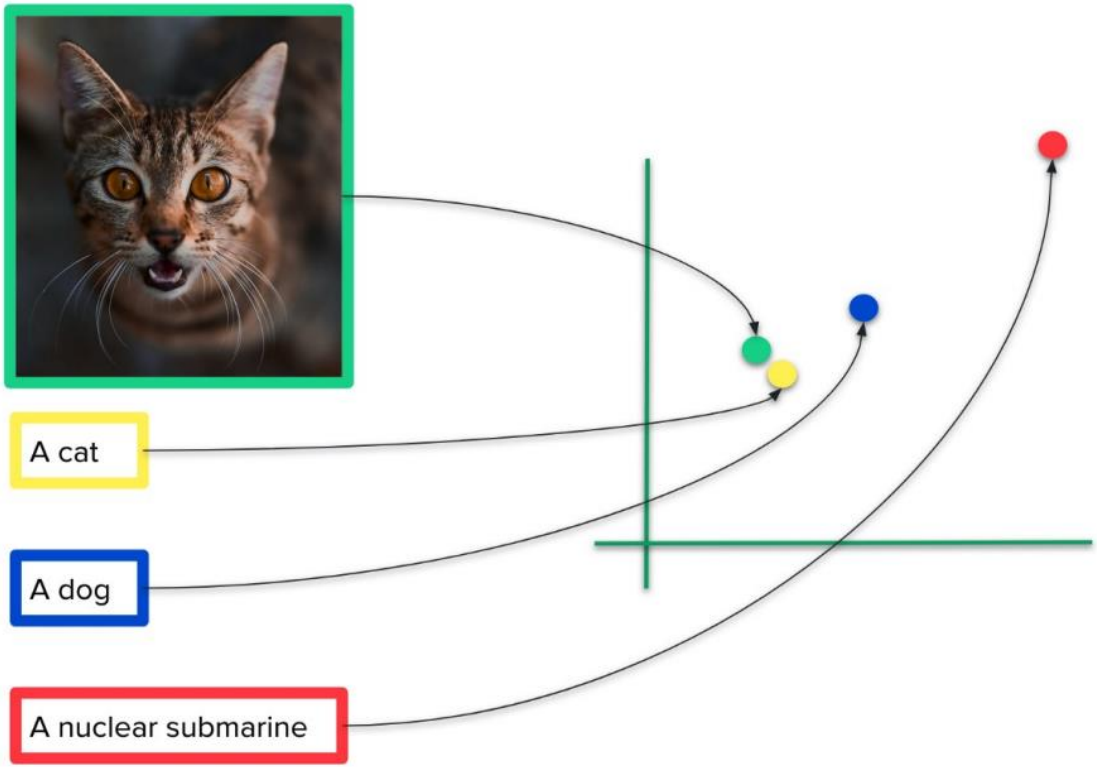
Session #18:
Self-Supervised Vision-Lang Models

Thursday, October 27
CSCI 601.771: Self-supervised Statistical Models


A cat
A dog
A nuclear submarine

The Toronto skyline with Google brain logo written in fireworks.

Teddy bears swimming at the Olympics 400m Butterfly event.

Stability AI, the startup behind Stable Diffusion, raises $101M

Kyle Wiggers @kyle_l_wiggers / 1:01 PM EDT • October 17, 2022

Comment

Stakeholder: Introduction

unCLIP: Two-stage model that generates image from text caption

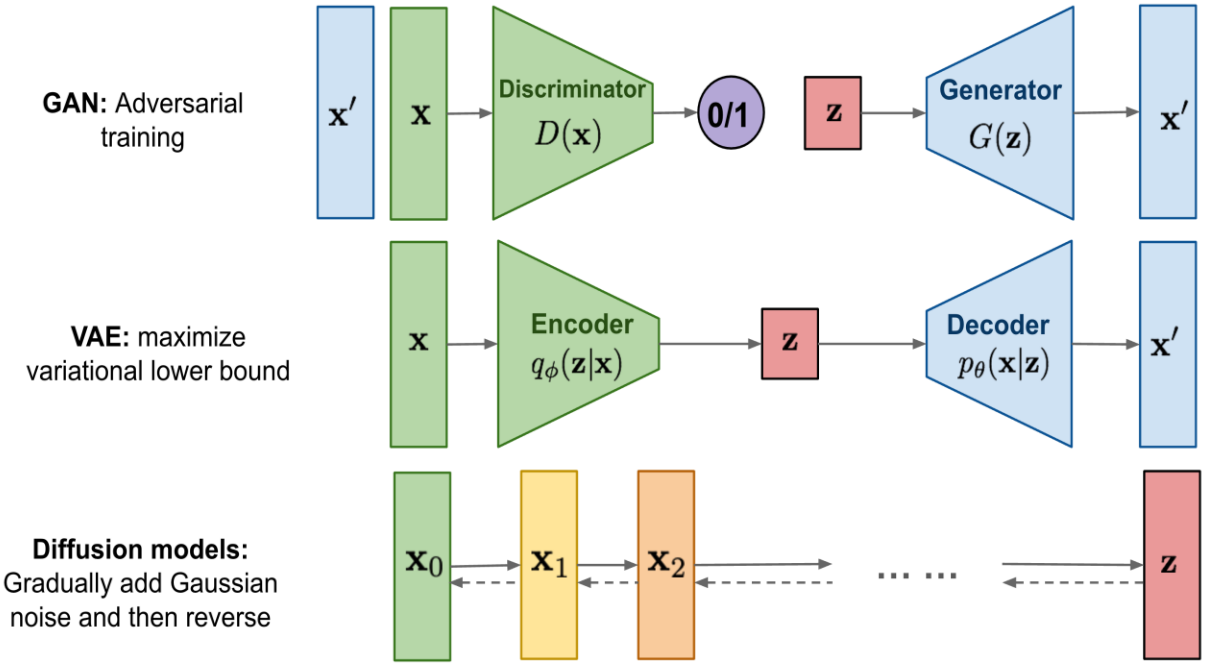
- CLIP: Model for image representation
- Diffusion model: Model for image generation

- Combined: text-conditional generative model
  - Prior model: Given text caption, generates CLIP image embeddings (diffusion > autoregressive)
  - Decoder (invert CLIP encoder): Given image embedding, produces image

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

🔑: Isabel, Vicky

Stakeholder: Related Work



Stakeholder: Method

Training data: (x,y) **Embeddings:** z_i, z_t

Prior: $P(z_i \mid y)$

Autoregressive: z_i converted into sequence of discrete codes and predicted autoregressively

- reduce dimensionality with PCA (transformer with causal attention mask)
- append to sequence: y, z_t, $z_i \cdot z_t$

Diffusion: z_i modeled using Gaussian distribution model

- decoder-only Transformer with causal attention mask
- append to sequence: y, z_t, diffusion timestep embedding, noised z_i

