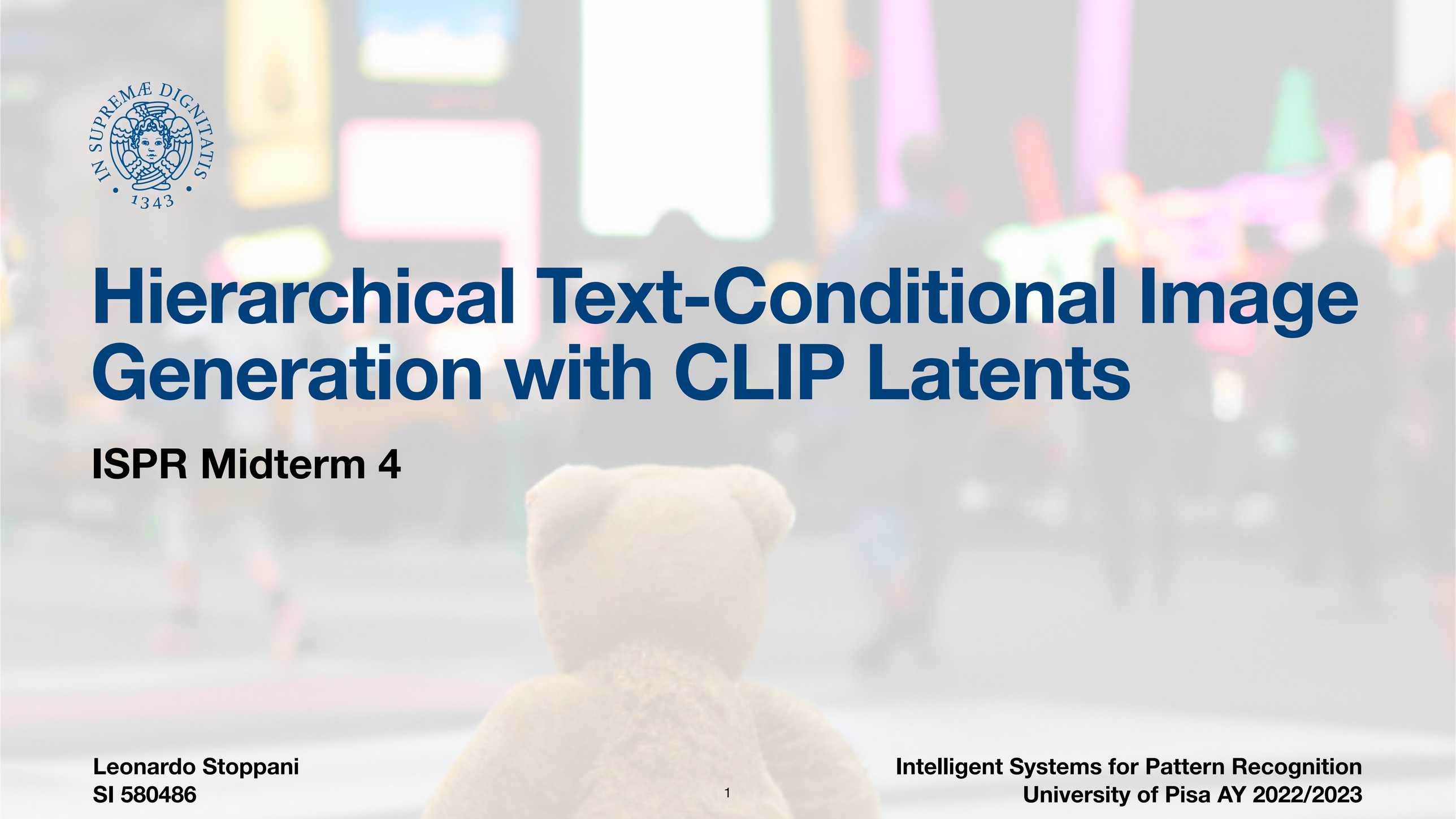


Hierarchical Text-Conditional Image Generation with CLIP Latents

ISPR Midterm 4

Leonardo Stoppani
SI 580486

Intelligent Systems for Pattern Recognition
University of Pisa AY 2022/2023

Introduction

Recent progress in computer vision driven by large dataset of captioned images

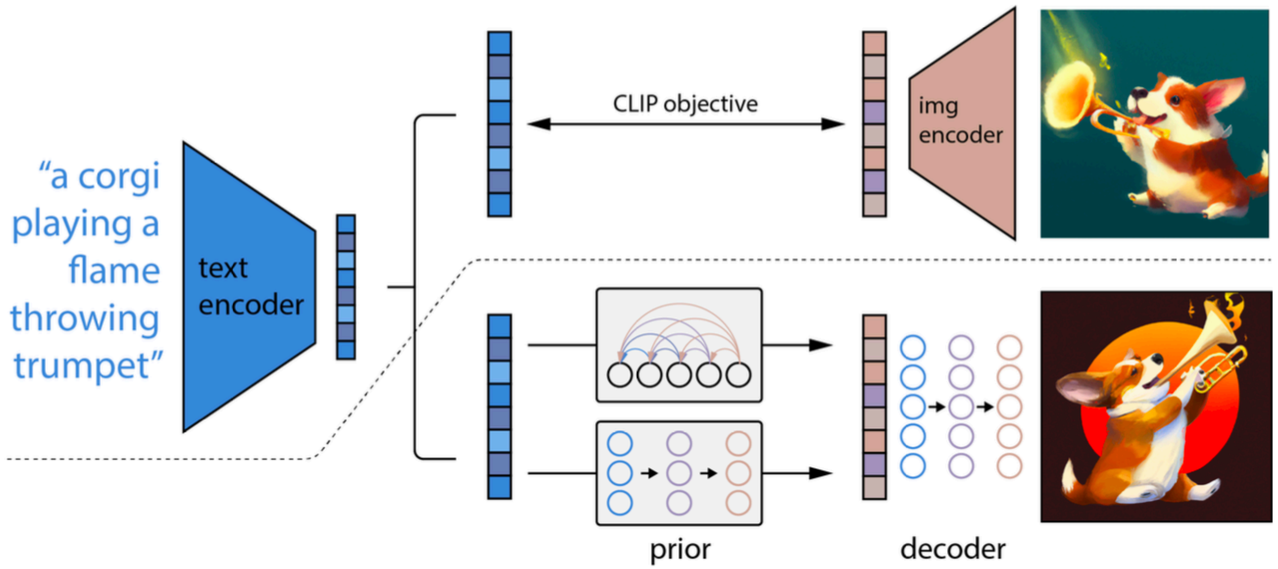
CLIP and Diffusion emerging as successful representation and generation learner for images

The paper “Hierarchical Text-Conditional Image Generation with CLIP Latents” by Aditya Ramesh et al. propose a novel approach to text-conditional image generation exploiting diffusion in CLIP latent space

unCLIP is a two-stage model capable of generating images by inverting the CLIP image encoder

Model - Description

The main idea is to exploit the CLIP latent space properties by conditioning the diffusion generation with the CLIP image embeddings



In order to achieve this:

- Train the CLIP model to learn a joint representation space for text and images
- Freeze the trained CLIP and generate image embeddings from text embeddings with a prior
- Generate the final image with diffusion conditioned on image embeddings

