Introduction to Text-to-Image Models

HARVEY MANNERING

Overview

1. Technical Details
2. Applications
3. Ethics Considerations

How Do Text-to-Image Models Work?

Text-to-Image Model

Text-to-image models typically work in two stages:

Text-to-Image Model

Text-to-image models typically work in two stages:

1. Encoding Text using CLIP

Text-to-Image Model

Text-to-image models typically work in two stages:

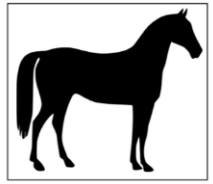
1. Encoding Text using CLIP
2. Generate an image using that CLIP encoding

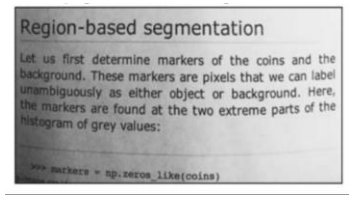
CLIP (Contrastive Language-Image Pre-Training)

a black-and-white silhouette of a horse

a page of text about segmentation

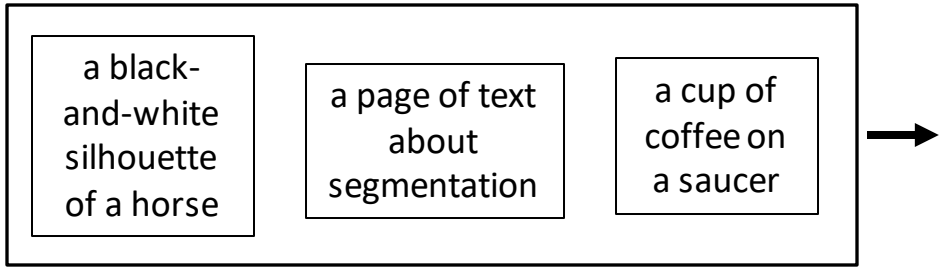
a cup of coffee on a saucer

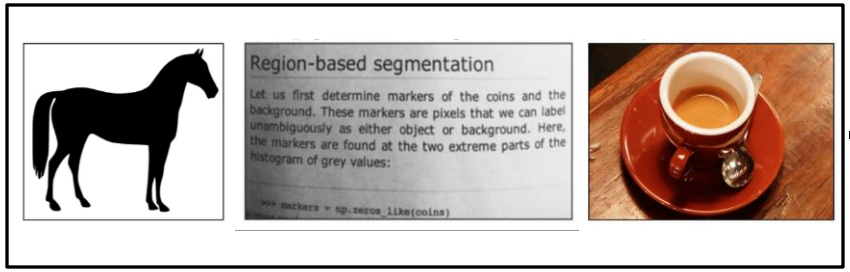




[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

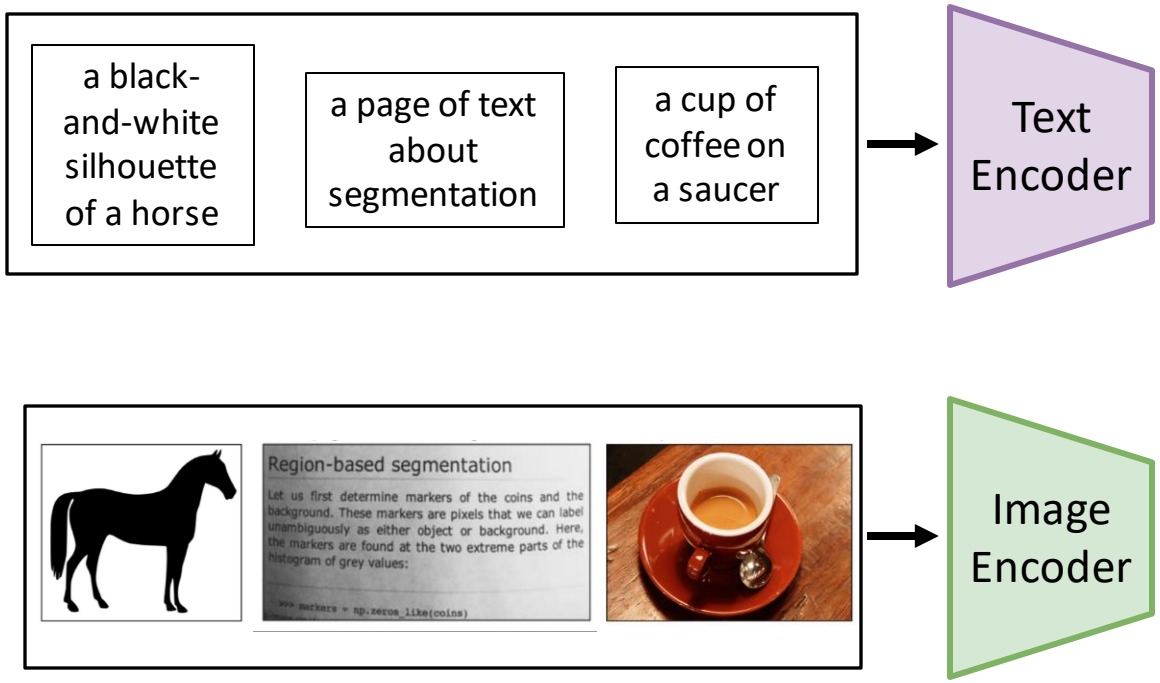
CLIP (Contrastive Language-Image Pre-Training)