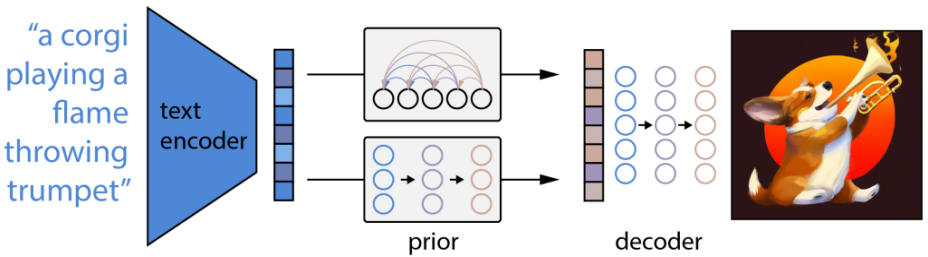
Decoder: $P(x \mid z_i, y)$

Generation (diffusion model): Modify GLIDE encoder to project CLIP embeddings 1) to existing timestep embedding and 2) into four tokens of context

- classifier-free guidance

High resolution (upsampler models): 64x64 > 256x256 > 1024x1024

- Gaussian blur | BSR degradation



: Isabel, Vicky

Stakeholder: Image Manipulation

Image *x* encoded as (z_i, x_T)

<u>Variation</u>

Same content but vary in:
shape and orientation

DDIM with η (=0 is deterministic)

Stakeholder: Image Manipulation

Interpolation: rotate embeddings with spherical interpolation

: Isabel, Vicky

Stakeholder: Image Manipulation

Text Diff: Norm vector for text diff between two text captions

- rotate between embedding and text diff with spherical interpolation

a photo of a victorian house → a photo of a modern house

a photo of an adult lion → a photo of lion cub

: Isabel, Vicky

Probing the CLIP Latent space

🔑: Isabel, Vicky

Decoder model allows us to visualize with the CLIP image encoder is seeing

Probing the CLIP Latent space

🔑: Isabel, Vicky

Use PCA on CLIP embeddings and progressively increase the dimension

Text to Image Generation - Human Evaluation

- Compare to GLIDE (SOTA in T2I Gen)
- Ask humans to give preference on the following:
  - Photorealism
  - Caption Similarity
  - Diversity

unCLIP Prior	Photorealism	Caption Similarity	Diversity
AR	$47.1\% \pm 3.1\%$	$41.1\% \pm 3.0\%$	$62.6\% \pm 3.0\%$
Diffusion	$48.9\% \pm 3.1\%$	$45.3\% \pm 3.0\%$	$70.5\% \pm 2.8\%$

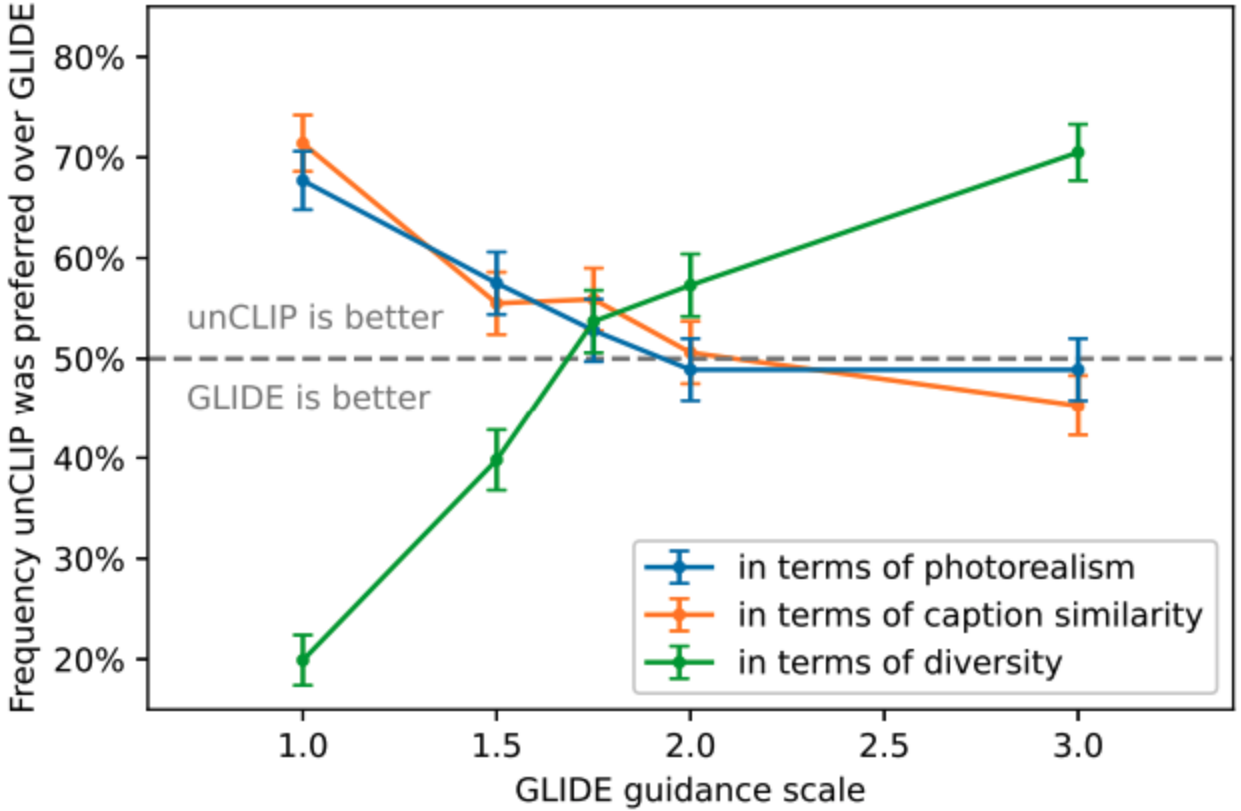
Probability humans preferred unCLIP over GLIDE