Model - Key Catch

$$P(x \mid y) = P(x, z_i \mid y) = P(x \mid z_i, y)P(z_i \mid y)$$

Training set as pairs (x, y) of images x and corresponding captions y

Given image x, z_i and z_t are its CLIP image and text embedding

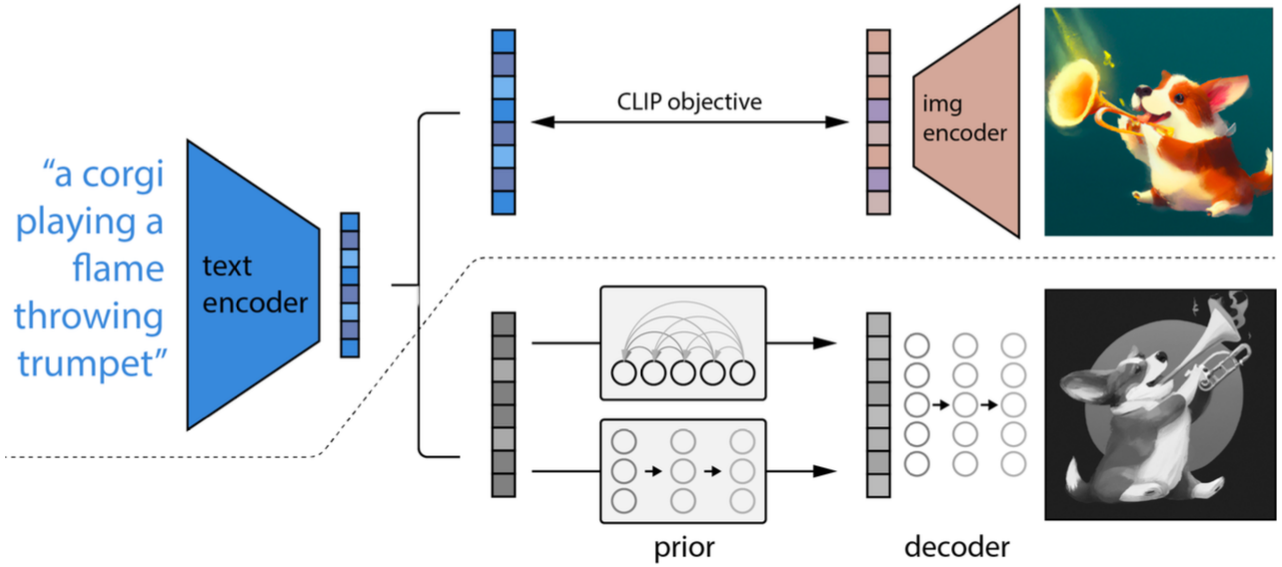
- $P(x \mid y) = P(x, z_i \mid y)$ holds for marginalisation on hidden variable z_i since embedding is unique for each image
- $P(x, z_i \mid y) = P(x \mid z_i, y)P(z_i \mid y)$ holds for chain rule

- $P(x \mid y)$ full stack generative model of images x given captions y -> **unCLIP**
- $P(x \mid z_i, y)$ probability of image x conditioned on CLIP z_i and caption y -> **Decoder**
- $P(z_i \mid y)$ probability of CLIP x_i conditioned on caption y -> **Prior**

Model - CLIP

CLIP (Contrastive Language–Image Pre-training) trains an image encoder and a text encoder in parallel to predict the correct image and caption pairings

CLIP objective is to minimise the differences between text and image encoding vectors