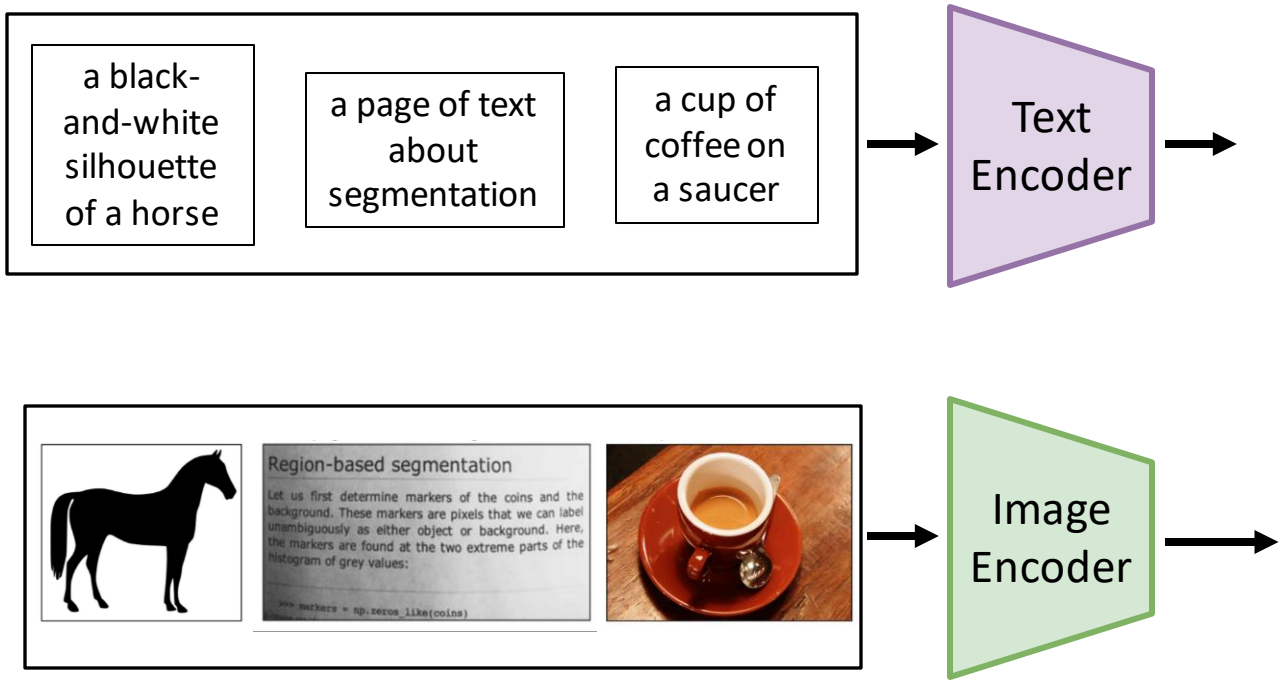
[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

CLIP (Contrastive Language-Image Pre-Training)



[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

CLIP (Contrastive Language-Image Pre-Training)



[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

CLIP (Contrastive Language-Image Pre-Training)

a black-and-white silhouette of a horse

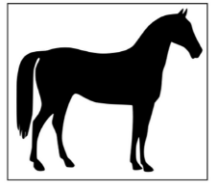
a page of text about segmentation

a cup of coffee on a saucer

Text Encoder

$$\begin{bmatrix} 0.01 & 0.00 & \dots & 0.02 \\ -0.01 & 0.03 & \dots & 0.04 \\ 0.01 & 0.01 & \dots & -0.01 \end{bmatrix}$$