🔑: Isabel, Vicky

Improved Diversity-Fidelity Trade-off with Guidance

🔑: Isabel, Vicky

Improved Diversity-Fidelity Trade-off with Guidance

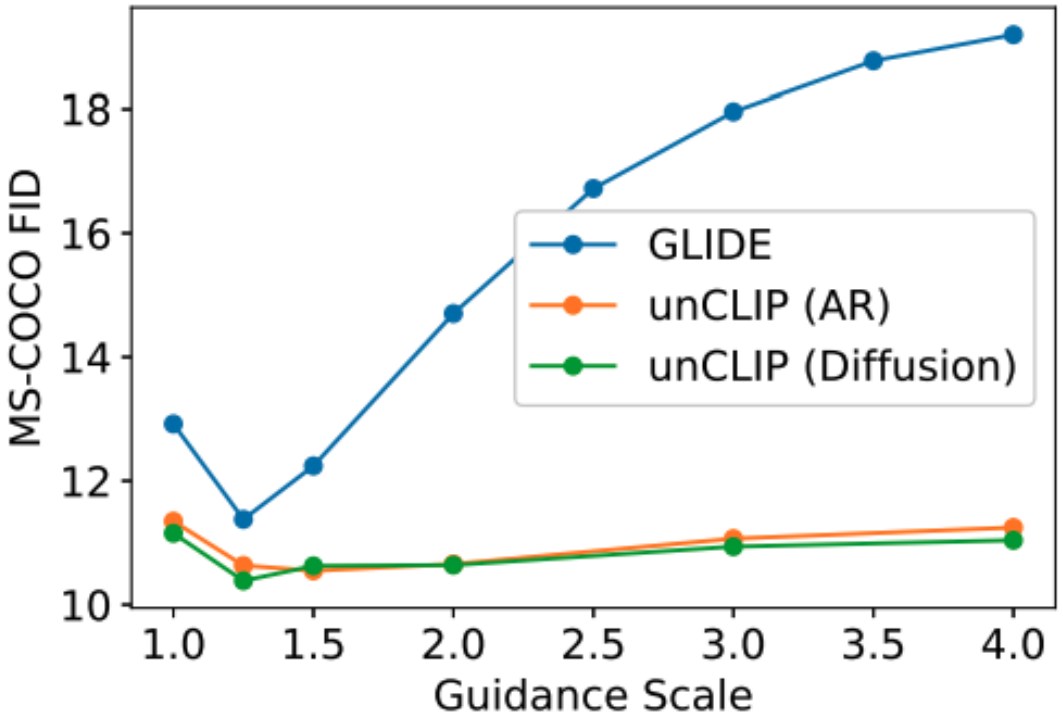


: Isabel, Vicky

Improved Diversity-Fidelity Trade-off with Guidance

Lower MS-COCO FID is better

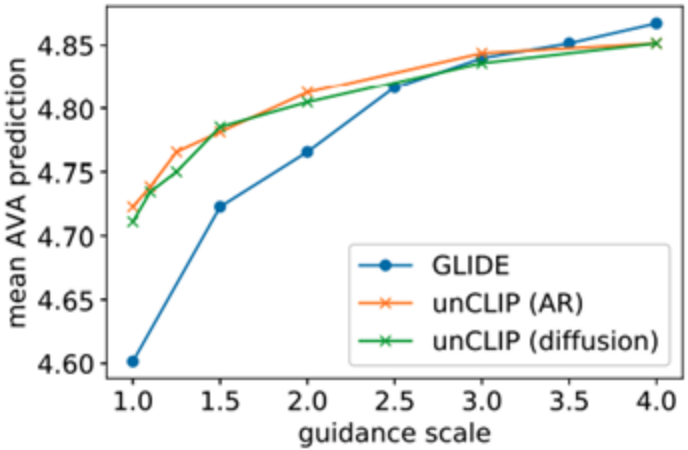
unCLIP gets news SOTA on MS-COCO FID

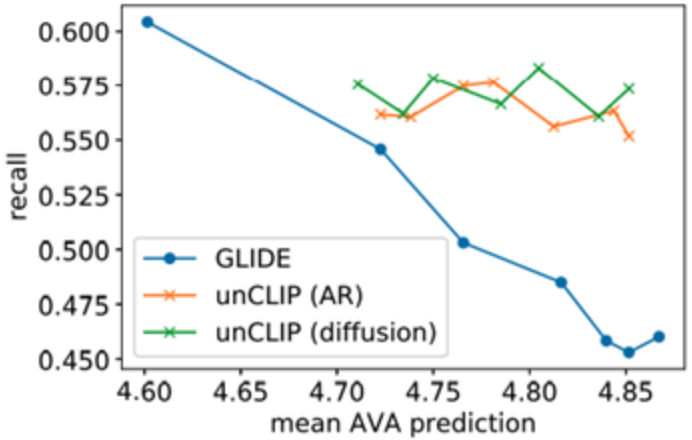


: Isabel, Vicky

Aesthetic Quality Comparison

1. Use GPT-3 to sample artistic captions
2. Train a model on AVA dataset to predict aesthetic judgements
3. Use model to evaluate aesthetic quality

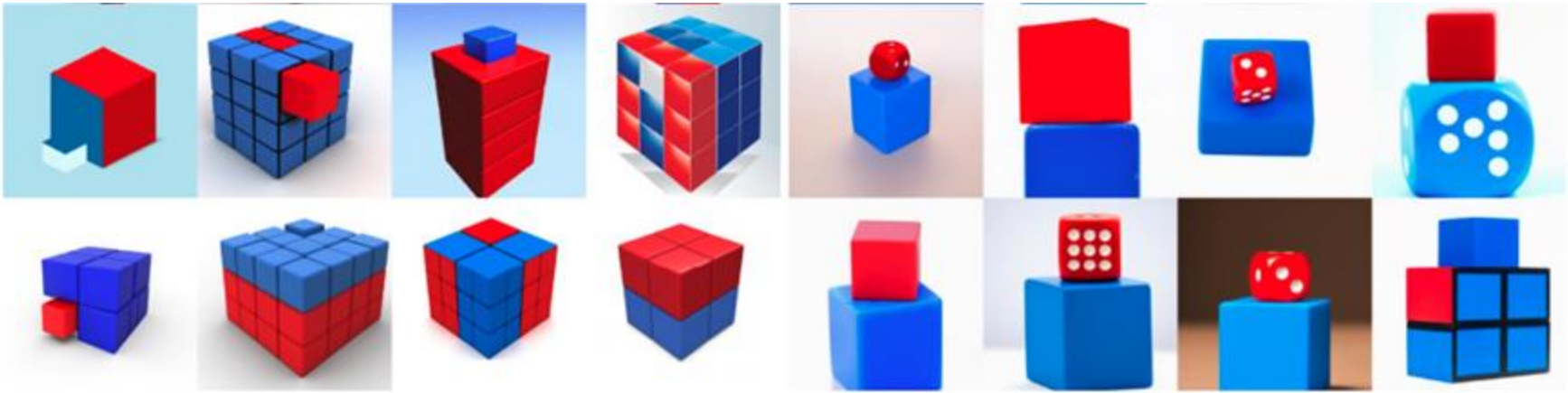




🔑: Isabel, Vicky

Limitations

unCLIP is worse than GLIDE at binding attributes to objects



(a) unCLIP (b) GLIDE

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

🔑: Isabel, Vicky

Limitations

unCLIP struggles to produce coherent text

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

🔑: Isabel, Vicky

Limitations

unCLIP produce low details for complex scenes

(b) A high quality photo of Times Square.

: Isabel, Vicky

Review: Strengths

1. Importance of the Prior
    1. Modelling CLIP image embedding is integral to good performance. Alternative considered and tested
2. unCLIP latent space has nice properties like interpolation and controllable variation
3. Visualization of the latent space of CLIP – a great tool for understanding the inner processes of image encoder.