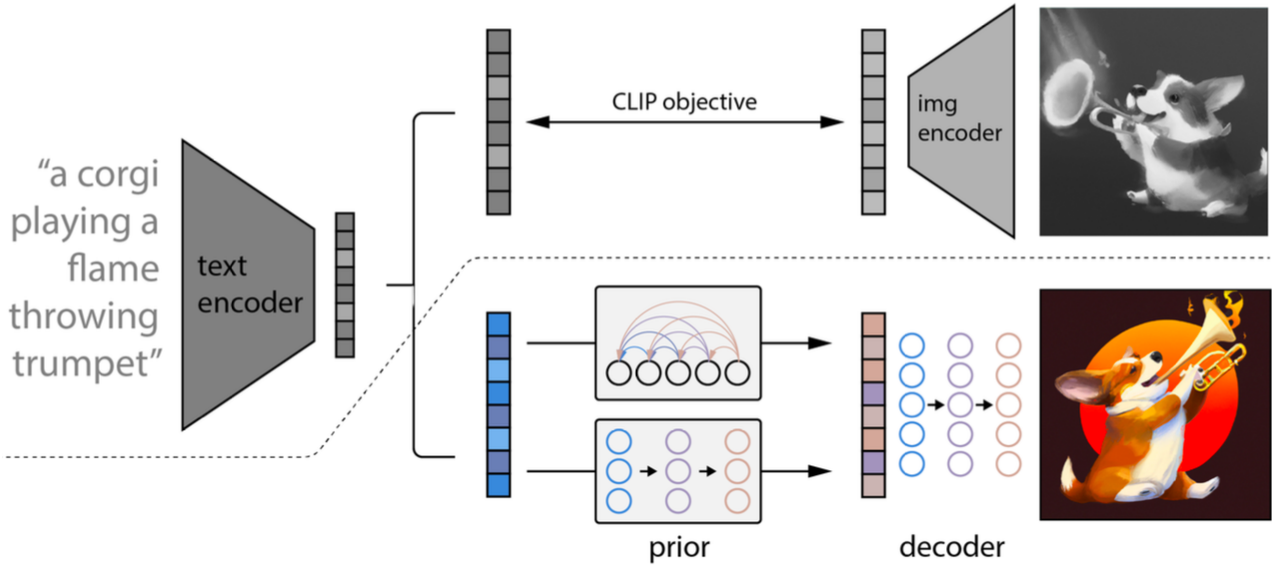
This approach allows to encode any image x into a bipartite representation (z_i, x_T) where latent z_i is the image CLIP representation and latent x_T encodes the residual information necessary for the decoder to reconstruct x

z_i is obtained by encoding the image with the CLIP image encoder, x_T by applying DDIM inversion to x conditioning on z_i

Model - Prior

Prior is a model which generates possible CLIP images embeddings from text embeddings and captions

Both Autoregressive and Diffusion models were tested as priors resulting in comparable performance with Diffusion being more compute-efficient



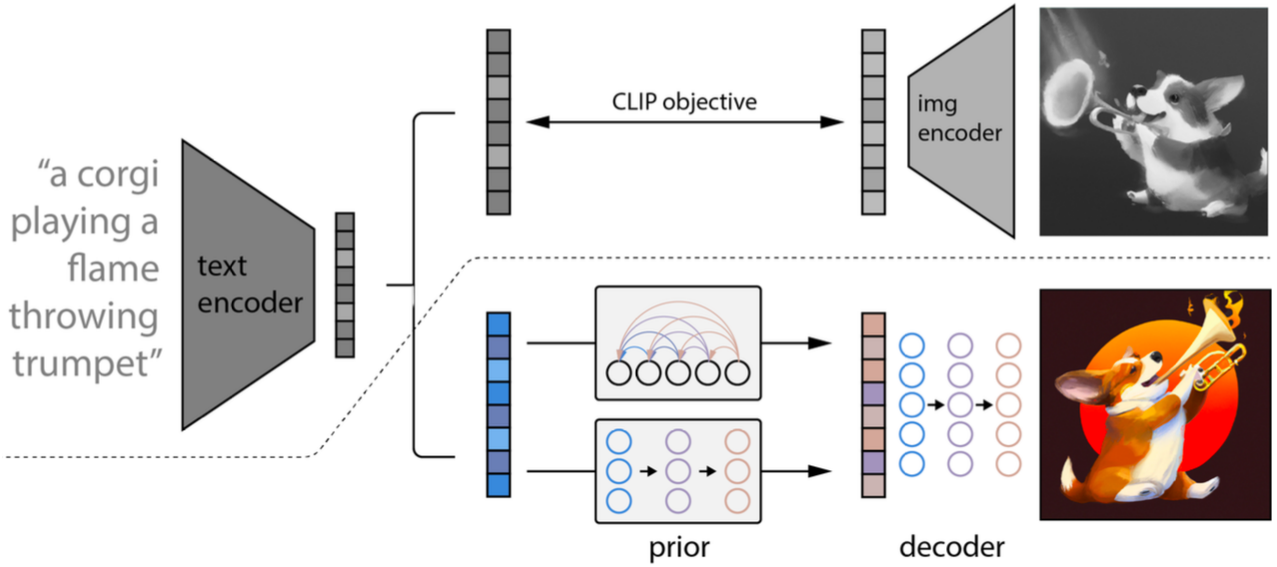
In order to increase the matching between caption and image, autoregressive prior was conditioned on a higher dot product $z_i \cdot z_t$

