Diffusion Models

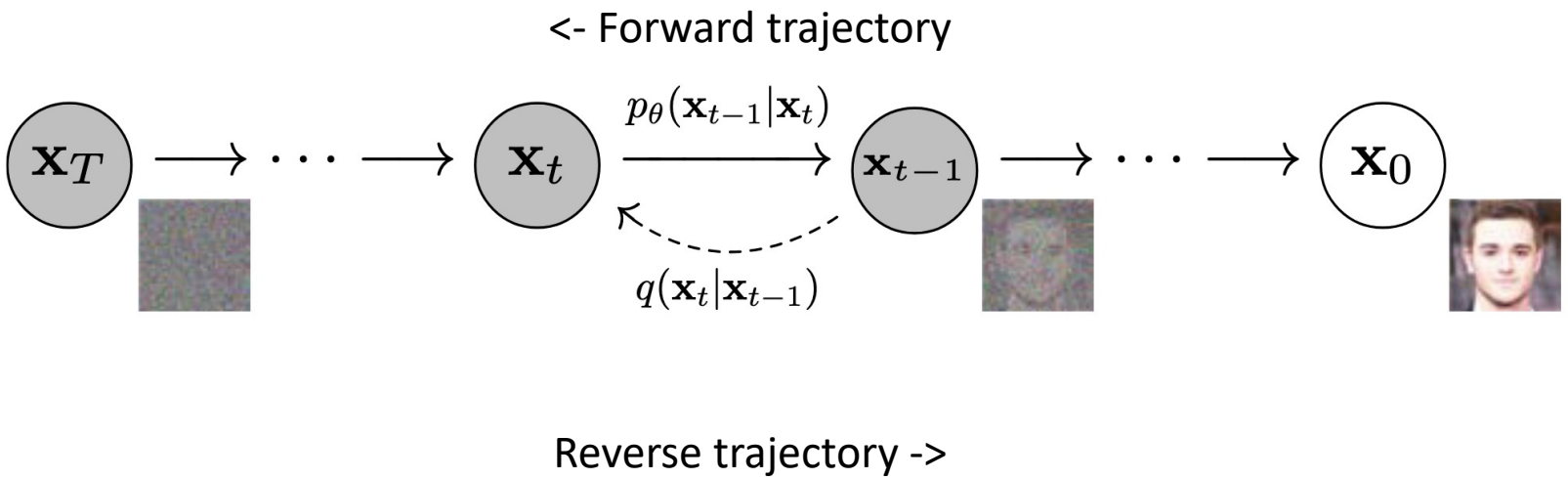
Forest Agostinelli
University of South Carolina

Outline

- Diffusion Models
- CLIP
- unCLIP

Diffusion Models

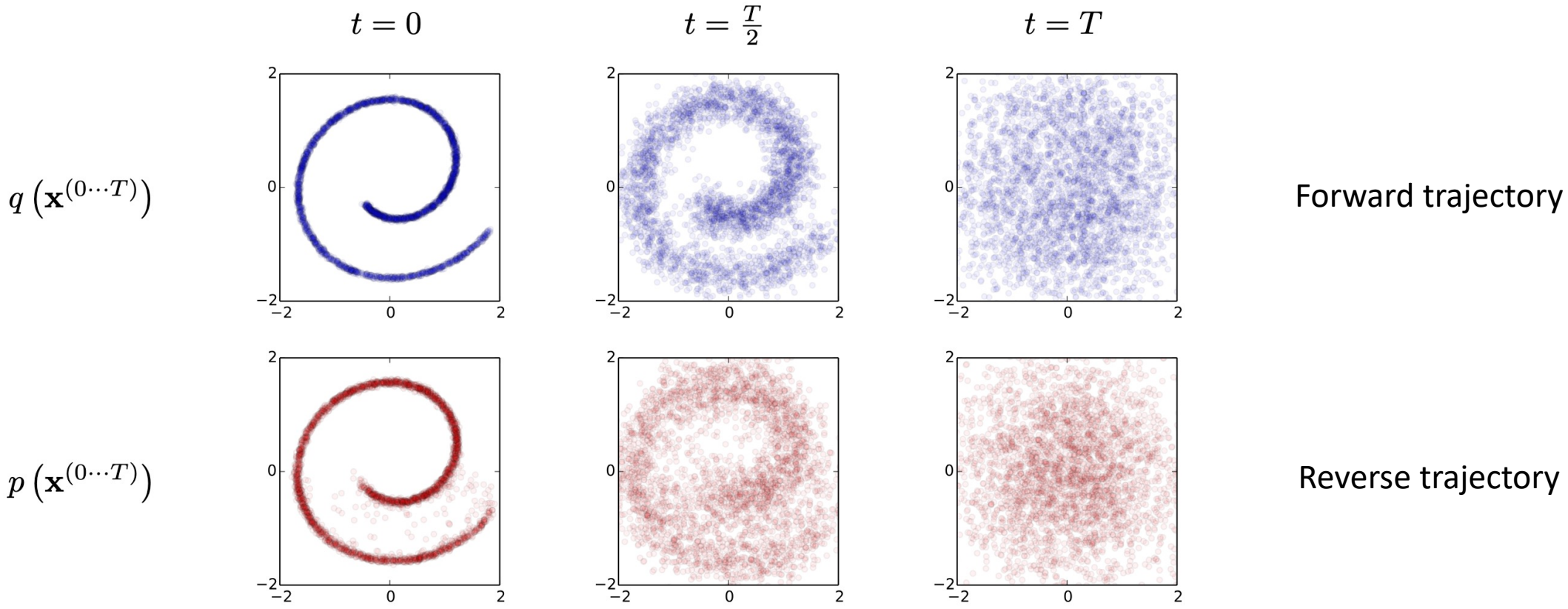
- The data that is to be modeled comes from some unknown distribution of data
- Diffusion models aim to learn to convert a known distribution, such as a Gaussian, into the data distribution over a sequence of steps
- **Forward Trajectory:** Transform the data distribution into a known distribution
- **Reverse Trajectory:** Transform the known distribution into the data distribution