Reviewer: Weaknesses

1. Too much unexplained concepts, hard to understand.
2. The model is weak with complex situations. e.g., producing coherent text and details in complex scenes.
3. Computation of the prior (AR & Diffusion) is poorly motivated:
    1. Autoregressive prediction of dimensionality-reduced CLIP image embeddings
        1. Assumptions of PCA met?
        2. Why autoregressive?
    2. Diffusion prior on sequence of [encoded text, CLIP text emb, emb for diffusion timestep, noised CLIP image emb, final emb to predict image embedding]

Reviewer: Weaknesses

- The label of each image is no longer a noun, but a sentence, so images that were forcibly divided into the same kind in the past have “infinitely fine-grained” labels.
  - Example: ImageNet labels images as "dogs," and with this paired example, one can learn the nuances of "dogs" in different environments and doing different things.
- The pairing of text and images is not strong enough. This is why the authors repeatedly stress collecting huge datasets, because they have to suppress the noise by means of big data.
- OpenAI has not released the data set, we only know that there are 500,000 queries, but we do not know what these queries are, such as simple descriptions or complex descriptions. If it is just a simple description, such as "black cat", "lecture room", etc.

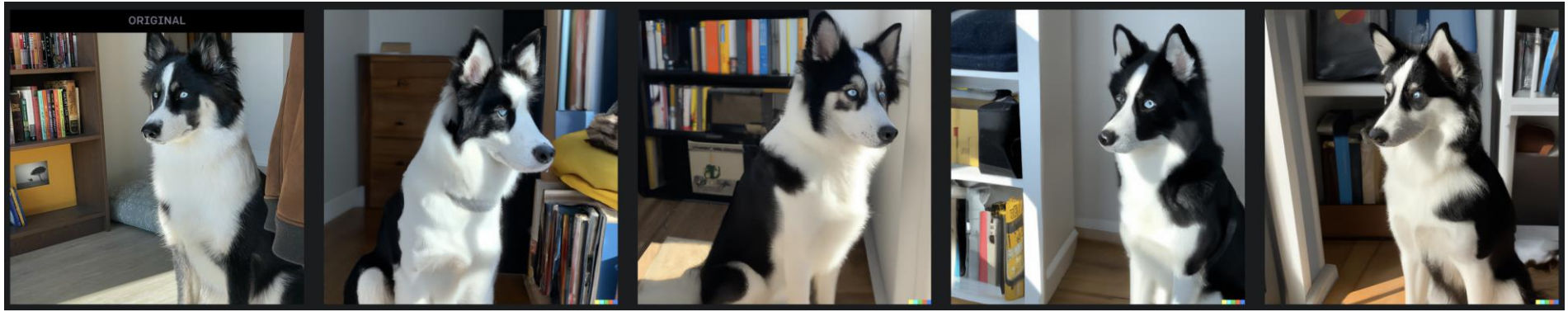
Empiricists

- DALL-E 2 is the only open-source implementation of unclip (https://labs.openai.com)
- Can you guess which one is the real puppy?

: Iliana, Aayush, Boyuan

Empiricists

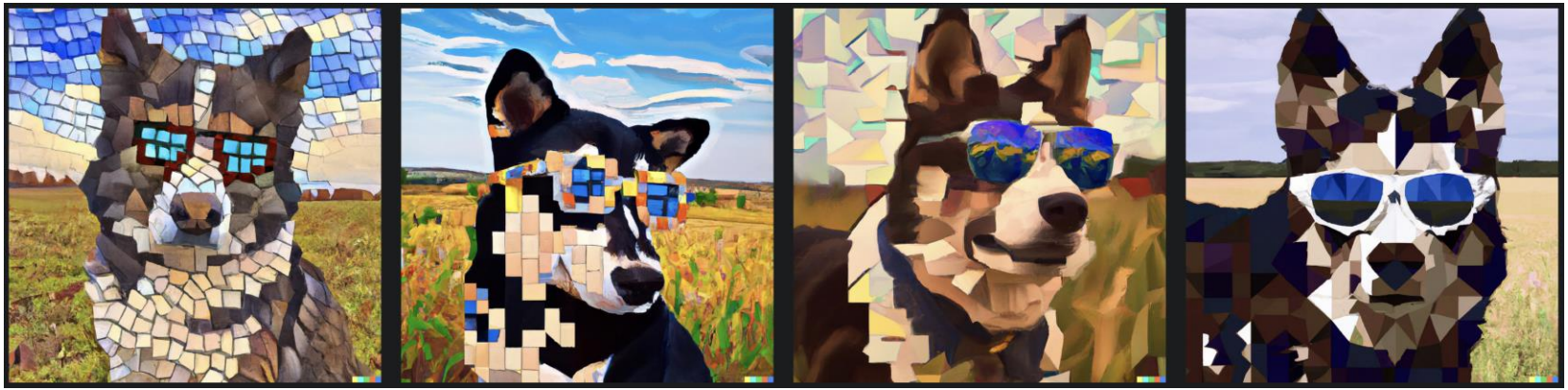
- Variations generated by the decoder diffusion model



: Iliana, Aayush, Boyuan

Empiricists

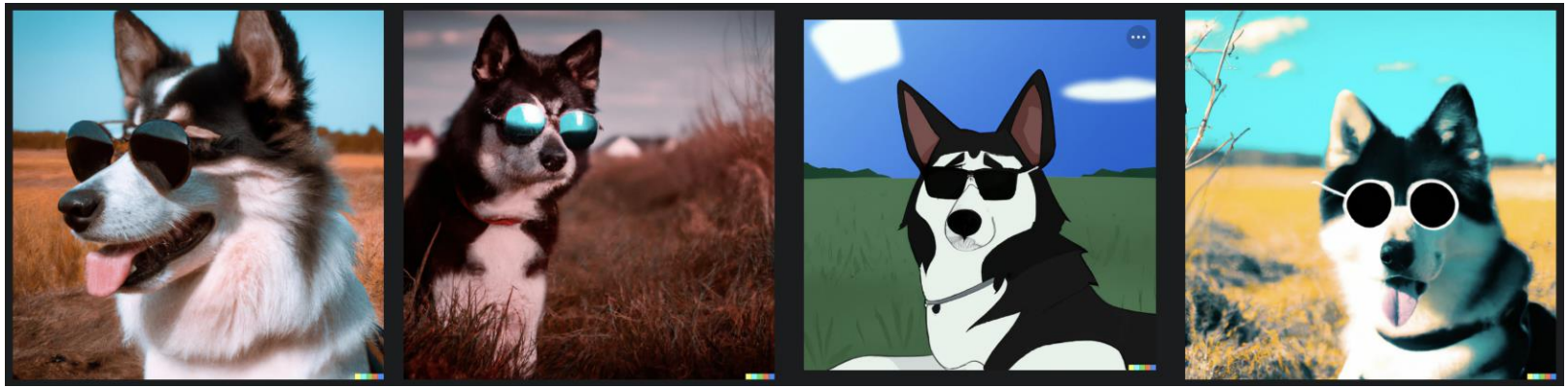
- A cubist painting of a Yakutian Laika dog in a field with sunglasses



