Diffusion models

Slides from Lana Lazebnik

Outline

- Conditional diffusion models
- Large-scale models
- Controlling and fine-tuning image generation
- Societal, ethical, and legal issues

Outline

- Conditional diffusion models

Class-conditioned DDPMs

- “We can sample with as few as 25 forward passes while maintaining FIDs comparable to BigGAN”

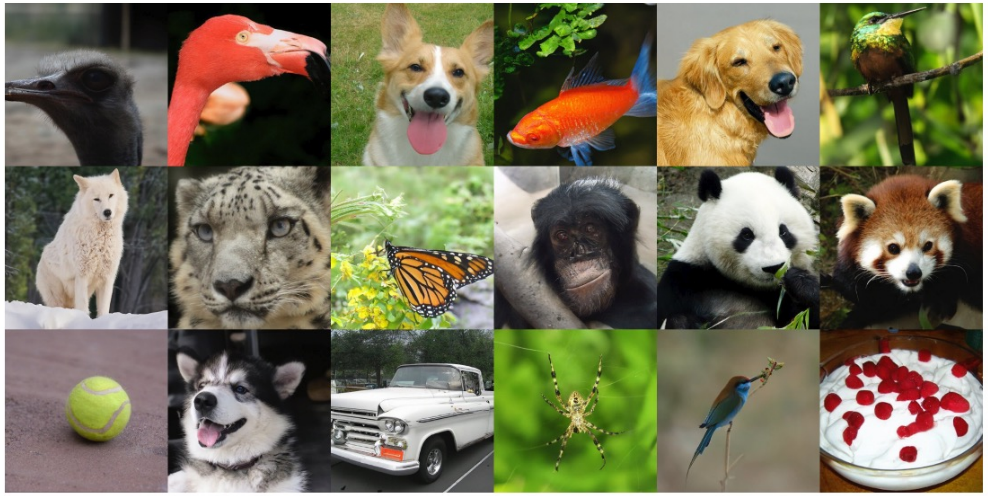


Figure 1: Selected samples from our best ImageNet 512×512 model (FID 3.85)

Abstract

We show that diffusion models can achieve image sample quality superior to the current state-of-the-art generative models. We achieve this on unconditional image synthesis by finding a better architecture through a series of ablations. For conditional image synthesis, we further improve sample quality with classifier guidance: a simple, compute-efficient method for trading off diversity for fidelity using gradients from a classifier. We achieve an FID of 2.97 on ImageNet 128×128, 4.59 on ImageNet 256×256, and 7.72 on ImageNet 512×512, and we match BigGAN-deep even with as few as 25 forward passes per sample, all while maintaining better coverage of the distribution. Finally, we find that classifier guidance combines well with upsampling diffusion models, further improving FID to 3.94 on ImageNet 256×256 and 3.85 on ImageNet 512×512. We release our code at `https://github.com/openai/guided-diffusion`.

P. Dhariwal and A. Nichol. [Diffusion Models Beat GANs on Image Synthesis](). NeurIPS 2021

Classifier guidance

- We can sample from the class-conditional density $p(x_t|c)$ with the help of a pre-trained classifier $p(c|x_t)$
- Bayes rule:

$$p(x_t|c) \propto p(c|x_t)p(x_t)$$

$$\log p(x_t|c) = \log p(c|x_t) + \log p(x_t) + \text{const.}$$

$$\nabla_{x_t} \log p(x_t|c) \ = \ \nabla_{x_t} \log p(c|x_t) \ + \ \nabla_{x_t} \log p(x_t)$$

conditional score function | obtained from classifier output | unconditional *score function* (pre-trained)

- To sample from class c, steer sample in the modified direction $\nabla_{x_t}[\log p(x_t) + w \log p(c|x_t)]$

Classifier-free guidance

- Instead of training an additional classifier, get an "implicit classifier" by jointly training a conditional and unconditional diffusion model: $p(c|x_t) \propto p(x_t|c)/p(x_t)$
- Both $p(x_t|c)$ and $p(x_t)$ are represented using the same network, trained by dropping out c with some probability (corresponding to the unconditional case)
- The modified score function corresponding to this implicit classifier is

$$\nabla_{x_t}[\log p(x_t) + w \log p(c|x_t)]$$
$$= \nabla_{x_t}[\log p(x_t) + w(\log p(x_t|c) - \log p(x_t))]$$
$$= \nabla_{x_t}[(1 - \mathrm{w})\log p(x_t) + w \log p(x_t|c)]$$

Sample is steered away from the unconditional distribution in the direction of the conditional one

J. Ho and T. Salimans. Classifier-Free Diffusion Guidance. arXIv 2021

Classifier-free guidance

Figure 1: Classifier-free guidance on the malamute class for a 64x64 ImageNet diffusion model. Left to right: increasing amounts of classifier-free guidance, starting from non-guided samples on the left.

J. Ho and T. Salimans. Classifier-Free Diffusion Guidance. arXIv 2021

Text-guided diffusion

- Instead of a class label, c can be an encoded text prompt, injected into the U-Net using *cross-attention*

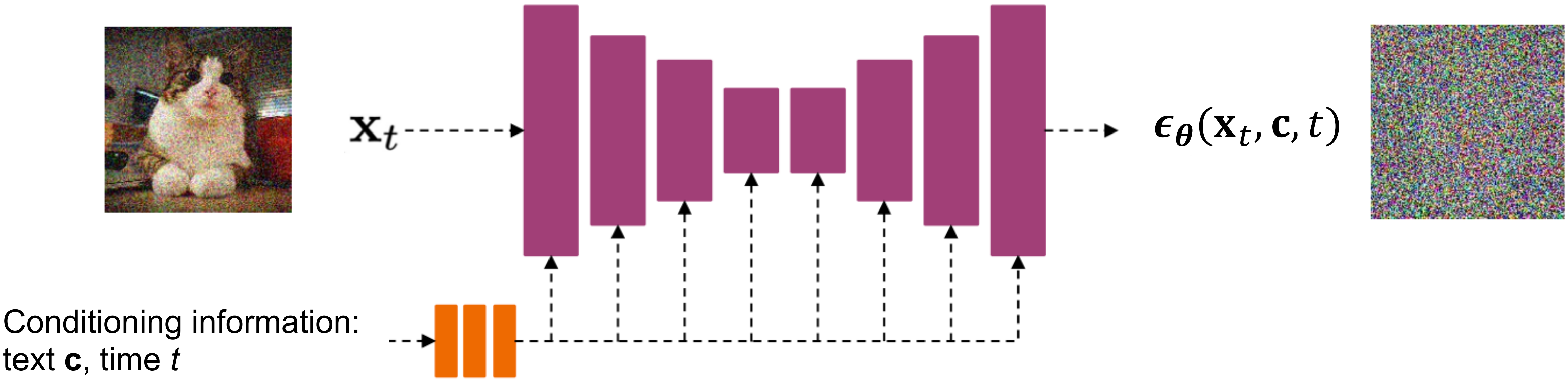


Image source

Text-guided diffusion

- Instead of a class label, c can be an encoded text prompt, injected into the U-Net using *cross-attention*