Sohl-Dickstein, Jascha, et al. "Deep unsupervised learning using nonequilibrium thermodynamics." International conference on machine learning. PMLR, 2015.
Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

Diffusion Example

- Each point is a sample from a 2-dimensional Swiss roll distribution

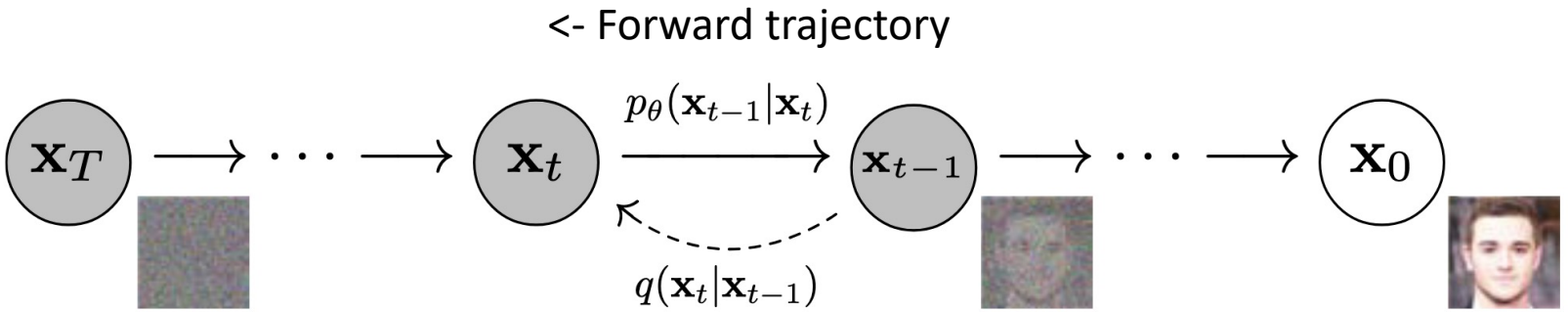


Comparison to Previous Methods

- Previous generative models had objectives that conflicted with one another
- Variational autoencoders
  - Forcing the latent representation to follow a Gaussian distribution could potentially conflict with ensuring the latent representation should be useful to the decoder
  - One must balance the reconstruction loss and KL divergence with hyperparameters
- Generative adversarial networks (GANs)
  - The generator and discriminator were in direct conflict with one another
  - This could lead to instability during training, including mode collapse
- Diffusion models do not have these conflicting objectives

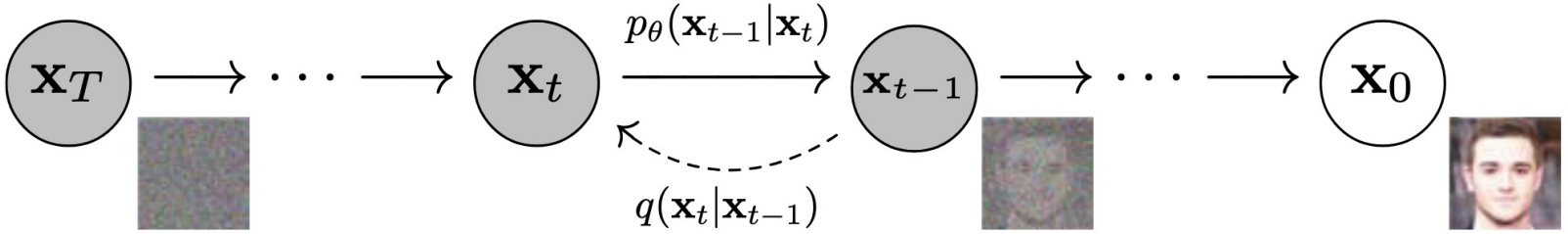
The Forward Trajectory

- For distributions, such as the Gaussian distribution, the diffusion rate can be controlled with a parameter
  - The diffusion rate controls how many steps, on average, a sample from the data distribution is converted to a sample from the known distribution
- If the diffusion rate is too high, then the problem becomes closer to that of learning a function that maps a sample from one distribution directly to another in a single step
  - Such as with generative adversarial networks or normalizing flows
- If the diffusion rate is too low, then sampling will be computationally expensive



The Reverse Trajectory

- If the known distribution is Gaussian, then the reverse trajectory need only compute the mean and variance of the gaussian given the x_t
- Training is performed by maximizing the log-likelihood of the data



