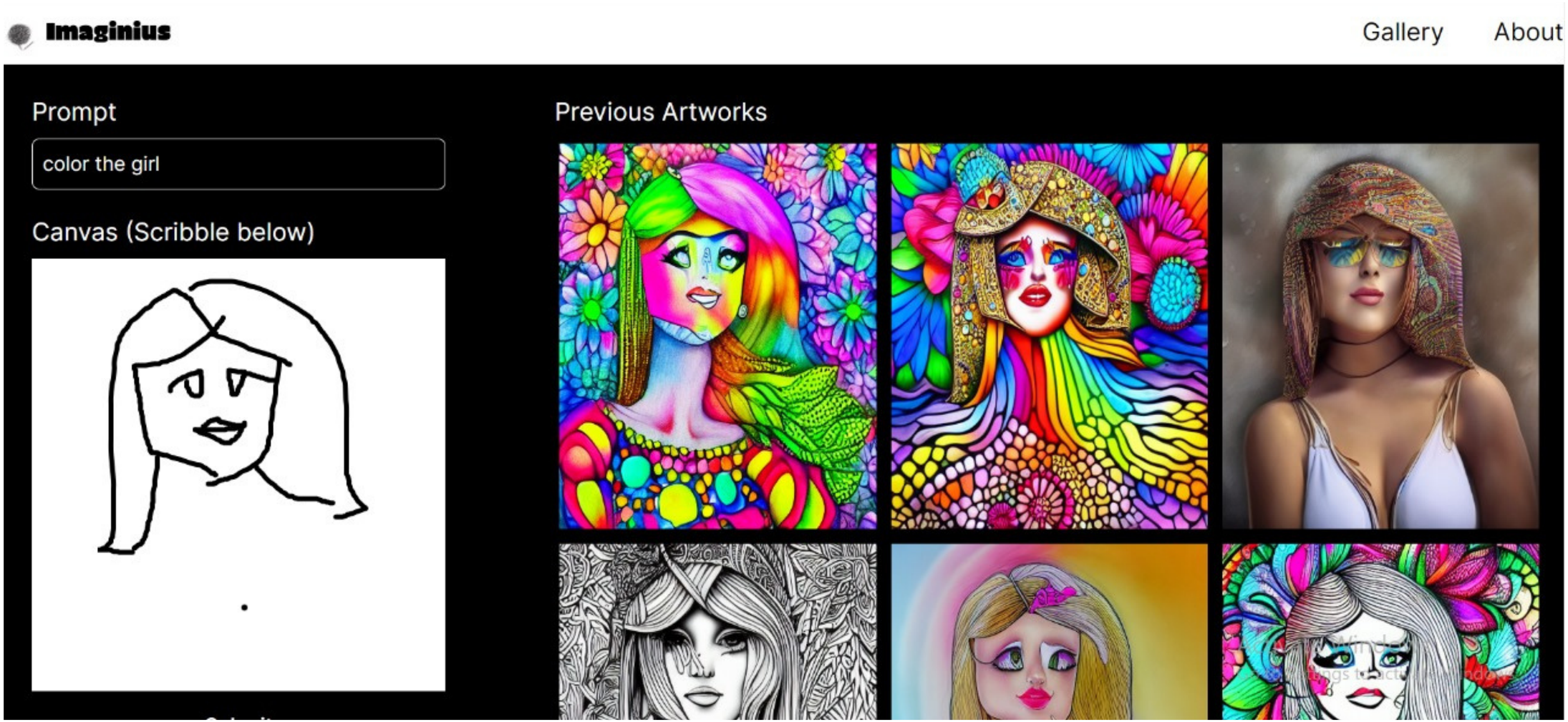


Models in ControlNet Stable Diffusion

MODEL NAME	CONTROL IMAGE OVERVIEW	CONTROL IMAGE EXAMPLE	GENERATED IMAGE EXAMPLE
Trained with canny edge detection	A monochrome image with white edges on a black background.		
Trained with Midas depth estimation	A grayscale image with black representing deep areas and white representing shallow areas.		
Trained with HED edge detection (soft edge)	A monochrome image with white soft edges on a black background.		

