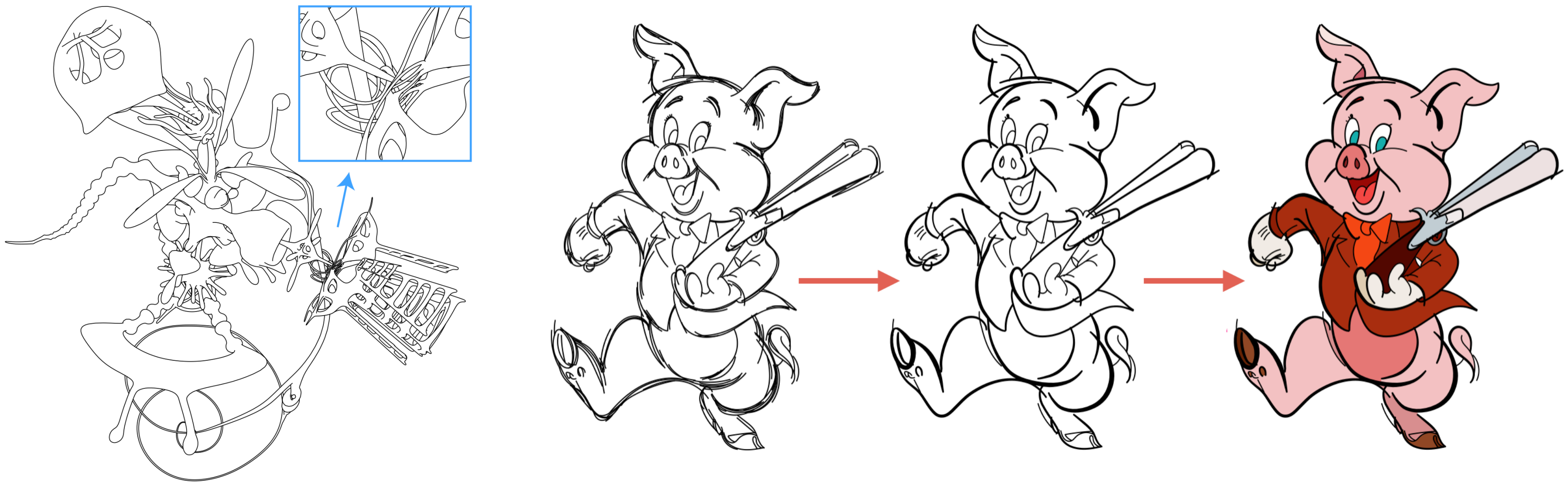
CSC317: Computer Graphics

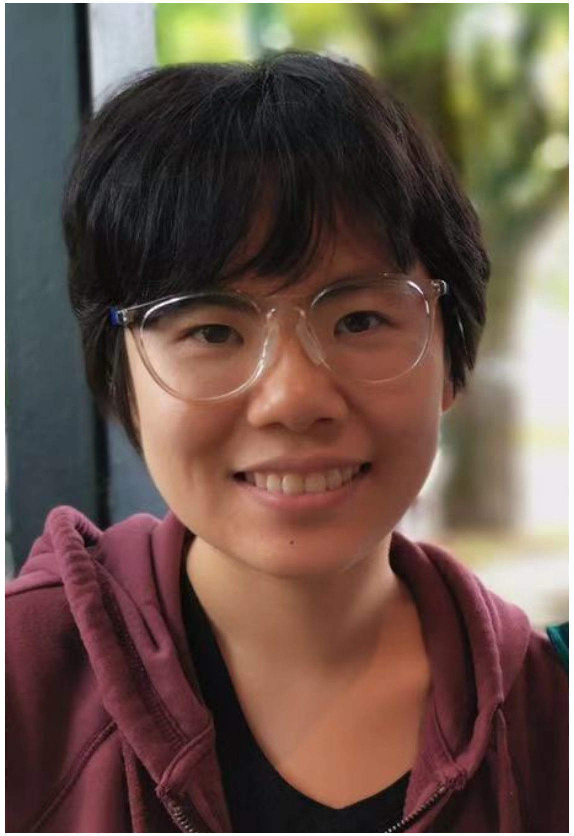
Lecture instructor: Chenxi Liu

About me

- Postdoc working with Prof. Alec Jacobson
- Completed my PhD program at UBC
- Worked on Vector sketch generation and processing



- Working on: Text-to-image generation + Vector graphics



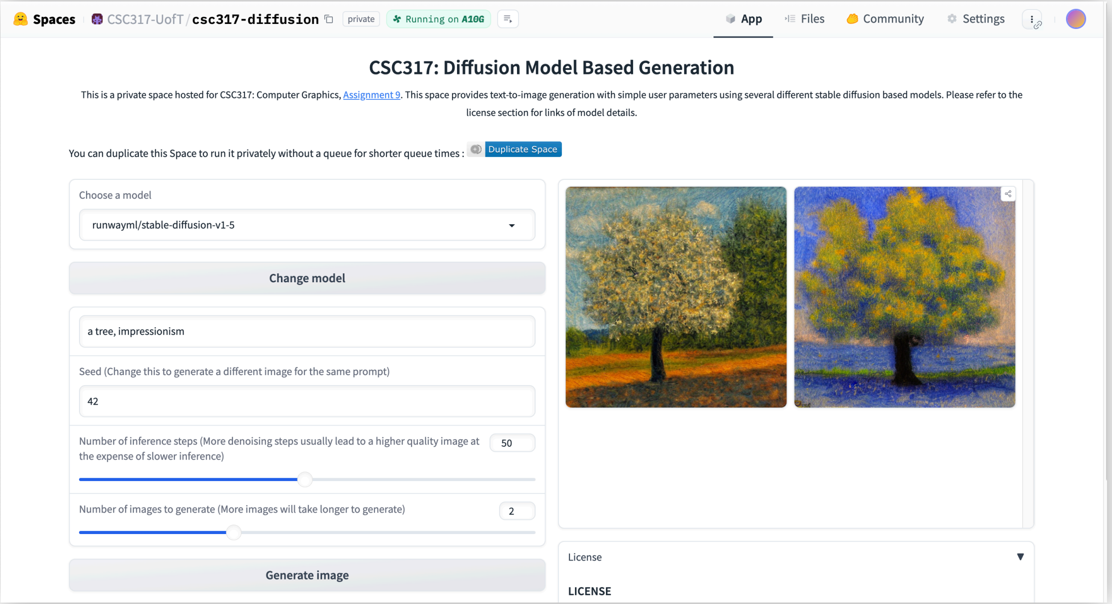
Contact:
chenxil.liu@utoronto.ca

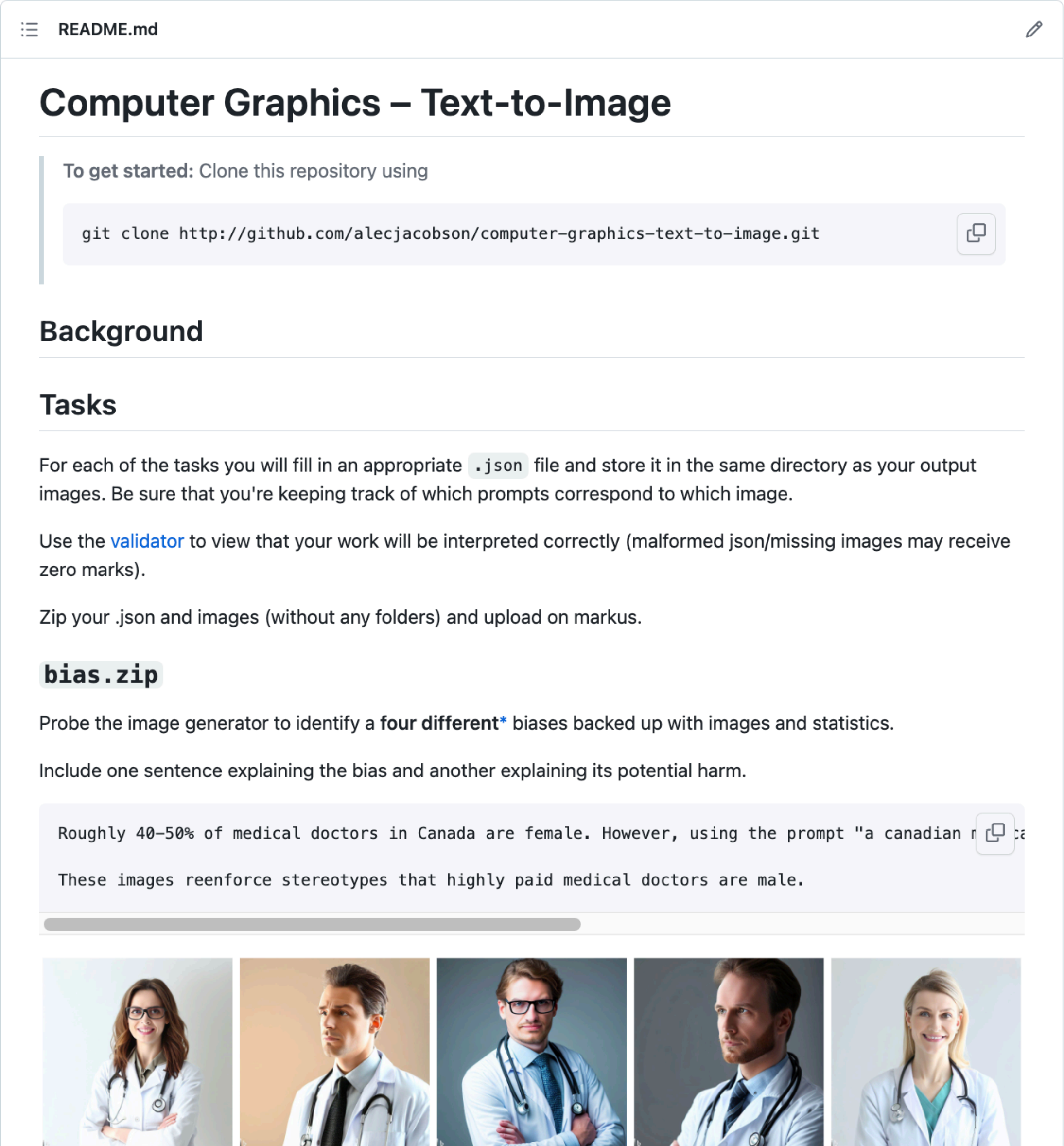
Announcement

Assignment 9 is out.

Deadline: **November 29**

Access to our private HuggingFace space for the duration of the assignment.





README.md

Computer Graphics – Text-to-Image

> **To get started:** Clone this repository using
>
> ```
> git clone http://github.com/alecjacobson/computer-graphics-text-to-image.git
> ```

Background

Tasks

For each of the tasks you will fill in an appropriate `.json` file and store it in the same directory as your output images. Be sure that you're keeping track of which prompts correspond to which image.

Use the validator to view that your work will be interpreted correctly (malformed json/missing images may receive zero marks).

Zip your .json and images (without any folders) and upload on markus.

`bias.zip`

Probe the image generator to identify a **four different*** biases backed up with images and statistics.

Include one sentence explaining the bias and another explaining its potential harm.

```
Roughly 40-50% of medical doctors in Canada are female. However, using the prompt "a canadian r
These images reinforce stereotypes that highly paid medical doctors are male.
```

Today: Text-to-Image Generation

What is Text-to-Image Generation?

A task where the goal is to **generate an image that corresponds to a given textual description**.

[A quick demo…]

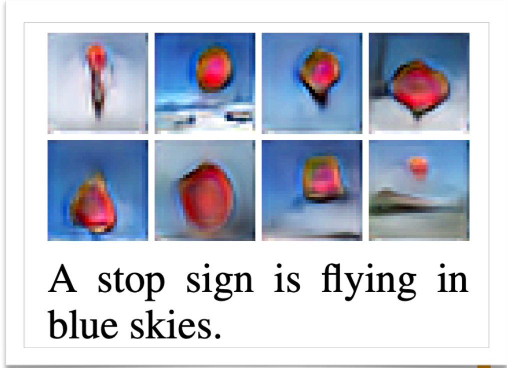
What is Text-to-Image Generation?

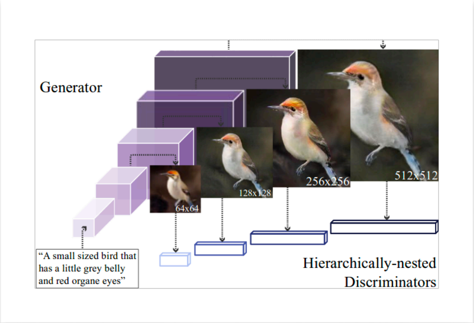
A task where the goal is to **generate an image that corresponds to a given textual description**.

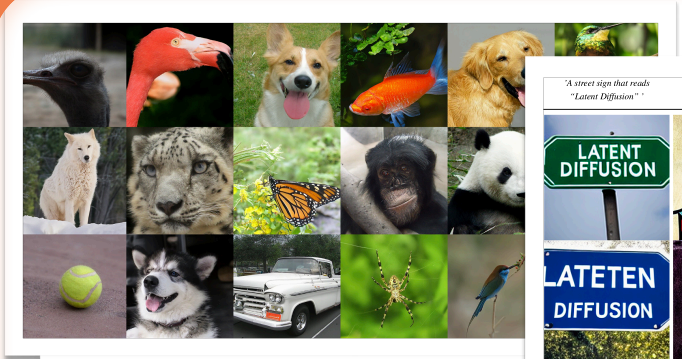
[A quick demo…]

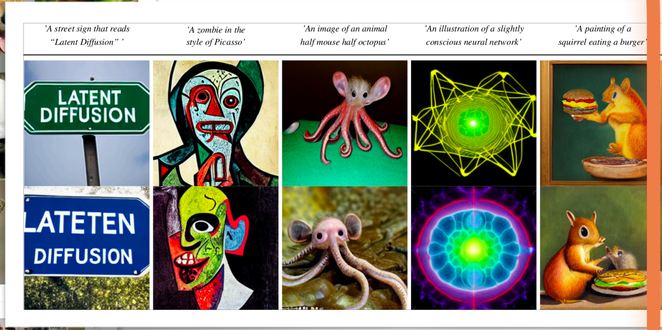
The term *computer graphics* describes any use of computers to create and manipulate images.

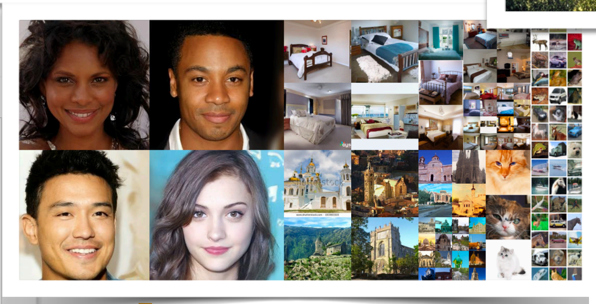
Timeline of Text-to-Image Generation











DALL·E 2

DreamStudio

2007

2016

GANs

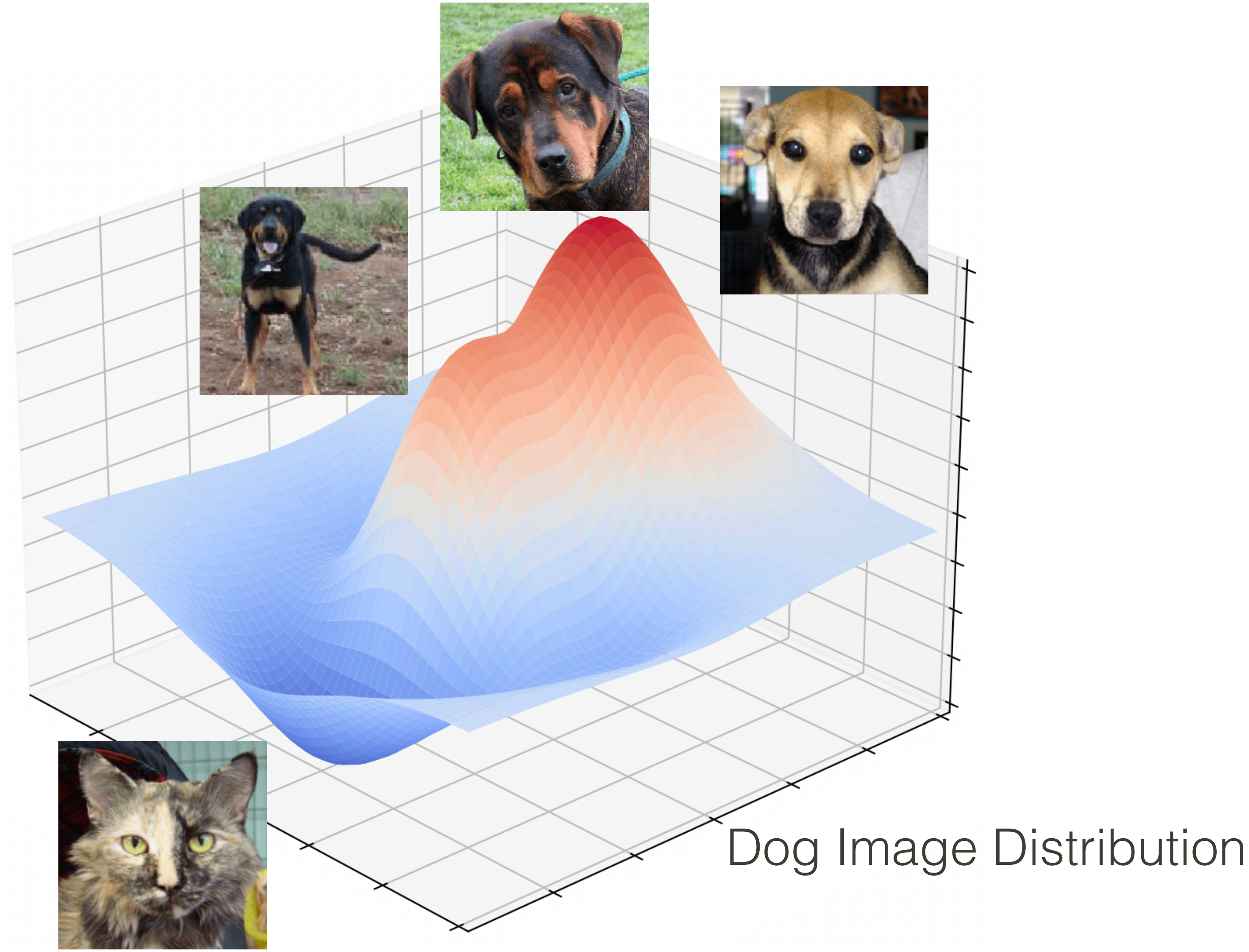
2020-2021

Diffusion Models
Commercial Text-to-Image Tools

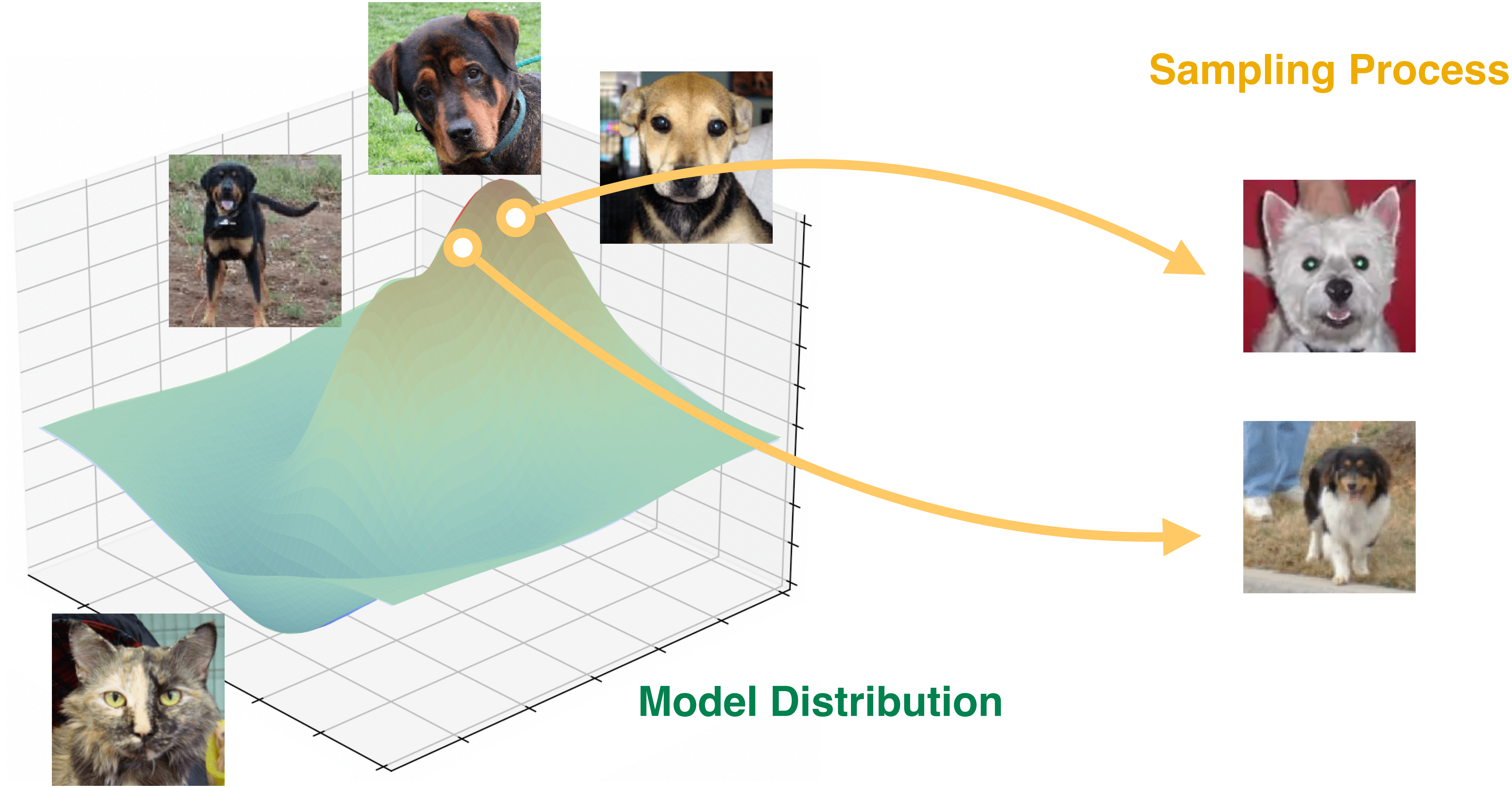
A Handwavy Introduction to Diffusion Model