: Iliana, Aayush, Boyuan

Empiricists

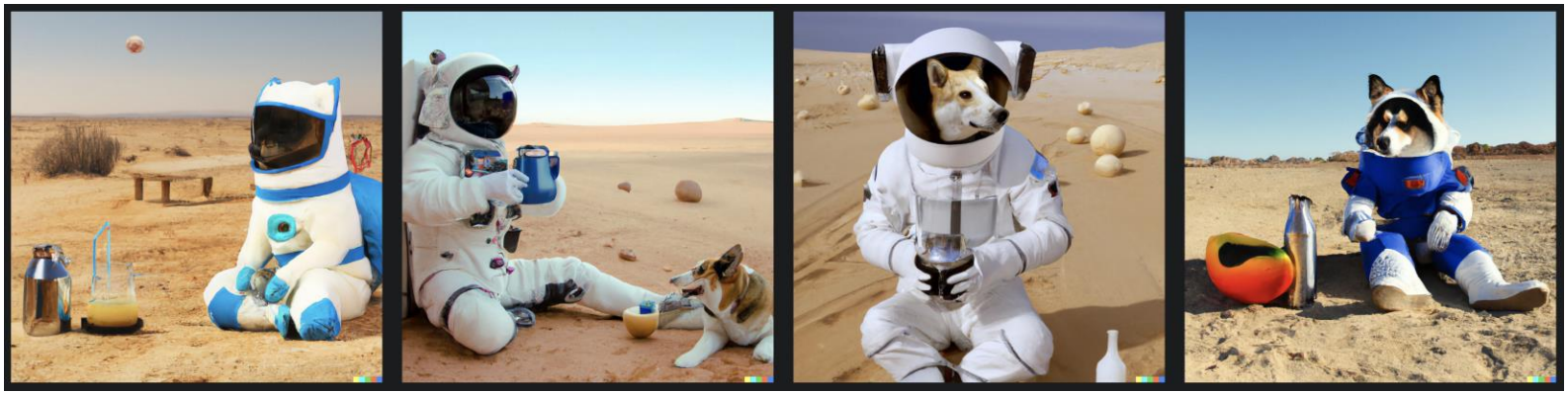
- A Yakutian Laika dog in a field with sunglasses in anime style