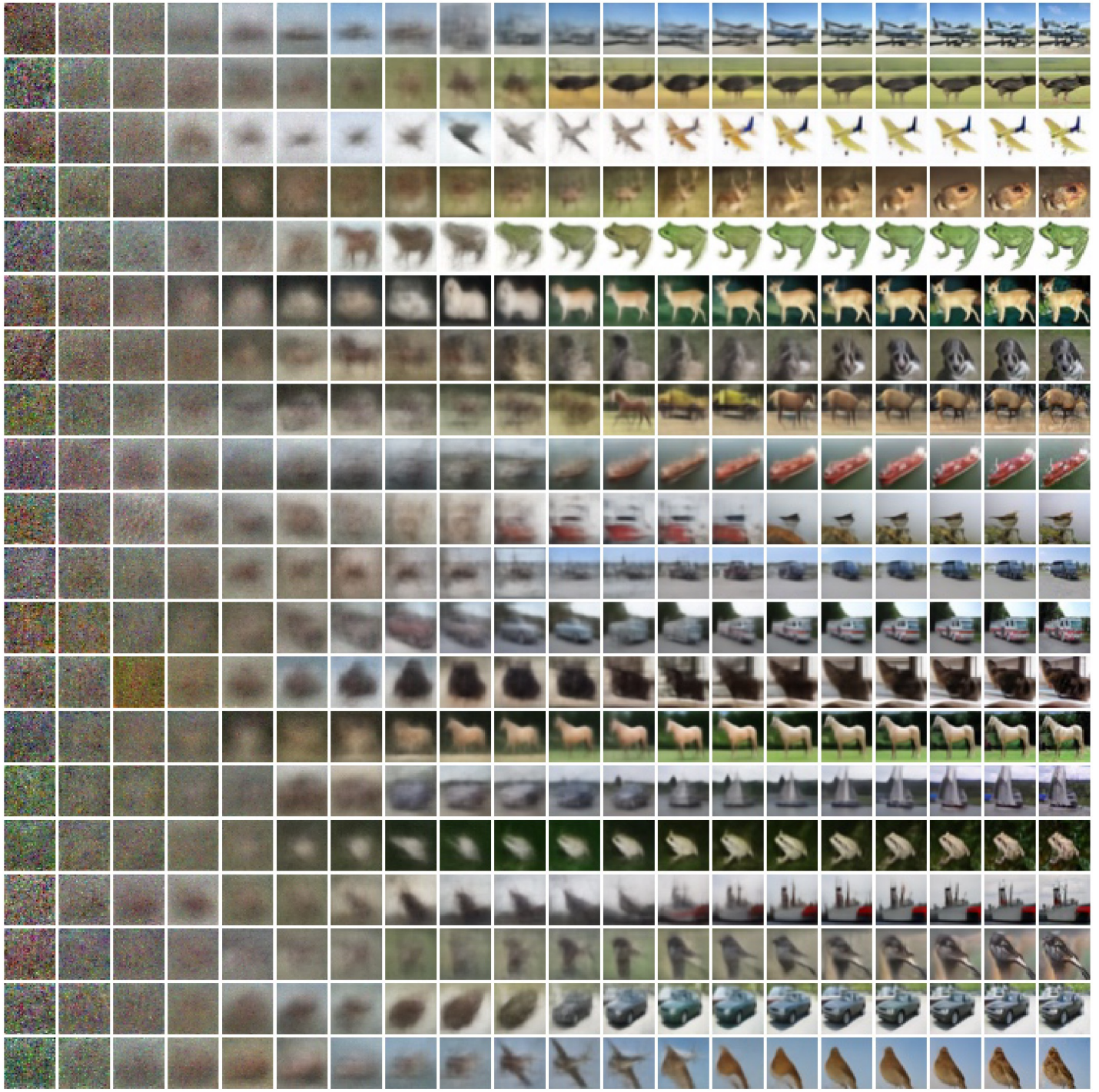
Reverse trajectory ->

Image Generation Across Time

- Sample from a Gaussian and run the reverse model

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$



High Quality Image Generation

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

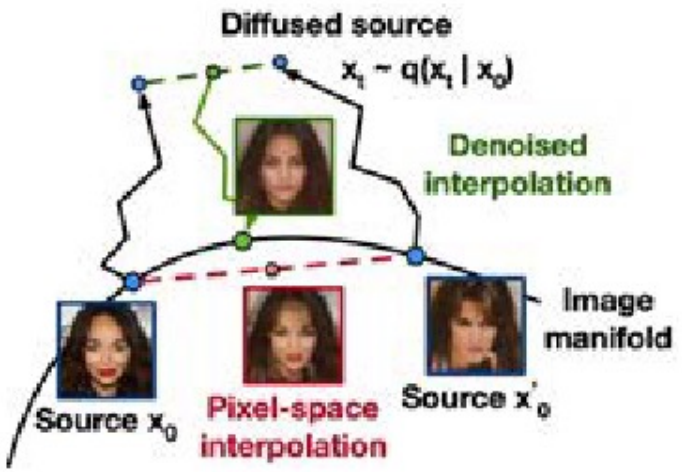
Image Generation

- There are some examples that are not as good

Semantic Meanings of Latents

- When conditioned on the same latent, especially closer to the final output, the generated images share similar attributes