Partially based on Yang Song's blog and talk

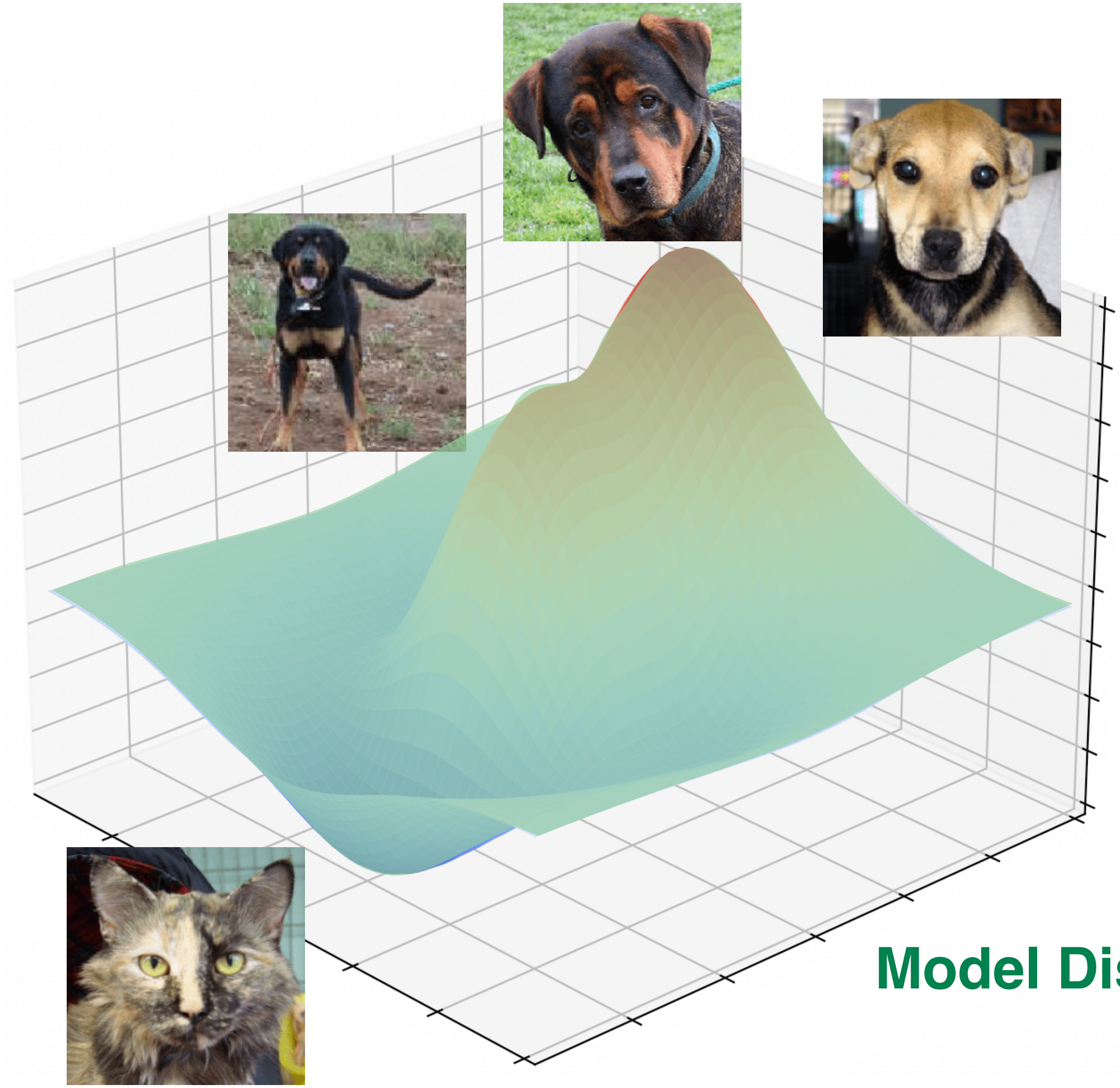
Image Distribution



Generation via Likelihood-Based Models

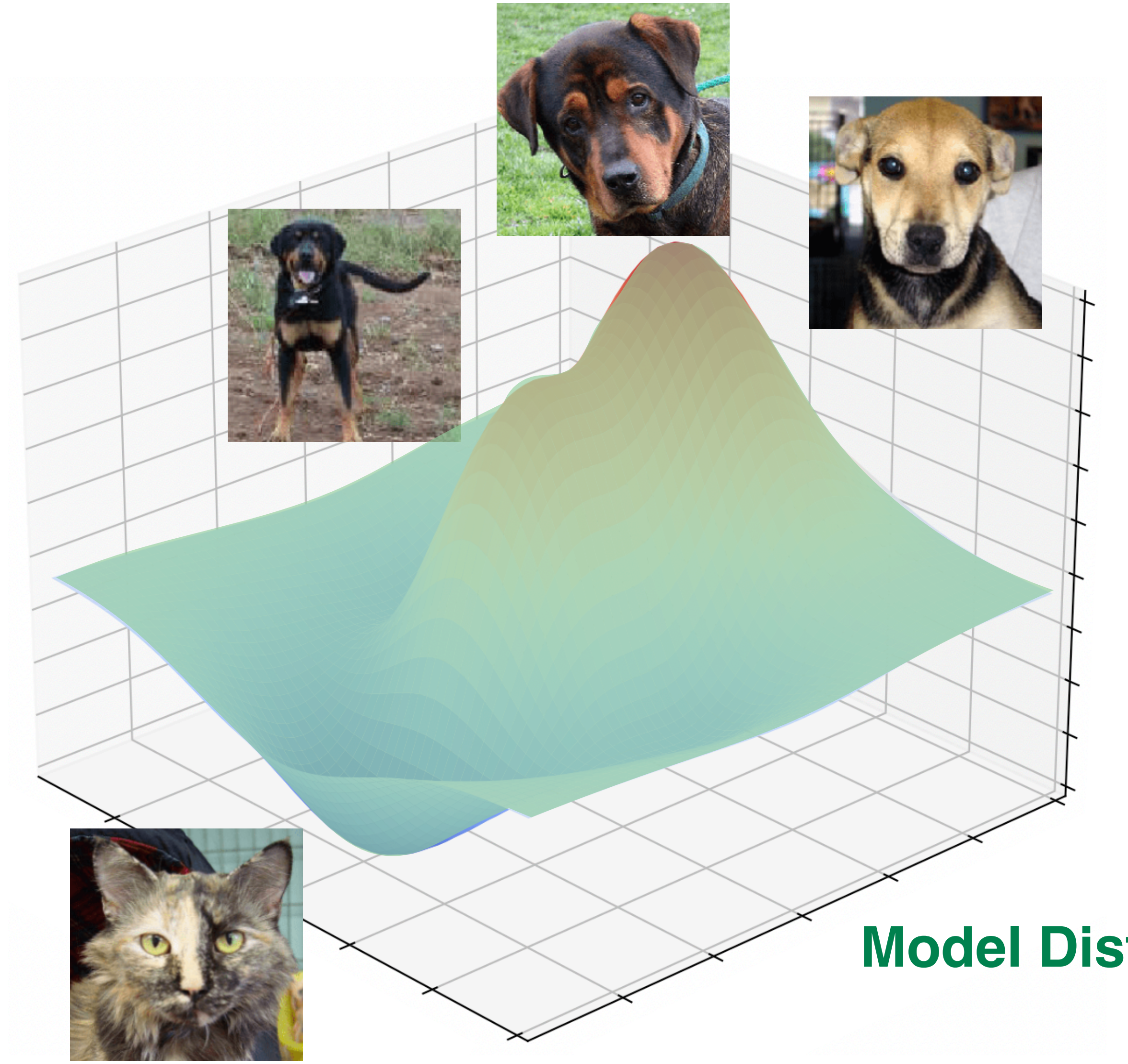


Fitting Likelihood-Based Models

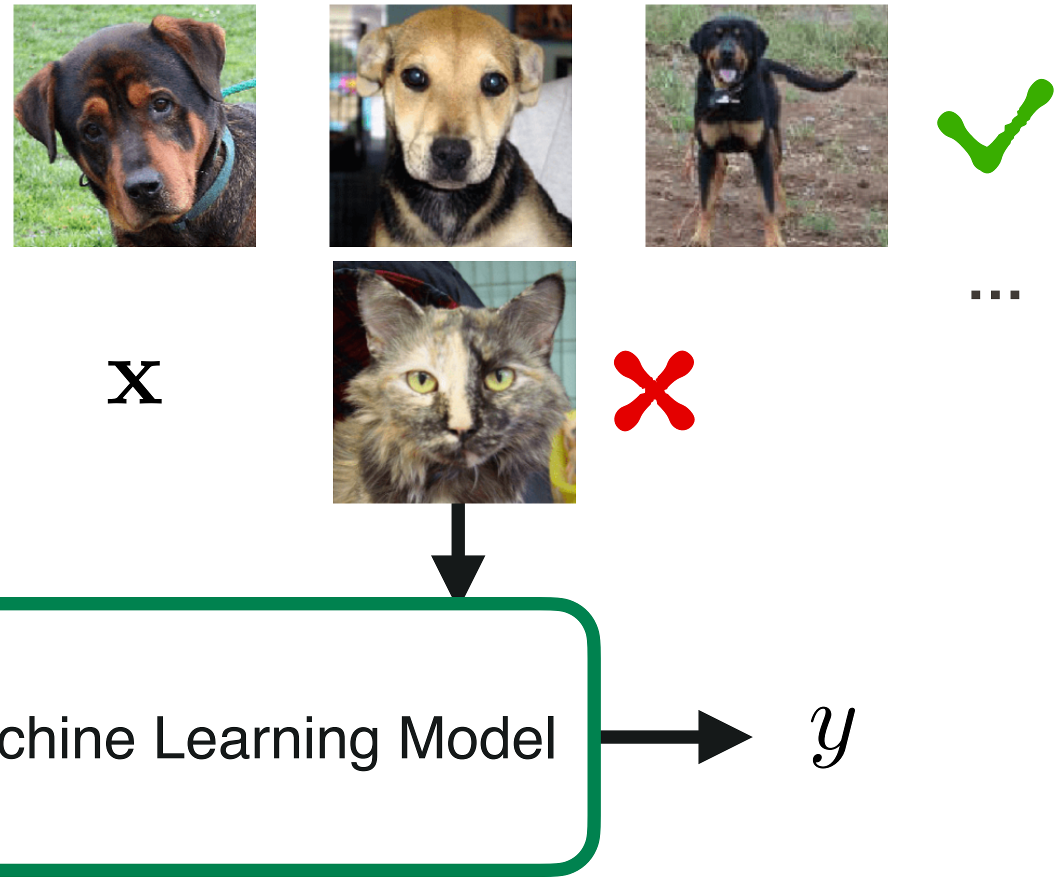


Model Distribution

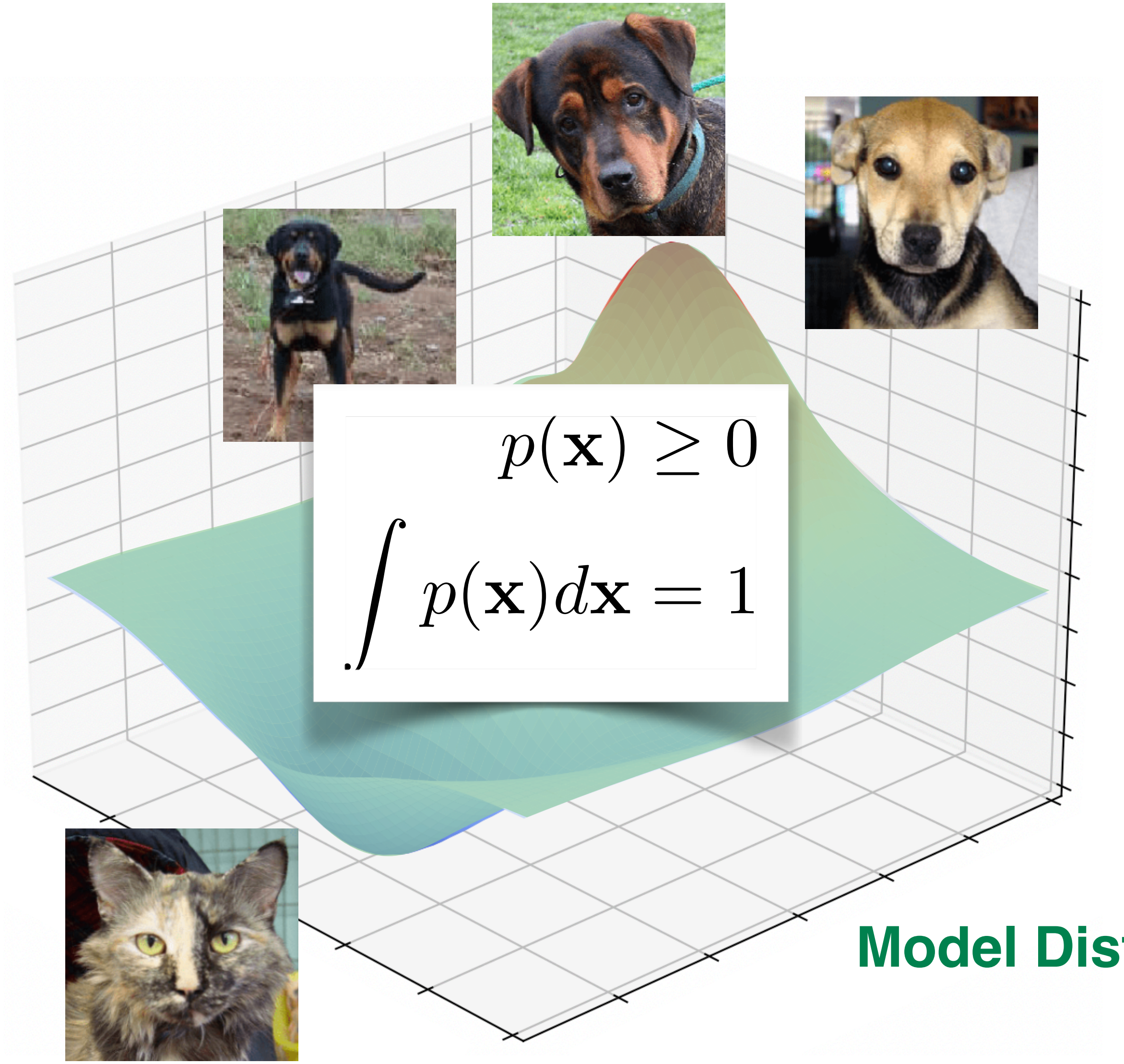
Fitting Likelihood-Based Models

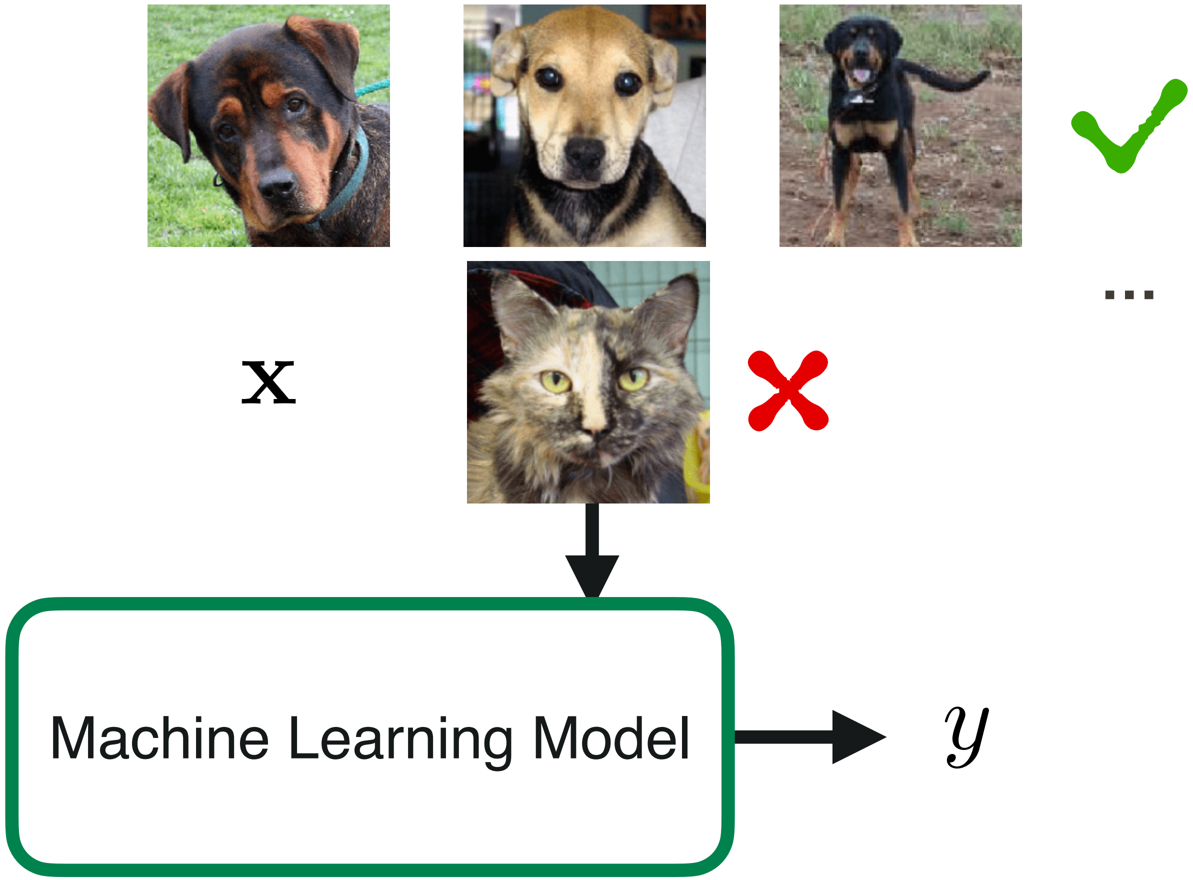


Model Distribution



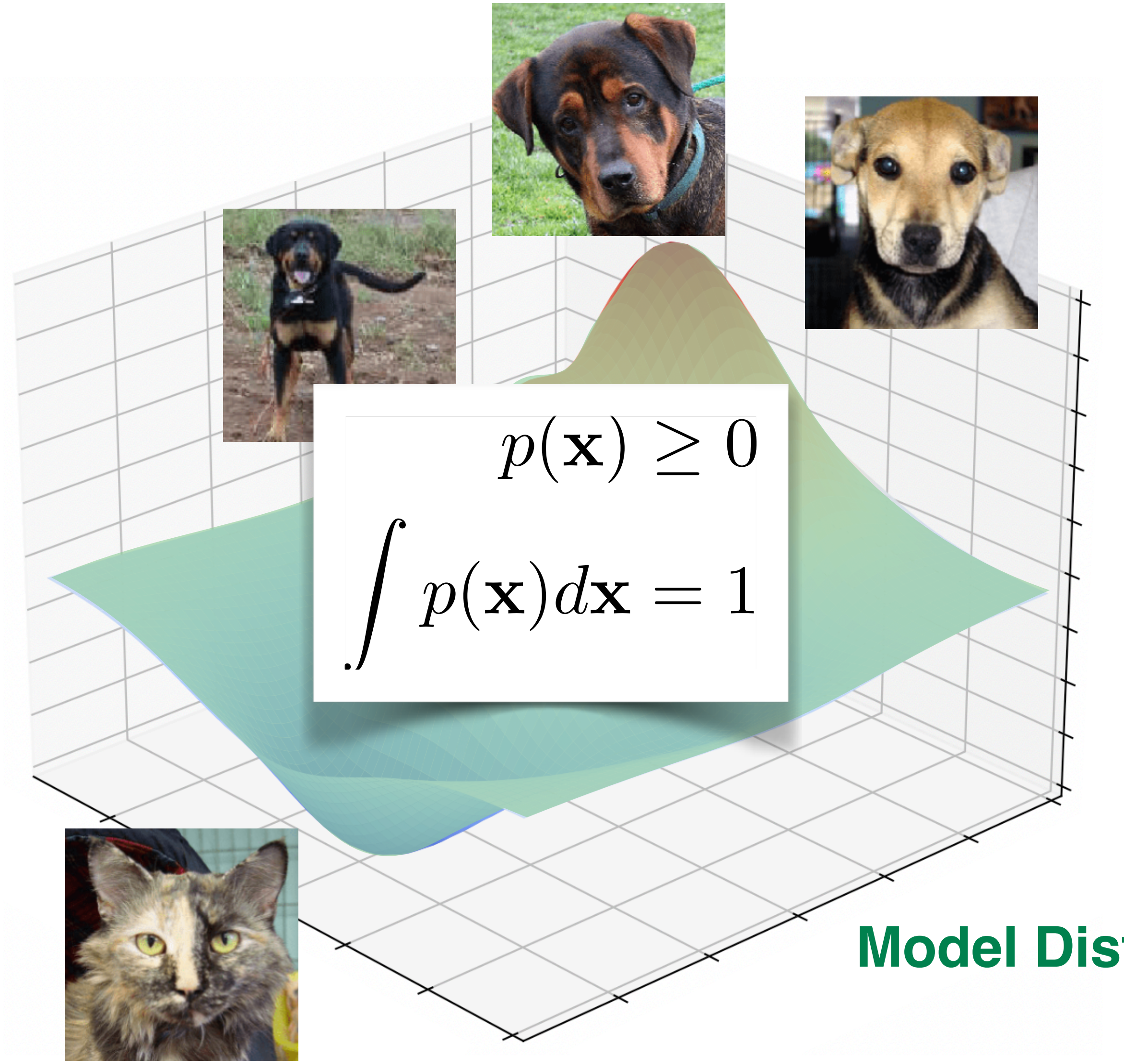
Fitting Likelihood-Based Models



