Shifted Diffusion for Text-to-image Generation

Yufan Zhou[1], BingChen Liu[2], Yizhe Zhu[2], Xiao Yang[2], Changyou Chen[1], Jinhui Xu[1]

[1]University at Buffalo, SUNY
[2]ByteDance

WED-AM-184

Introduction

We propose Shifted Diffusion, a novel diffusion model which generates image embeddings from text.

By integrating prior knowledge of pre-trained CLIP model into the diffusion process, we can enhance the accuracy of generating image embeddings.

With Shifted Diffusion, we can

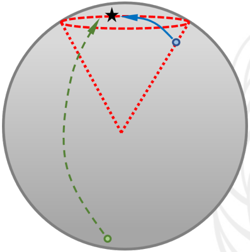
- Improve text-to-image generation models by introducing an extra image embedding input;
- Train or fine-tune text-to-image generation models on image-only dataset, without heavy workload of human captioning;

Shifted Diffusion

We generate CLIP image embedding using diffusion models.

It has been shown that the effective output space of CLIP encoder is restricted to a narrow cone.

Instead of generating embeddings from Gaussian noise (green arrow) like previous methods, we propose to generate embeddings from random embedding (blue arrow).



Shifted Diffusion

Specifically, we design the diffusion process to be

$$q(\mathbf{z}^t \mid \mathbf{z}^{t-1}) = \mathcal{N}(\mathbf{z}^t; \sqrt{1-\beta_t}\,\mathbf{z}^{t-1} + \mathbf{s}_t, \beta_t \boldsymbol{\Sigma}),$$

which has an extra shift term compared to baseline diffusion

$$q(\mathbf{z}^t \mid \mathbf{z}^{t-1}) = \mathcal{N}(\mathbf{z}^t; \sqrt{1-\beta_t}\,\mathbf{z}^{t-1}, \beta_t \mathbf{I}),$$

Shifted Diffusion

We can show that

$$q(\mathbf{z}^t \mid \mathbf{z}^0) = \mathcal{N}(\mathbf{z}^t; \sqrt{\bar{\alpha}_t}\, \mathbf{z}^0 + \sum_{i=1}^{t} \mathbf{s}_i \sqrt{\bar{\alpha}_t / \bar{\alpha}_i}, (1 - \bar{\alpha}_t)\boldsymbol{\Sigma}),$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_i)$. Specifically, we set $\mathbf{s}_t = (1 - \sqrt{1 - \beta_t})\boldsymbol{\mu}$, which leads to

$$q(\mathbf{z}^t \mid \mathbf{z}^0) = \mathcal{N}(\mathbf{z}^t; \sqrt{\bar{\alpha}_t}\, \mathbf{z}^0 + (1 - \sqrt{\bar{\alpha}_t})\boldsymbol{\mu}, (1 - \bar{\alpha}_t)\boldsymbol{\Sigma}).$$

Shifted Diffusion

We can derive a posterior distribution

$$
\begin{aligned}
q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0) &= \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\nu}, \boldsymbol{\Lambda}), \\
\boldsymbol{\nu} &= \gamma(\mathbf{z}_t - \mathbf{s}_t) + \eta\, \mathbf{z}_0 + \tau(1 - \sqrt{\bar{\alpha}_{t-1}})\boldsymbol{\mu}, \\
\boldsymbol{\Lambda} &= (1 - \bar{\alpha}_{t-1})\beta_t \boldsymbol{\Sigma}/(1 - \bar{\alpha}_t),
\end{aligned}
$$

where

$$
\begin{aligned}
\gamma &= (1 - \bar{\alpha}_{t-1})\sqrt{\alpha_t}/(1 - \bar{\alpha}_t), \\
\eta &= \beta_t \sqrt{\bar{\alpha}_{t-1}}/(1 - \bar{\alpha}_t), \\
\tau &= \beta_t/(1 - \bar{\alpha}_t).
\end{aligned}
$$

$\boldsymbol{\mu}, \boldsymbol{\Sigma}$ denote mean and covariance matrix of random image embedding.