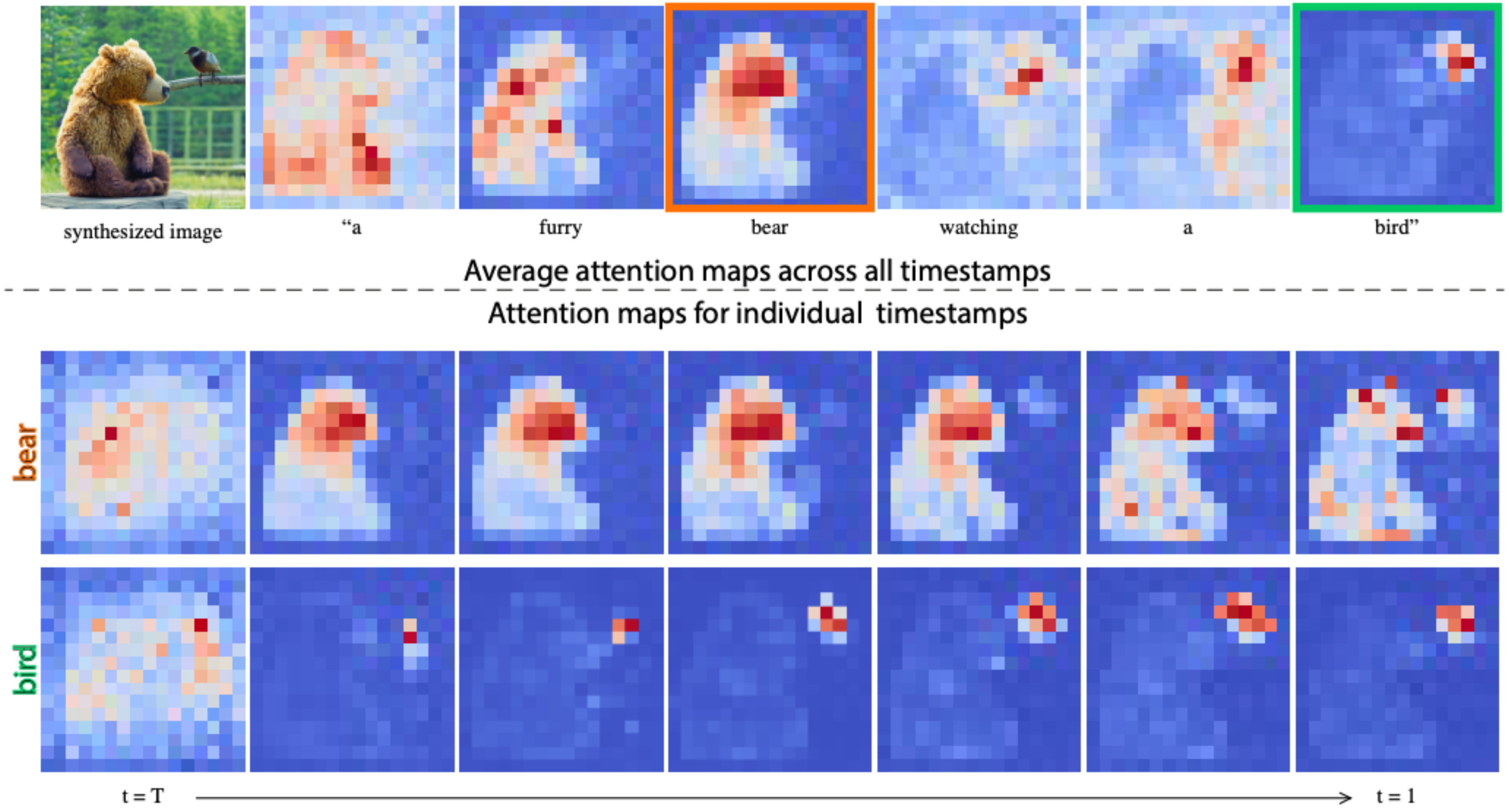


Image source

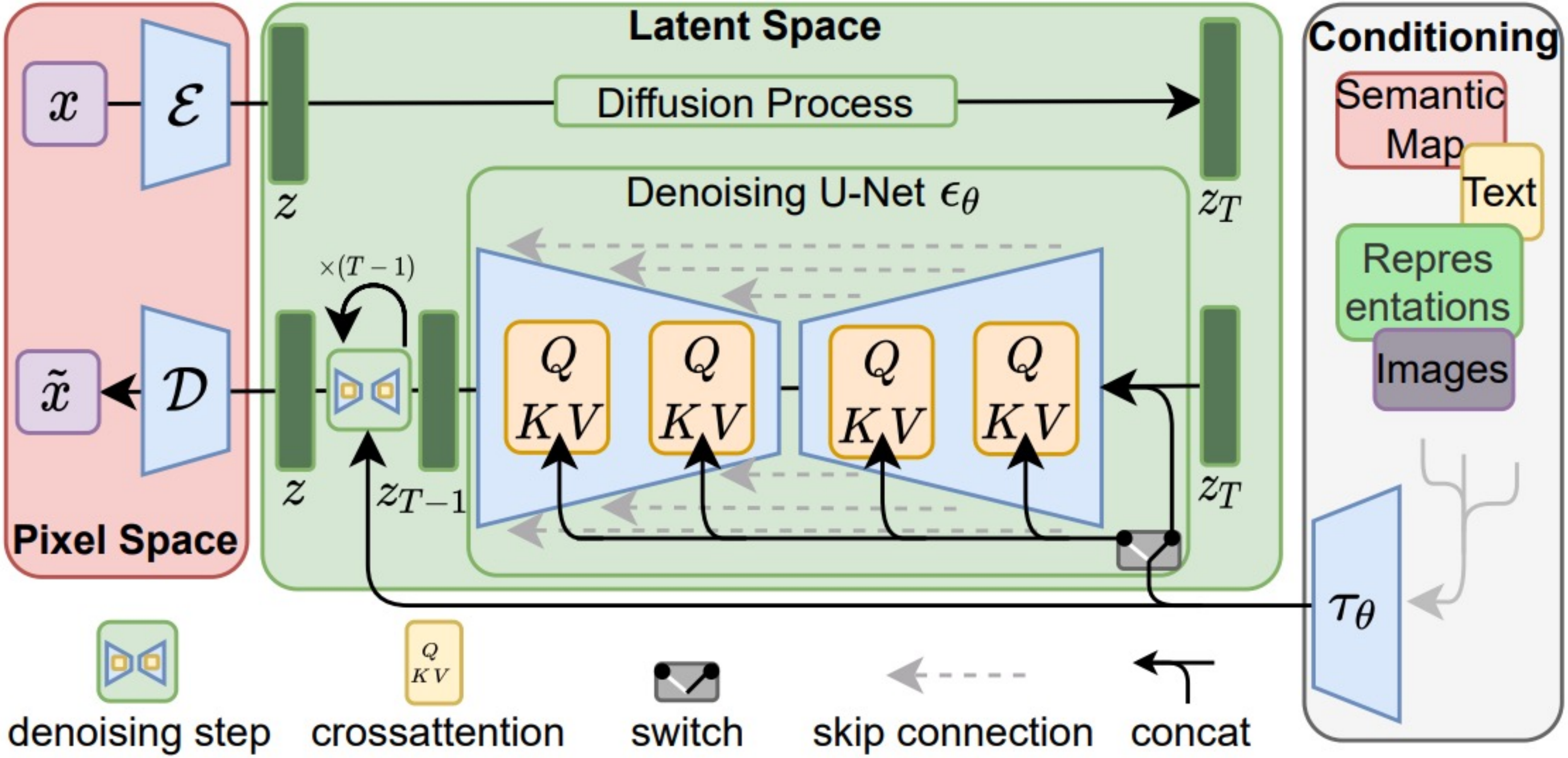
Text-guided diffusion

- Instead of a class label, c can be an encoded text prompt, injected into the U-Net using *cross-attention*
- Classifier-free guidance works the same way as before, by training both conditional and unconditional models using text dropout
- CLIP guidance: steer samples in the direction of $\nabla_{x_t} \mathrm{CLIP}(x_t, c)$
- Note: both classifier and CLIP must be *noise-aware* (trained on noised images)

Outline

- Conditional diffusion models
- Large-scale models

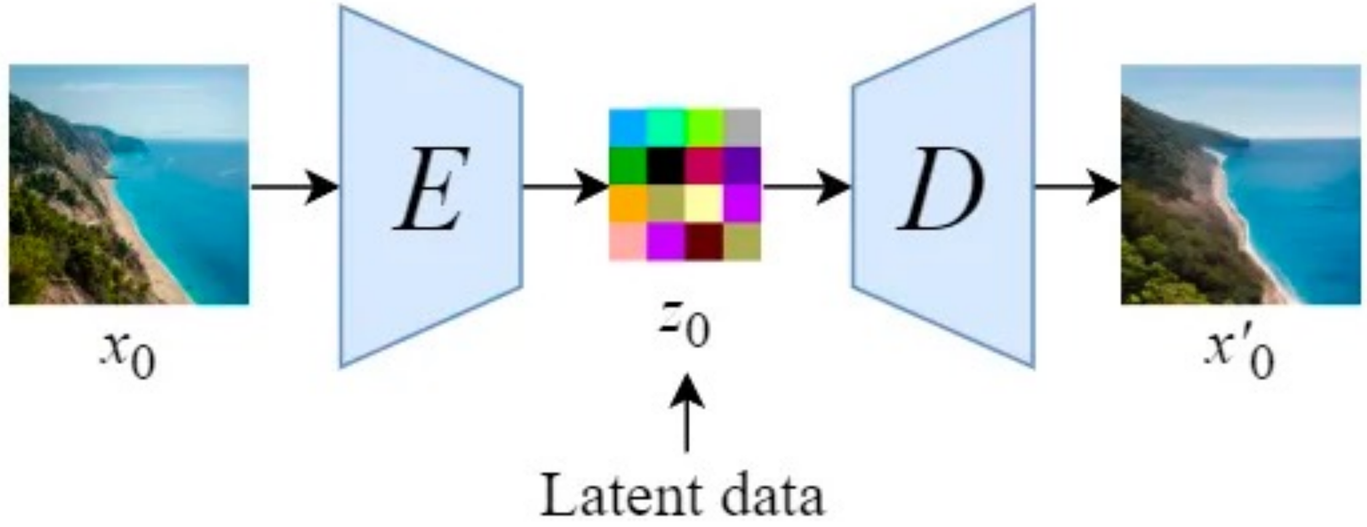
Latent diffusion model (basis of Stable Diffusion)