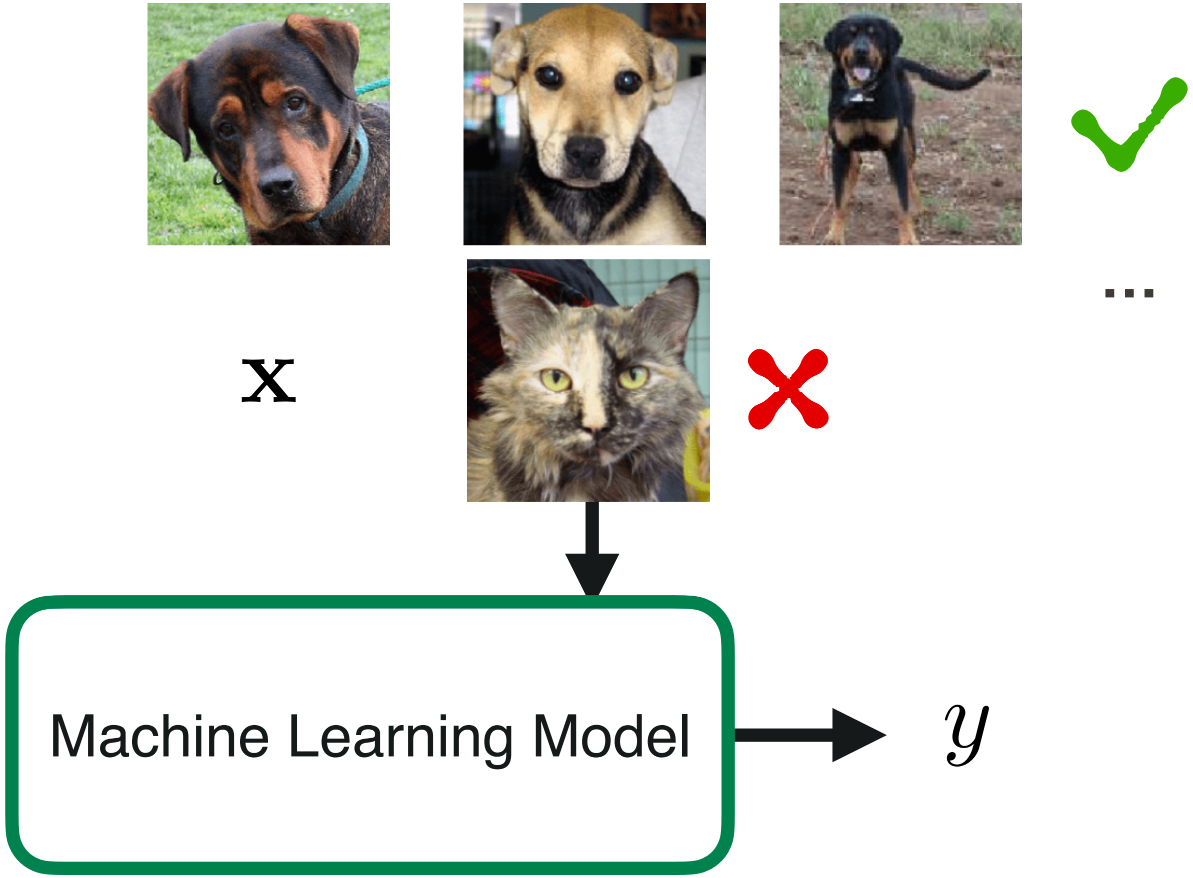


Model Distribution

Fitting Likelihood-Based Models

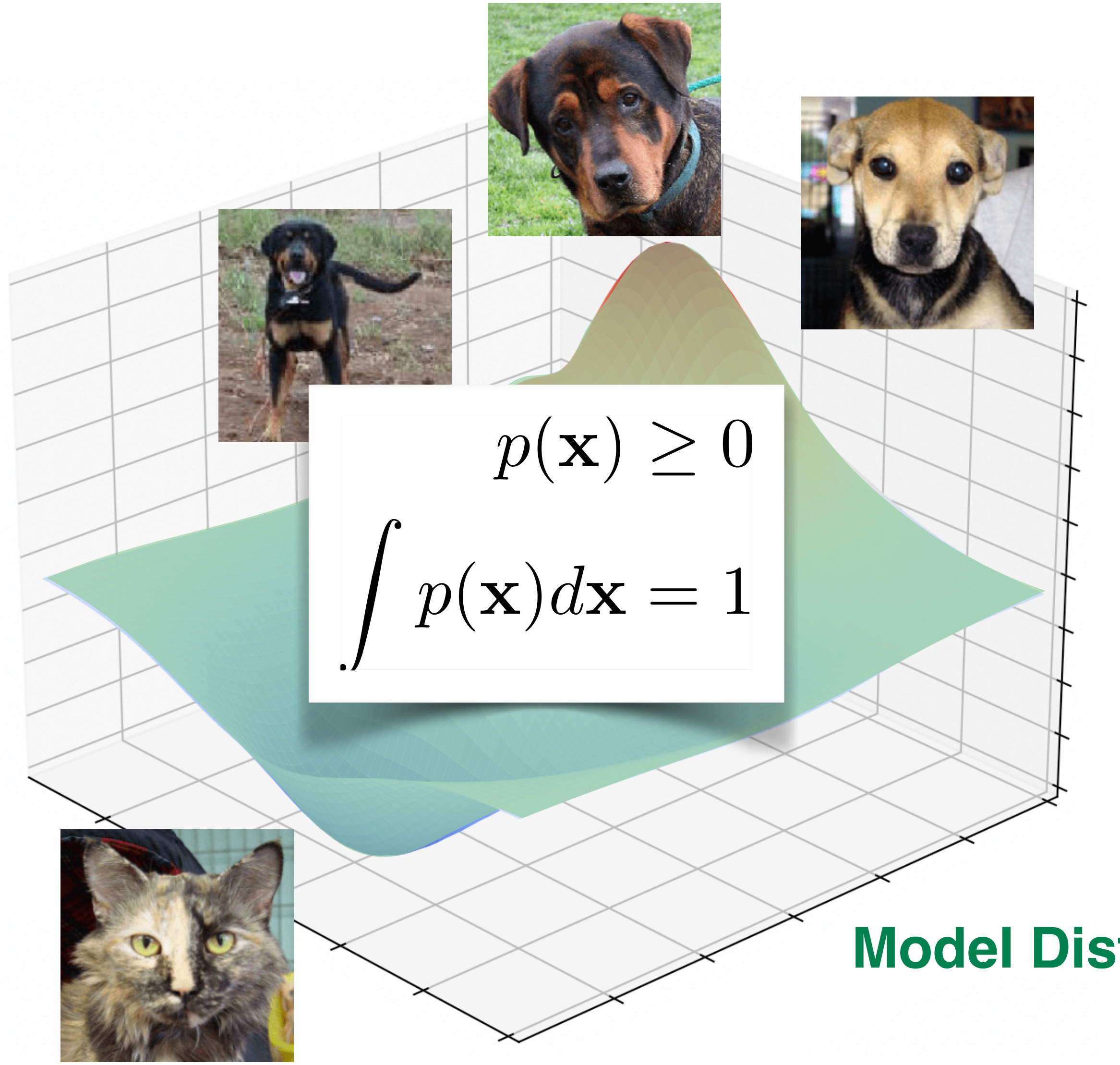


Model Distribution

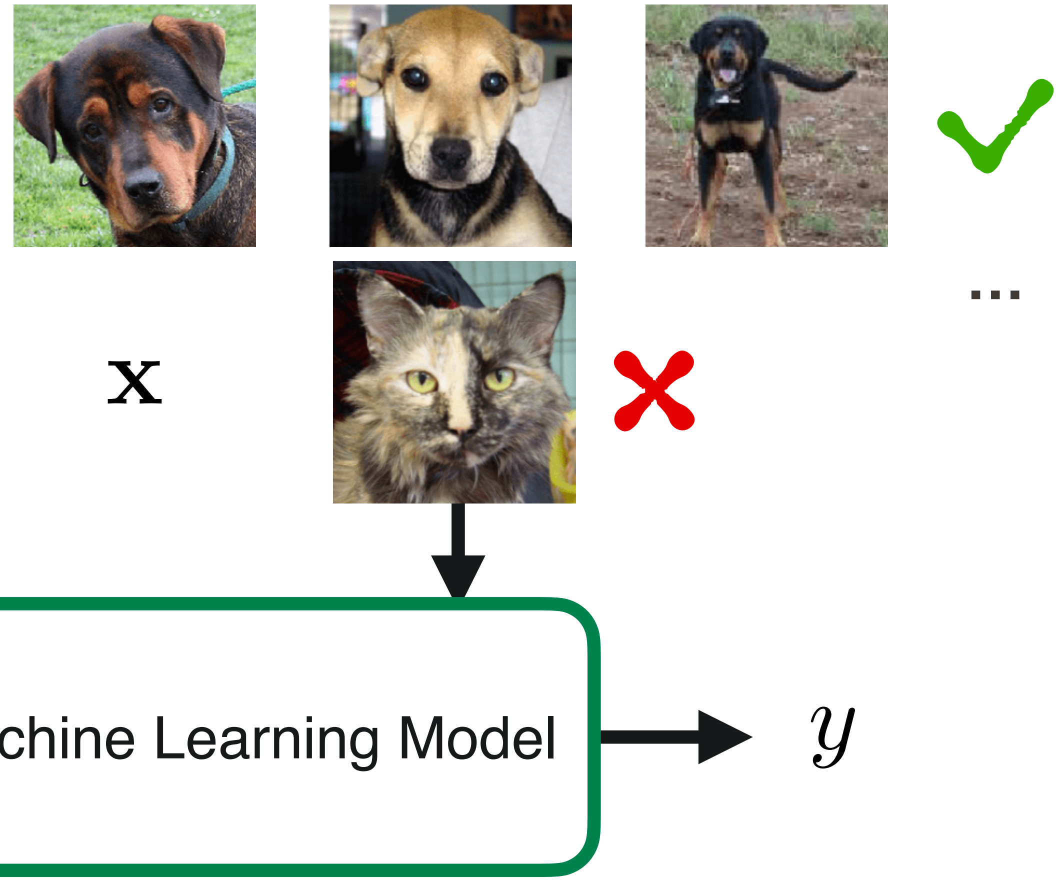


$$Z = \int y \, d\mathbf{x}$$

Fitting Likelihood-Based Models

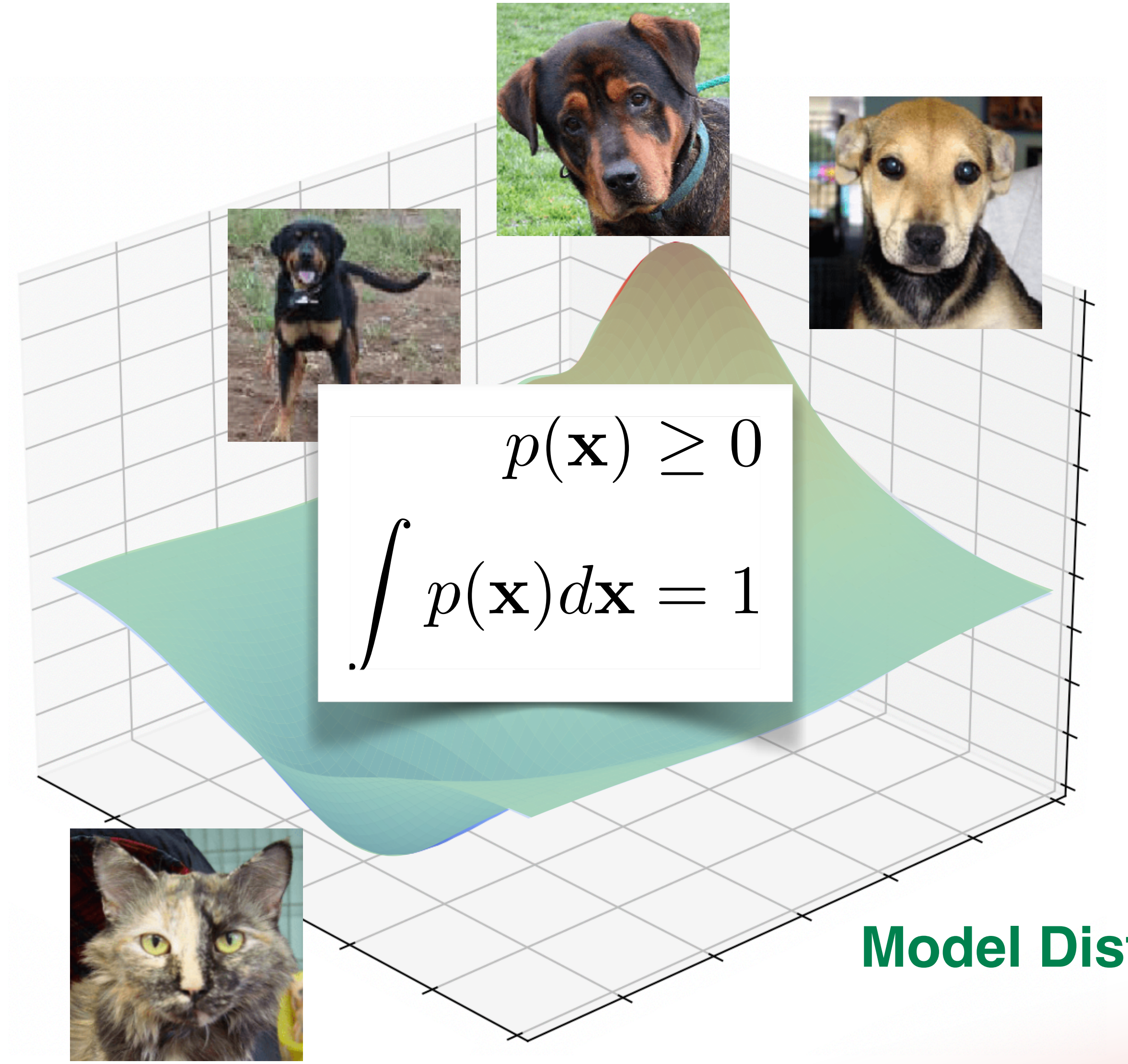


Model Distribution

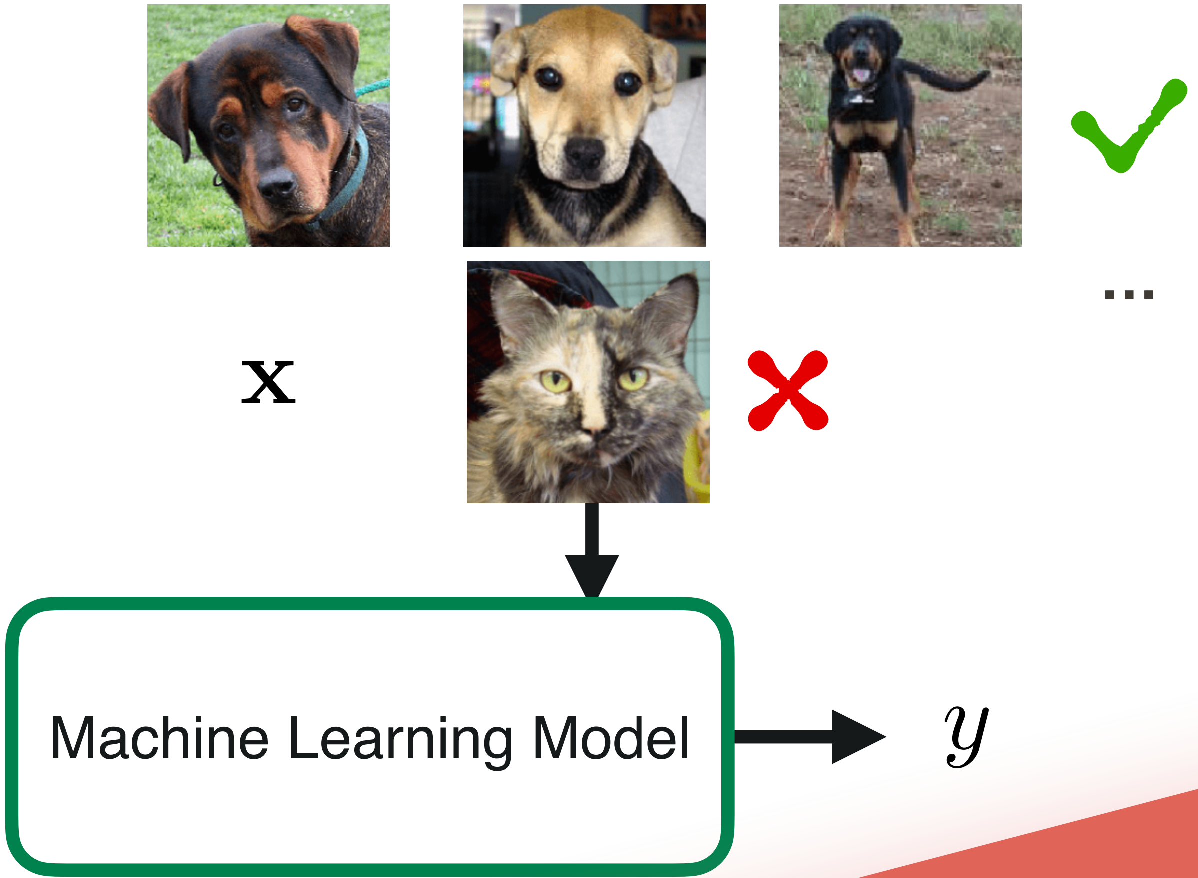


$$p(\mathbf{x}) = \frac{y}{Z}$$

Fitting Likelihood-Based Models



Model Distribution



Intractable!