Interpolation



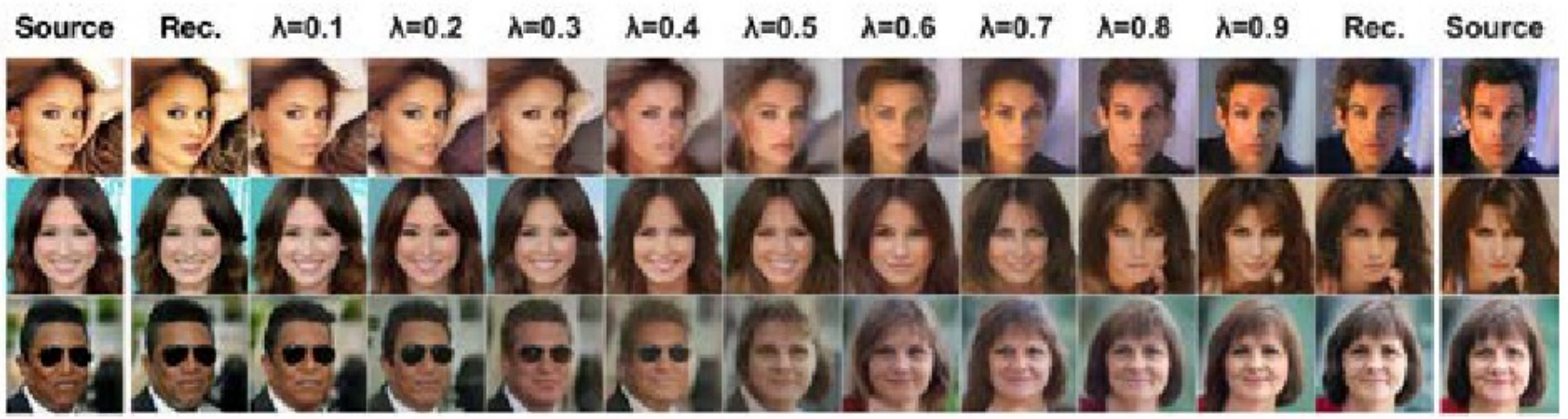


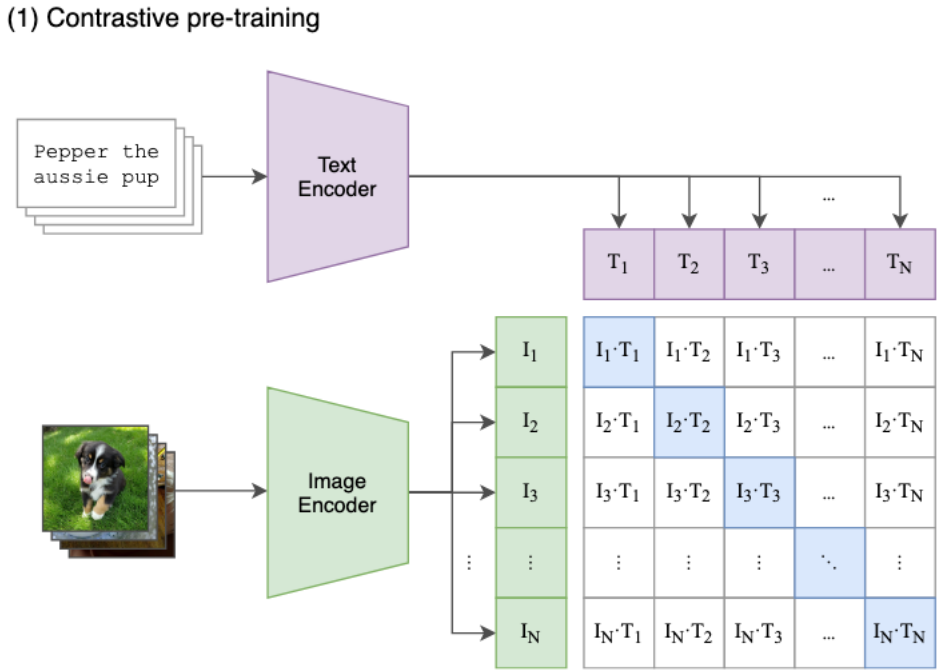
Figure 8: Interpolations of CelebA-HQ 256x256 images with 500 timesteps of diffusion.

Outline

- Diffusion Models
- CLIP
- unCLIP

CLIP

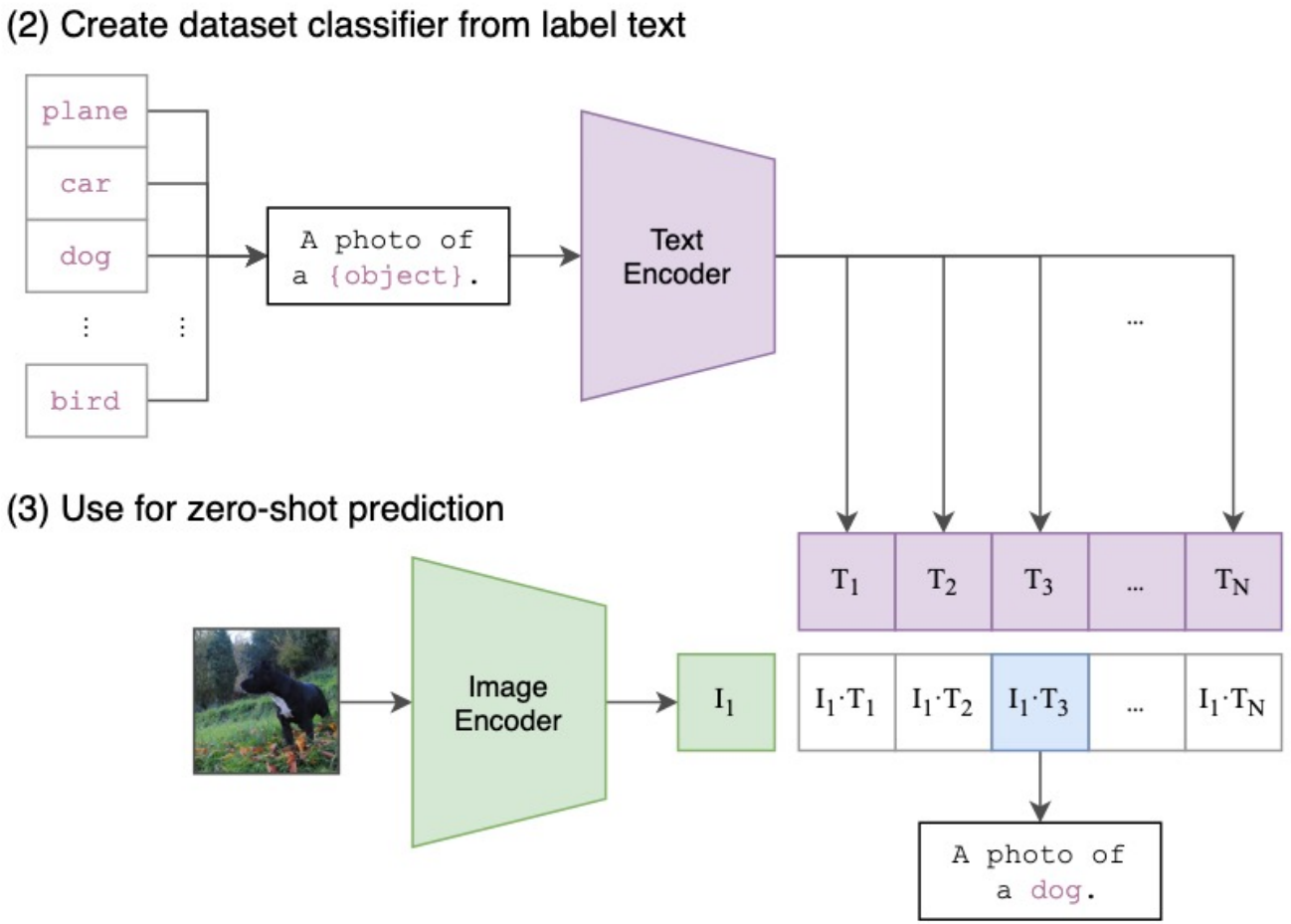
- Contrastive Learning Image Pretraining (CLIP)
- During training, an image and text encoder are trained to predict the correct pairing for image and text amongst a batch of data
- Therefore, the examples within a batch are explicitly contrasted with one another
- The matching text and image should have a high similarity score (i.e. similar embeddings) while the rest should have a low similarity score