$Dimensions : 3 \times 512$



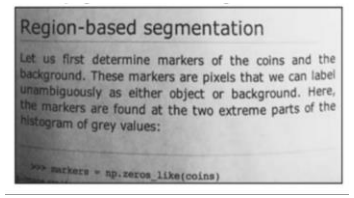


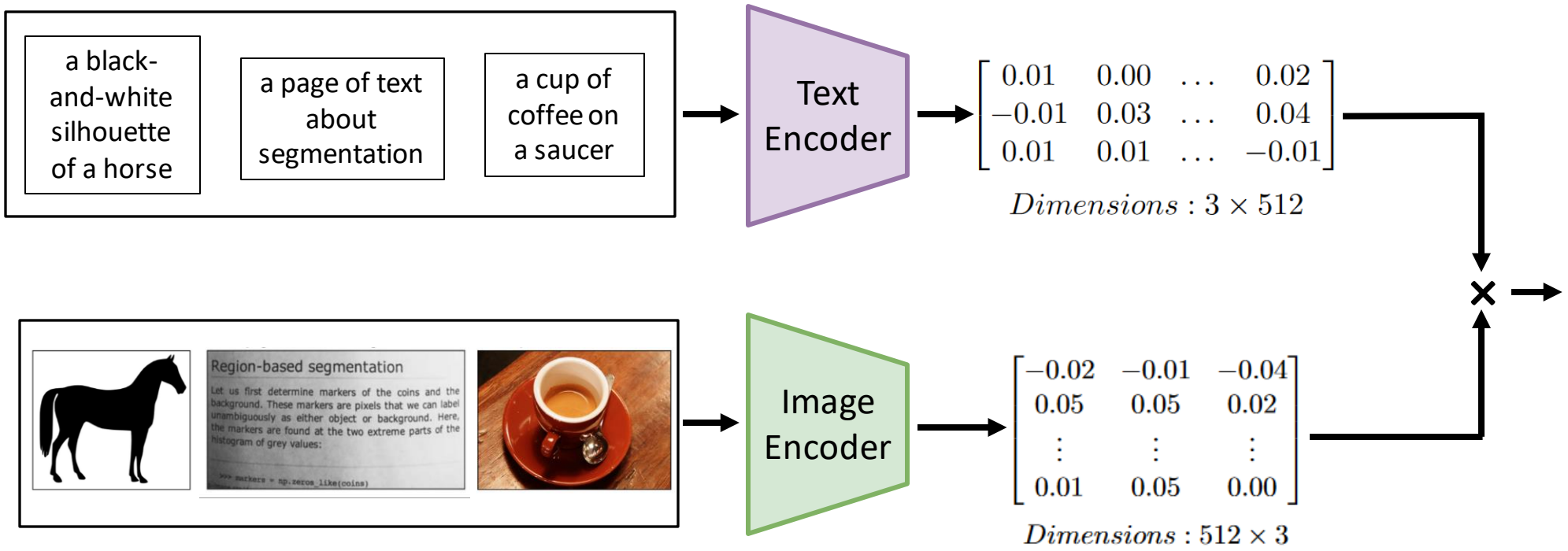
Image Encoder

$$\begin{bmatrix} -0.02 & -0.01 & -0.04 \\ 0.05 & 0.05 & 0.02 \\ \vdots & \vdots & \vdots \\ 0.01 & 0.05 & 0.00 \end{bmatrix}$$