: Iliana, Aayush, Boyuan

Empiricists

- A Yakutian Laika dog in a space suit sipping a piña colada in the desert



: Iliana, Aayush, Boyuan

Empiricists: Typographic attacks

: Iliana, Aayush, Boyuan

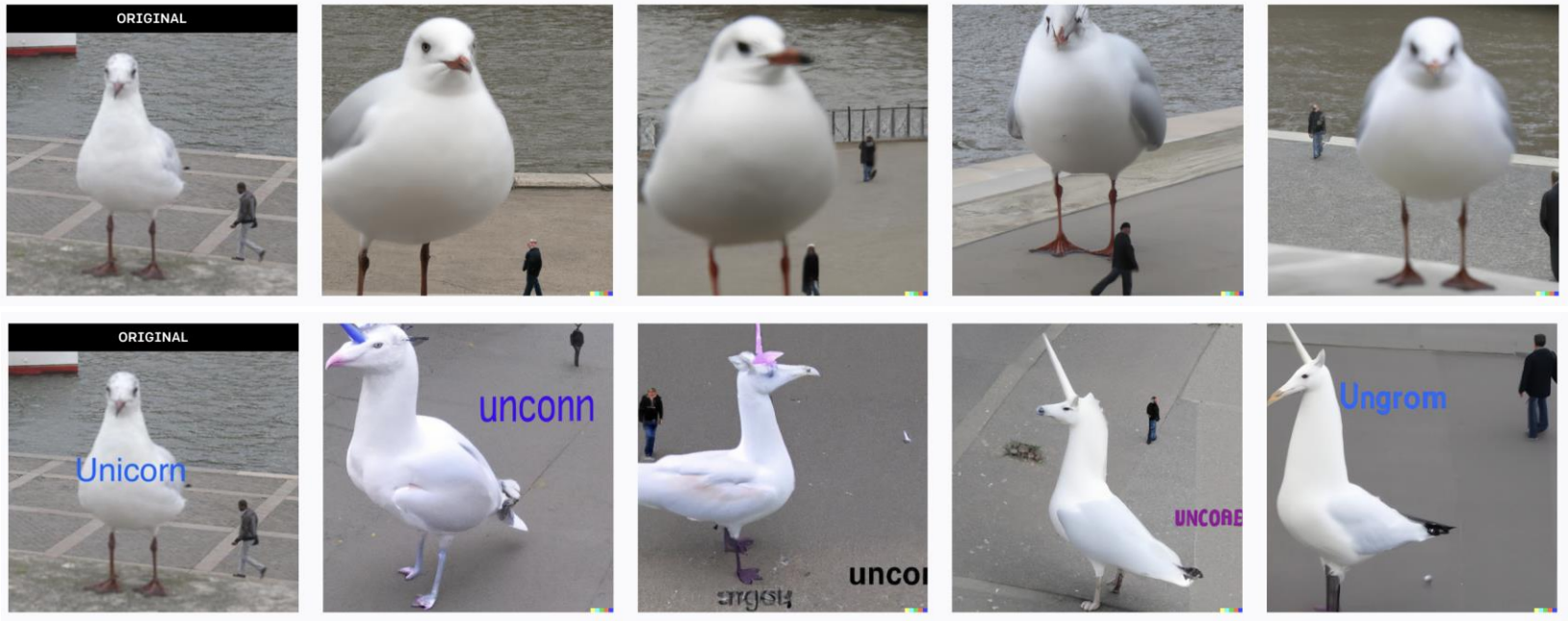
Empiricists: Typographic attacks

: Iliana, Aayush, Boyuan

Empiricists: Typographic attacks

: Iliana, Aayush, Boyuan

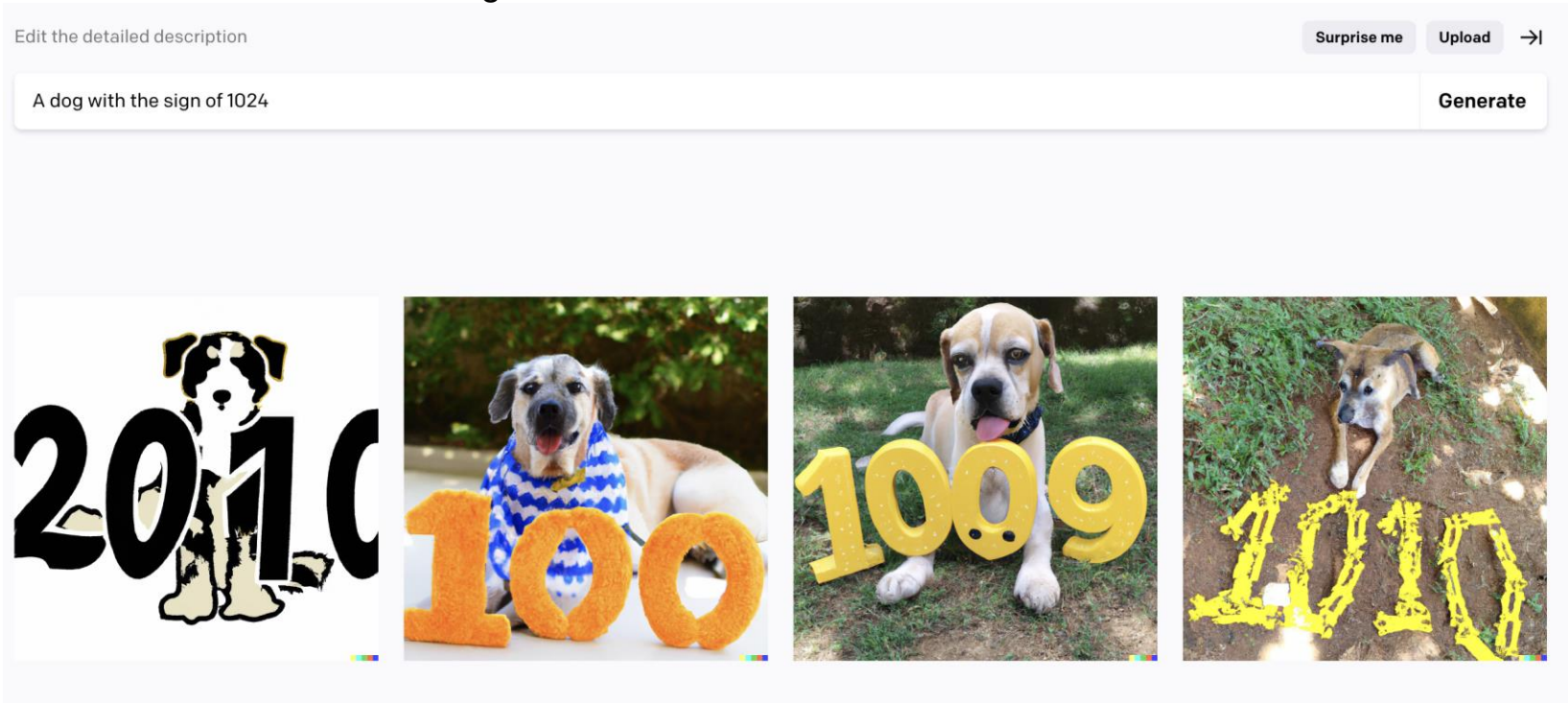
Empiricists: Typographic attacks



: Iliana, Aayush, Boyuan

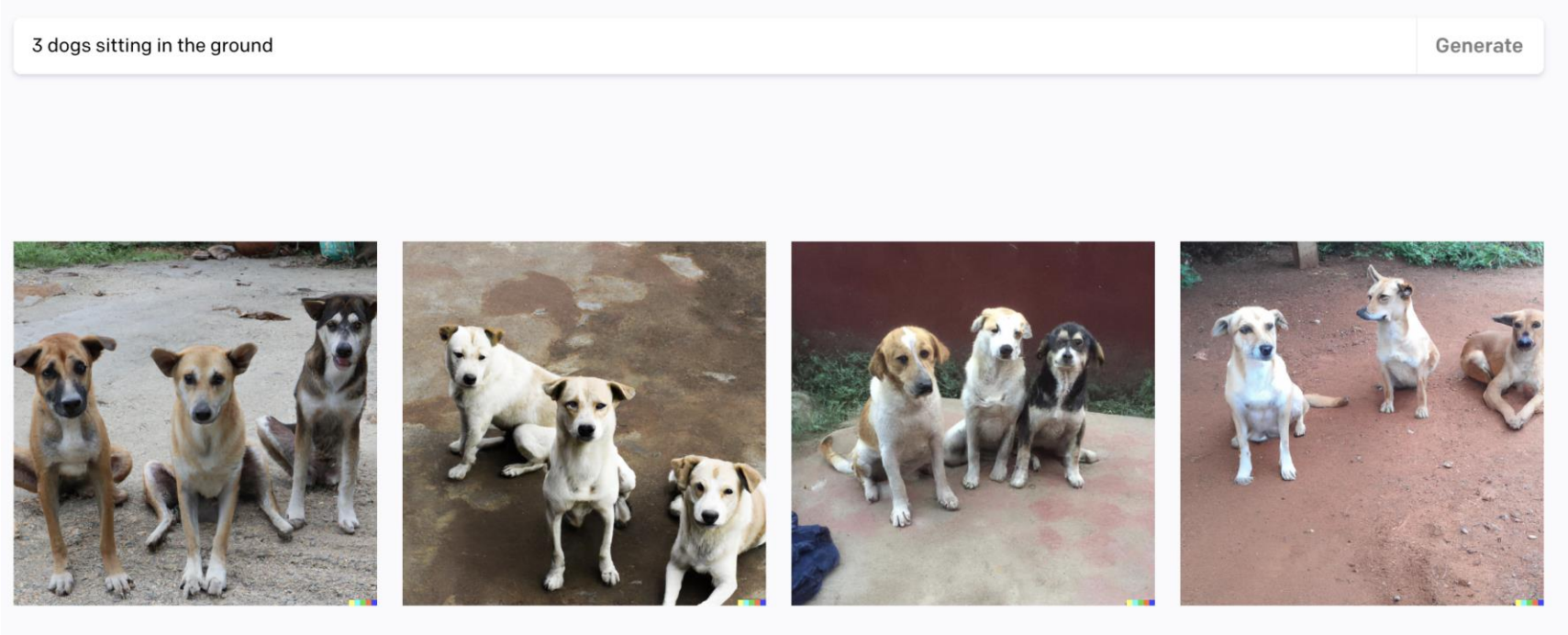
Empiricists

Limitation1: Number Understanding



Empiricists

Limitation2: Counting



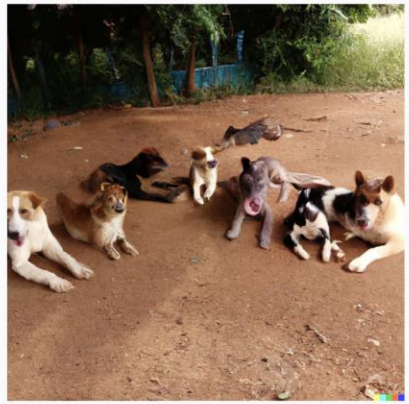
: Iliana, Aayush, Boyuan

Empiricists

Limitation2: Counting

10 dogs sitting in the ground | Generate



: Iliana, Aayush, Boyuan

Empiricists

Limitation 2: Bad Details

Example: A photo of the time square

: Iliana, Aayush, Boyuan

Empiricists

Limitation 2: Bad Details

Example: A photo of the time square