Shifted Diffusion

Because we have close-form expression of $q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0)$, $q(\mathbf{z}^t \mid \mathbf{z}^0)$.

The diffusion loss

$$\mathbf{L}_{\boldsymbol{\theta}} = \mathbb{E}_q\{\mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}_T \mid \mathbf{z}_0) \| p(\mathbf{z}_T)) - \log p_{\boldsymbol{\theta}}(\mathbf{z}_0 \mid \mathbf{z}_1) + \sum_{t>1} \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}_0) \| p_{\boldsymbol{\theta}}(\mathbf{z}_{t-1} \mid \mathbf{z}_t))\},$$

now has closed-form solution which can be easily optimized by methods such as gradient descent.

Experiments

Shifted diffusion allows for efficient adaptation by successfully fine-tuning a pre-trained text-to-image generation model using an image-only dataset.

This is crucial as pre-trained models often struggle to meet specific requirements, and supervised fine-tuning can be challenging as creating image-text pairs can be labor-intensive.

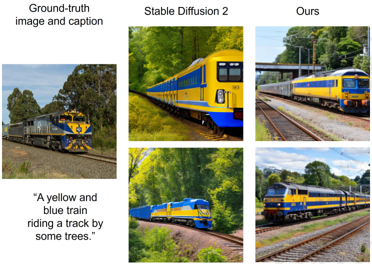


Figure: Pre-trained Stable Diffusion 2 vs. our fine-tuned model.

Experiments

Shifted diffusion is better than baseline diffusion, lower FID scores and higher CLIP similarities are achieved.

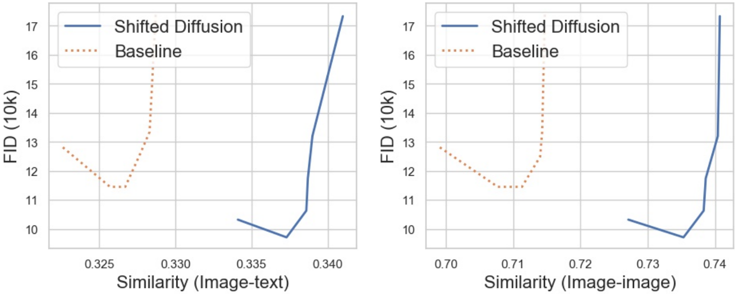
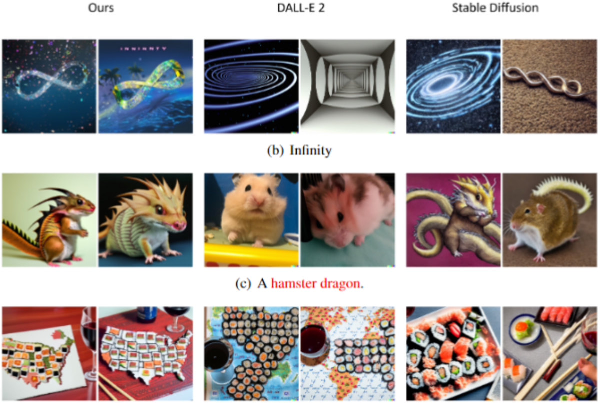


Figure: Comparison between baseline and Shifted Diffusion on MS-COCO, evaluated with the same fine-tuned Stable Diffusion 2 model.

Experiments

Some generated examples on standard text-to-image generation. We use Shifted Diffusion to introduce an extra image embedding input for text-to-image generation model.



(b) Infinity

(c) A hamster dragon.

(d) A map of the United States made out sushi. It is on a table next to a glass of red wine.

Experiments

Ours	DALL-E 2	Stable Diffusion

(e) A portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. A full moon over the city of Los Angeles is in the background at night.

(f) A cute wooden owl statue holding a large globe of the Earth above its head.

(g) A statue of Abraham Lincoln wearing an opaque and shiny astronaut's helmet. The statue sits on the moon, with the planet Earth in the sky.

Thank You