$Dimensions : 512 \times 3$

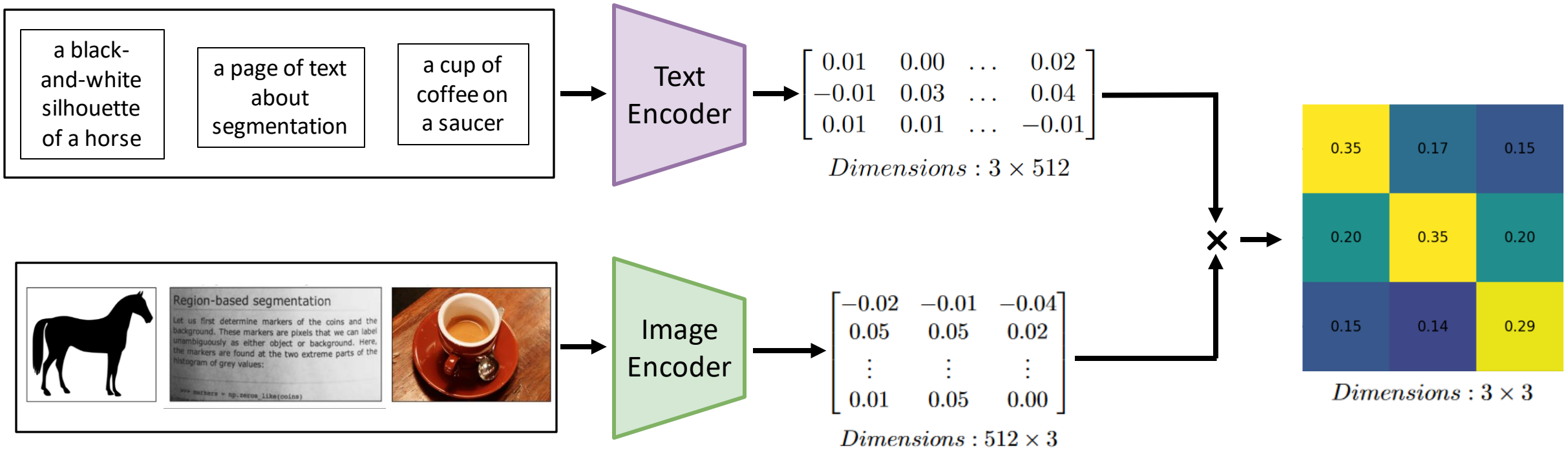
[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

CLIP (Contrastive Language-Image Pre-Training)



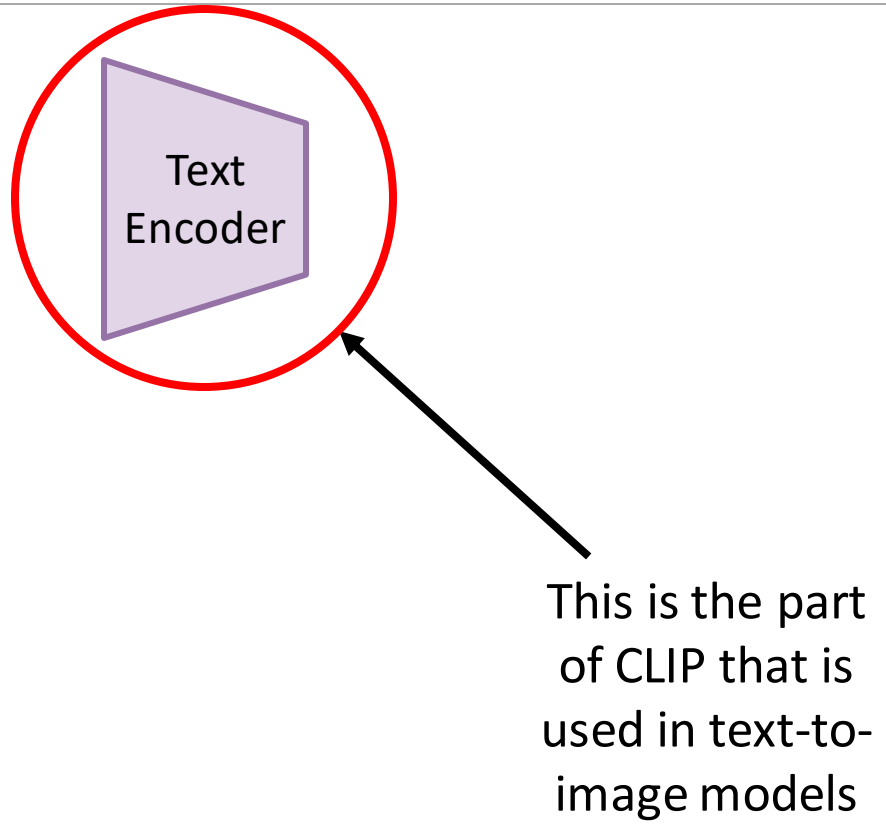
[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

CLIP (Contrastive Language-Image Pre-Training)



[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

CLIP (Contrastive Language-Image Pre-Training)