Fitting Likelihood-Based Models

$$p(\mathbf{x}) = \frac{y}{Z}$$

$$\log p(\mathbf{x}) = \log y - \log Z$$

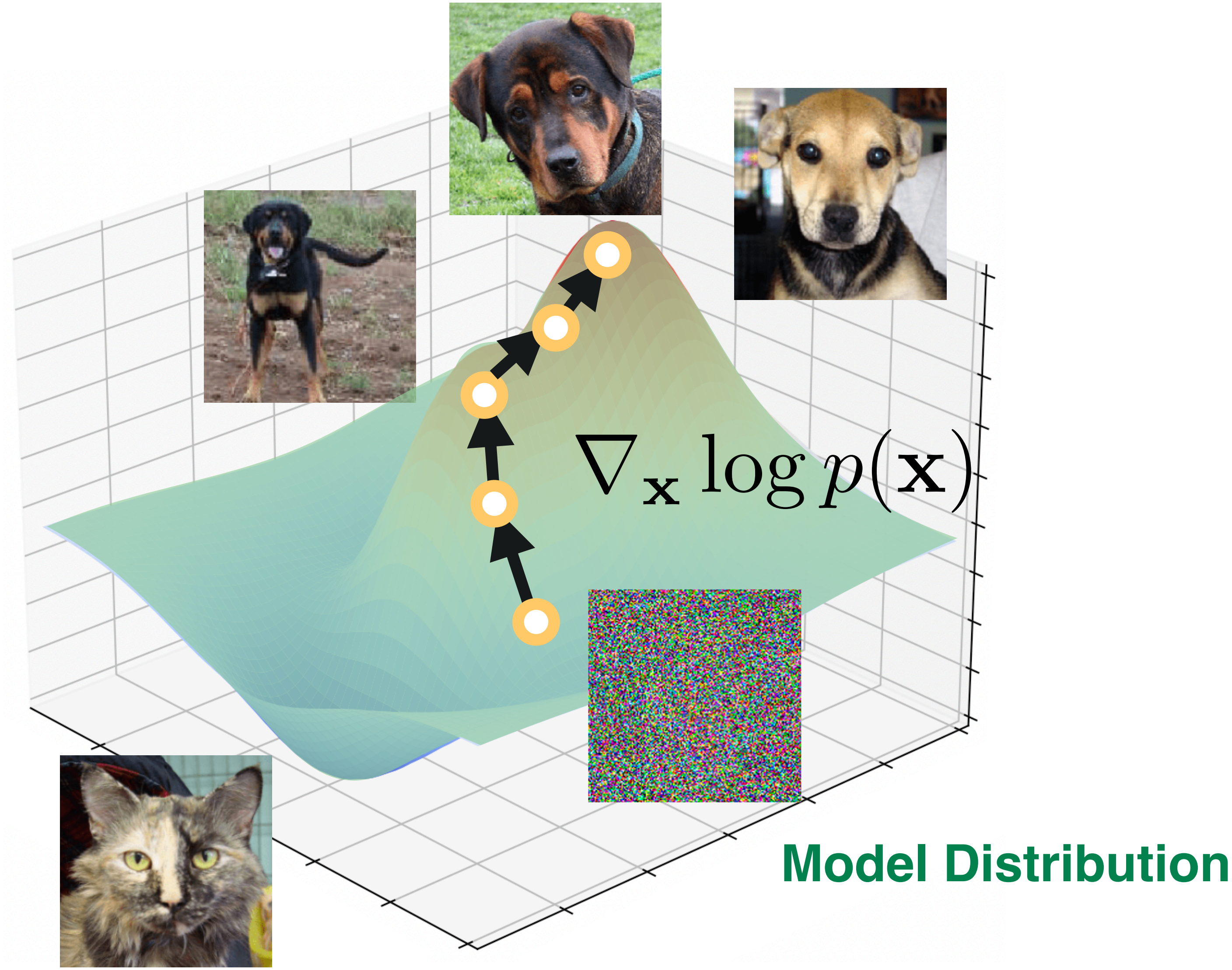
0!

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}) = \nabla_{\mathbf{x}} \log y - \cancel{\nabla_{\mathbf{x}} \log Z}$$

Fit this instead

Machine Learning Model

Fitting Likelihood-Based Models



Why is this called diffusion model?

Forward Diffusion Process

1. Sample a random noise image $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

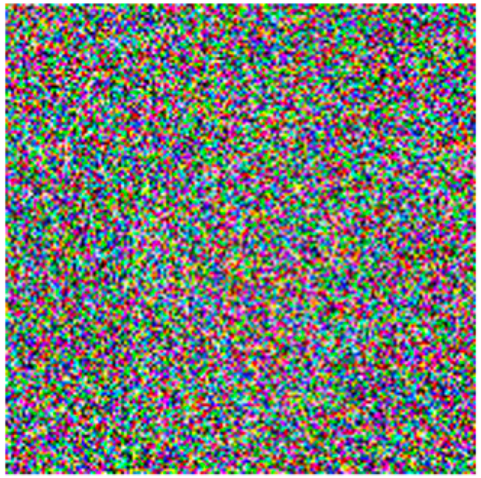
$\mathbf{x}^0$ ϵ $\mathbf{x}^T$

2. Add noise by blending. *[This is a designed procedure/a schedule]*

Reverse Diffusion Process

1. Sample a random noise image $\mathbf{x}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

How do we get this clean image?



$\mathbf{x}^T$

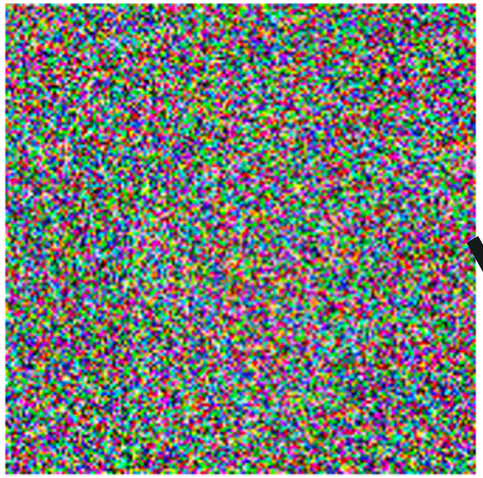
$\longrightarrow$

$\mathbf{x}^0$

Reverse Diffusion Process

1. Sample a random noise image $\mathbf{x}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

How do we get this clean image?



$\mathbf{x}^T$

$\epsilon_\theta(\mathbf{x}^T, T)$

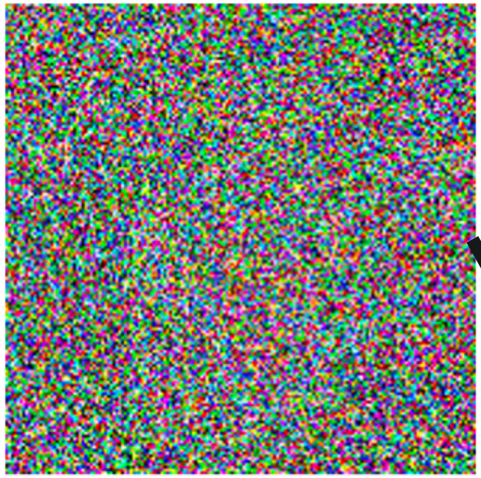
Machine Learning Model

$\mathbf{x}^0$

Reverse Diffusion Process

1. Sample a random noise image $\mathbf{x}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

How do we get this clean image?

$\sqrt{\alpha_t}$  $-\ \sqrt{1-\alpha_t}\ \epsilon_\theta(\mathbf{x}^T, T) =$ 

$\mathbf{x}^T$ $\mathbf{x}^{T-1}$

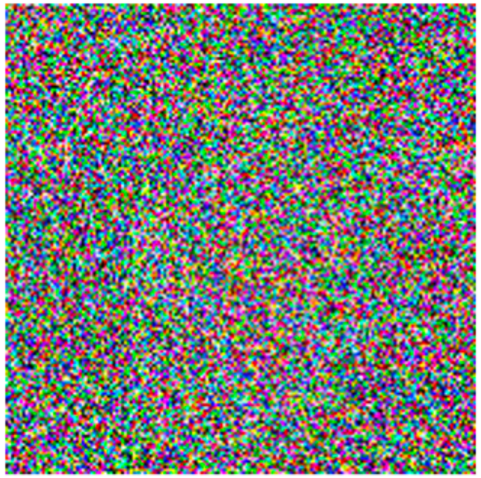
Machine Learning Model

* Note there are extra weighting terms

Reverse Diffusion Process

1. Sample a random noise image $\mathbf{x}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

2. Denoise



$\mathbf{x}^T$

$\mathbf{x}^{T-1}$

Reverse Diffusion Process

1. Sample a random noise image $\mathbf{x}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

2. Denoise

$\mathbf{x}^{T-1}$

$\longrightarrow$

$\mathbf{x}^{T-2}$

Reverse Diffusion Process

1. Sample a random noise image $\mathbf{x}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

2. Denoise

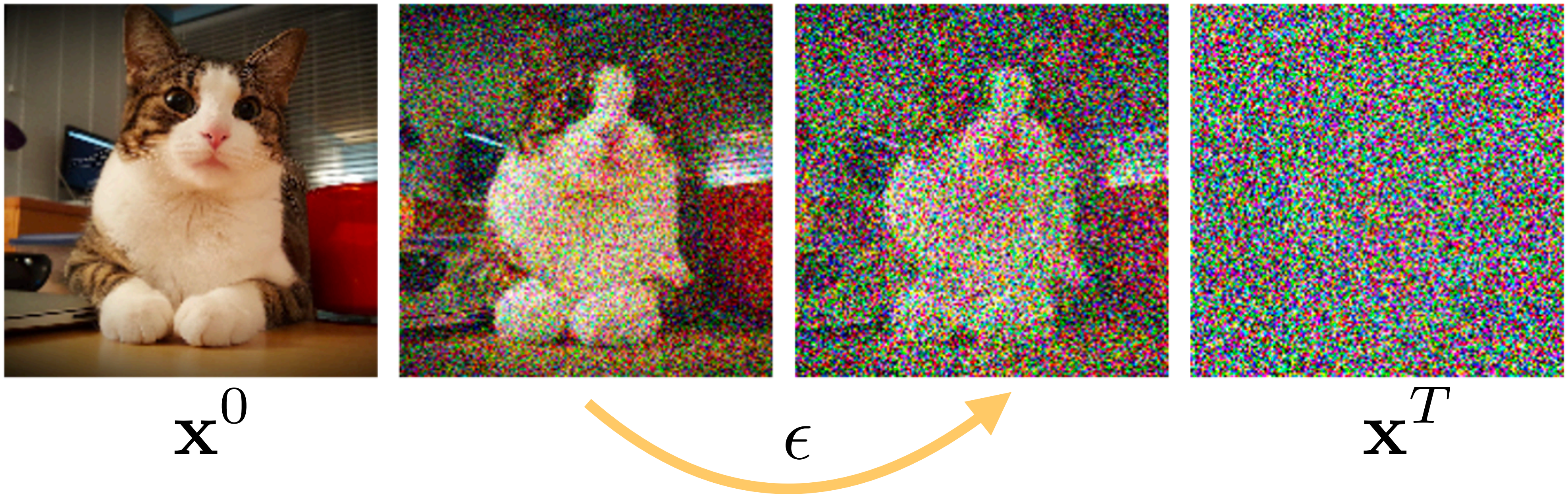
$\mathbf{x}^{T-2}$

$\longrightarrow$

$\mathbf{x}^0$

Model Fitting

Machine Learning Model?



$$\mathbb{E}_{t,\mathbf{x}^0,\epsilon_t}||\epsilon_\theta(\mathbf{x}^t, t) - \epsilon_t||^2$$

Connection between Two Definitions

Want to fit $\nabla_{\mathbf{x}} \log p(\mathbf{x})$
with $s_\theta(\mathbf{x})$

Connection between Two Definitions

Want to fit $\nabla_{\mathbf{x}} \log p(\mathbf{x})$
with $s_\theta(\mathbf{x})$

$p(\mathbf{x})$ is unknown!

Connection between Two Definitions

Want to fit $\nabla_{\mathbf{x}} \log p(\mathbf{x})$

with $s_\theta(\mathbf{x})$

$p(\mathbf{x})$ is unknown!

$$\mathbf{x} \xrightarrow{q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})} \tilde{\mathbf{x}}$$

Connection between Two Definitions

Want to fit $\nabla_{\mathbf{x}} \log p(\mathbf{x})$
with $s_\theta(\mathbf{x})$

$p(\mathbf{x})$ is unknown!

$\mathbf{x} \xrightarrow{q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})} \tilde{\mathbf{x}}$

$$\mathbb{E}_{\tilde{\mathbf{x}},\mathbf{x}} ||s_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x})||^2$$

Connection between Two Definitions

Want to fit $\nabla_{\mathbf{x}} \log p(\mathbf{x})$
with $s_\theta(\mathbf{x})$

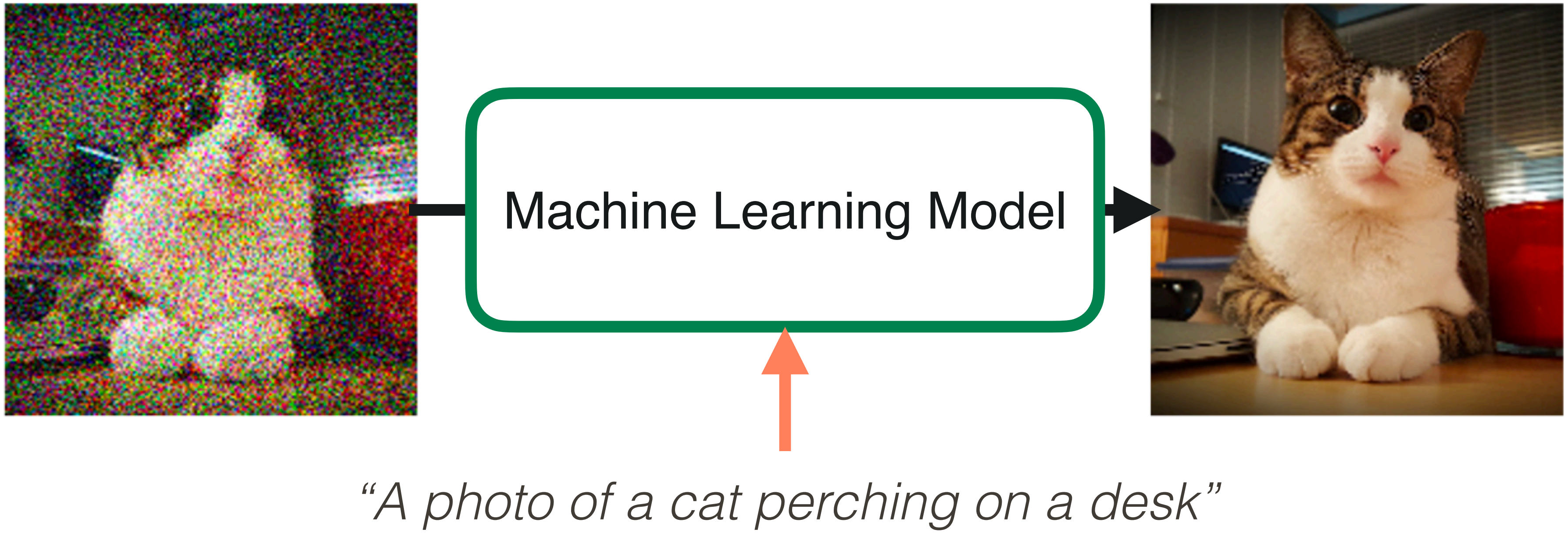
$p(\mathbf{x})$ is unknown!

$$\mathbf{x} \xrightarrow{q_\sigma(\tilde{\mathbf{x}}|\mathbf{x})} \tilde{\mathbf{x}}$$

$$\mathbb{E}_{\tilde{\mathbf{x}},\mathbf{x}} || s_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x}) ||^2$$

$$\mathbb{E}_{t,\mathbf{x}^0,\epsilon_t} || \epsilon_\theta(\mathbf{x}^t, t) - \epsilon_t ||^2$$

Conditional v.s Unconditional

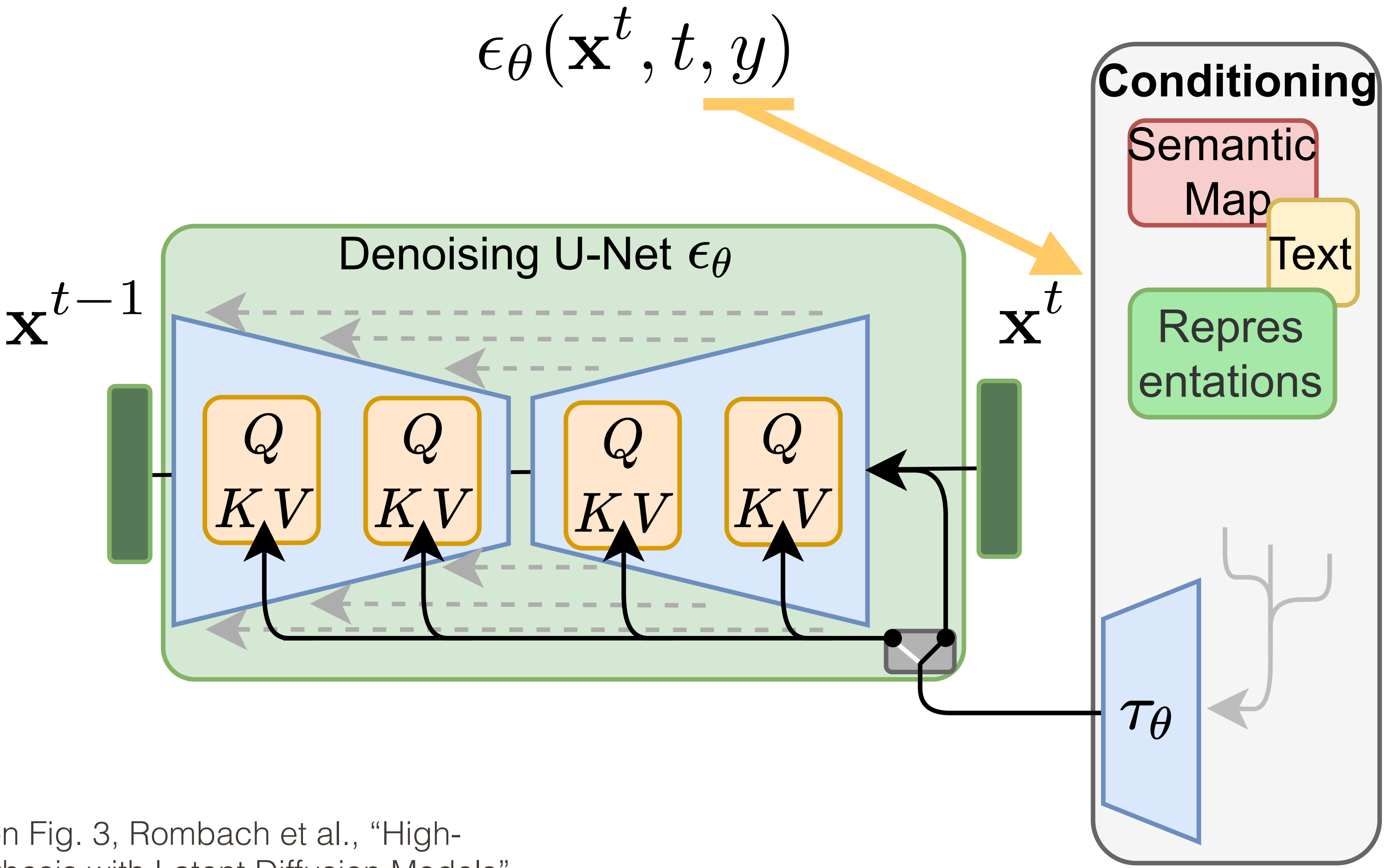


- More control for users.