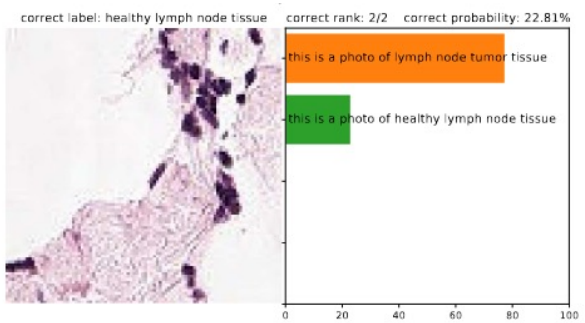
Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

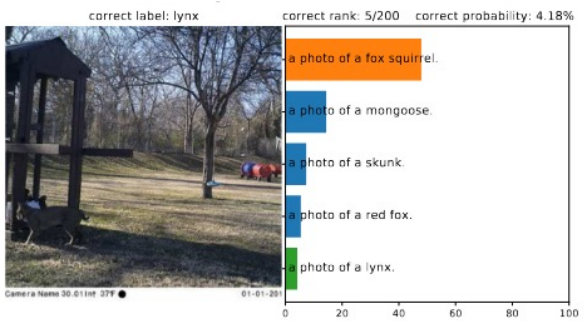
Zero-Shot Prediction

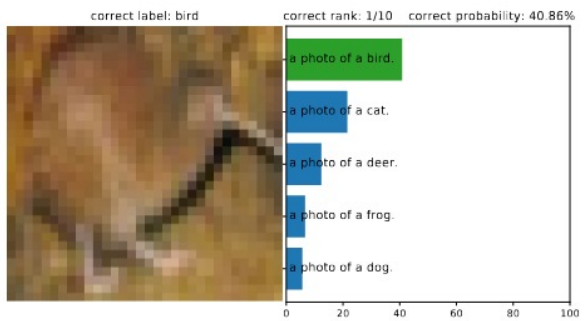
- After training, new tasks can be done with CLIP for a single image by feeding it a sequence of text prompts and seeing which one matches best

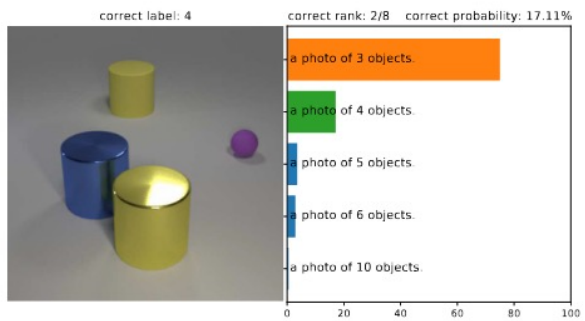


Zero-Shot Prediction

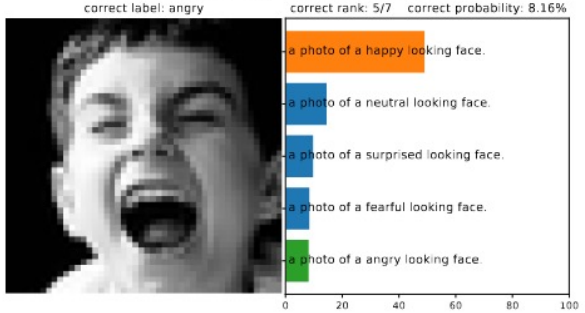






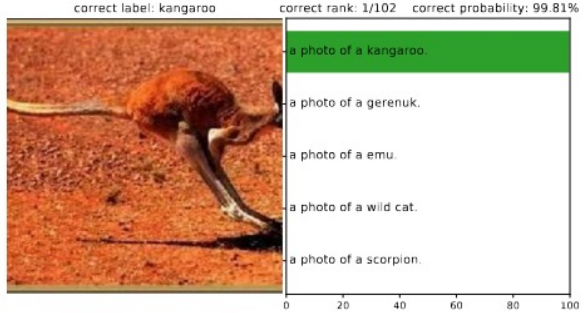


Facial Emotion Recognition 2013 (FER2013)