[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

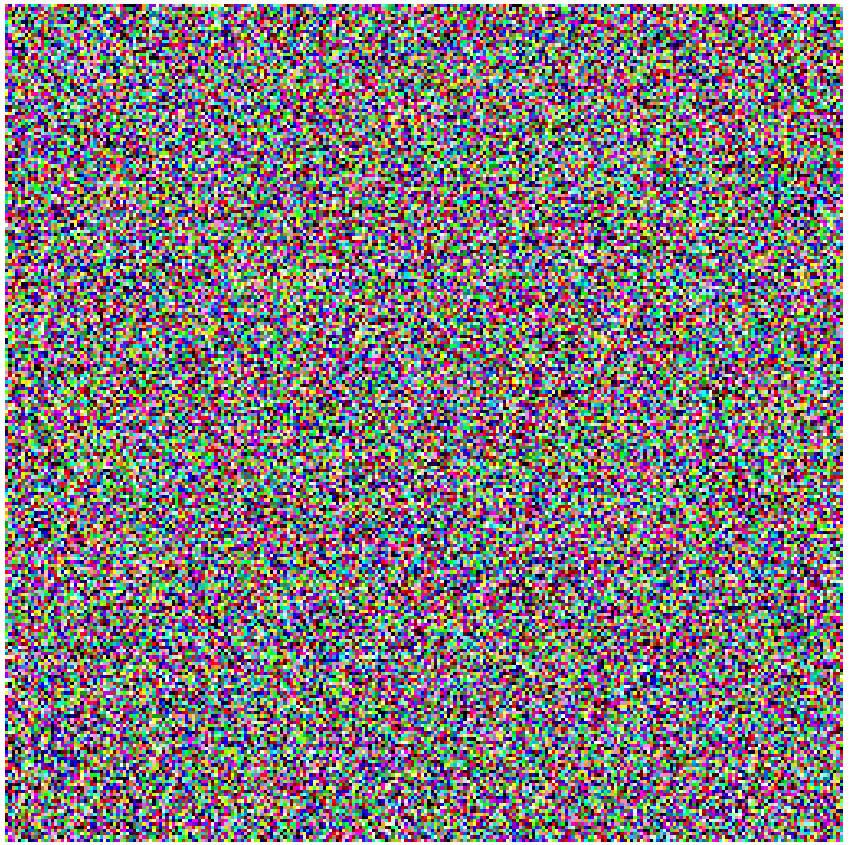
Diffusion

Diffusion generates realistic images from random noise

It works by training a denoising neural network

When generating the image, we start from pure noise then repeatedly denoise and renoise our image

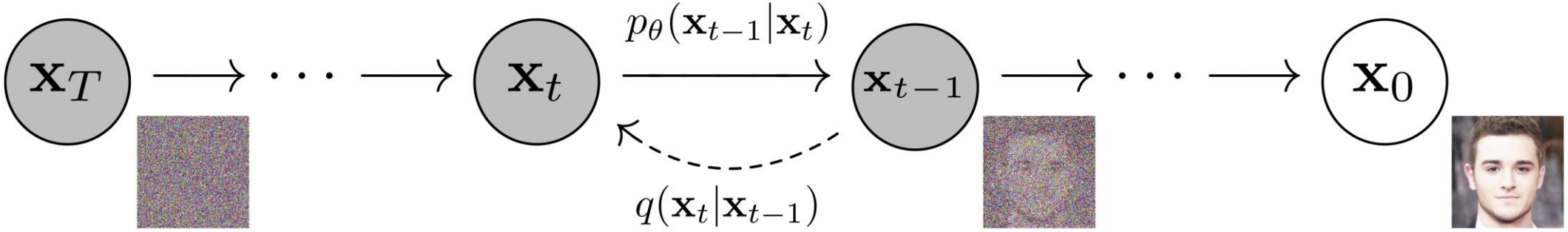
Diffusion is very slow and very expensive



[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

Diffusion

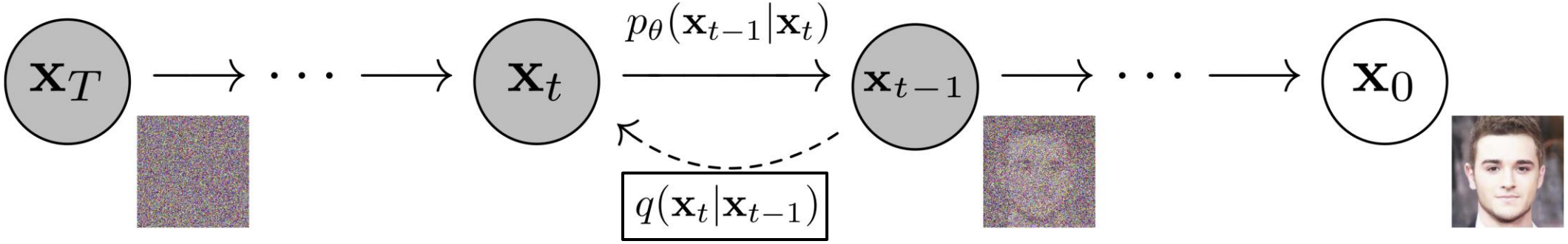
We can model diffusion as a Markov chain, when every step of the diffusion process is dependent only on the previous step