R. Rombach et al. [High-Resolution Image Synthesis with Latent Diffusion Models](). CVPR 2022

Latent diffusion model (basis of Stable Diffusion)

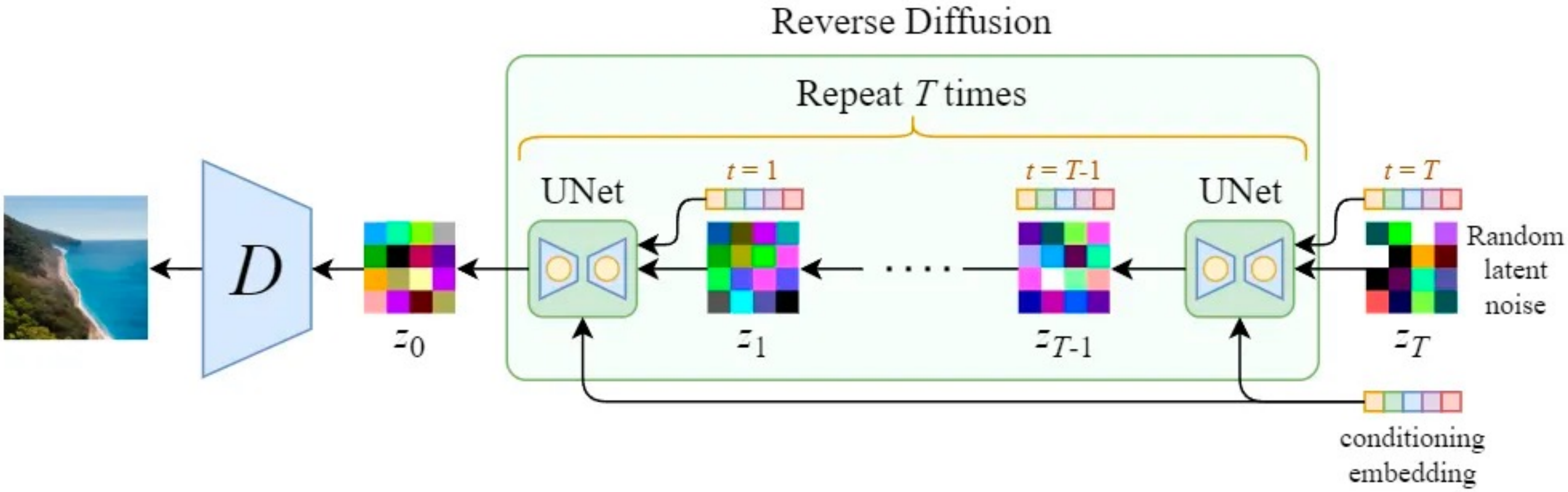
- Key idea: train a separate *encoder* and *decoder* to convert images to and from a lower-dimensional latent space, run conditional diffusion model in latent space



https://medium.com/@steinsfu/stable-diffusion-clearly-explained-ed008044e07e

Latent diffusion model (basis of Stable Diffusion)

- Key idea: train a separate *encoder* and *decoder* to convert images to and from a lower-dimensional latent space, run conditional diffusion model in latent space

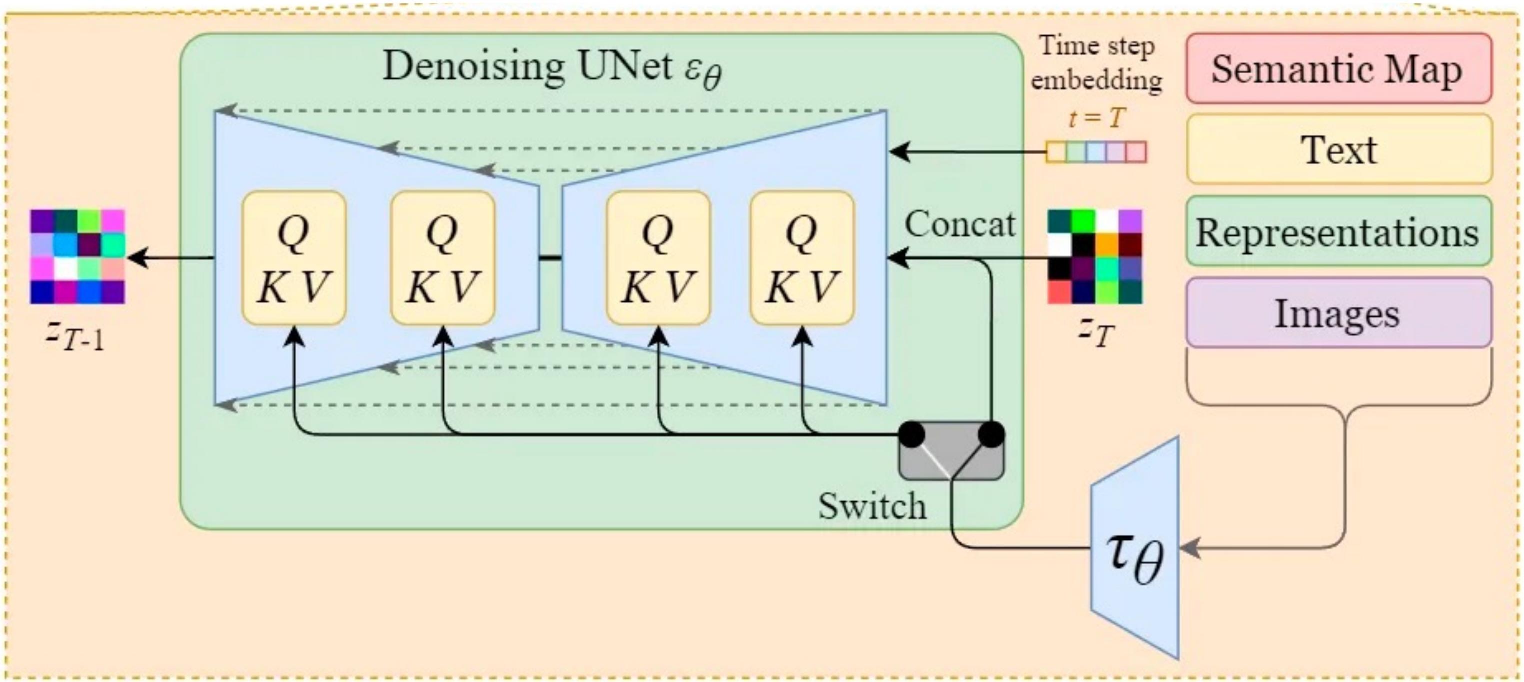


https://medium.com/@steinsfu/stable-diffusion-clearly-explained-ed008044e07e

Latent diffusion model (basis of Stable Diffusion)