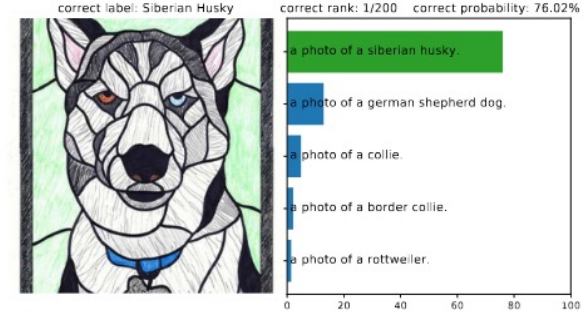


UCF101

Caltech-101



ImageNet-R (Rendition)



Comparison to Explicit Classification

- Although CLIP is not explicitly trained to classify images, it still compares similarly to baselines that were trained with explicit classification

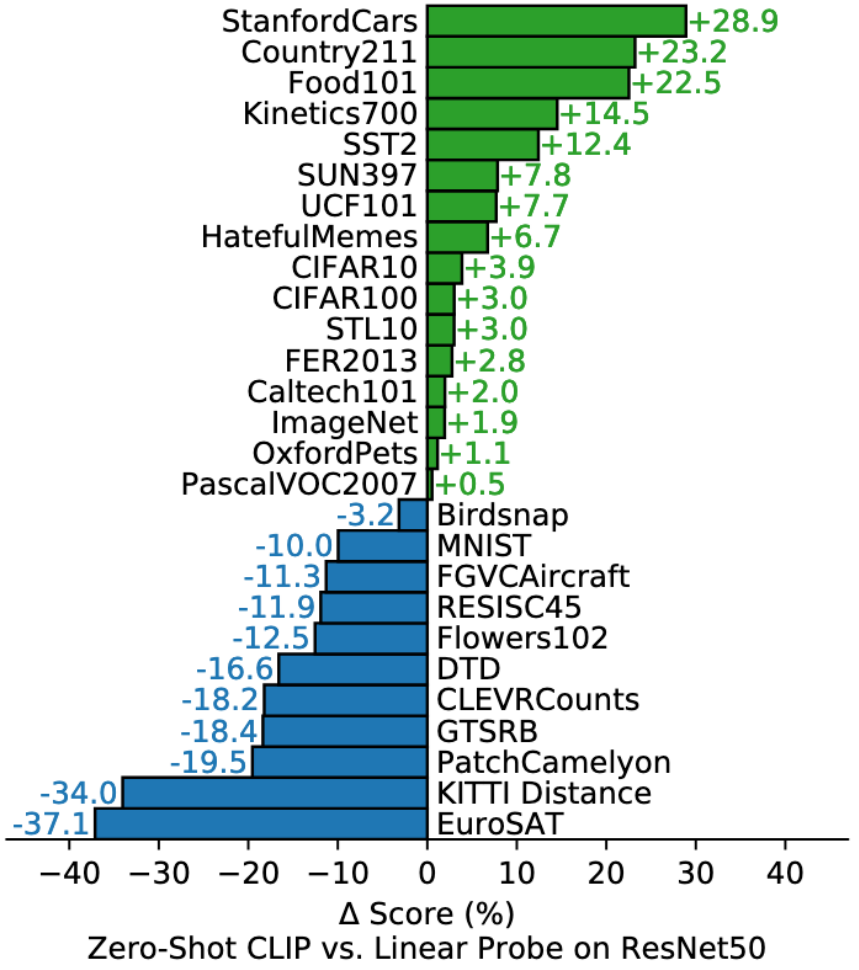


Figure 4. **Zero-shot CLIP is competitive with a fully supervised baseline.** Across a 27 dataset eval suite, a zero-shot CLIP classifier outperforms a fully supervised linear classifier fitted on ResNet50 features on 16 datasets, including ImageNet.

Outline

- Diffusion Models
- CLIP
- unCLIP

unCLIP

- CLIP is first trained so that there is a shared embedding space for text and images
- Then, based on a text embedding, this is used to obtain a prior for a diffusion decoder
- The diffusion decoder then produces the final image