[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

Diffusion

We can model diffusion as a Markov chain, when every step of the diffusion process is dependent only on the previous step

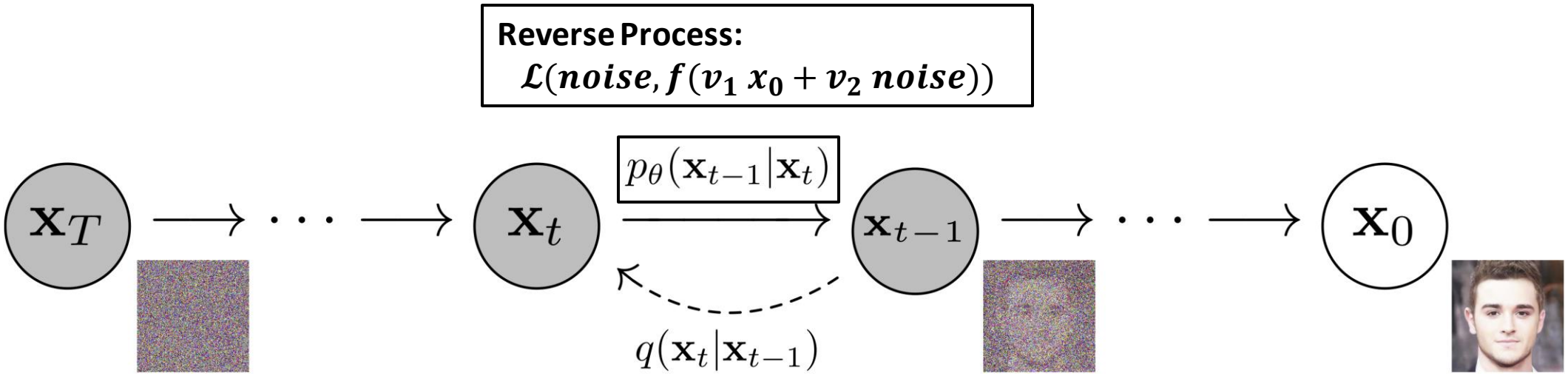


Forward Process:
$v_1\, x_0 + v_2\, noise$

[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

Diffusion

We can model diffusion as a Markov chain, when every step of the diffusion process is dependent only on the previous step



[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

Stable Diffusion Model

Step 1: Train an autoencoder to shrink the image to a smaller latent space representation

Step 2: Train the diffusion model in the latent space for lower computational cost