While for diffusion prior two sample z_i were generated and the one with highest dot product with z_t were selected

Model - Decoder

Decoder is a diffusion model trained by adding CLIP image embeddings to the existing tilmestep embedding

To generate high resolution images two diffusion upsamples were trained, one upsampling from 64x64 to 256x256 another to 1024x1024



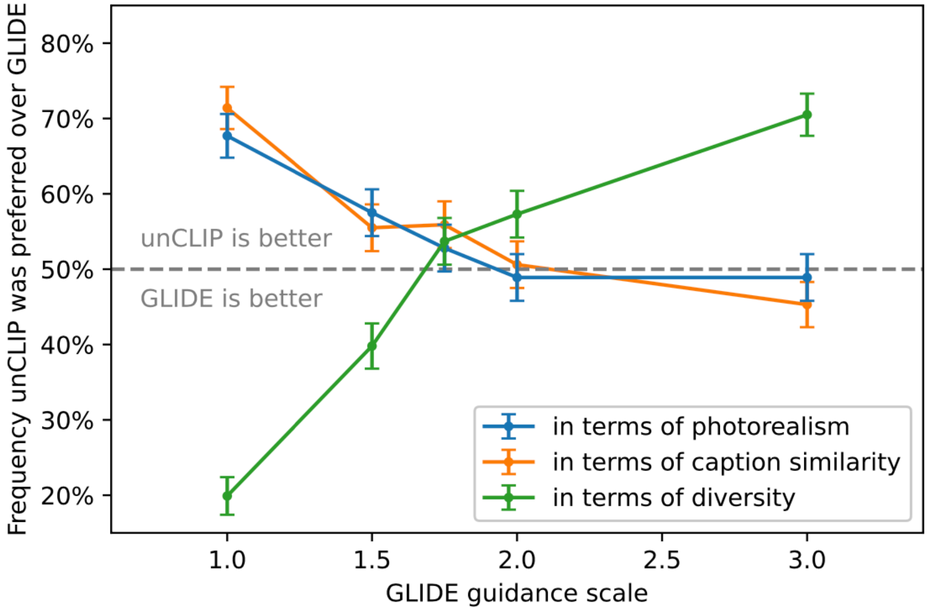
At inference time the model was applied directly at the target resolution, since it had already gained the ability to generalises to the higher resolution

True conditional distribution $P(x \mid y)$ can be sampled by first sampling z_i using the prior and then sampling x using the decoder

Key Results

Importance of the Prior underlined by comparing images generated by diffusion conditioned on caption (GLIDE), text embeddings and priors image embeddings. Worst result with just caption conditioning, reasonable with text embeddings, best with unCLIP full stack.

unCLIP yields greater diversity than **GLIDE** for comparable results in terms of photorealism and caption similarity



unCLIP achieves a new **SoTA** on MS-COCO validation set with a FID of 10.39 when sampling with the diffusion prior

Interestingly guiding unCLIP does **not decrease Recall** w.r.t. GLIDE, while still improving aesthetic quality

Considerations

Bordes et al. work also involved training diffusion models conditioned on CLIP image representation

The work in this paper **differs** mainly for the use of multimodal CLIP representation rather than image-only

A **key advantage** of using CLIP is that it embeds images and text to the same latent space, together to the bipartite representation enables different kinds of manipulations like variations, interpolations and text diffs

Although conditioning image generation on CLIP embeddings improve diversity, this approach comes with **limitations**

unCLIP is worse than GLIDE at binding attributes to the objects and struggle at producing coherent text and details in complex scenes

Last issue could be alleviate by training decoder at a higher base resolution at the cost of higher training and inference computational time