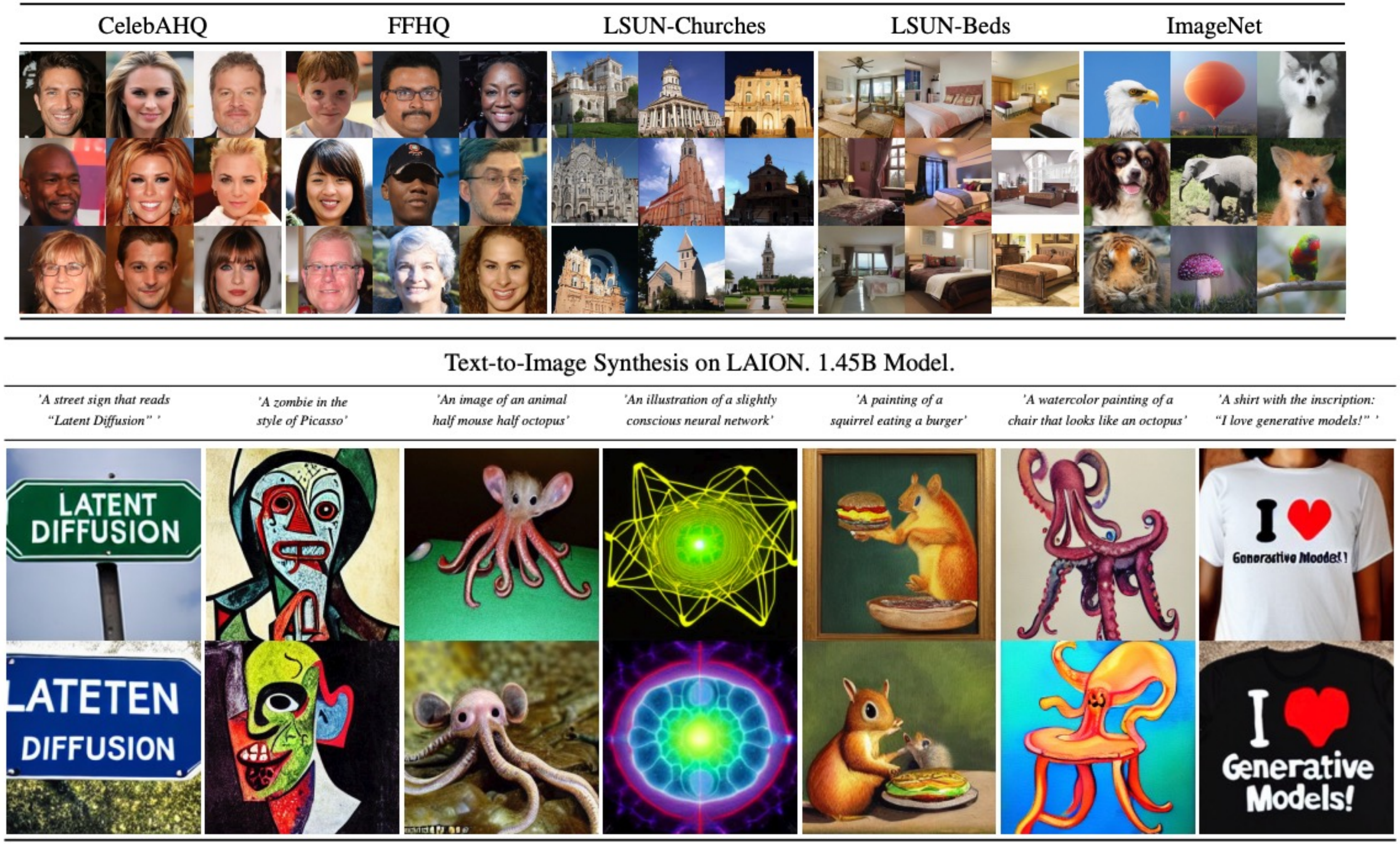
- Key idea: train a separate *encoder* and *decoder* to convert images to and from a lower-dimensional latent space, run conditional diffusion model in latent space

Close-up of U-Net: Conditioning information incorporated using cross-attention



https://medium.com/@steinsfu/stable-diffusion-clearly-explained-ed008044e07e

Latent diffusion model (basis of Stable Diffusion)



Google Imagen (not public)

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Teddy bears swimming at the Olympics 400m Butter-fly event.

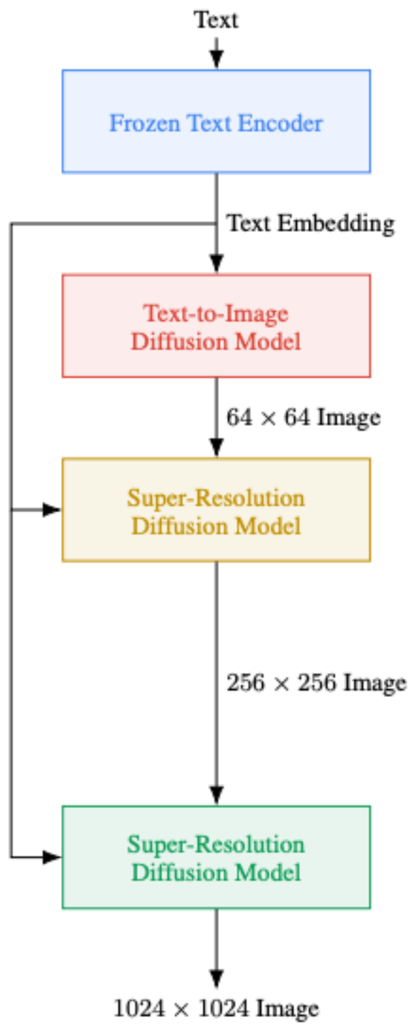
A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

C. Saharia et al. [Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding](). NeurIPS 2022

Google Imagen: Details

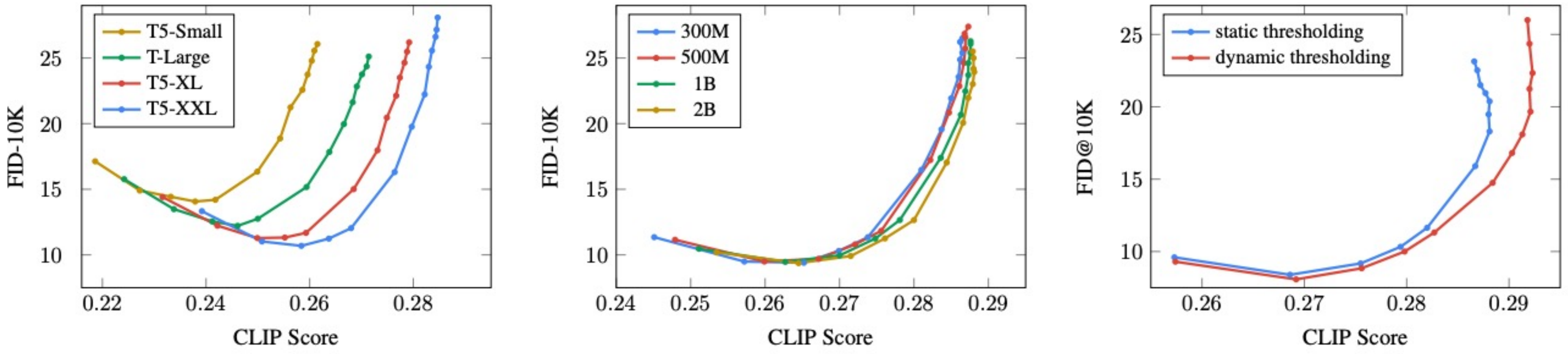
- Text encoder is a large language model (4.6B parameters) trained on text only
- Diffusion model to generate at 64x64, upsample to 256x256, then 1024x1024
  - Architecture: *efficient U-Net* (2B parameters): more parameters at lower resolutions, convolutions *after* downsampling and *before* upsampling
  - Classifier-free guidance with a *dynamic thresholding* technique, enabling good generation quality with high guidance weights
  - Training dataset: 460M image-text pairs (internally collected), 400M pairs from the [LAION dataset]



C. Saharia et al. [Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding]. NeurIPS 2022

Google Imagen: Evaluation

- Impact of model size, implementation choices



(a) Impact of encoder size. (b) Impact of U-Net size. (c) Impact of thresholding.

Curves are obtained by varying guidance weight

FID evaluated on COCO dataset by sampling prompts and generating images using the same prompts

Google Imagen: Evaluation

- Human evaluation on DrawBench (set of 200 prompts)