MODEL NAME	CONTROL IMAGE OVERVIEW	CONTROL IMAGE EXAMPLE	GENERATED IMAGE EXAMPLE
Trained with M-LSD line detection	A monochrome image composed only of white straight lines on a black background		
Trained with normal map	A normal mapped image.		
Trained with OpenPose bone image	A OpenPose bone image.		
Trained with human scribbles	A hand-drawn monochrome image with white outlines on a black background.		
Trained with semantic segmentation	An ADE20K's segmentation protocol image		

We have used **ControlNet + Stable Diffusion** with **Scribble**



Different Models working for same prompt

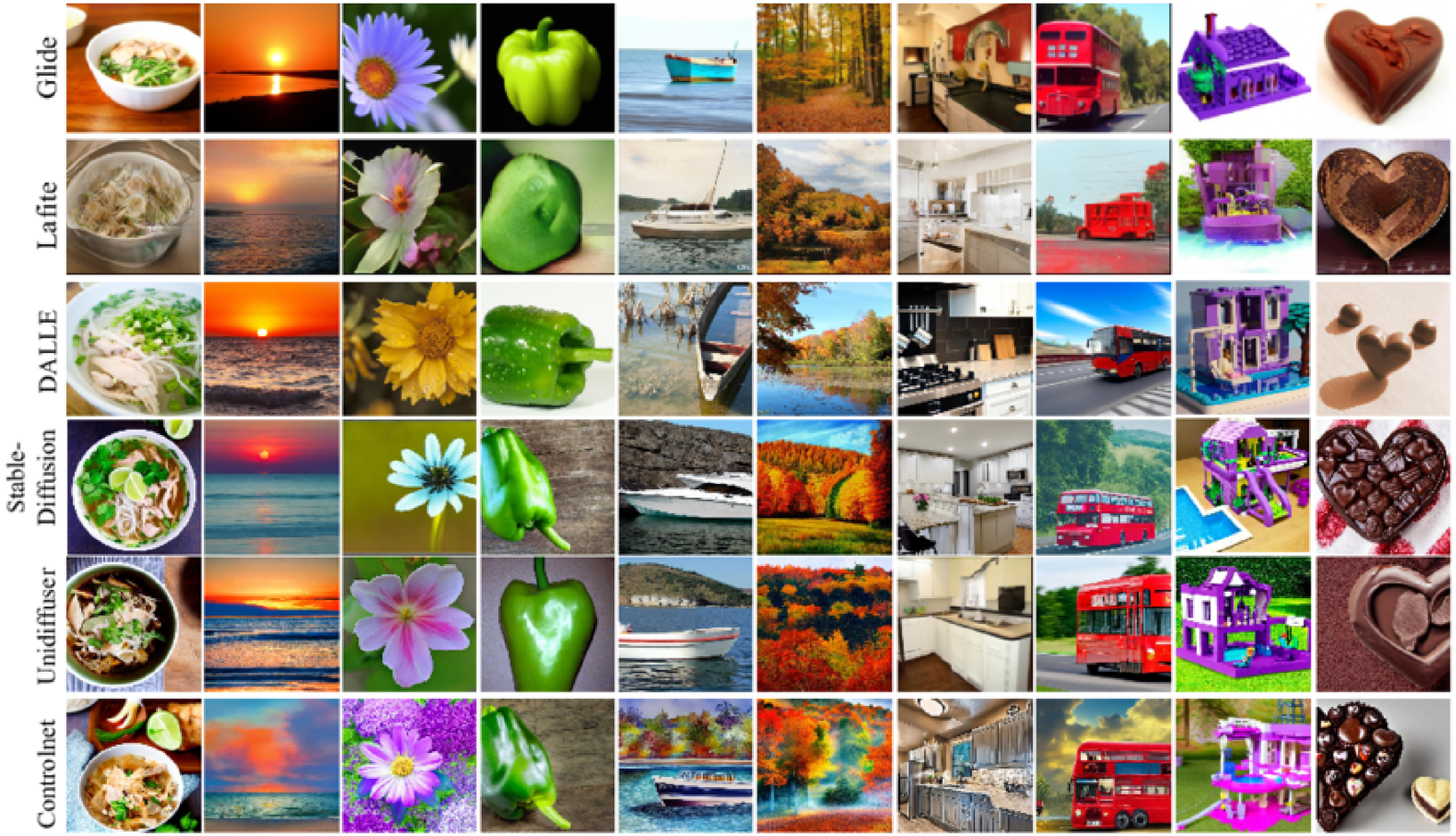


Illustration of the images from the perspectives of quality, authenticity, and text-image correspondence. (a) 10 high quality examples and 10 low quality examples of the images generated by the prompt of "a corgi". (b) 10 high authenticity and 10 low authenticity examples of images generated by the prompt of "a girl". (c) 10 high text-image correspondence and 10 low correspondence examples of images generated by the prompt of "a grandmother reading a book to her grandson and granddaughter"

AI-based image generation has been appliedto various fields. However, AI Generated Images (AIGIs) may have someunique distortions compared to natural images, thus many generatedimages are not qualified for real-world applications. Consequently, it isimportant and significant to study subjective and objective Image QualityAssessment (IQA) methodologies for AIGIs.

Evaluation

Inception Score (IS): The Inception Score measures the quality and diversity of generated images. It uses a pre-trained Inception model to compute a score that balances image quality and diversity. A higher IS generally indicates better results.

Frechet Inception Distance (FID) : FID measures the similarity between the distribution of real images and generated images in feature space, using the Inception model's activations. Lower FID scores are indicative of better quality and diversity.

Precision and Recall: Precision and recall metrics can assess the relevance of generated images to the input text. Precision measures how many generated images are relevant, while recall measures how many relevant images are generated.

LSUN-Churches 256 × 256			
Method	FID ↓	Prec. ↑	Recall ↑
DDPM [30]	7.89	-	-
ImageBART [21]	7.32	-	-
PGGAN [39]	6.42	-	-
StyleGAN [41]	4.21	-	-
StyleGAN2 [42]	3.86	-	-
ProjectedGAN [76]	**1.59**	0.61	0.44
*LDM-8** (ours, 200-s)	4.02	**0.64**	**0.52**