  In contrast, unconditional generative model would need to use random seeds to control the output.

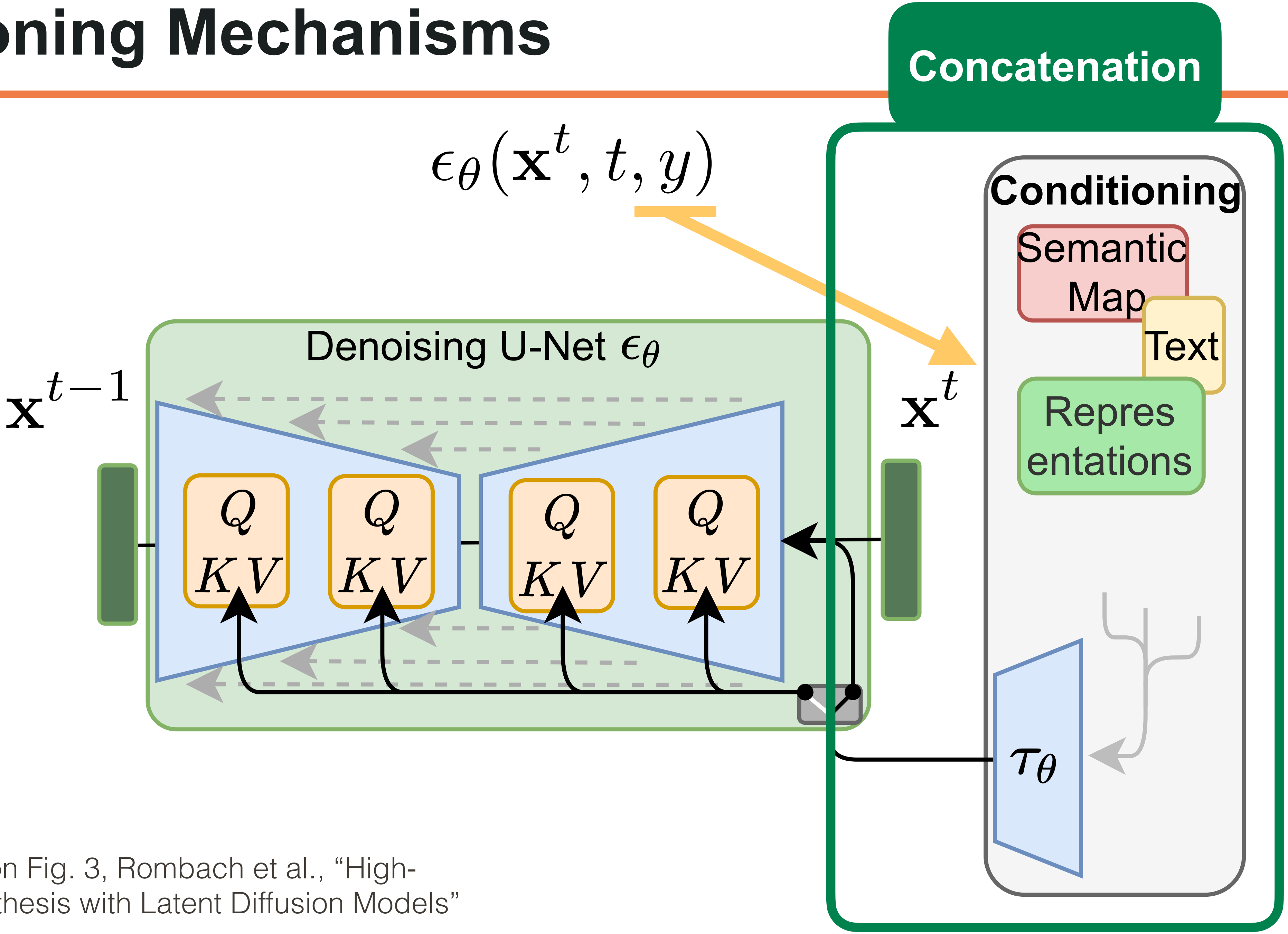
- Empirically, conditional generative models are **easier to train and perform better** than unconditional ones.

Conditioning Mechanisms



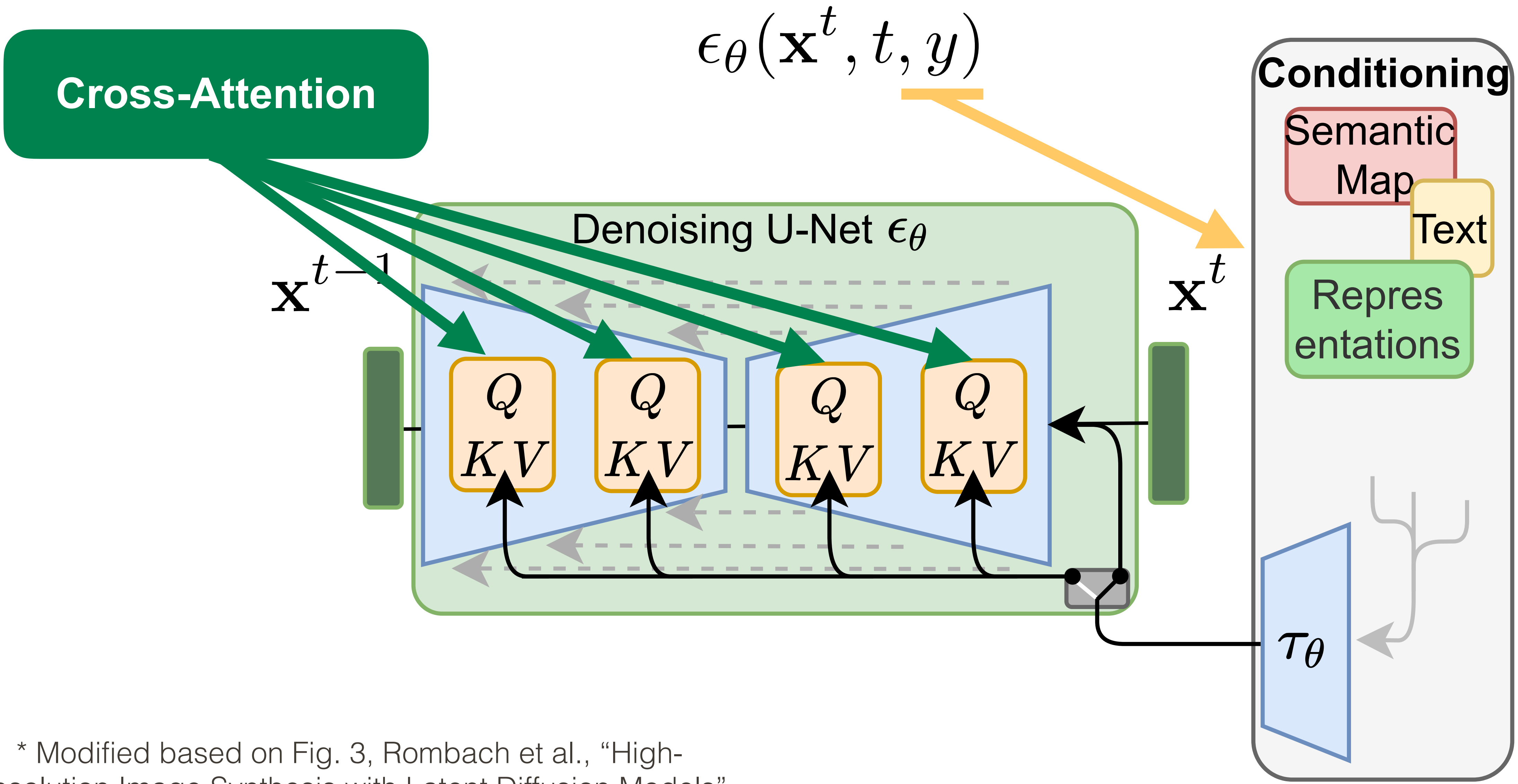
* Modified based on Fig. 3, Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models"

Conditioning Mechanisms



* Modified based on Fig. 3, Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models"

Conditioning Mechanisms



* Modified based on Fig. 3, Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models"

Cross-Attention

Q: Query	K: Key	V: Value
"$\mathbf{x}^t$"	"y"	"y"

Cross-Attention

Q: Query K: Key V: Value

$``\mathbf{x}^t"$ $``y"$ $``y"$

$$\mathrm{softmax}(\frac{QK^T}{\sqrt{d}})V$$

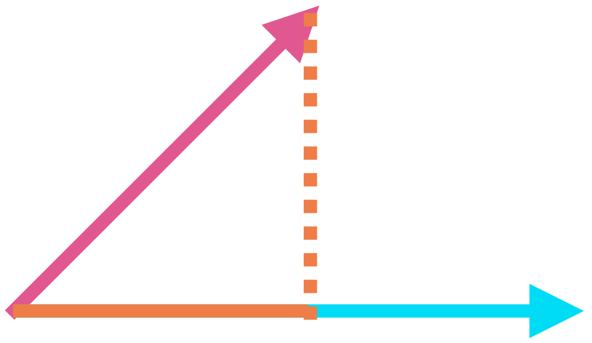
Cross-Attention

Q: Query K: Key V: Value

"$\mathbf{x}^t$" "y" "y"

Similarity Score

$$\mathrm{softmax}\left(\frac{\boxed{QK^T}}{\sqrt{d}}\right)V$$



Cross-Attention

Q: Query K: Key V: Value

“$\mathbf{x}^t$” “y” “y”

Normalization

$$\boxed{\mathrm{softmax}(}\frac{QK^T}{\sqrt{d}})V$$

Cross-Attention

Q: Query

“$\mathbf{x}^t$”

K: Key

“y”

V: Value

“y”

$$\mathrm{softmax}(\frac{QK^T}{\sqrt{d}})\boxed{V}$$

Condition is kept in places where **condition and image is similar**.

Cross-Attention

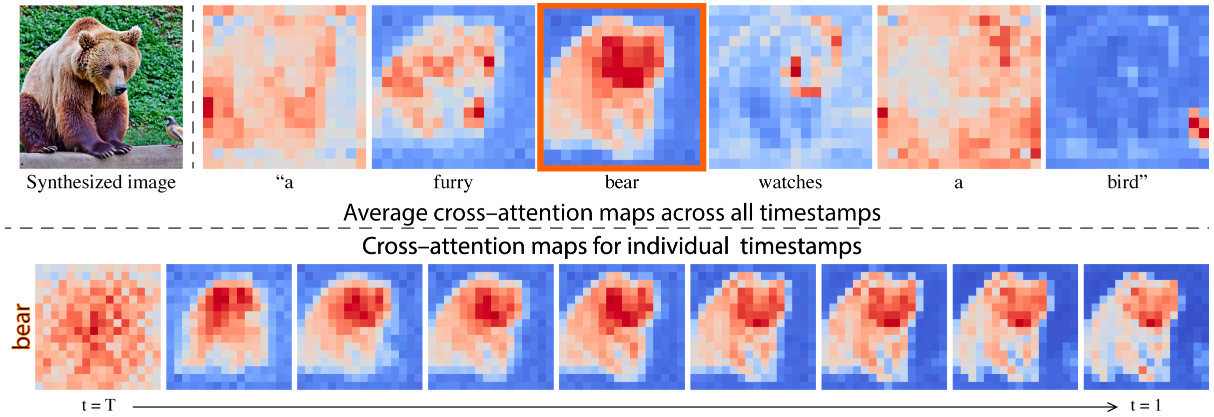
Q: Query	K: Key	V: Value
"$\mathbf{x}^t$"	"y"	"y"

$$\text{"}\mathbf{x}^t\text{"} + \mathrm{softmax}(\frac{QK^T}{\sqrt{d}})V$$

Cross-Attention

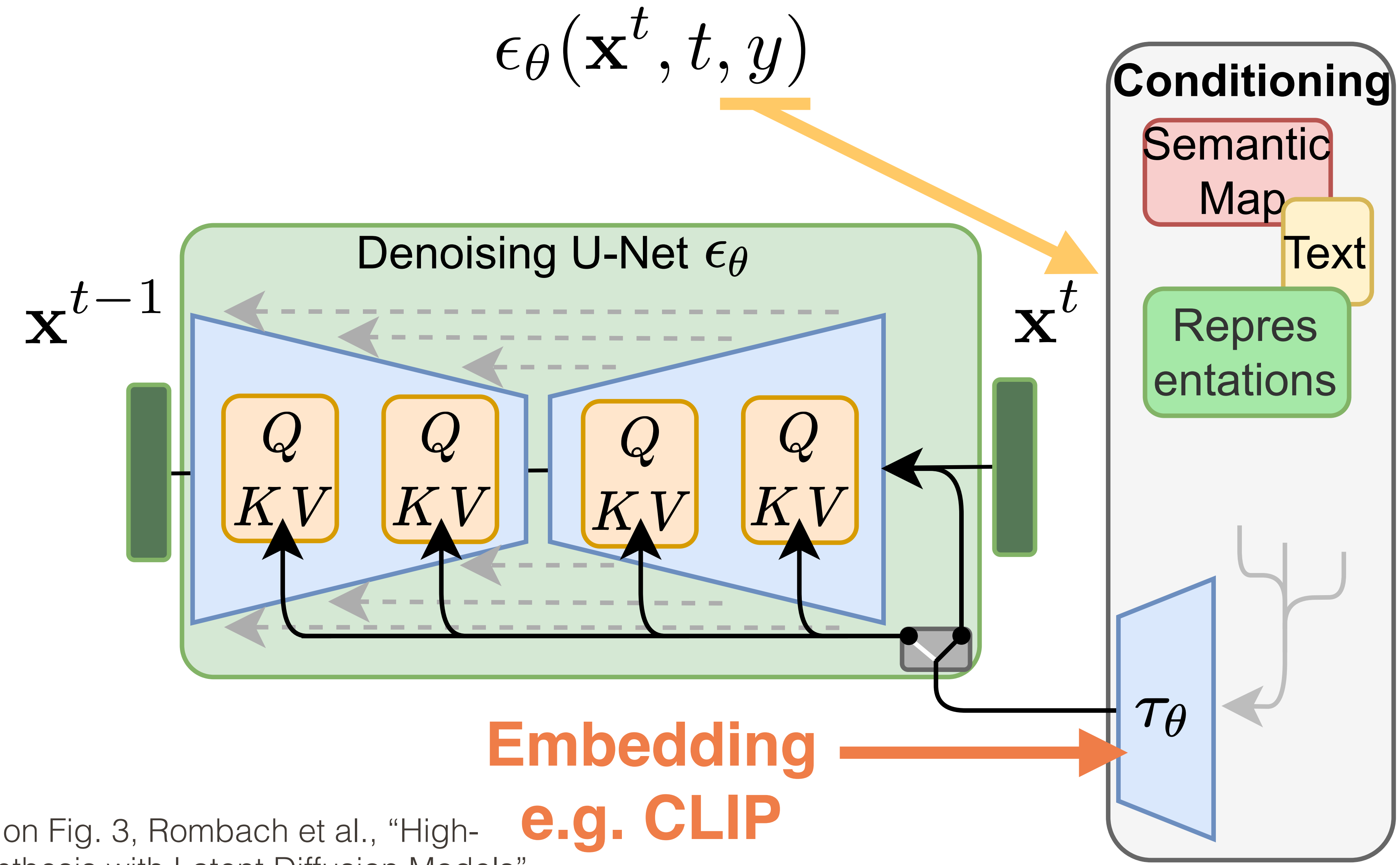
Figure 1: The original synthesized image and three DAAM maps for "monkey," "hat," and "walking," from the prompt, "monkey with hat walking."

Tang et al., "What the DAAM: Interpreting Stable Diffusion Using Cross Attention"



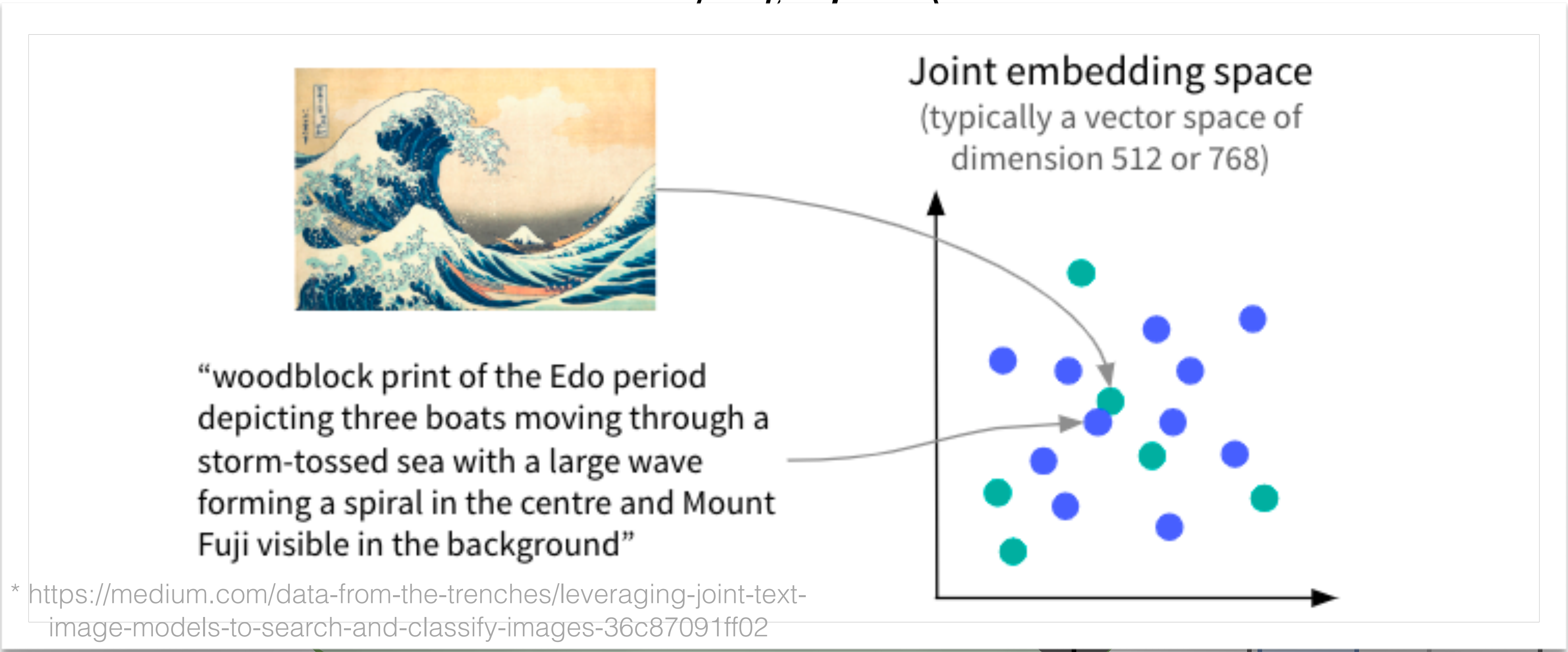
Hertz et al., "Prompt-to-Prompt Image Editing with Cross-Attention Control"

Text Embedding

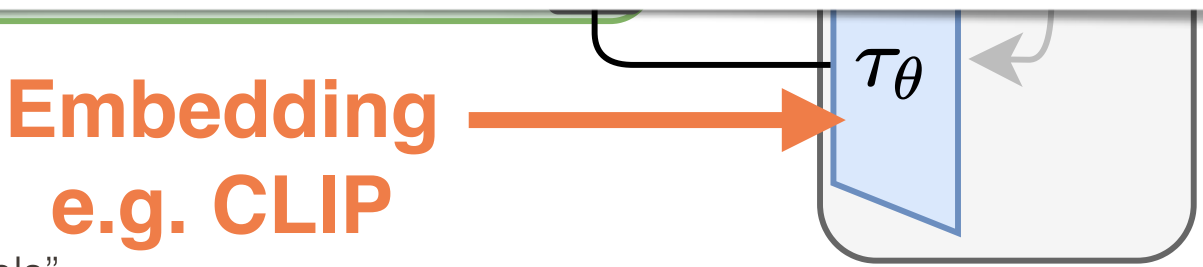


* Modified based on Fig. 3, Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models"

Text Embedding



* https://medium.com/data-from-the-trenches/leveraging-joint-text-image-models-to-search-and-classify-images-36c87091ff02