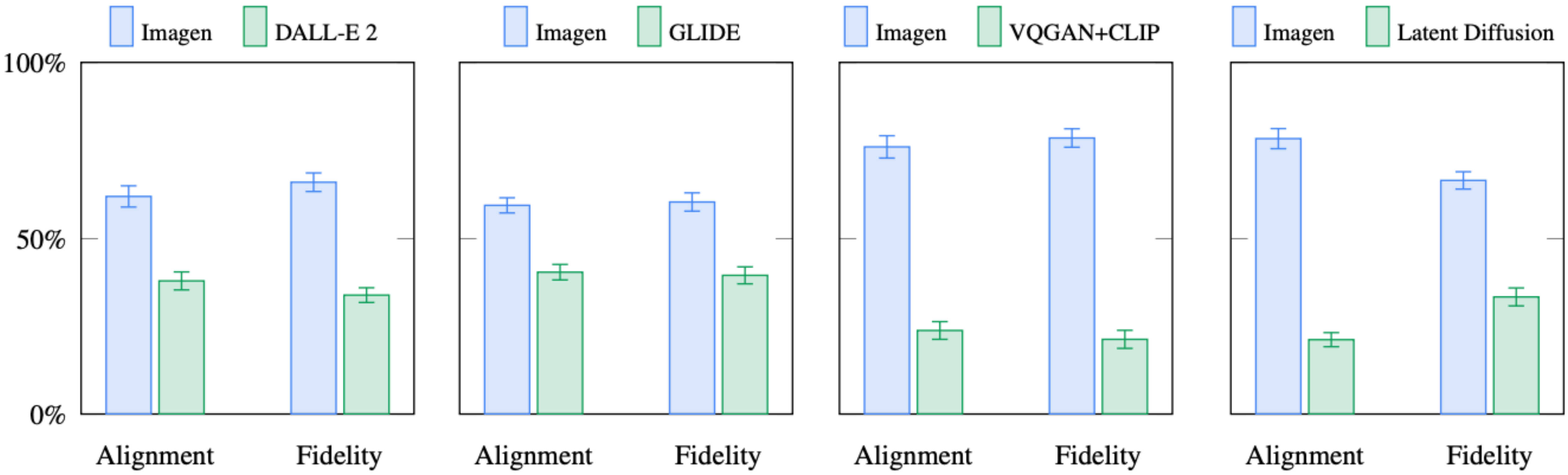
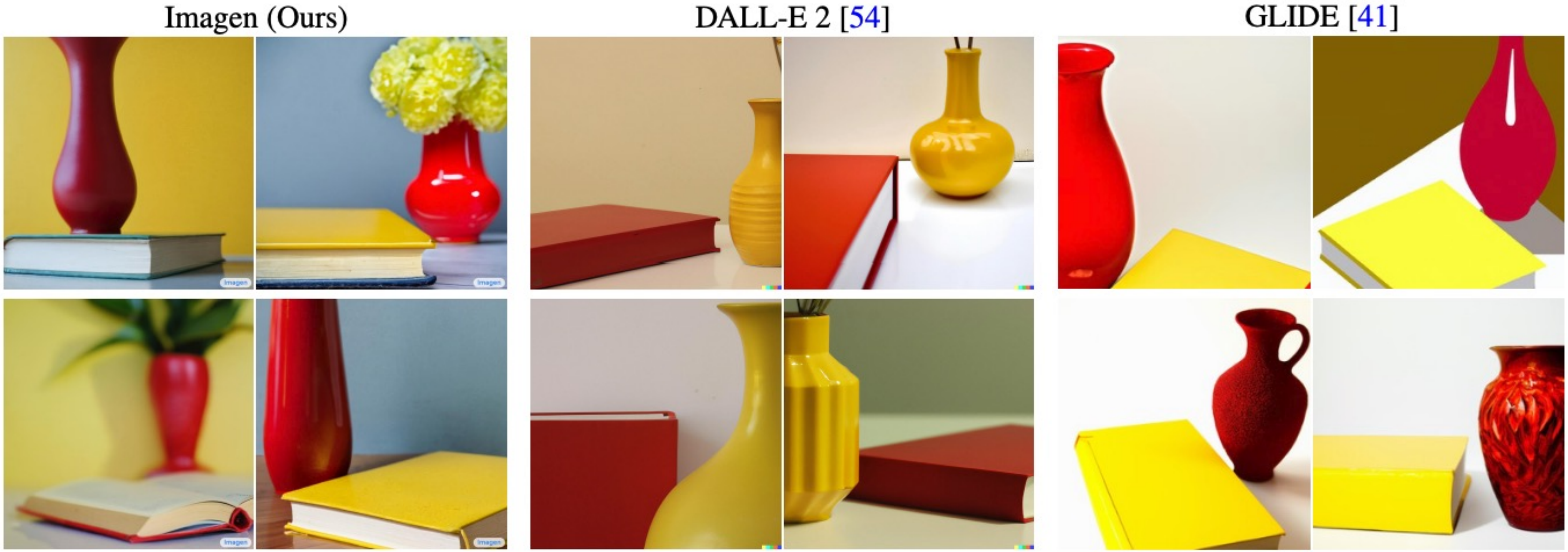


Imagen vs. DALL-E 2 vs. GLIDE



“A yellow book and a red vase”

Imagen vs. DALL-E 2 vs. GLIDE

"A black apple and a green backpack"

"We observe that GLIDE is better than DALL-E 2 in assigning the colors to the objects."

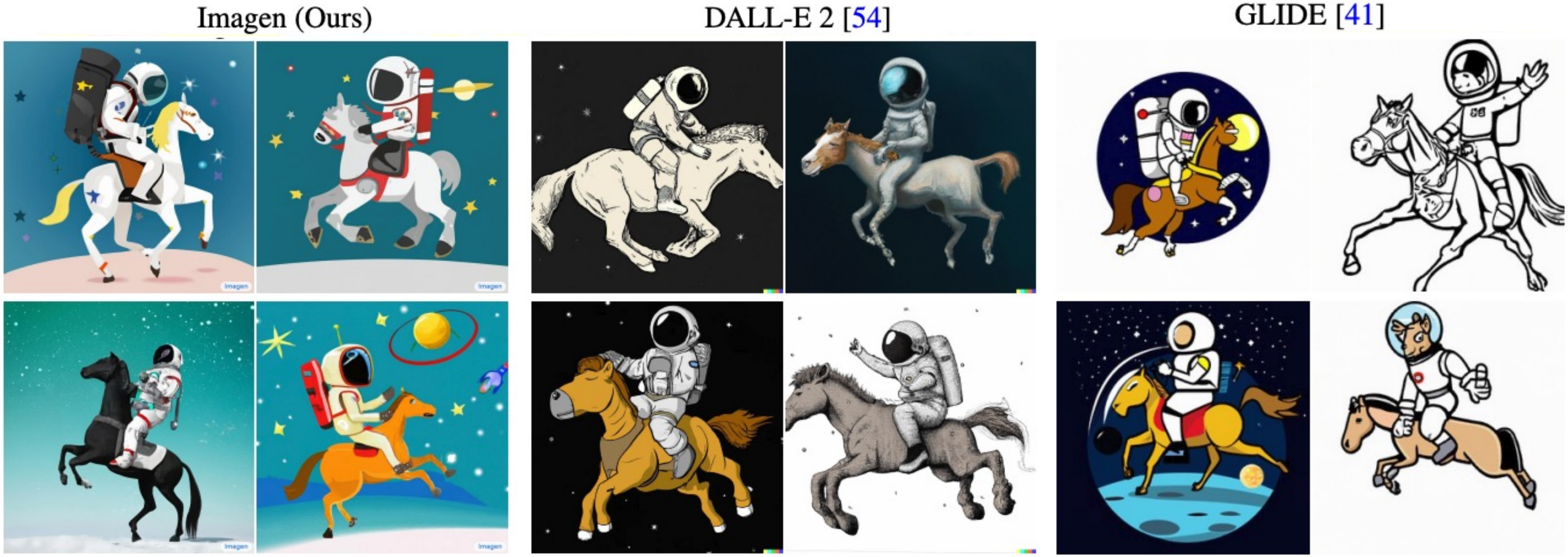
Imagen vs. DALL-E 2 vs. GLIDE

“A storefront with Text to Image written on it”

Imagen vs. DALL-E 2 vs. GLIDE

“A panda making latte art”