256×256

Image Encoder

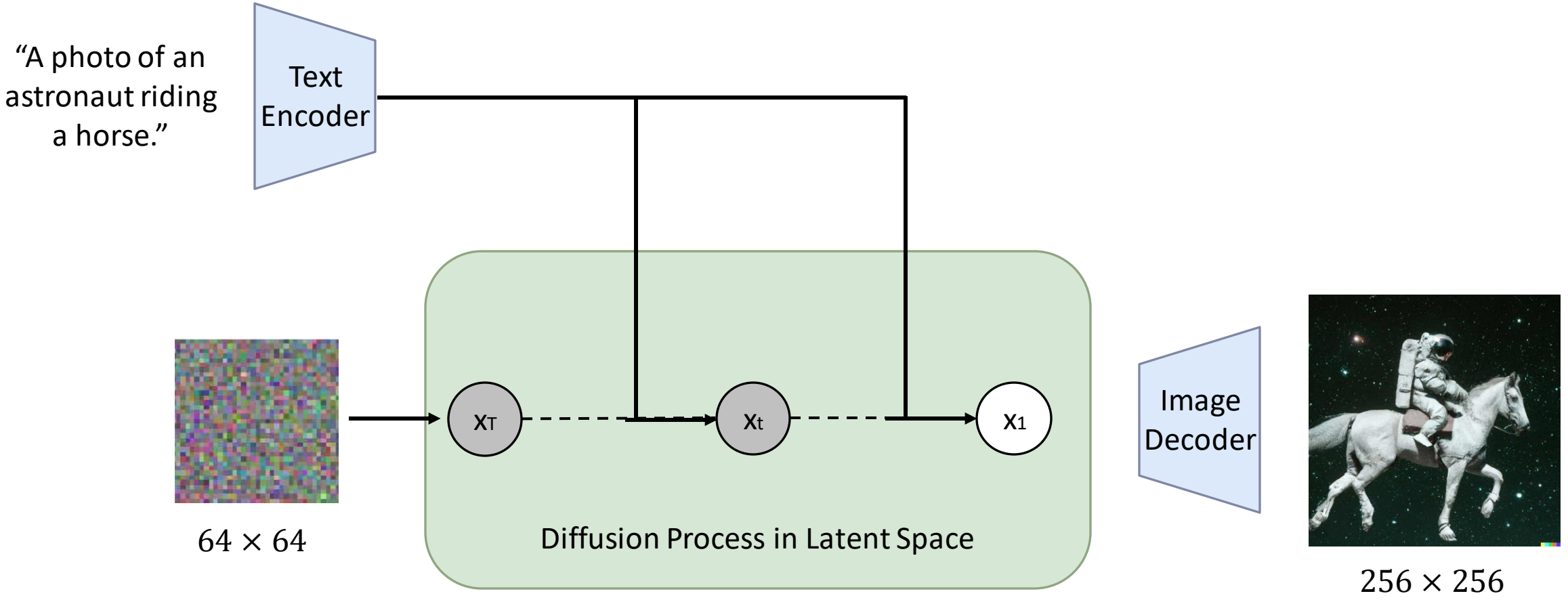
Latent Space
64×64

Image Decoder

256×256

[3] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Stable Diffusion Model



[3] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

How Can We Use Text-to-Image Models?

Real World Applications

AI generated art (this picture caused controversy after winner the Colorado state fair's emerging digital artists award in 2022)

Drafting art works

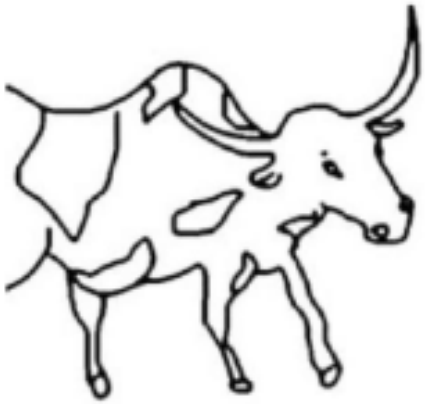
Story boarding

Image editing

ControlNet

Allows stable diffusion to be finetuned so it can take in extra conditions

Input (User Scribble)



User Prompt

"a robot ox on moon, UE5 rendering, ray tracing"

[4] Zhang, Lvmin, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *arXiv preprint arXiv:2302.05543* (2023).

ODISE

Segmentation can be done using internal features from stable diffusion

Open-vocabulary, meaning there is no predefined list of categories

Input Image

ODISE Segmentation

[5] Xu, Jiarui, et al. "Open-vocabulary panoptic segmentation with text-to-image diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

DreamBooth

Enables a text-to-image model to represent a specific subject

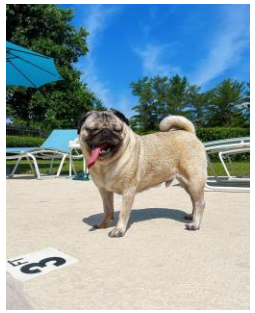
Start with a pretrained stable diffusion model, then finetune on 3 – 5 images of a specific animal, landmark, person, landscape etc.

To prevent stable diffusion from forgetting old concept we also use prior preserving loss

[6] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

DreamBooth

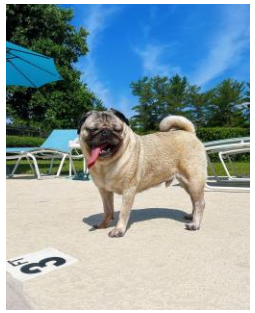
Finetuning dataset

[6] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

DreamBooth

Finetuning dataset

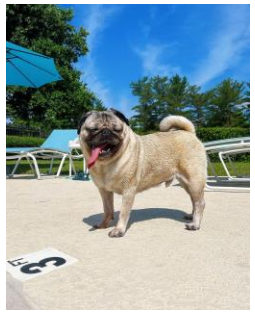
Finetuning prompt

```
photo of
zwx dog
```

