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

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

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
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

decoders
Method
unCLIP/DALL-E-2 architecture

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

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

- Encoded Text C
- CLIP Text embeddings $\tilde{z}_t$
- Transformer
- Diffusion Timestep $t - 1$
- Denoised CLIP image embeddings $\tilde{z}_i^{(t-1)}$
- Transformer
- Diffusion Timestep $t$
- Noised CLIP image embeddings $\tilde{z}_i^{(t)}$
- CLIP Image embeddings $\tilde{z}_i$
Training

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

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

* \( y \) 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

Diffusion Timestep $t$

CLIP Image embedding $\tilde{z}_i$
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

Training the decoder with CLIP encoder

CLIP Image Encoder (frozen) → decoder → MSE Loss
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 $\mathbf{y}$.
Image Manipulations
What is Latent space

Interpolation in Latent Space
Bipartite latent representation \((z_i, x_t)\)

Encode with CLIP image encoder

DDIM inversion [1]

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

“a photo of a victorian house” $\rightarrow$ “a photo of a modern house”

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

Compute Difference CLIP Embeddings

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

Apply text diff direction to the image embedding
Typographic Attacks

**Attack:**

**Clip Image Prediction:**

- **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%

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

- Condition decoder on captions alone
  - [X]

- Condition decoder on Caption + text embedding impersonating image embeddings
  - [X]

- Prior + CLIP image embedding
  - [✓]
GLIDE vs unCLIP (MS-COCO)

MS-COCO - standard evaluation:

- Zero-shot FID score 10.39 - beats GLIDE & DALL-E in MS-COCO

<table>
<thead>
<tr>
<th>Model</th>
<th>FID</th>
<th>Zero-shot FID</th>
<th>Zero-shot FID (filt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AttnGAN (Xu et al., 2017)</td>
<td>35.49</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>DM-GAN (Zhu et al., 2019)</td>
<td>32.64</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>DF-GAN (Tao et al., 2020)</td>
<td>21.42</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>DM-GAN + CL (Ye et al., 2021)</td>
<td>20.79</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>XMC-GAN (Zhang et al., 2021)</td>
<td>9.33</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>LAFITE (Zhou et al., 2021)</td>
<td>8.12</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>Make-A-Scene (Gafni et al., 2022)</td>
<td>7.55</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>DALL-E (Ramesh et al., 2021)</td>
<td>~28</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>LAFITE (Zhou et al., 2021)</td>
<td>26.94</td>
<td>10.39</td>
<td></td>
</tr>
<tr>
<td>GLIDE (Nichol et al., 2021)</td>
<td>12.24</td>
<td>10.39</td>
<td>12.89</td>
</tr>
<tr>
<td>Make-A-Scene (Gafni et al., 2022)</td>
<td>11.84</td>
<td>10.39</td>
<td>11.08</td>
</tr>
<tr>
<td>unCLIP (AR prior)</td>
<td>10.63</td>
<td>10.39</td>
<td>11.08</td>
</tr>
<tr>
<td>unCLIP (Diffusion prior)</td>
<td>10.39</td>
<td>10.39</td>
<td>10.87</td>
</tr>
</tbody>
</table>
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
Complex Scene
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.