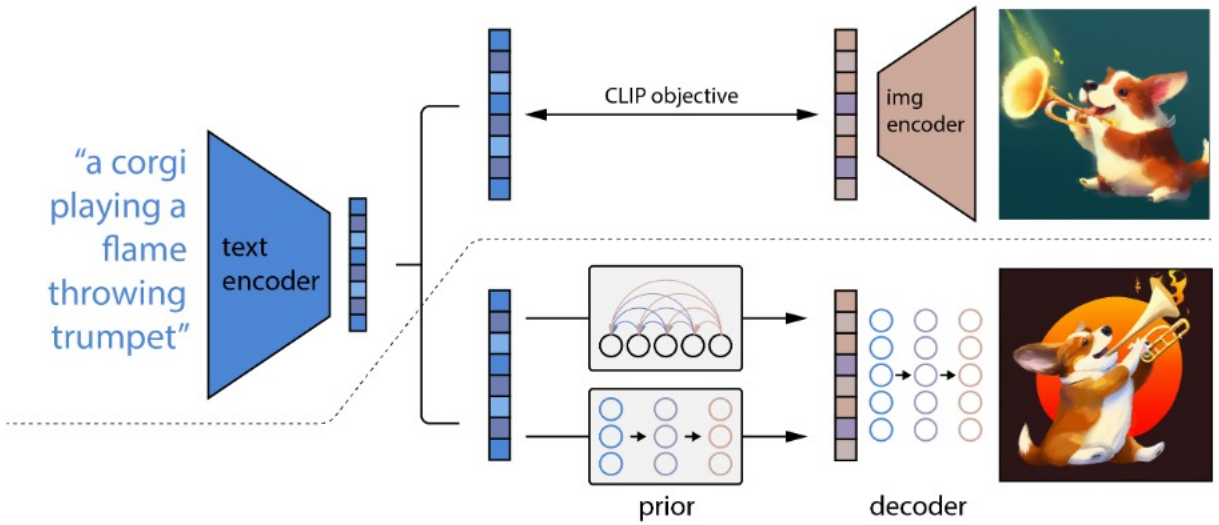


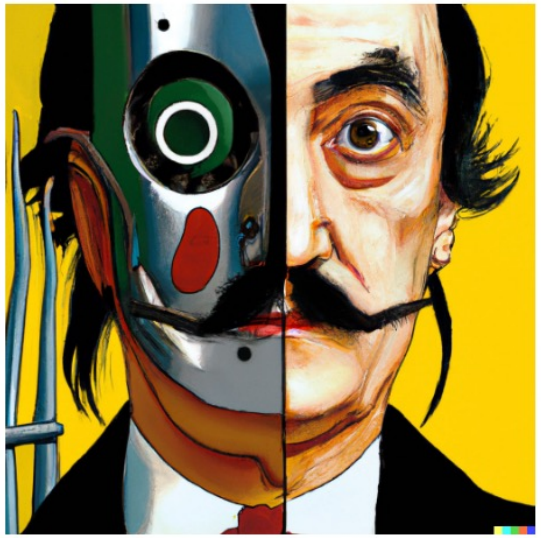
Figure 2: A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or diffusion prior to produce an image embedding, and then this embedding is used to condition a diffusion decoder which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 1.2 (2022): 3.

Producing High-Resolution Images

- High-resolution images are produced via up-sampling
- 64x64 to 256 x 256 to 1024 x 1024

unCLIP Examples



vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

Producing Variations of an Image

- Since the text and image embeddings are trained to be similar to one another, one can embed an image and give this as the prior for the diffusion model

Figure 3: Variations of an input image by encoding with CLIP and then decoding with a diffusion model. The variations preserve both semantic information like presence of a clock in the painting and the overlapping strokes in the logo, as well as stylistic elements like the surrealism in the painting and the color gradients in the logo, while varying the non-essential details.

Interpolation

- We can interpolate between images in the latent space and give this to the diffusion decoder

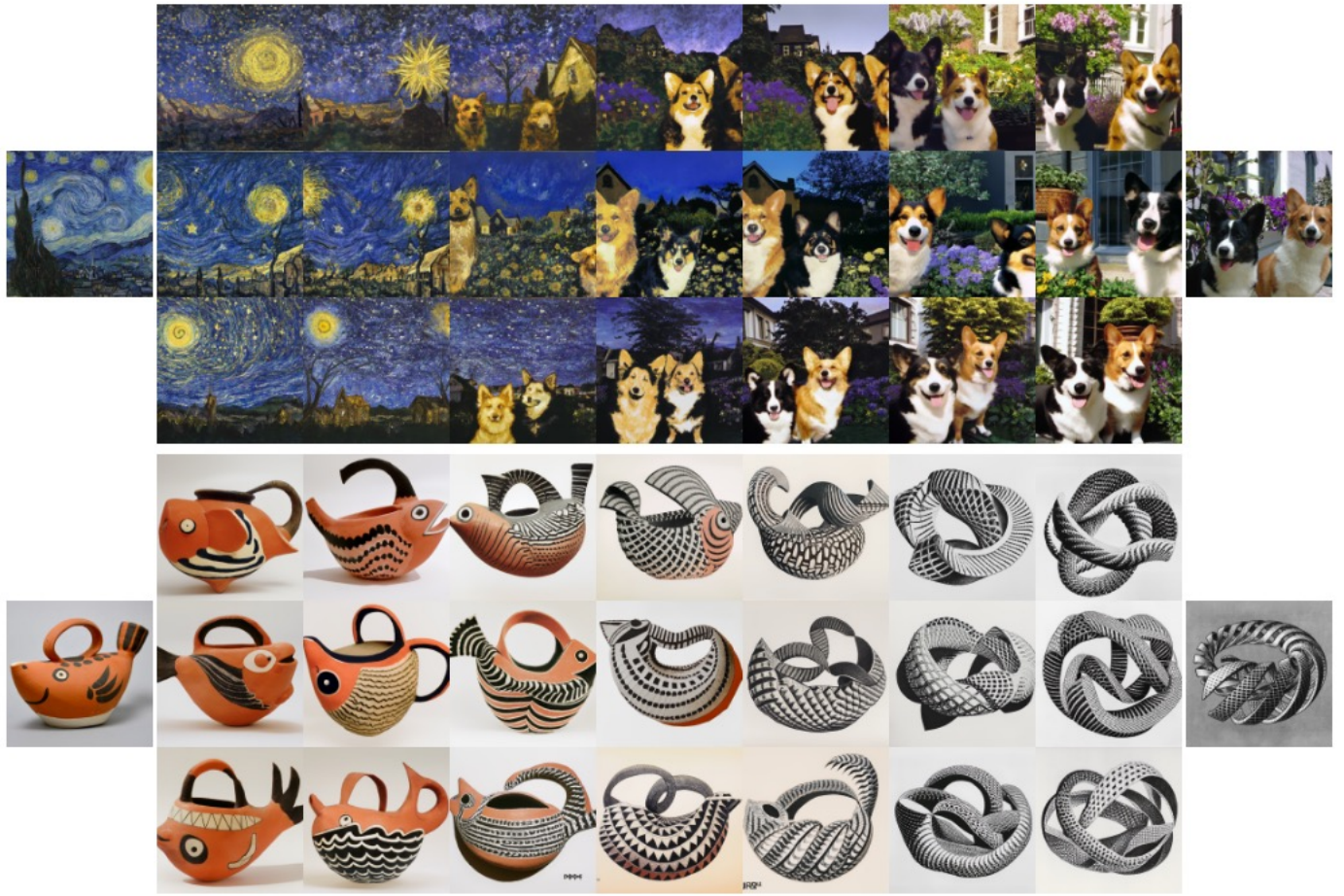


Figure 4: Variations between two images by interpolating their CLIP image embedding and then decoding with a diffusion model. We fix the decoder seed across each row. The intermediate variations naturally blend the content and style from both input images.