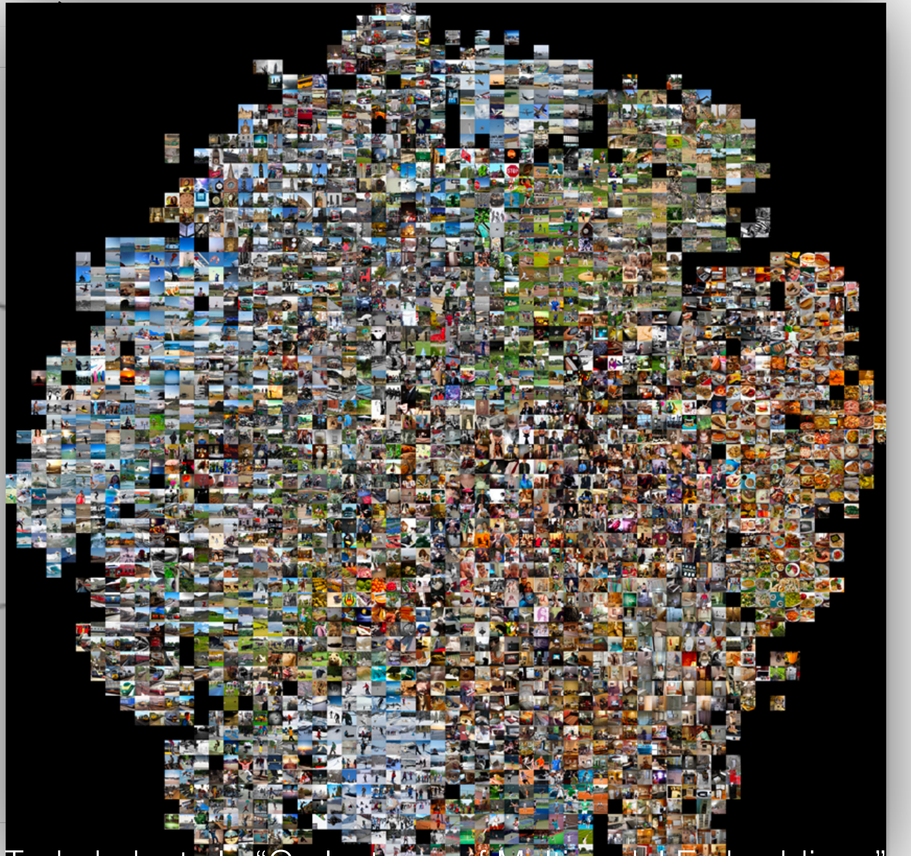


* Modified based on Fig. 3, Rombach et al., “High-Resolution Image Synthesis with Latent Diffusion Models”

Text Embedding

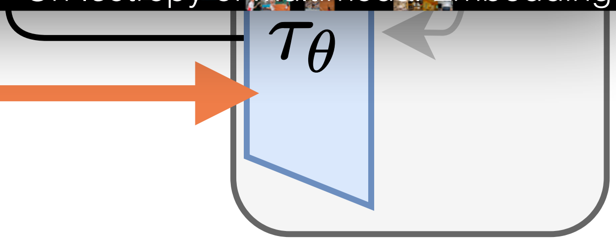
"woodblock print of the Edo period depicting three boats moving through a storm-tossed sea with a large wave forming a spiral in the centre and Mount Fuji visible in the background"

* https://medium.com/data-from-the-trenches/leveraging-joint-text-image-models-to-search-and-classify-images-36c87091ff02



Tyshchuk et al., "On Isotropy of Multimodal Embeddings"

Embedding
e.g. CLIP



* Modified based on Fig. 3, Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models"

Questions?