Imagen vs. DALL-E 2 vs. GLIDE



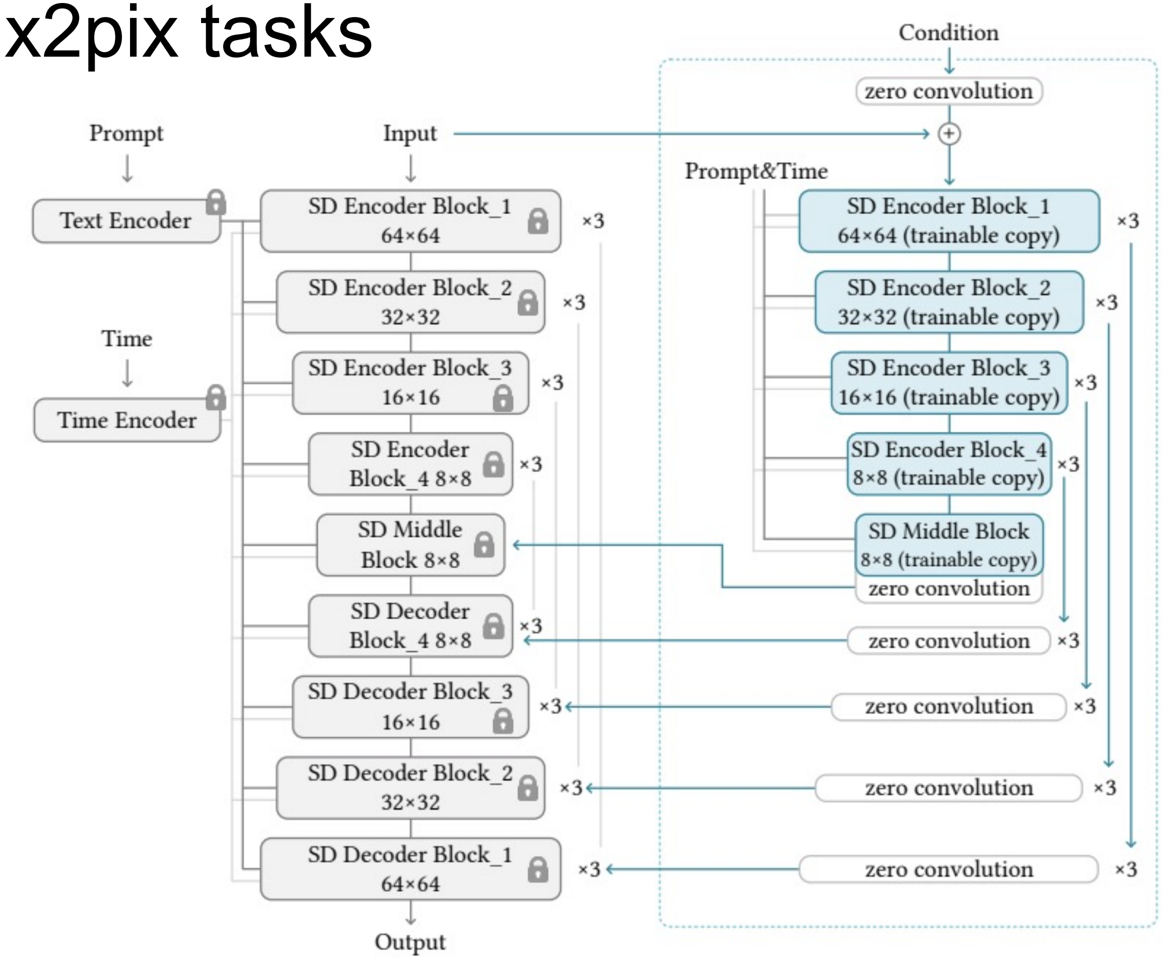
"A horse riding an astronaut"

Outline

- Conditional diffusion models
- Large-scale models
- Controlling and fine-tuning image generation

ControlNet

- Add a trainable “wrapper” around a pre-trained DM to fine-tune it for pix2pix tasks