Interpolation

- We can do a similar interpolation with text

Figure 5: Text diffs applied to images by interpolating between their CLIP image embeddings and a normalised difference of the CLIP text embeddings produced from the two descriptions. We also perform DDIM inversion to perfectly reconstruct the input image in the first column, and fix the decoder DDIM noise across each row.

Typographic Attacks

- When asking CLIP about the label of an image, having conflicting labels in the same picture can confuse the zero-shot classifier

Granny Smith: 100%
iPod: 0%
Pizza: 0%

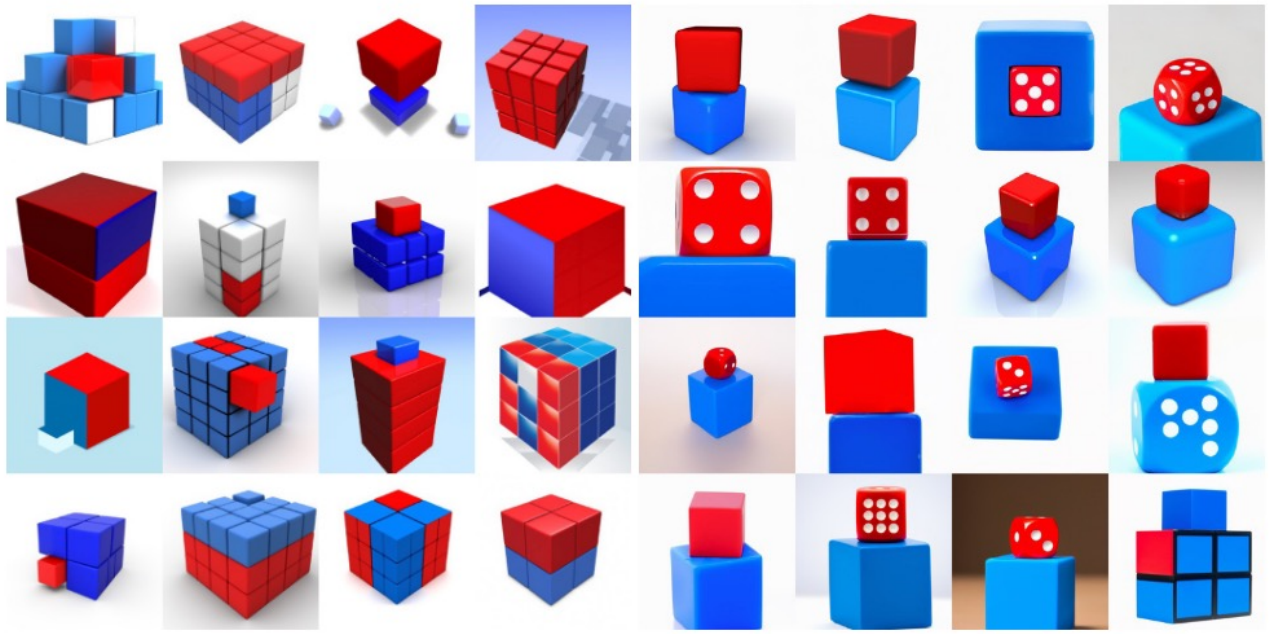
Granny Smith: 0.02%
iPod: 99.98%
Pizza: 0%

Granny Smith: 94.33%
iPod: 0%
Pizza: 5.66%

Figure 6: Variations of images featuring typographic attacks [20] paired with the CLIP model's predicted probabilities across three labels. Surprisingly, the decoder still recovers Granny Smith apples even when the predicted probability for this label is near 0%. We also find that our CLIP model is slightly less susceptible to the "pizza" attack than the models investigated in [20].

Limitations: Explicit Relational Reasoning

- I, personally, believe that explicit search is needed in such cases
  - After a certain point, a neural network will not be able to give the correct answer with a single forward pass
- Open AI's o1 says that it explicitly "thinks" before it gives and answer
  - I believe "thinking" here refers to search



(a) unCLIP (b) GLIDE

Figure 14: Samples from unCLIP and GLIDE for the prompt "a red cube on top of a blue cube".

Limitations: Explicit Relational Reasoning

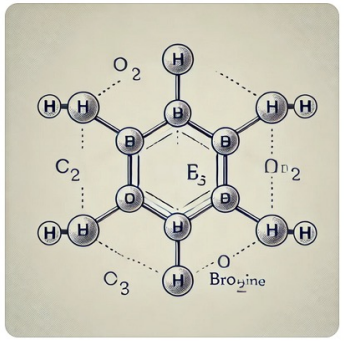
Prompt: Draw a picture of a Rubik's cube after turning the white face clockwise and the red face counter clockwise

Here is the illustration of the Rubik's Cube with the white face turned clockwise and the red face turned counterclockwise. Let me know if you'd like any modifications or a different perspective!

Prompt: Produce an image of the 24-puzzle where the sum of all rows are even

Here is an image of a 24-puzzle arranged in a 5x5 grid with all rows summing to even numbers. Let me know if there are any adjustments you'd like!

Prompt: A picture of a lewis structure of a Benzene ring that has a bond with a bromine.



Here's the Lewis structure of a benzene ring bonded to a bromine atom. Let me know if you'd like any modifications!

Limitations: Writing

- I believe the difficulties with text stem from a similar issue as relational reasoning

Figure 16: Samples from unCLIP for the prompt, "A sign that says deep learning."