Hierarchical Text-Conditional Image Generation with CLIP Latents

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, Mark Chen

Group 6:
Jeffrey Chan, Qingyuan Li, Kevin Samms, Zhaoning Wang
Outline

- **Background/Motivation**
- **Method**
  - Overall Method
  - Prior
  - Decoder
- **Image Manipulation**
- **Text-to-Image Generation Analysis**
  - Why the prior matters?
  - GLIDE vs Dalle-2/unCLIP (Human Evaluation)
  - Diversity-Fidelity Trade-off with Guidance
- **Limitations**
Background/Motivation
Text to Image Generation

“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”
Conditioned Diffusion Model

\[ \hat{\epsilon}_\theta(x_t | y) = \epsilon_\theta(x_t | \emptyset) + s \cdot (\epsilon_\theta(x_t | y) - \epsilon_\theta(x_t | \emptyset)) \]

Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." Advances in Neural Information Processing Systems 34 (2021): 8780-8794.
Nichol, Alex, et al. "Glide: Towards photorealistic image generation and editing with text-guided diffusion models." arXiv preprint arXiv:2112.10741 (2021).
(1) Contrastive pre-training

(2) Create dataset classifier from label text

(3) Use for zero-shot prediction

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

\[ \hat{\mu}_\theta(x_t|c) = \mu_\theta(x_t|c) + s \cdot \Sigma_\theta(x_t|c) \nabla_{x_t} (f(x_t) \cdot g(c)) \]

from GLIDE: classifier-free guidance > CLIP guidance

Encoded Text

Diffusion Timestep \( t \)

Noised image \( x_t \)

CLIP Text Encoder

 ADM

Compute Gradients

CLIP Gradient

\[ \nabla_{x_t} (f(x_t) \cdot g(c)) \]
How use CLIP more effectively to improve generations?

“A motorcycle parked in a parking space next to another motorcycle.”

This work

Image Embeddings

Text Embeddings

CLIP Text Encoder

Better Generations

Not Great Generations

decoder

9
Method
unCLIP/DALL-E-2 architecture

- **Prior**
  - Given CLIP Text encoder output (text embedding) $\mathbf{y}$, generate corresponding Image Embedding $\mathbf{z}_i$

- **Decoder**
  - Produces the image from Image embedding $\mathbf{z}_i$
Prior

● Autoregressive (AR) prior:
  ○ AR models predict a sequence of data on a previous data sequence
  ○ Use a transformer to predict Image embedding sequence from the Text embedding sequence.

● Diffusion prior:
  ○ Diffusion model on CLIP Image Embedding
  ○ Input:
    ■ Encoded text
    ■ CLIP text embedding
    ■ Timestep
    ■ Noised CLIP Image Embedding
Diffusion Prior

CLIP Image embeddings $\tilde{z}_i$

Transformer

Diffusion Timestep $t - 1$

Denoised CLIP image embeddings $\tilde{z}_i^{(t-1)}$

SAME

Transformer

Encoded Text $C$

CLIP Text embeddings $\tilde{z}_t$

Diffusion Timestep $t$

Noised CLIP image embeddings $\tilde{z}_i^{(t)}$
Training

- Using CLIP to get input and ground-truth while training the prior.
Training Loss

\[ L_{\text{prior}} = \mathbb{E}_{t \sim [1,T], z_i^{(t)} \sim q_t} \left[ \| f_\theta(z_i^{(t)}, t, y) - z_i \|_2^2 \right] \]

* \( y \) is the combination of encoded text \( C \) and CLIP Text Embedding \( \tilde{z}_t \)
Decoder

- Diffusion model based on GLIDE
  - GLIDE uses a transformer to embedding the input text
  - Dall-E-2 put CLIP embedding into the process

- Upsampler
  - Used to generate higher-resolution Images
  - No conditioning, and no guidance
Decoder U-Net detail

32x32 -> 16x16 -> 8x8 -> 4x4 -> 8x8 -> 16x16 -> 32x32

CLIP Image embedding $\tilde{z}_i$

Diffusion Timestep $t$
Convolution Blocks

GLIDE Encoded Text

CLIP Image embeddings

Residual Block

Attention Layer

Residual Block

Diffusion Timestep $t$
Upsampler

2 unconditional off-the-shelf upsamplers to create images in higher resolution

[1] Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. arXiv:2102.09672, 2021.
[2] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image Super-Resolution via Iterative Refinement. arXiv:arXiv:2104.07636, 2021.
Training the decoder with CLIP encoder

CLIP Image Encoder (frozen)

MSE Loss

decoder
Inference

- **Prior**
  - Convert the CLIP Text Embedding to CLIP Image Embedding $\mathcal{Z}$

- **Decoder**
  - Produces the image from Image embedding $\mathcal{Z}$ and optionally with text embedding $\mathcal{Y}$.
Image Manipulations
What is Latent space

Interpolation in Latent Space
Bipartite latent representation $(\mathbf{Z}_i, \mathbf{X}_t)$

Encode with CLIP image encoder

DDIM inversion [1]

[1] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233, 2021
Variation

Input Image:

Generation:

Fix $z_i$

Vary $X_t$
Interpolation

Modify image embedding:

\[ z_{i\theta} = \text{slerp}(z_{i1}, z_{i2}, \theta) \]
Text Diff

\[ \mathbf{z}_i \mathbf{z}_d = \text{norm}(\mathbf{z}_t - \mathbf{z}_{t_0}) \]

Compute Difference CLIP Embeddings

\[ z_\theta = \text{slerp}(z_i, z_d, \theta) \]

Apply text diff direction to the image embedding

"a photo of a victorian house" → "a photo of a modern house"
Typographic Attacks

Attack:

Clip Image Prediction:

Generation Image Embedding:
Text-to-Image Generation Analysis
Why the prior matters?

- Condition decoder on captions alone: ✗
- Condition decoder on Caption + text embedding impersonating image embeddings: ✗
- Prior + CLIP image embedding: ✔
MS COCO FID SCORE

![Graph showing MS COCO FID scores for different guidance scales for GLIDE, unCLIP (AR), and unCLIP (Diffusion).]
GLIDE vs unCLIP

(MS-COCO)

MS-COCO - standard evaluation:

- Zero-shot FID score 10.39 - beats GLIDE & DALL-E in MS-COCO
GLIDE vs unCLIP
(Human Evaluations)

FID not always in agreement with human evaluation

Photorealism → winner: GLIDE - by small margin; 48.9%CI

Caption Similarity → winner: GLIDE - by small margin; 45.3%CI

Sample Diversity (4 x 4 grid) → winner: unCLIP stack by wide margin; 70.5%CI
Diversity-Fidelity Trade-off with Guidance

unCLIP has better diversity and relatively good fidelity

GLIDE is better

Image aesthetics improved for both unCLIP and GLIDE
GLIDE vs unCLIP

Aesthetic Quality

Result:
- Guidance improves GLIDE, and CLIP decoder (negative effect on CLIP prior)
- GLIDE sacrifices Recall for aesthetic quality improvement, unCLIP does not
Limitation of the model
Attribute Binding

- Suffer prompt where it must bind two separate objects (cubes) to two separate attributes (colors).
- Reconstructions mix up objects and attributes

“a red cube on top of a blue cube”.
A sign that says deep learning
Conclusion

- Image embedding creates better generation than text embeddings.
- CLIP embedding $Z_i$ holds image content information; meanwhile $X_t$ holds the style of image generation.
- Diffusion prior (Text-to-Image embeddings) increases the fidelity of image generation.
- unCLIP has limitations with attribute binding, text generation, and complex scenes.