Empiricists

Limitation 2: Bad Details

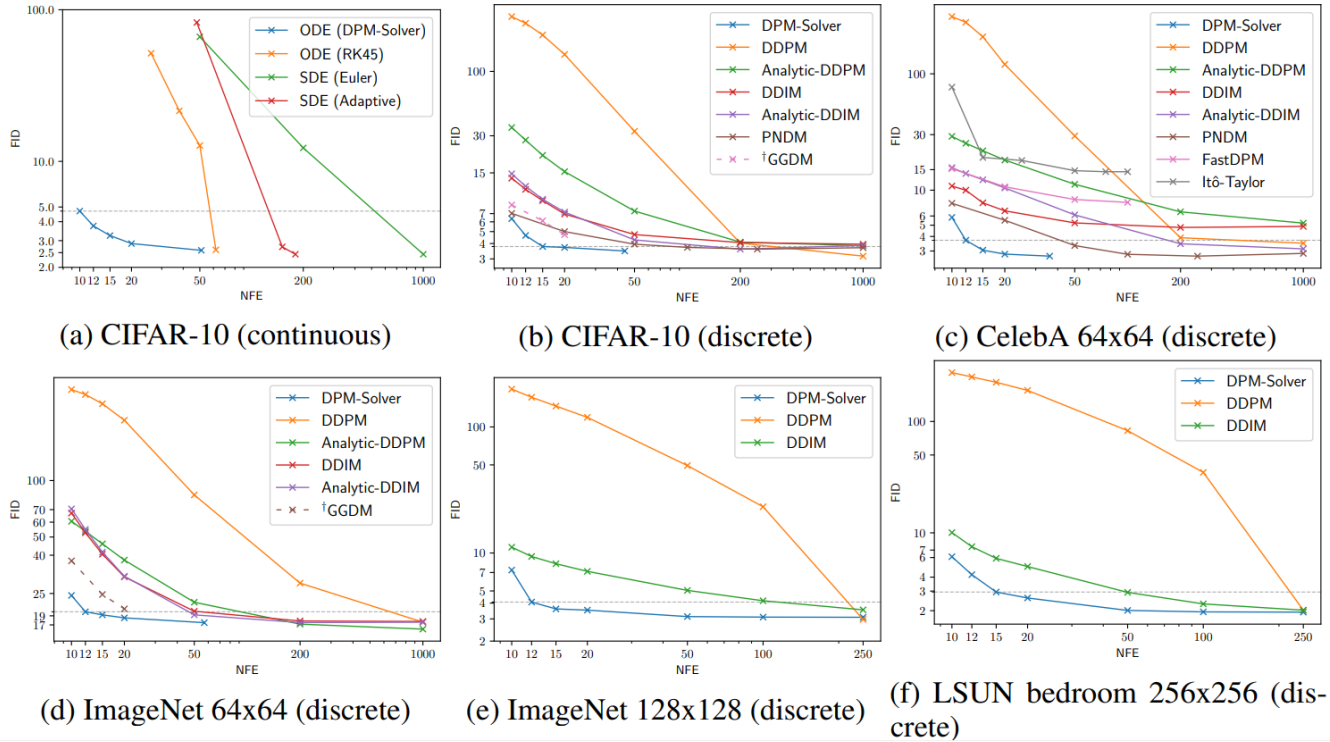
Example: A photo of the time square

Visionary

- Practicality
- Theory
- Application

Practicality

1.Speed-up in sampling



(a) CIFAR-10 (continuous) (b) CIFAR-10 (discrete) (c) CelebA 64x64 (discrete)

(d) ImageNet 64x64 (discrete) (e) ImageNet 128x128 (discrete) (f) LSUN bedroom 256x256 (discrete)

(2)Speed-up in training?

A remaining open problem…

Relax Constraints?

- Replace Langevin dynamics with?
  - Replace gaussian noise with more complex perturbation?
  - Hierarchical VAE is more powerful -> but less stable and harder to train?

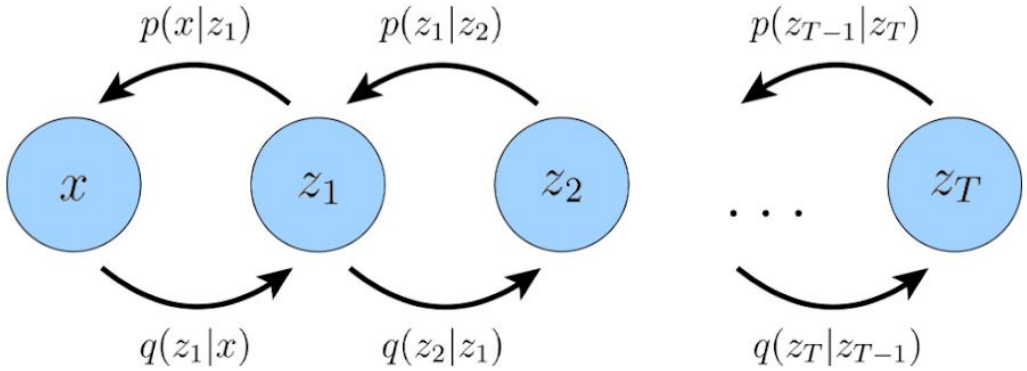
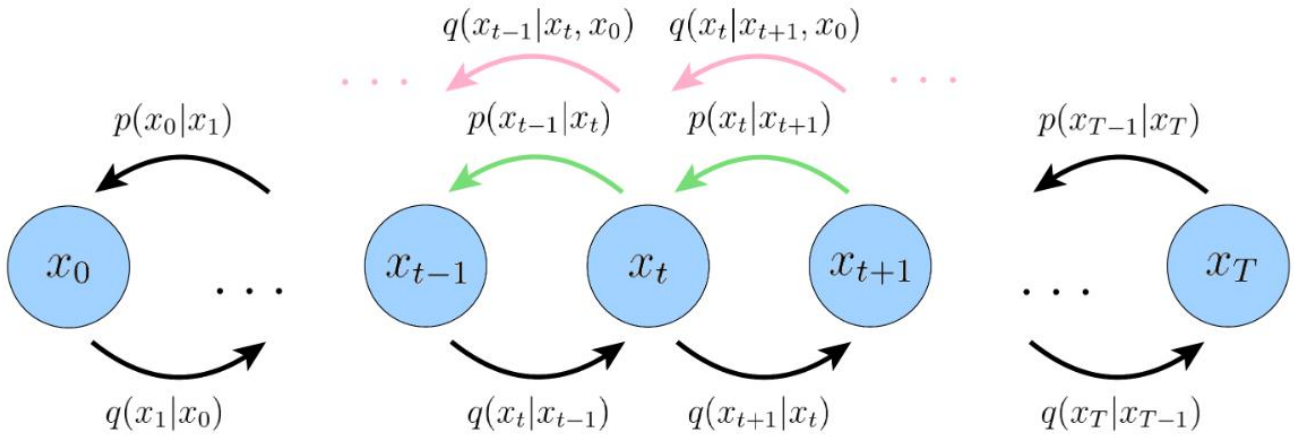


Figure 2: A Markovian Hierarchical Variational Autoencoder with T hierarchical latents. The generative process is modeled as a Markov chain, where each latent $\boldsymbol{z}_t$ is generated only from the previous latent $\boldsymbol{z}_{t+1}$.