[6] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

DreamBooth

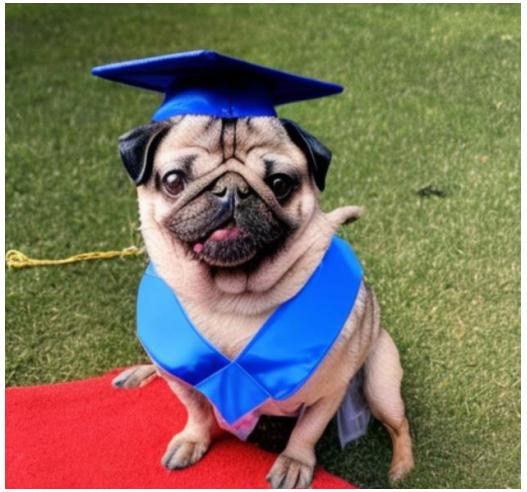
Finetuning dataset

Finetuning prompt

```
photo of
zwx dog
```

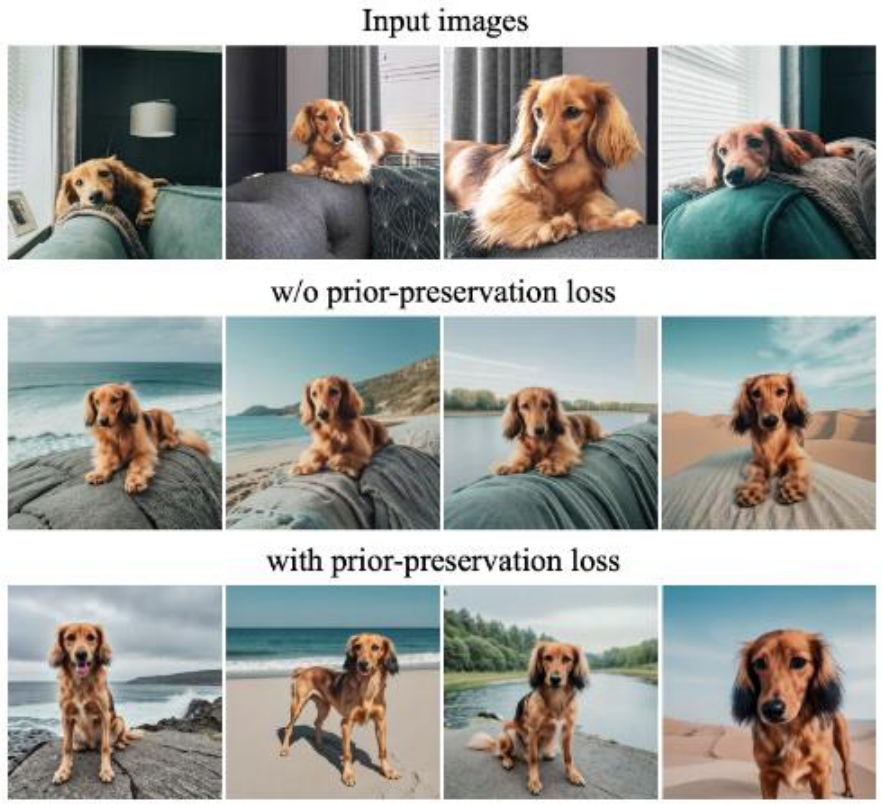
Inference



```
photo of zwx dog
graduating with honours
```

[6] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

DreamBooth



Prior preserving loss prevent stable diffusion from forgetting old concepts and prevent over fitting

It is simply implemented by finetuning on outputs from the pretrained stable diffusion

This prevents the model from drifting too far from its initial state

[6] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

What Are The Ethical Considerations?

Algorithmic Bias

Software engineers are ≈90% male, however one study found stable diffusion depicted software engineers as male 100% of the time.

This suggests that stereotypes are being amplified by text-to-image models

Better curation of training data is needed

Transforming a user's input prompt may also be helpful

[7] Bianchi, Federico, et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023.

Digital Art or Digital Forgery?

For small datasets, diffusion models can just copy the training data

Overfitting and repeated data in the training set can also cause content replication

Even if image are not directly copied, copyrighted data exist within training data

[8] Somepalli, Gowthami, et al. "Diffusion art or digital forgery? investigating data replication in diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

References

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

[3] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[4] Zhang, Lvmin, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *arXiv preprint arXiv:2302.05543* (2023).

[5] Xu, Jiarui, et al. "Open-vocabulary panoptic segmentation with text-to-image diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[6] Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[7] Bianchi, Federico, et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023.

[8] Somepalli, Gowthami, et al. "Diffusion art or digital forgery? investigating data replication in diffusion models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

[9] Wu, Chenfei, et al. "Visual chatgpt: Talking, drawing and editing with visual foundation models." *arXiv preprint arXiv:2303.04671* (2023).