Advanced Diffusion-Model-Based Editing Tools

Image-to-Image Methods

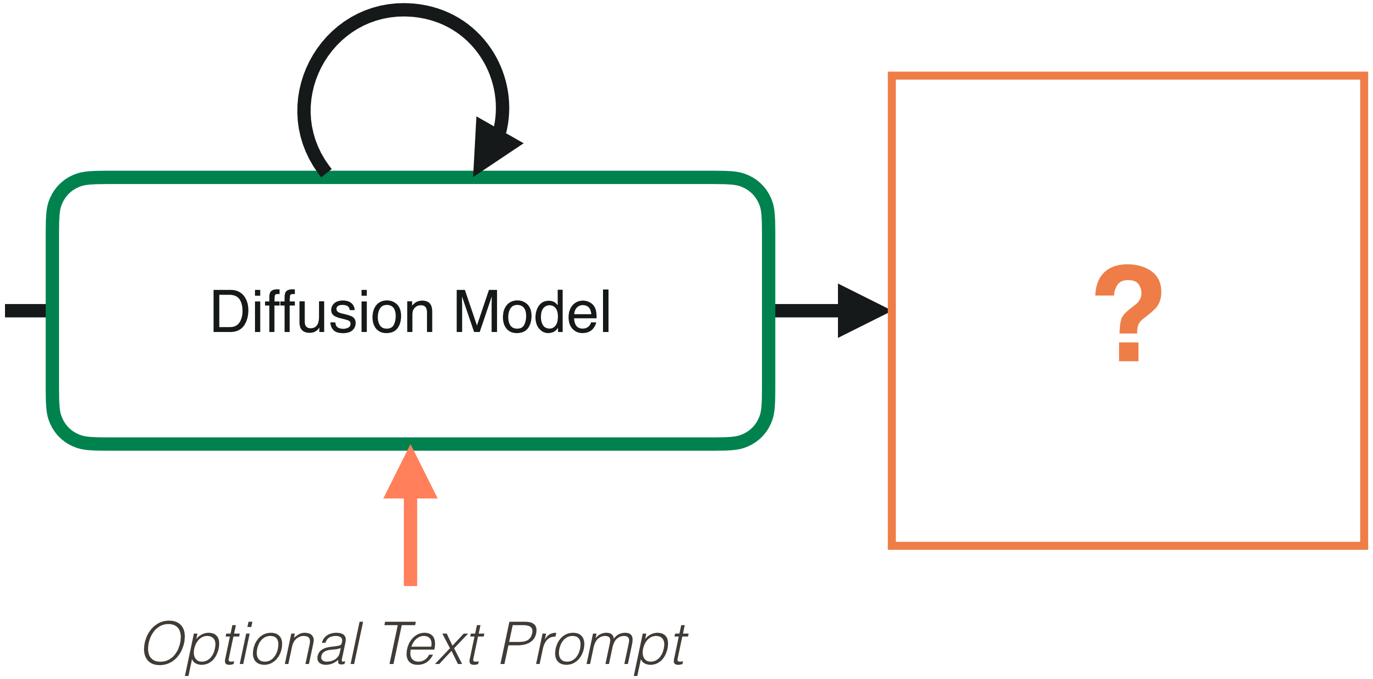


Image-to-Image Methods: Image Variances

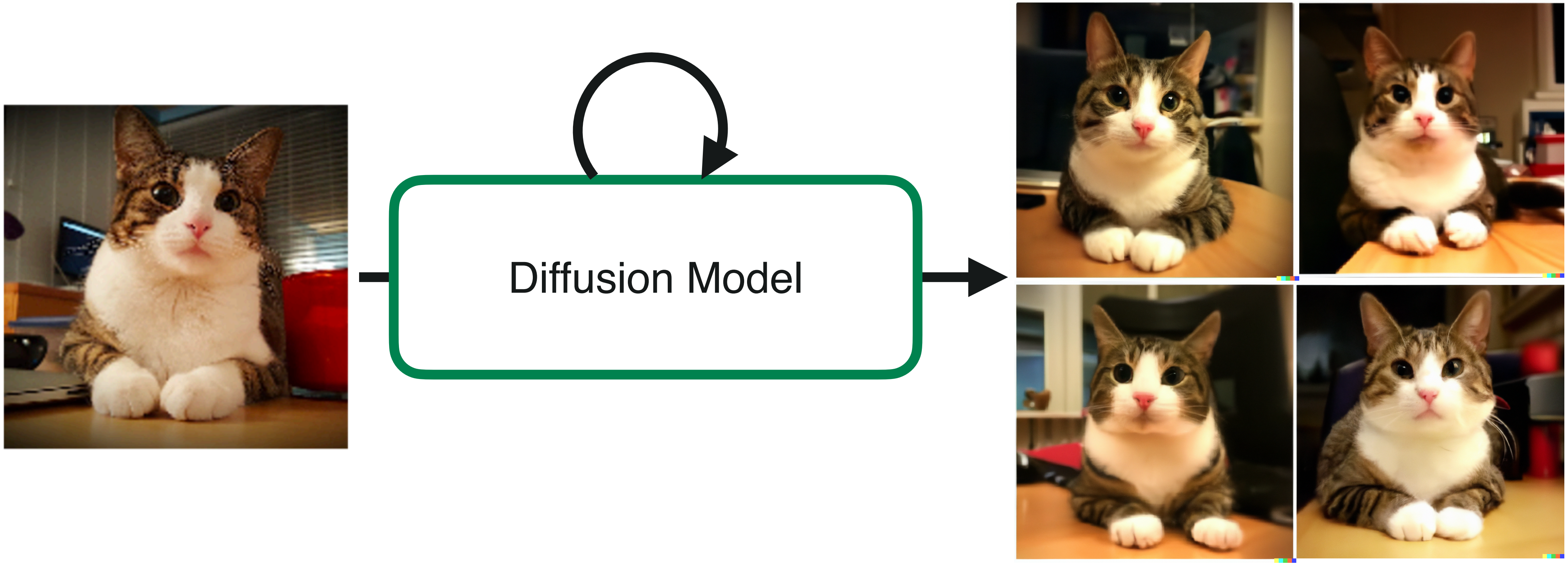


Image-to-Image Methods: Image Variances

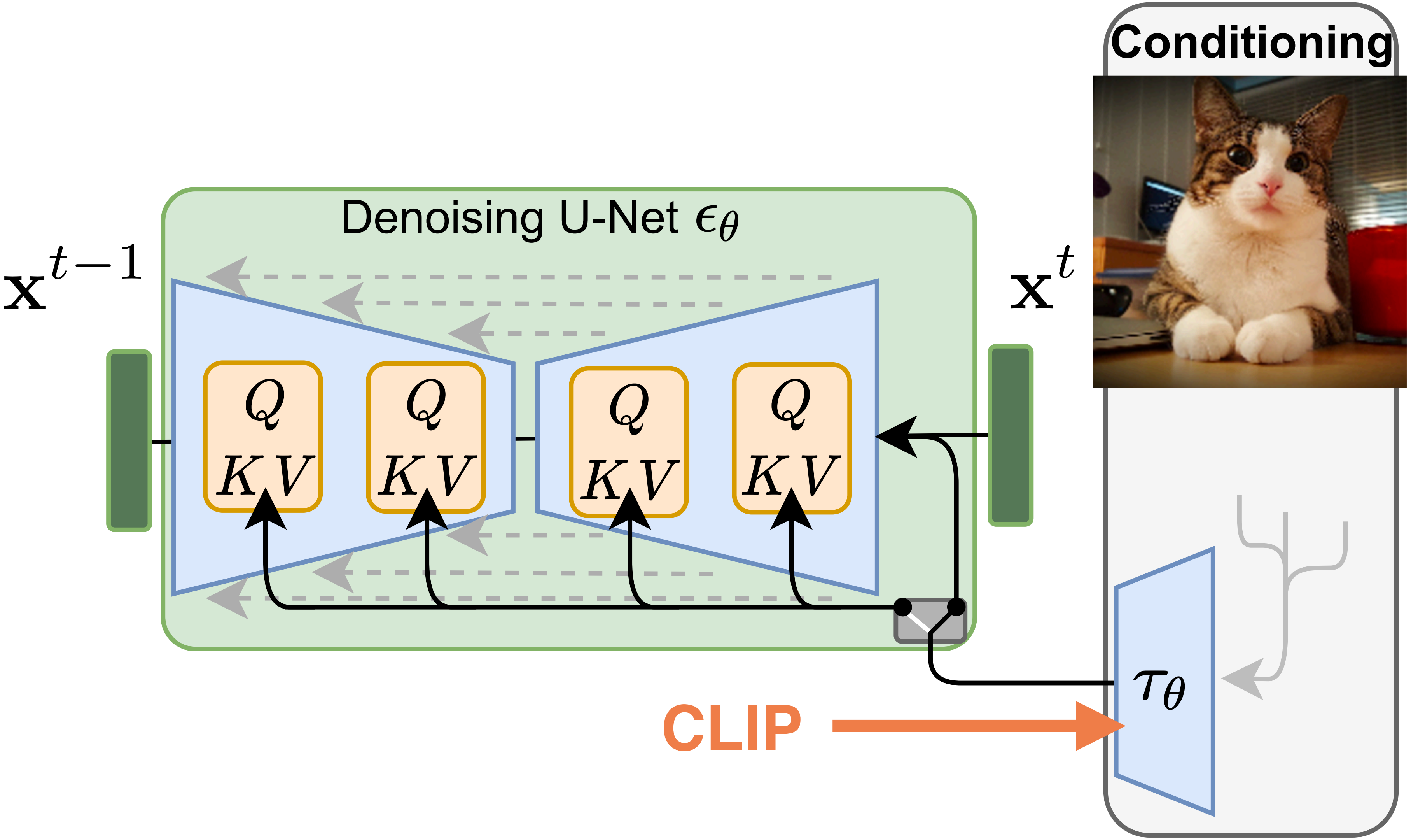


Image-to-Image Methods: ControlNet

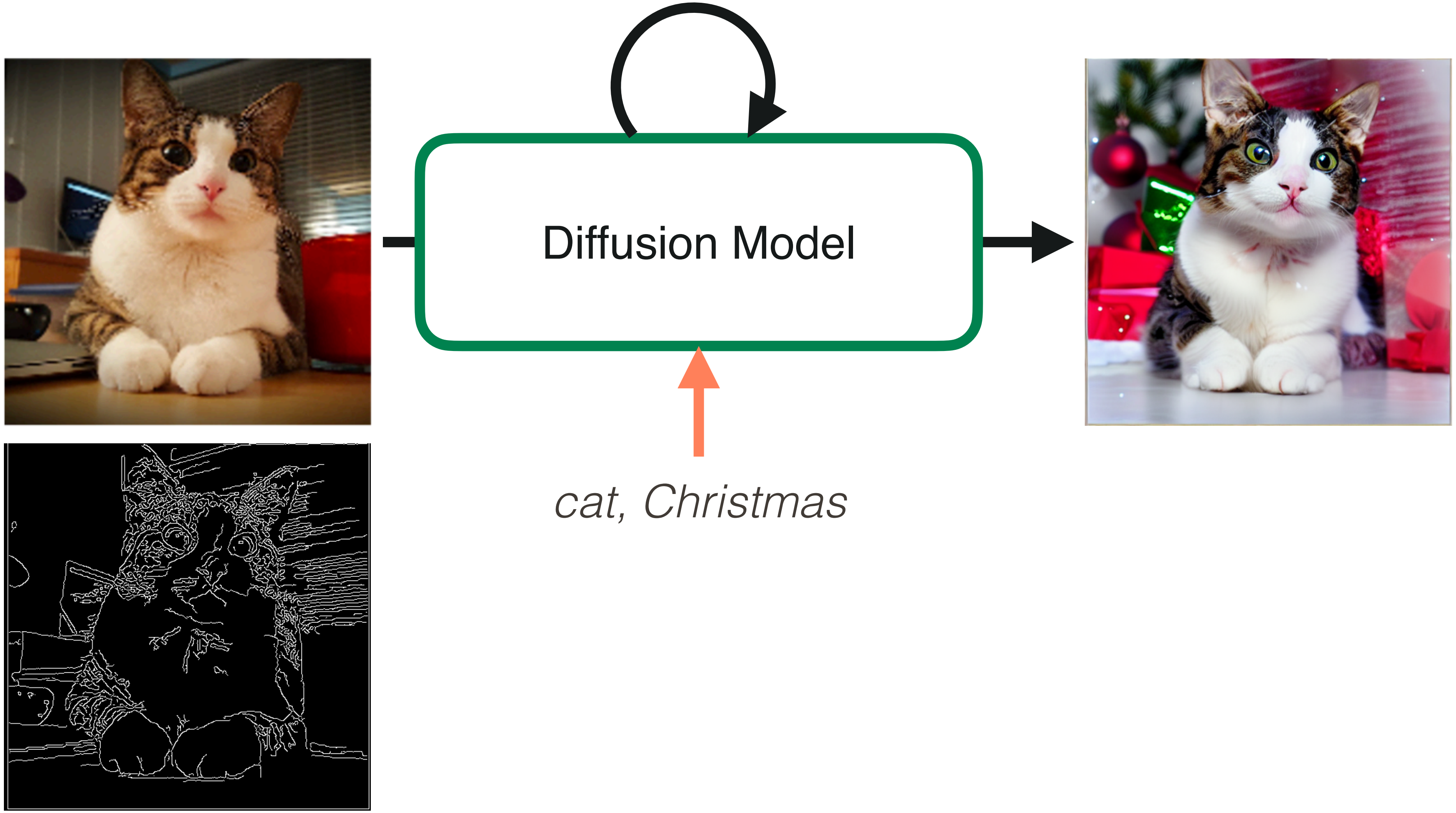


Image-to-Image Methods: Inpaint/Outpainting

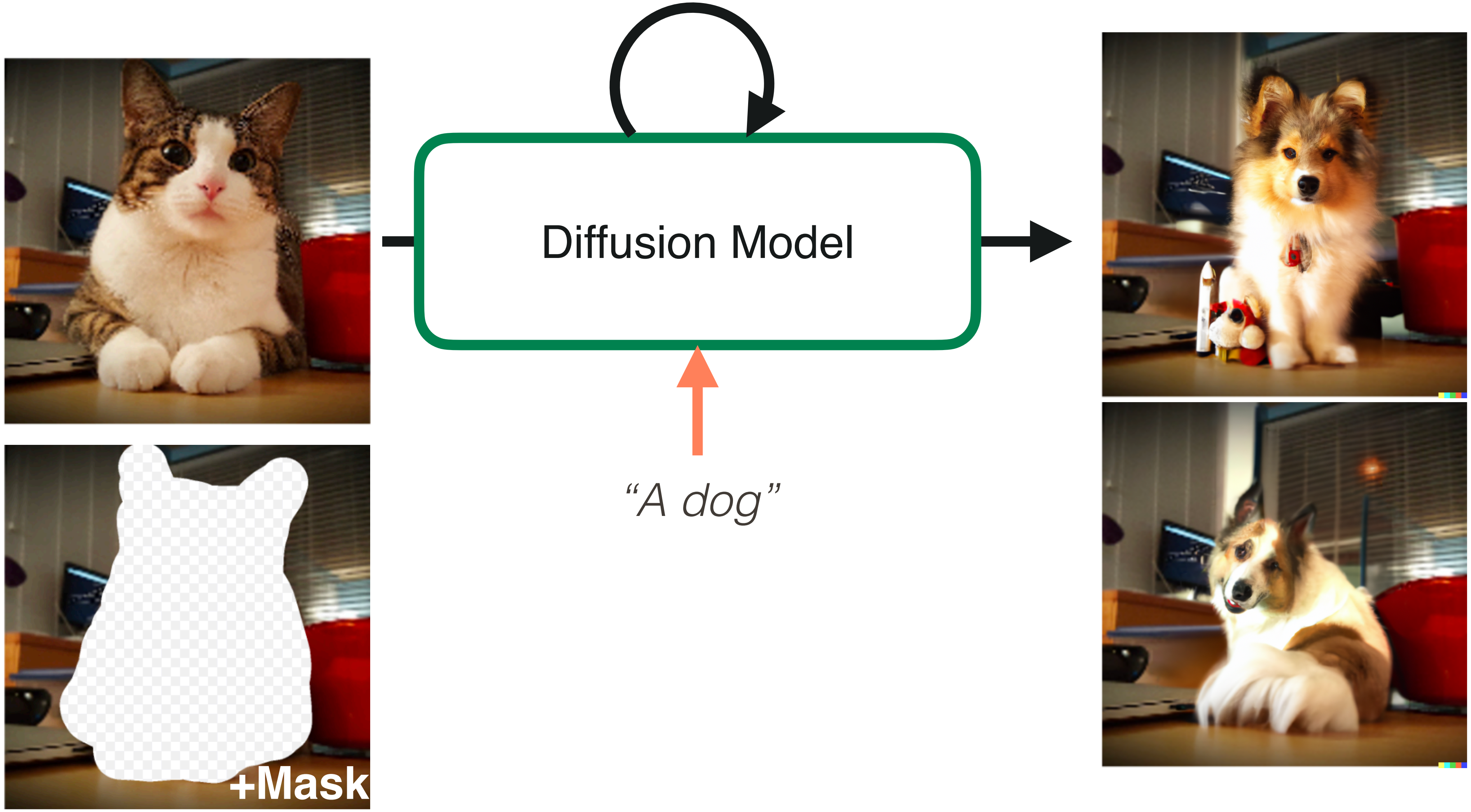
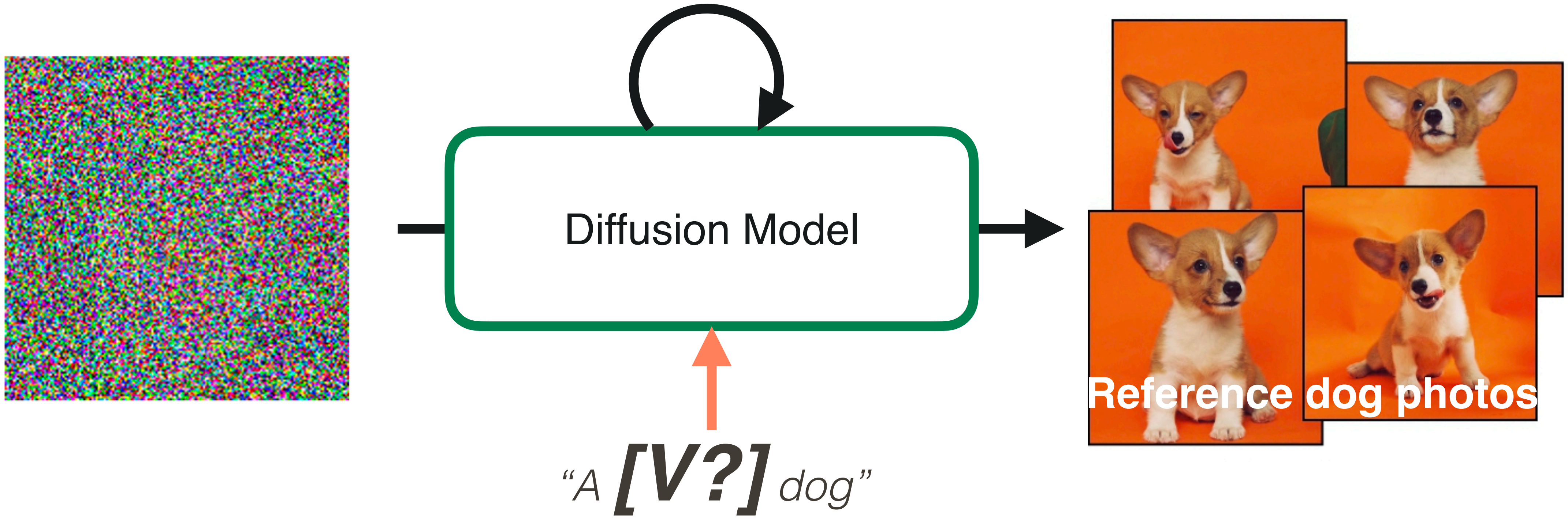
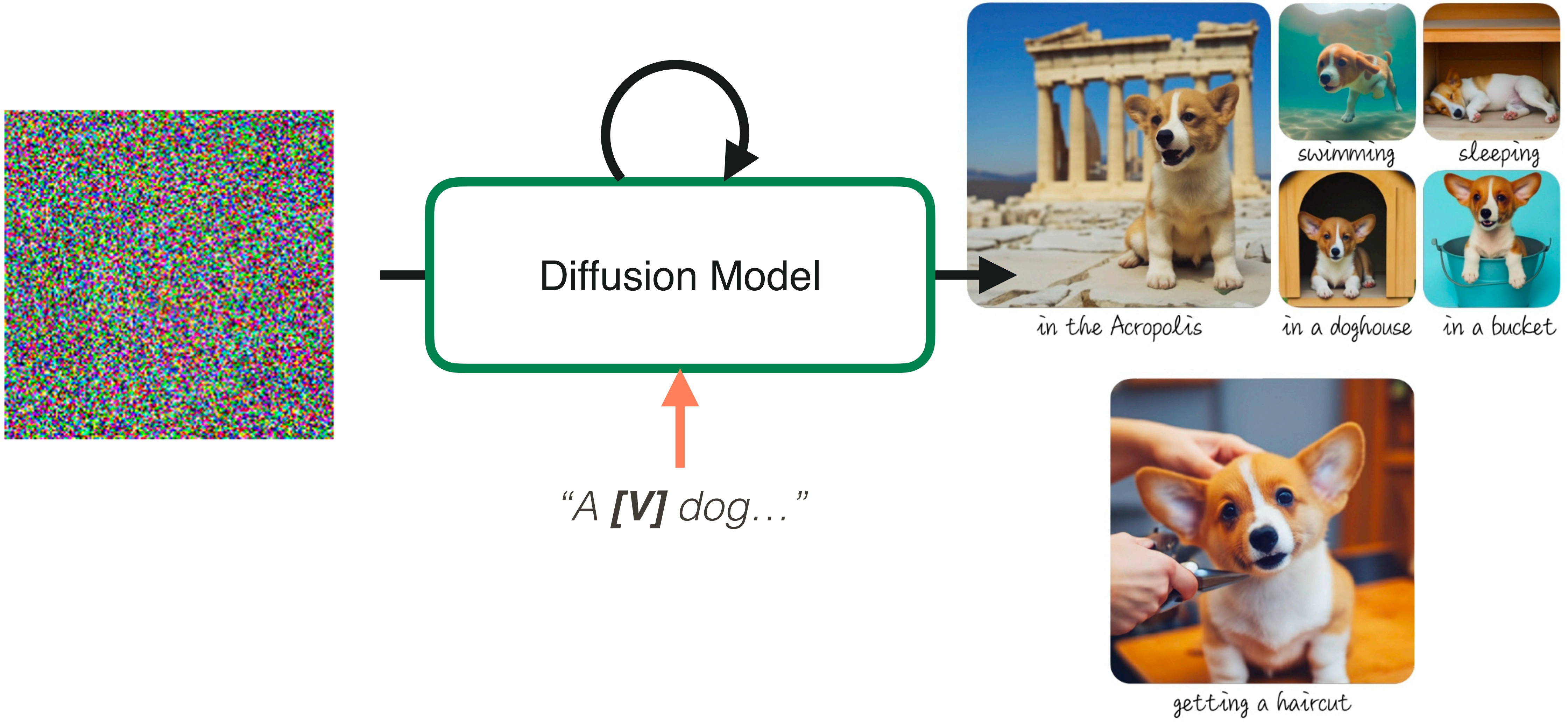


Image-to-Image Methods: Identity



* Based on Fig. 1, Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation"

Image-to-Image Methods: Identity



* Based on Fig. 1, Ruiz et al., "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation"

Strengths & Weaknesses

What it can do

1. Quick generation of complex, high-quality realistic and artistic images; good for creative exploration

From Aaron Hertzmann's blog

"cats in devo hats"

What it can do

2. Integrating specific styles

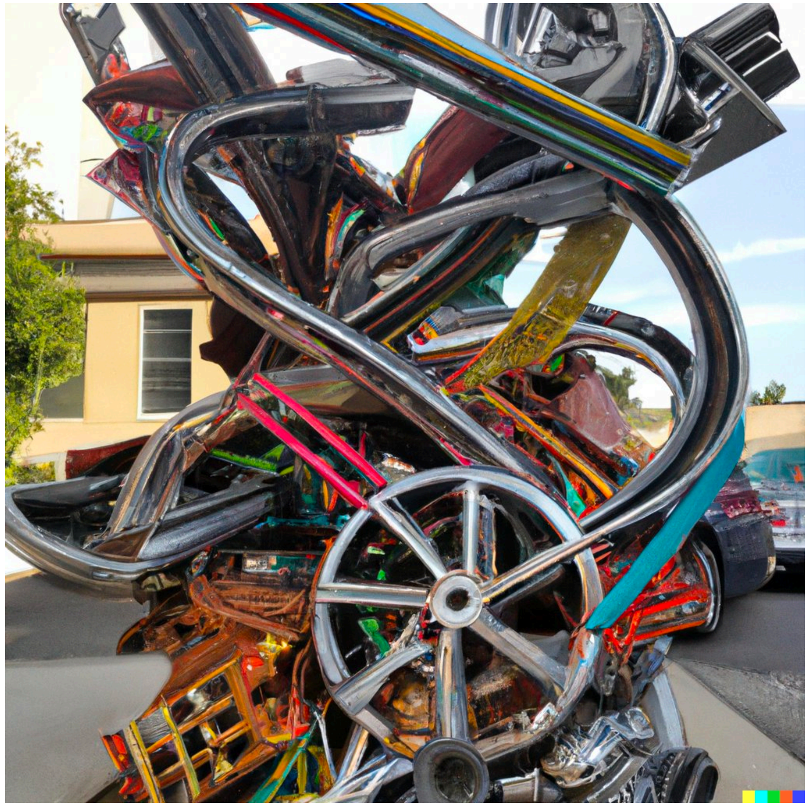
From Aaron Hertzmann's blog

"jacob lawrence painting of san francisco"

What it can do

2. Integrating specific styles

From Aaron Hertzmann's blog



"frank stella sculpture made of car parts"

What it can do

2. Integrating specific styles

"a monk riding a snail, medieval illuminated manuscript"

What it can't do

1. Following specific instructions (especially when the scene is complex):
    - Composition

"a young dark-haired boy resting in bed, and a grey-haired older woman sitting in a chair beside the bed underneath a window with sun streaming through, Pixar style digital art"

What it can't do

1. Following specific instructions (especially when the scene is complex):
    - Composition
    - Generating multiple objects
    - Coloring multiple objects

A yellow bowl and a blue cat

Catastrophic Neglect

One or more subjects are not generated

A yellow bow and a brown bench

Incorrect *Attribute Binding*

Attributes (e.g., color) not matched correctly to subject

Chefer et al., "Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models"

What it can’t do

2. Being reasonably unbiased

“lawyer”, April 6, 2022

“DALL·E 2 Preview - Risks and Limitations” by OpenAI

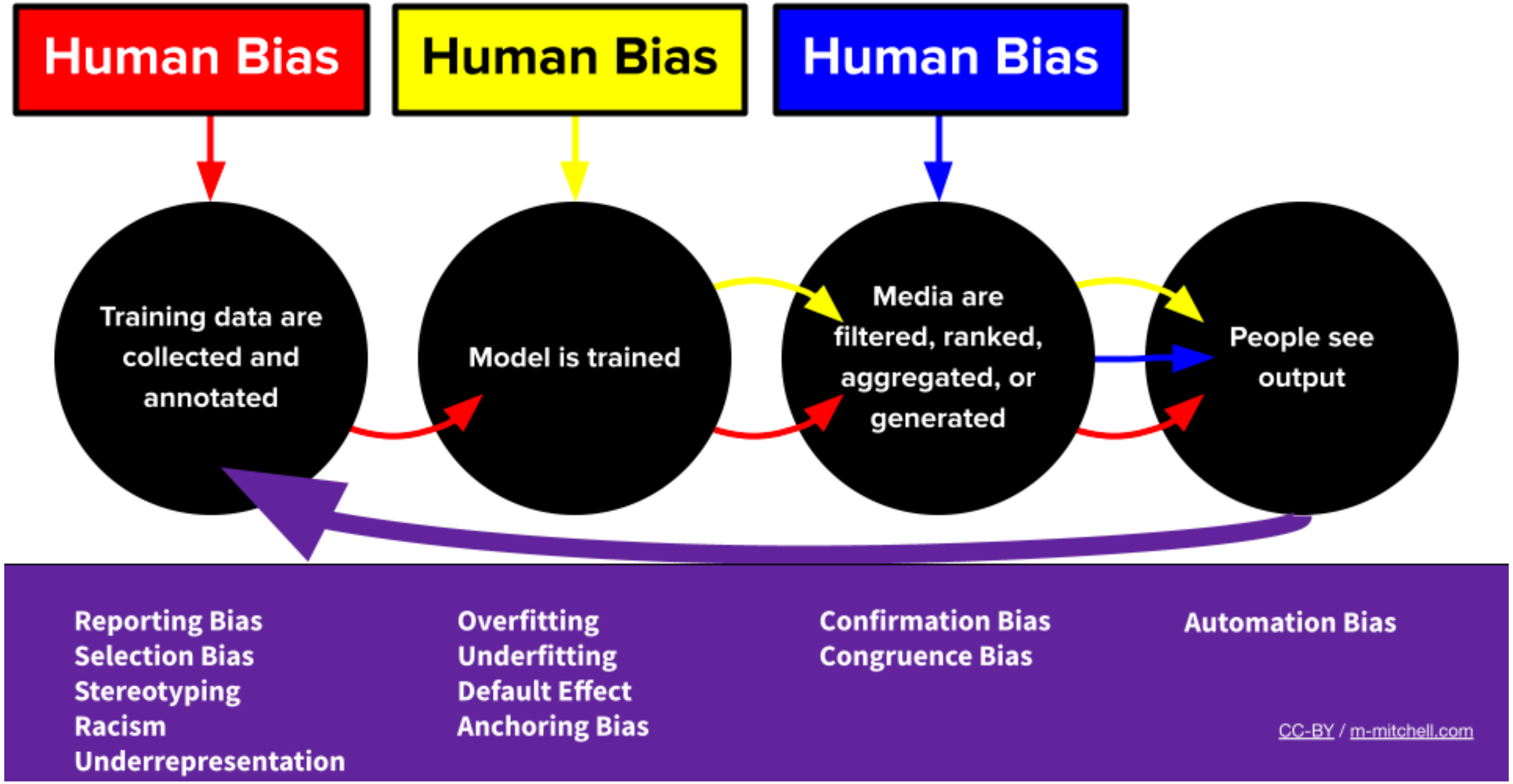
What it can't do

2. Being reasonably unbiased

"nurse", April 6, 2022

"DALL·E 2 Preview - Risks and Limitations" by OpenAI

Where is bias from



The Bias ML Pipeline by Meg
https://huggingface.co/meg

How to be better

Bias can never be fully removed.

Heavily based on “Machine Learning in development: Let's talk about bias!” By HuggingFace

How to be better: Addressing Bias

1. Task definition stage
    - How ML techniques are integrated into the system? Is a ML model biased in a given use case?
    - What is the optimization objective?

How to be better: Addressing Bias

2. Dataset selection and curation stage: A significant source of bias

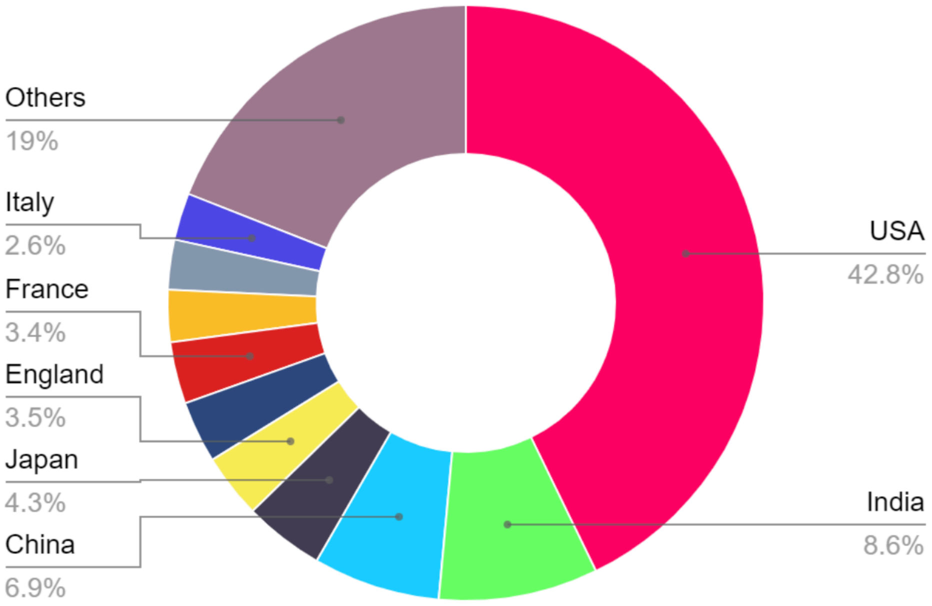


Fig. 1: LAION5B training images: 10 most frequently occurring countries

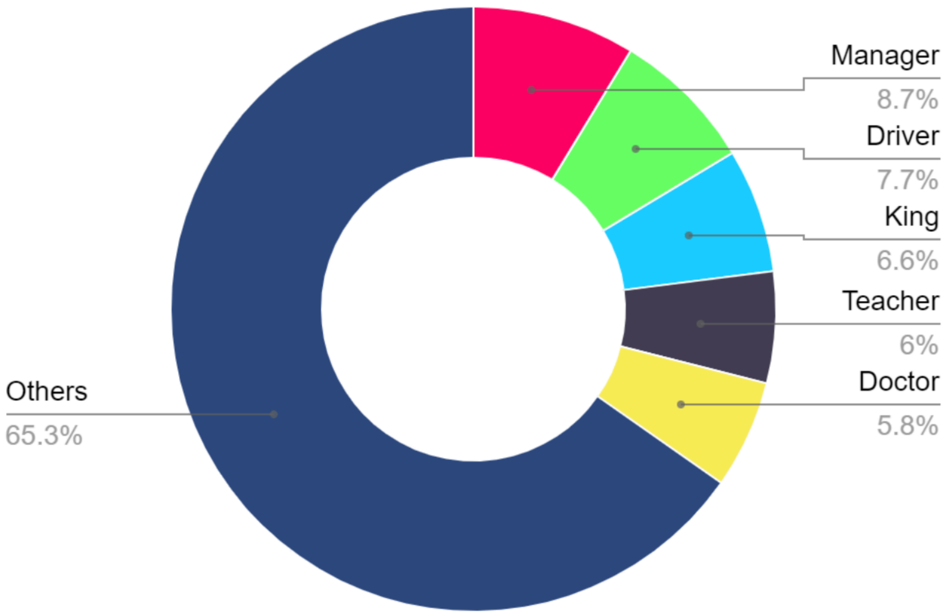


Fig. 2: LAION5B most frequent job titles, showing unusually large numer of monarchs

Cultural bias in AI models: a worked example with Barack Obama by Benjamin Peterson

How to be better: Addressing Bias

2. Dataset selection and curation stage
    - Where is the data from? How was the dataset curated? What is the context?
    - Measure the data. Any harmful associations?
    - Document the dataset.
    - Choose the dataset with least bias related harm. Iteratively improve the dataset.

How to be better: Addressing Bias

3. Model selection and training stage
   - Visualize model outputs.
   - Evaluate against benchmark.
   - Document the model.