Relax Constraints?

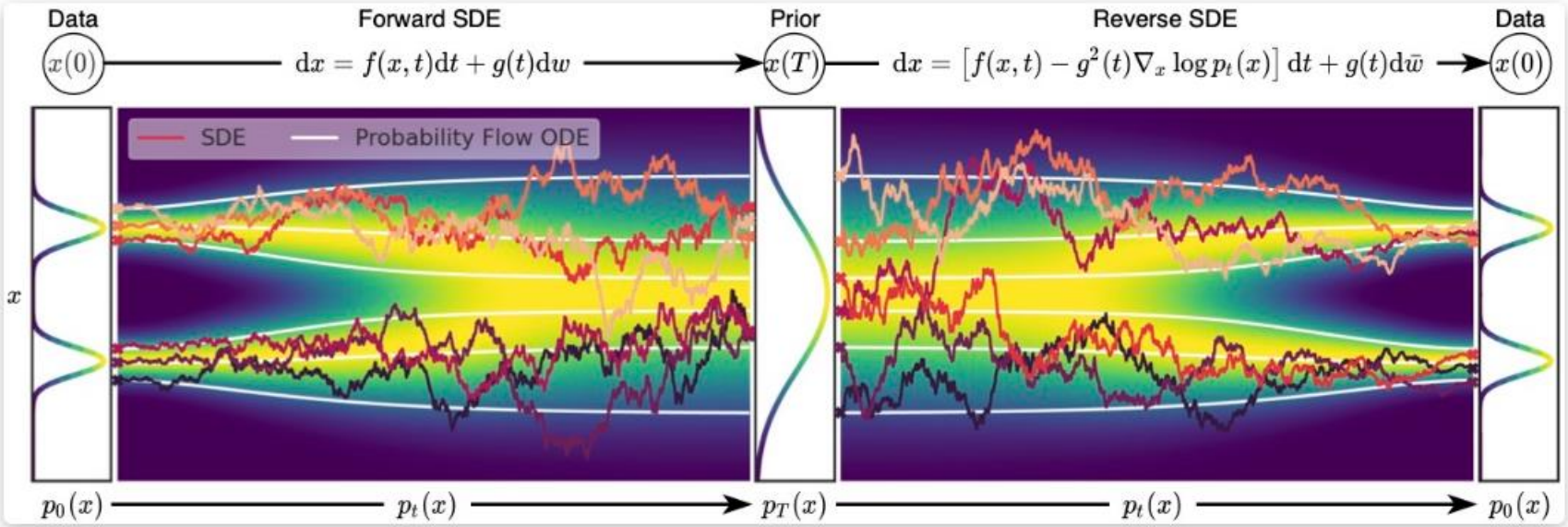
- In variational diffusion model, xo's information is also used
- Can/should we relax the markov property further?

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0) \parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}} \quad (58)$$

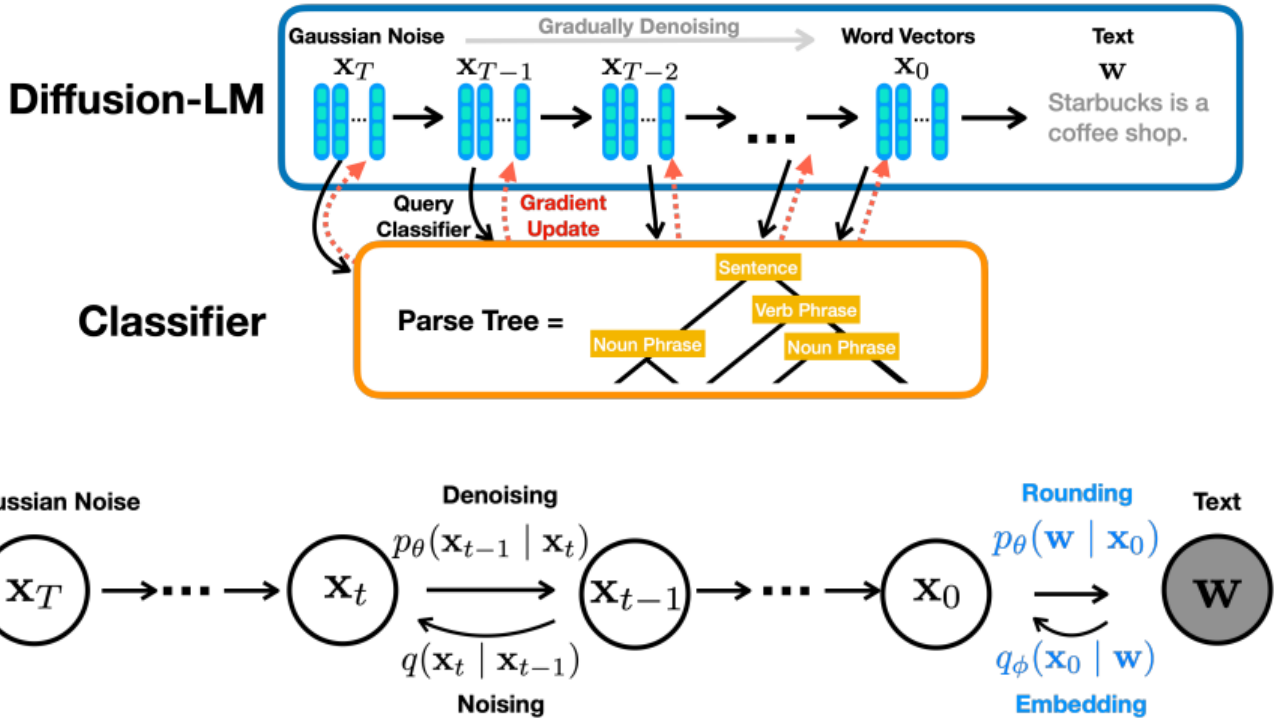


Other potential work

- Better sampling strategies
- Discrete -> Continuous process
  - Stochastic differential equations (SDE)
  - Probability flow ODE



Application-Diffusion for Language?



Application – Diffusion for Video?

- Video prediction
- Text-conditioned video generation