Model Performance

To assess the quality of images created by generative models, it is common to use the Fréchet inception distance (FID) metric. In a nutshell, FID calculates the distance between the feature vectors of real images and generated images. On the COCO benchmark, Imagen currently achieved the best (lowest) zero-shot FID score of 7.27, outperforming DALL·E 2 with a 10.39 FID score.

What is the inception score?

The inception score (IS) is a mathematical algorithm used to measure or determine the quality of images created by generative AI through a generative adversarial network (GAN). The word "inception" refers to the spark of creativity or initial beginning of a thought or action traditionally experienced by humans. The score produced by the IS algorithm can range from zero (worst) to infinity (best).

The inception score algorithm measures two factors:

Quality : How good the generated image is. Generated images should be believable or realistic as if a real person painted a picture or took a photograph. For example, if the AI produces images of cats, each image should include a clearly identifiable cat. If the object is not clearly identifiable as a cat, the corresponding IS will be low.

Diversity : How diverse the generated image is. Generated images should have high randomness (entropy), meaning that the generative AI should produce highly varied images. Diversity. How diverse the generated image is. Generated images should have high randomness (entropy), meaning that the generative AI should produce highly varied images.

Generative AI developers use the inception score as a measure of image quality

Inception score vs. Fréchet inception distance

Another metric used to evaluate the quality of AI-generated images is the Fréchet inception distance. FID was introduced in 2017 and has generally superseded inception score as the preferred measure of generative image model performance.

The principal difference between IS and FID is the comparative use and evaluation of real images, referred to as "ground truth." This allows FID to analyze real images alongside computer-generated images in a bid to better simulate human perception. By comparison, IS only evaluates computer-generated images.

Although FID has generally edged out IS as the preferred quality metric for GANs, FID has also been shown to demonstrate some statistical bias, and does not always accurately reflect human perception.

Thank You