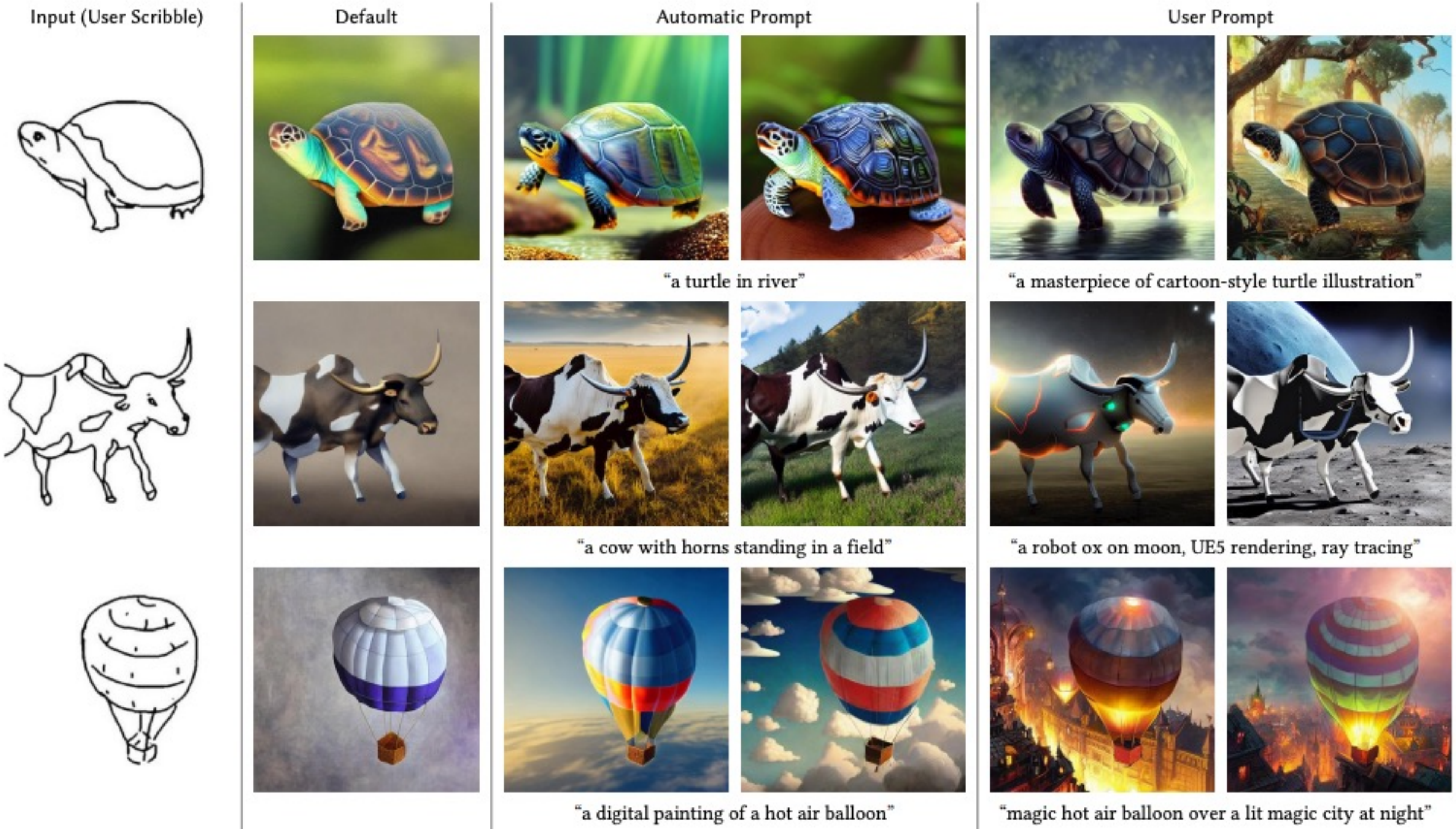
(a) Stable Diffusion (b) ControlNet

L. Zhang and M. Agrawala. [Adding Conditional Control to Text-to-Image Diffusion Models](). arXiv 2023

ControlNet

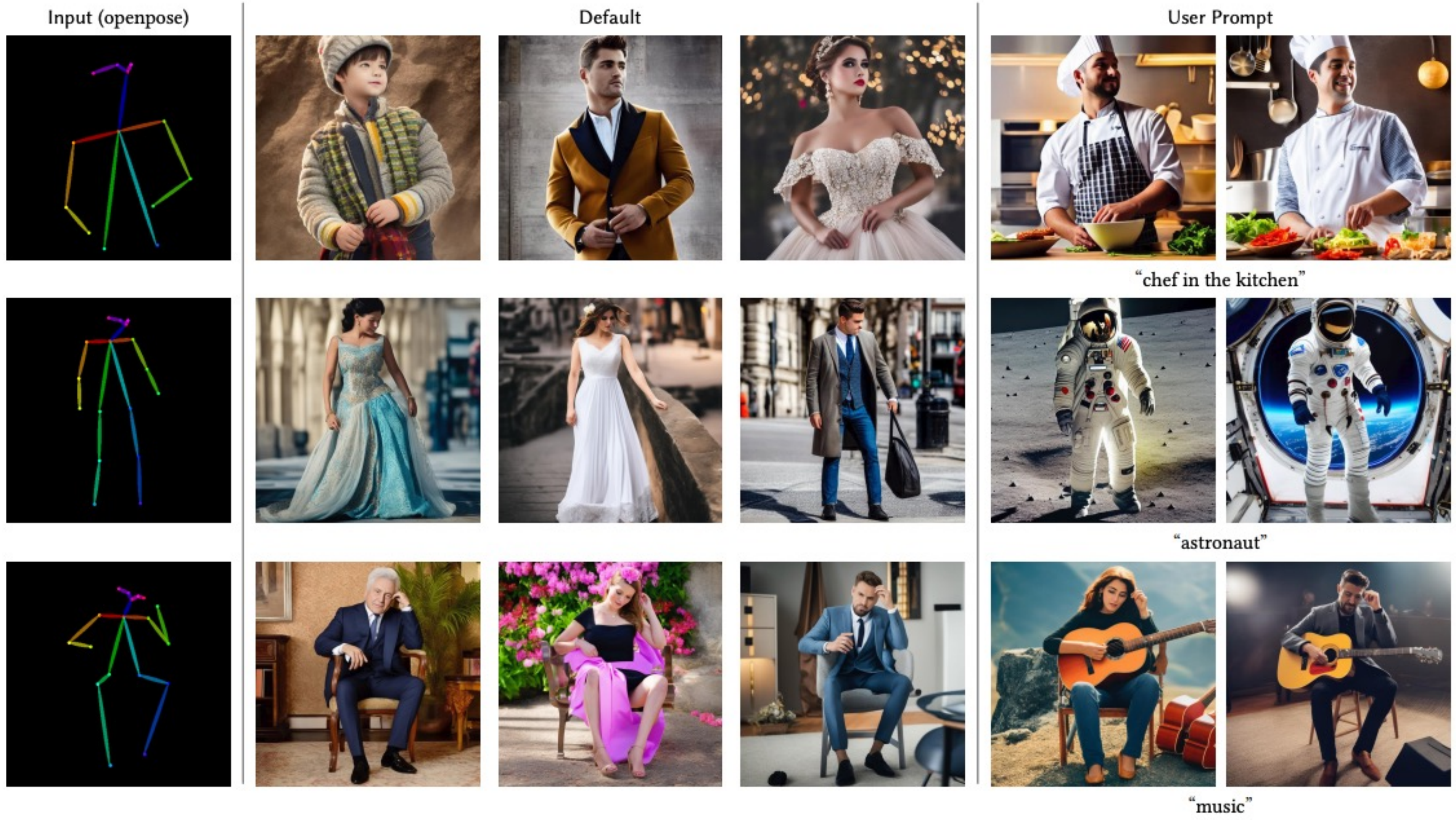
L. Zhang and M. Agrawala. [Adding Conditional Control to Text-to-Image Diffusion Models](). arXiv 2023

ControlNet



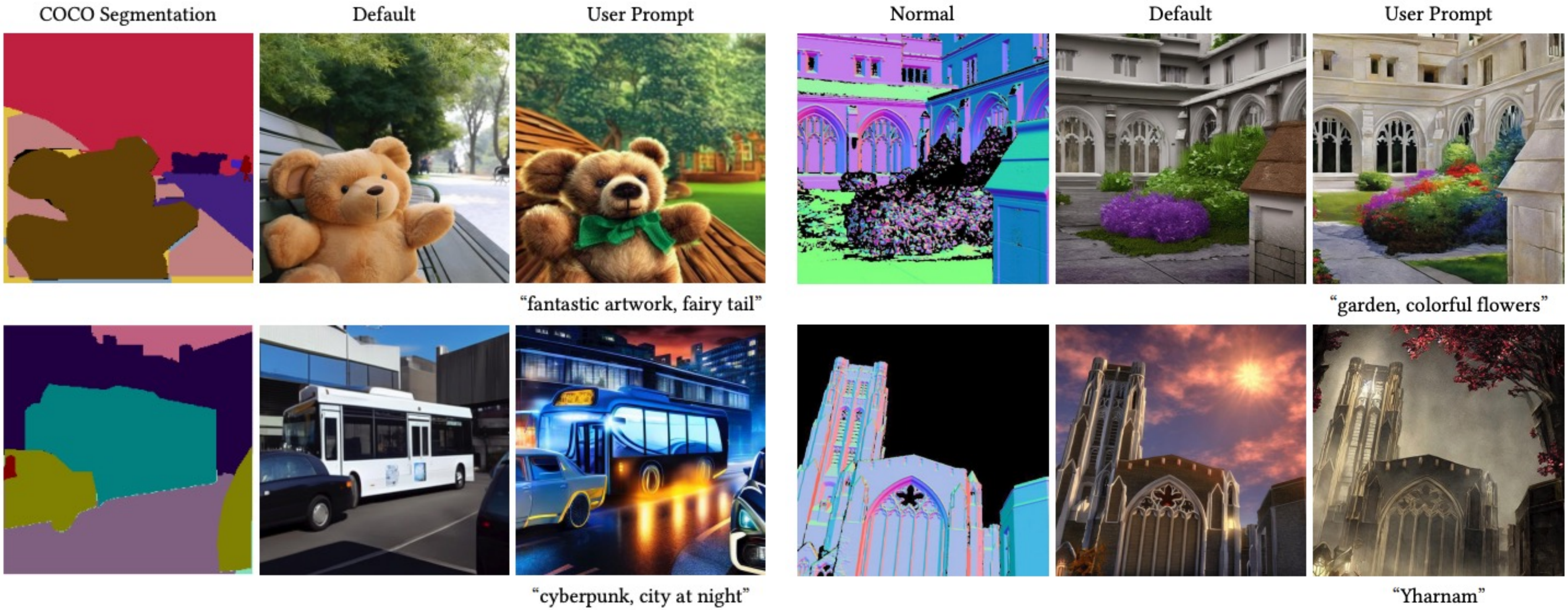
L. Zhang and M. Agrawala. [Adding Conditional Control to Text-to-Image Diffusion Models](). arXiv 2023

ControlNet



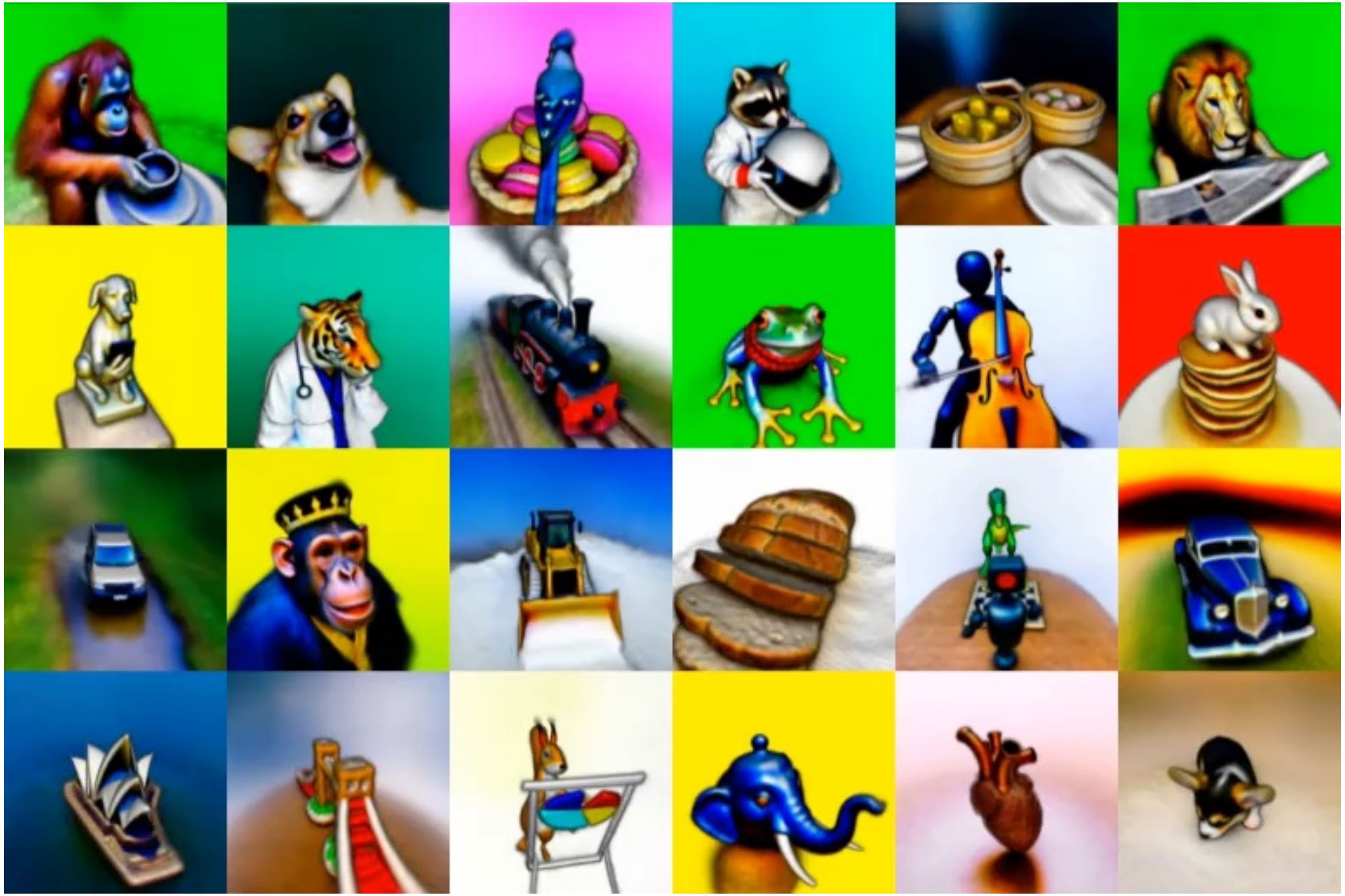
L. Zhang and M. Agrawala. [Adding Conditional Control to Text-to-Image Diffusion Models](). arXiv 2023