Generative Models & Artists

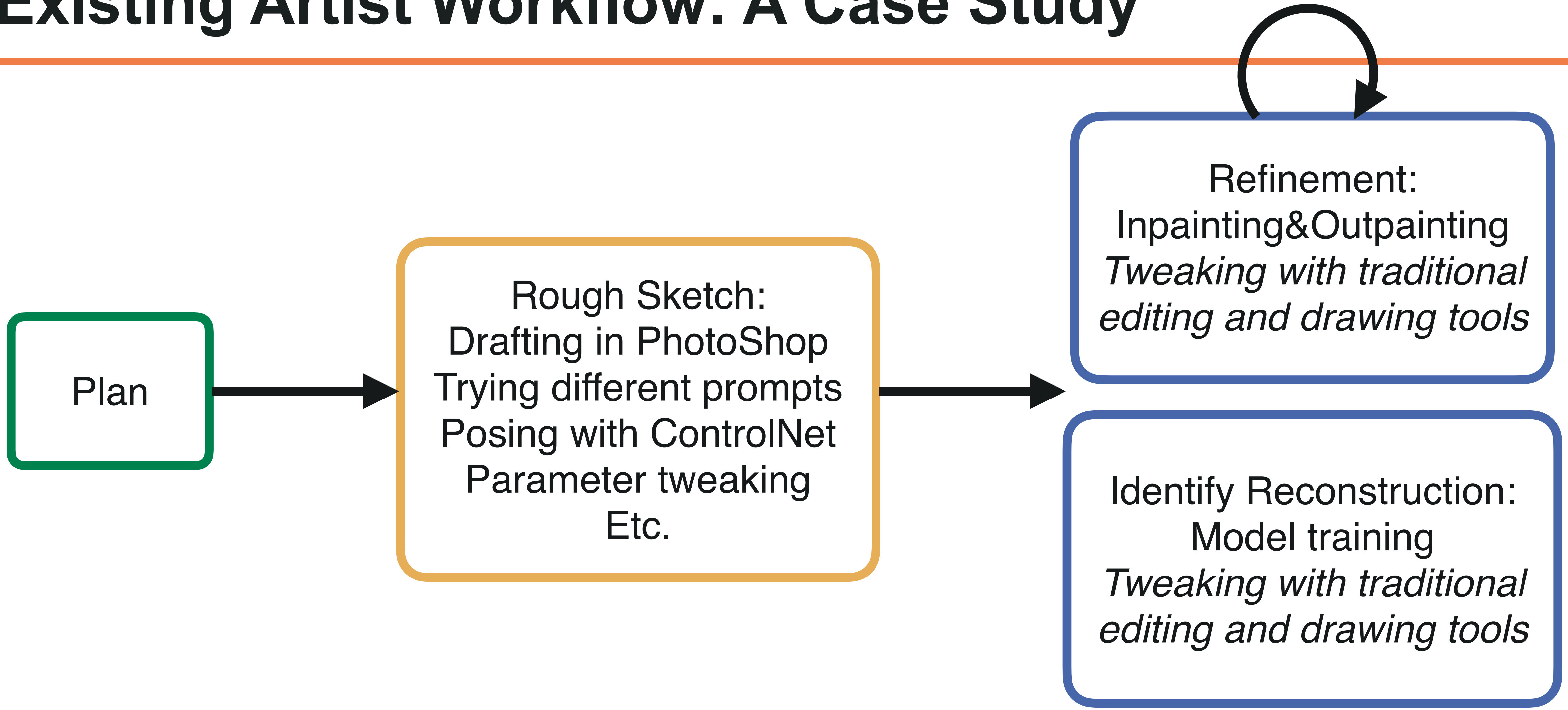
This is a fast evolving topic with many debats and open questions.

Warning: Contents may no longer be the state-of-art or relevant; and nothing presented should be taken as the “fact”.

Existing Artist Workflow: A Case Study

"An AI artist explains his workflow"
https://www.youtube.com/watch?v=K0ldxCh3cnI

Existing Artist Workflow: A Case Study



"An AI artist explains his workflow"

https://www.youtube.com/watch?v=K0ldxCh3cnI

Open Questions: How to protect artists?

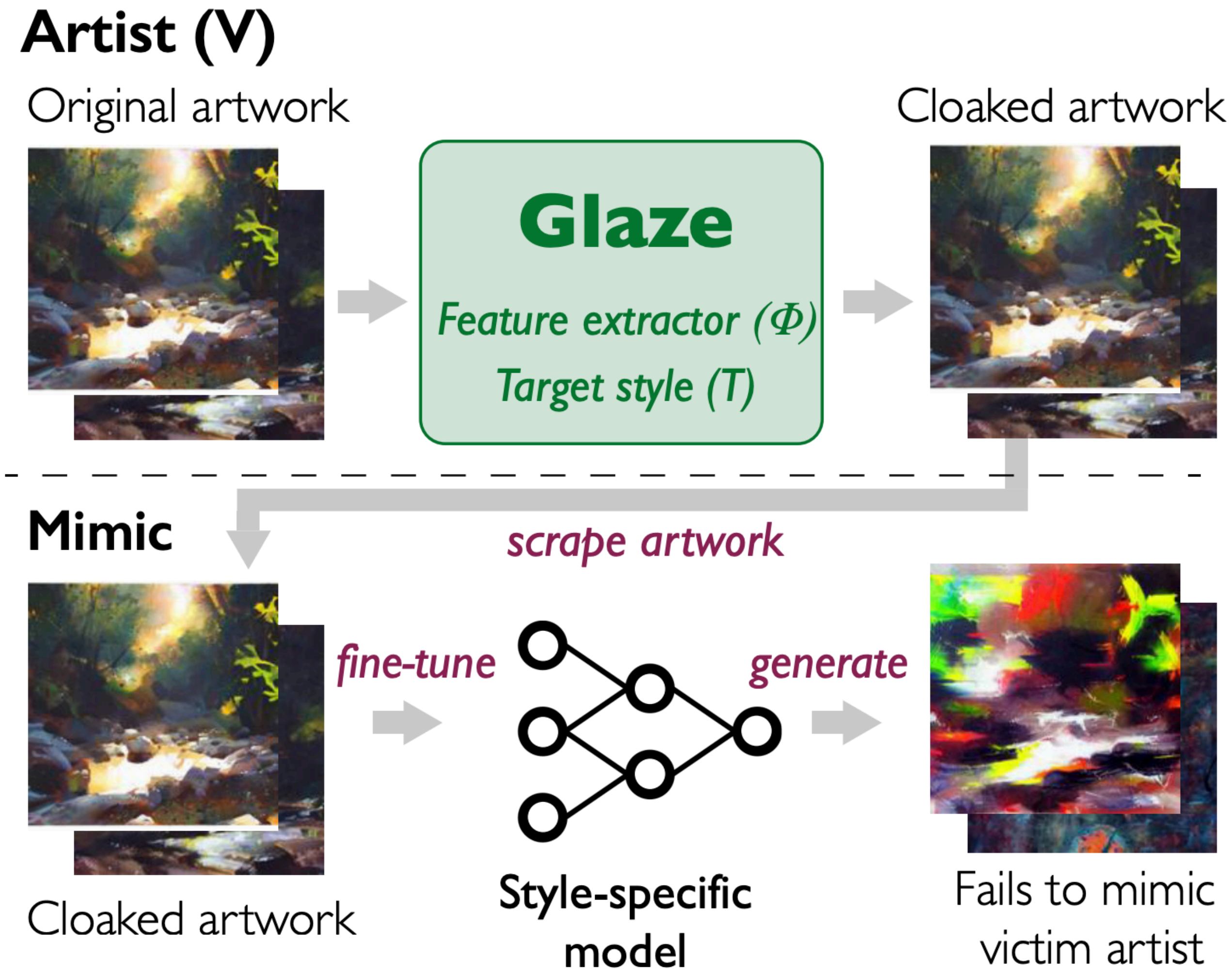


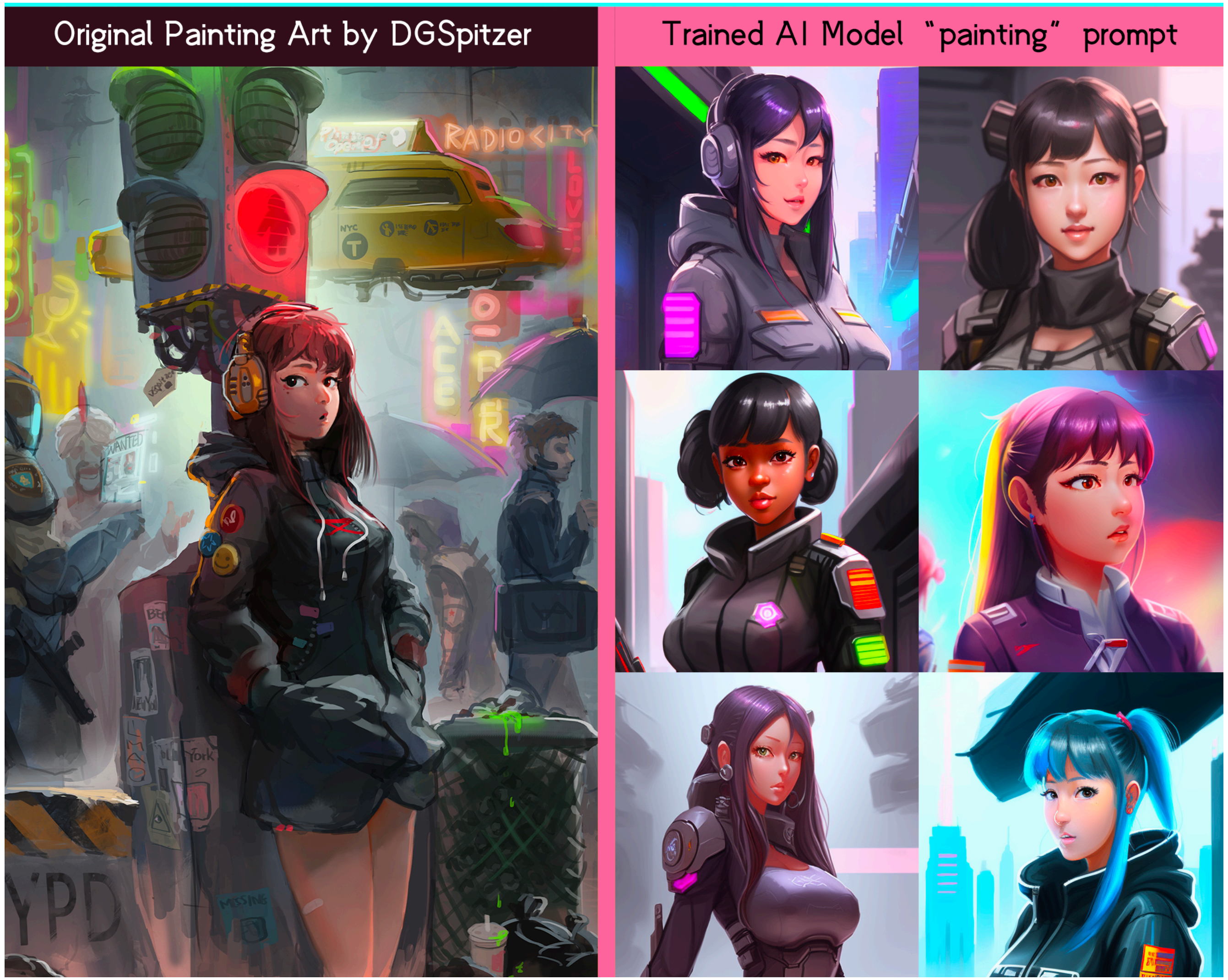
Fig. 5, Shan et al., "Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models"

Open Questions: How to attribute artists?

A significant concern of most participants, surprisingly, is not just the existence of AI art, but rather scraping of existing artworks without permission or compensation.

As one participant stated: “If artists are paid to have their pieces be used and asked permission, and if people had to pay to use that AI software with those pieces in it, I would have no problem.”

— *Shan et al., “Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models”*



A model trained by an artist (DGSpitzer) on own drawings

Assignment Overview

Grading Policy

This assignment is graded subjectively.

We will be lenient.

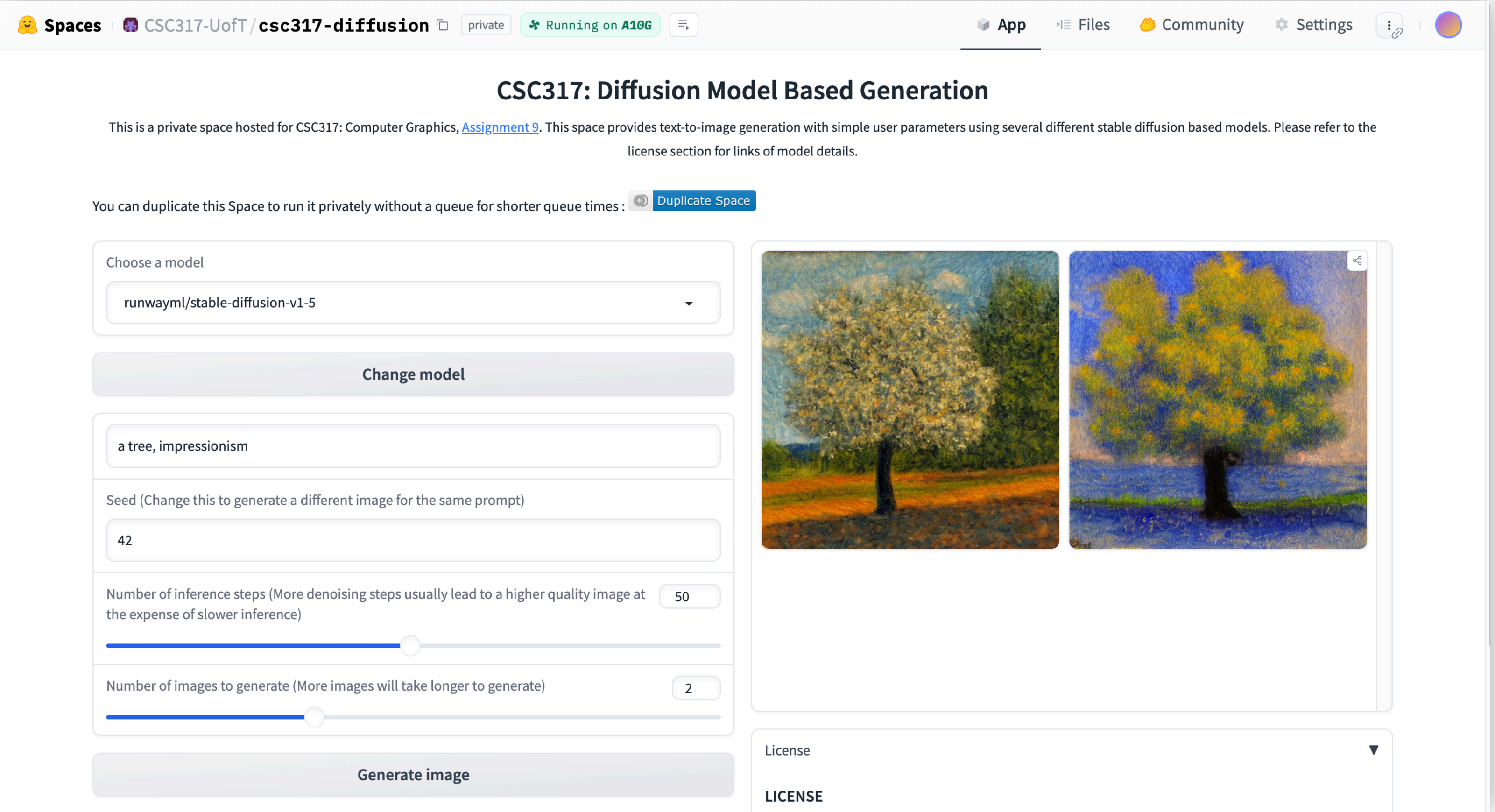
* You can request remarking if you question the mark

Privately-Hosted Generator

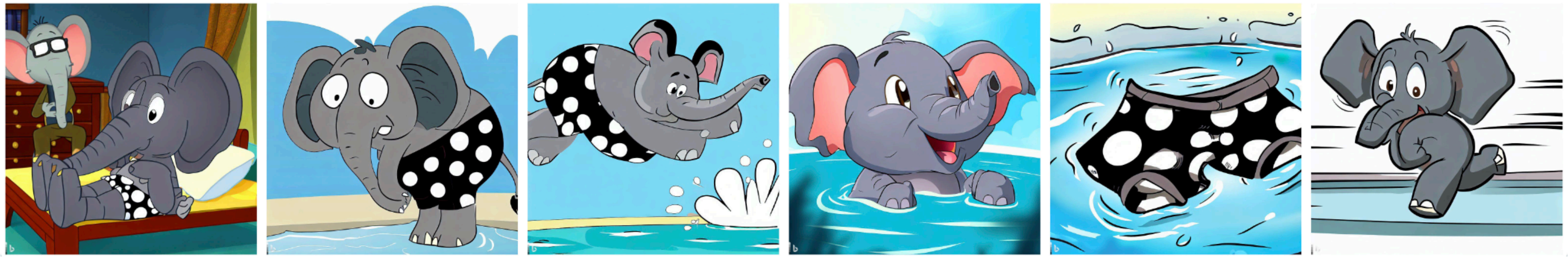
You will receive an invitation email by the end of today.

Contact us if:

1. You don’t receive the email;
2. The queuing becomes too bad.
   (We’ll switch to better GPU before deadline)



Format



```
{
  "pairs": [
    {
      "prompt": "[replace with prompt #1]",
      "model": "[SDv1.5|SDv2.1|SDXL]",
      "seed": 42,
      "image-path": "image-1.jpg"
    },
    {
      "prompt": "[replace with prompt #2]",
      "model": "[SDv1.5|SDv2.1|SDXL]",
      "seed": 42,
      "image-path": "image-2.jpg"
    },
    {
      "prompt": "[replace with prompt #3]",
      "model": "[SDv1.5|SDv2.1|SDXL]",
      "seed": 42,
      "image-path": "image-3.jpg"
    },
    {
      "prompt": "[replace with prompt #4]",
      "model": "[SDv1.5|SDv2.1|SDXL]",
      "seed": 42,
      "image-path": "image-4.jpg"
    },
    {
      "prompt": "[replace with prompt #5]",
      "model": "[SDv1.5|SDv2.1|SDXL]",
      "seed": 42,
      "image-path": "image-5.jpg"
    },
    {
      "prompt": "[replace with prompt #6]",
      "model": "[SDv1.5|SDv2.1|SDXL]",
      "seed": 42,
      "image-path": "image-6.jpg"
    }
  ],
  "story": "[replace with description of the story taking place.]"
}
```

[FEED ME]

Drag & drop
either all files or a .zip
for a specific task

Be Reasonable

Only use the generator for this assignment.
Only submit images generated by our setup.

Awards

We'll pick and frame THREE "open-ended" or "story" images

Awards

We'll fund the author of the best image to SIGGRAPH next year!

Thank you!
Questions?

Further Readings

Intro

- Zhu, Xiaojin, et al. "A text-to-picture synthesis system for augmenting communication." *AAAI*. Vol. 7. 2007. https://pages.cs.wisc.edu/~jerryzhu/pub/ttp.pdf
- Mansimov, Elman, et al. "Generating images from captions with attention." *arXiv preprint arXiv:1511.02793* (2015). https://arxiv.org/pdf/1511.02793.pdf
- Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851. https://hojonathanho.github.io/diffusion/
- Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." Advances in neural information processing systems 34 (2021): 8780-8794. https://arxiv.org/abs/2105.05233
- Reed, Scott, et al. "Generative adversarial text to image synthesis." International conference on machine learning. PMLR, 2016. https://proceedings.mlr.press/v48/reed16.pdf

A Handwavy Introduction to Diffusion Model

- Weng, Lilian. (Jul 2021). What are diffusion models? Lil'Log. https://lilianweng.github.io/posts/2021-07-11-diffusion-models/
- Song, Yang. (May 2021). Generative Modeling by Estimating Gradients of the Data Distribution. Yang Song's blog. https://yang-song.net/blog/2021/score/
- Yang Song's tutorial video. https://www.youtube.com/watch?v=wMmqCMwuM2Q

A Handwavy Introduction to Diffusion Model

- Vaclav Kosar. Cross-Attention in Transformer Architecture. https://vaclavkosar.com/ml/cross-attention-in-transformer-architecture
- Tang, Raphael, et al. "What the daam: Interpreting stable diffusion using cross attention." arXiv preprint arXiv:2210.04885 (2022). https://github.com/castorini/daam
- Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022). https://prompt-to-prompt.github.io/

Advanced Diffusion-Model-Based Editing Tools

- Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023. https://github.com/lllyasviel/ControlNet
- 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. https://dreambooth.github.io/

Strengths & Weaknesses

- Aaron Hertzmann's blog. Creative Explorations with DALL-E 2. https://aaronhertzmann.com/2022/05/25/dall-e.html
- Miranda Dixon-Luinenburg. What DALL-E 2 can and cannot do. https://www.lesswrong.com/posts/uKp6tBFStnsvrot5t/what-dall-e-2-can-and-cannot-do
- Chefer, Hila, et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models." ACM Transactions on Graphics (TOG) 42.4 (2023): 1-10. https://github.com/yuval-alaluf/Attend-and-Excite

Strengths & Weaknesses

- OpenAI. (Jul 2022). DALL·E 2 Preview - Risks and Limitations. https://github.com/openai/dalle-2-preview/blob/main/system-card.md#bias-and-representation
- Jernite, Yacine, et al. (Aug 2023).Hugging Face Ethics and Society Newsletter 2: Let's Talk about Bias!. Hugging Face Blog. https://huggingface.co/blog/ethics-soc-2

Generative Models & Artists

- “An AI artist explains his workflow” https://www.youtube.com/watch?v=K0ldxCh3cnI
- Shan, Shawn, et al. "Glaze: Protecting artists from style mimicry by text-to-image models." arXiv preprint arXiv:2302.04222 (2023). https://arxiv.org/abs/2302.04222
- DGSpitzer Art Diffusion. https://huggingface.co/DGSpitzer/DGSpitzer-Art-Diffusion