ControlNet



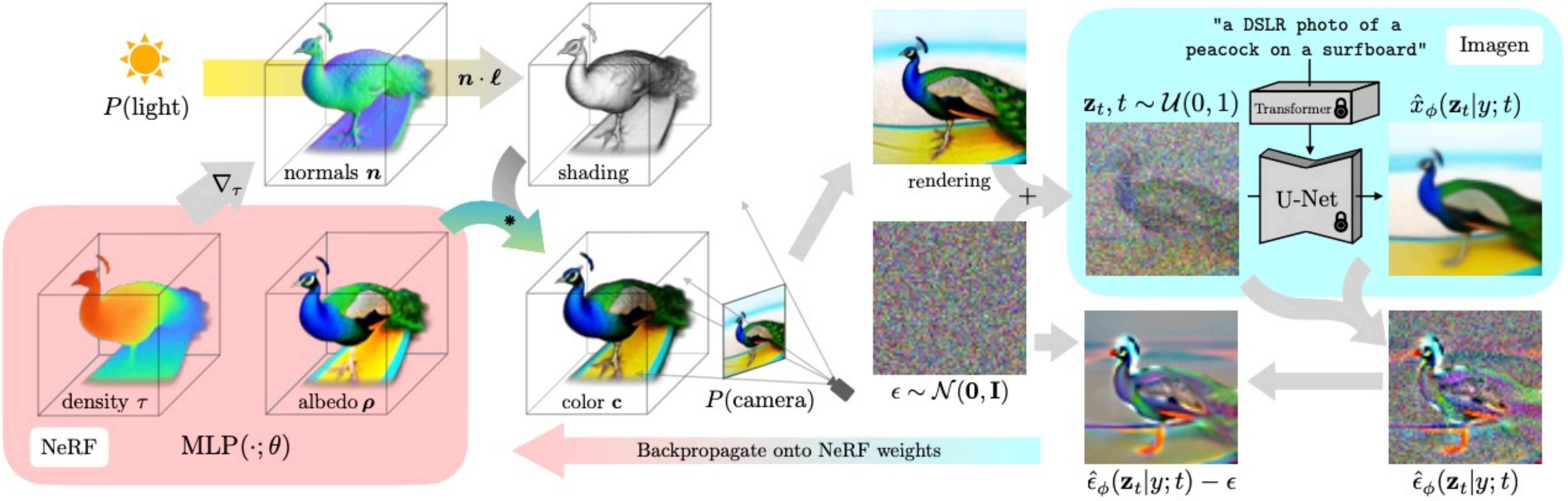
L. Zhang and M. Agrawala. [Adding Conditional Control to Text-to-Image Diffusion Models](). arXiv 2023

Connecting 2D to 3D: DreamFusion



B. Poole, A. Jain, J. Barron, B. Mildenhall. [DreamFusion: Text-to-3D using 2D Diffusion](). arXiv 2022

Connecting 2D to 3D: DreamFusion



B. Poole, A. Jain, J. Barron, B. Mildenhall. [DreamFusion: Text-to-3D using 2D Diffusion](). arXiv 2022

Outline

- Conditional diffusion models
- Large-scale models
- Controlling and fine-tuning image generation
- Societal, ethical, and legal issues

Societal, ethical, and legal issues

- Closed or open?
- Safe or unsafe?
- Potential for generating DeepFakes and misinformation
- Dataset image rights
- Artists’ rights
- The nature of creativity

In the news

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content

An image created by Stable Diffusion showing a recreation of Getty Images' watermark. Image: The Verge / Stable Diffusion

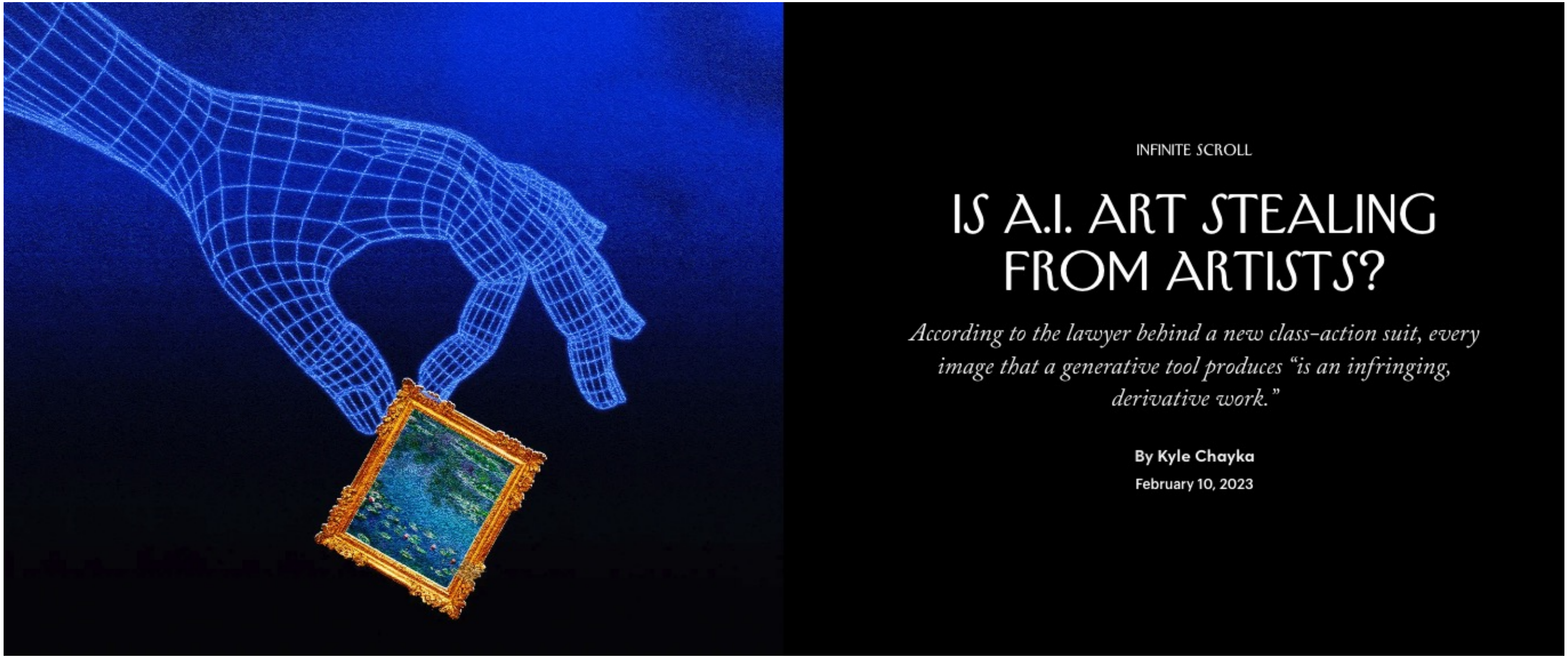
/ Getty Images claims Stability AI 'unlawfully' scraped millions of images from its site. It's a significant escalation in the developing legal battles between generative AI firms and content creators.

By JAMES VINCENT

Jan 17, 2023, 4:30 AM CST | 18 Comments / 18 New

https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit

In the news



https://www.newyorker.com/culture/infinite-scroll/is-ai-art-stealing-from-artists

In the news

Fake Trump arrest photos: How to spot an AI-generated image

This image looks realistic, but take a closer look at Trump's right arm and neck

https://www.bbc.com/news/world-us-canada-65069316

In the news

Midjourney Bans AI Images of Chinese President Xi Jinping

APR 03, 2023 MATT GROWCOOT

https://petapixel.com/2023/04/03/midjourney-bans-ai-images-of-chinese-president-xi-jinping/