LDEdit: Towards Generalized Text Guided Image Manipulation via Latent Diffusion Models

Paramanand Chandramouli, Kanchana Vaishnavi Gandikota
University of Siegen

LDEdit- Introduction

Goal To develop a fast and flexible approach to open domain image manipulation from text prompts.

Our Solution Adapt pretrained text-to-image latent diffusion model to perform text guided manipulations using DDIM sampling.

Advantages

- Faster manipulation in lower dimensional latent space.
- DDIM sampling ensures a near cycle-consistency between source & target.

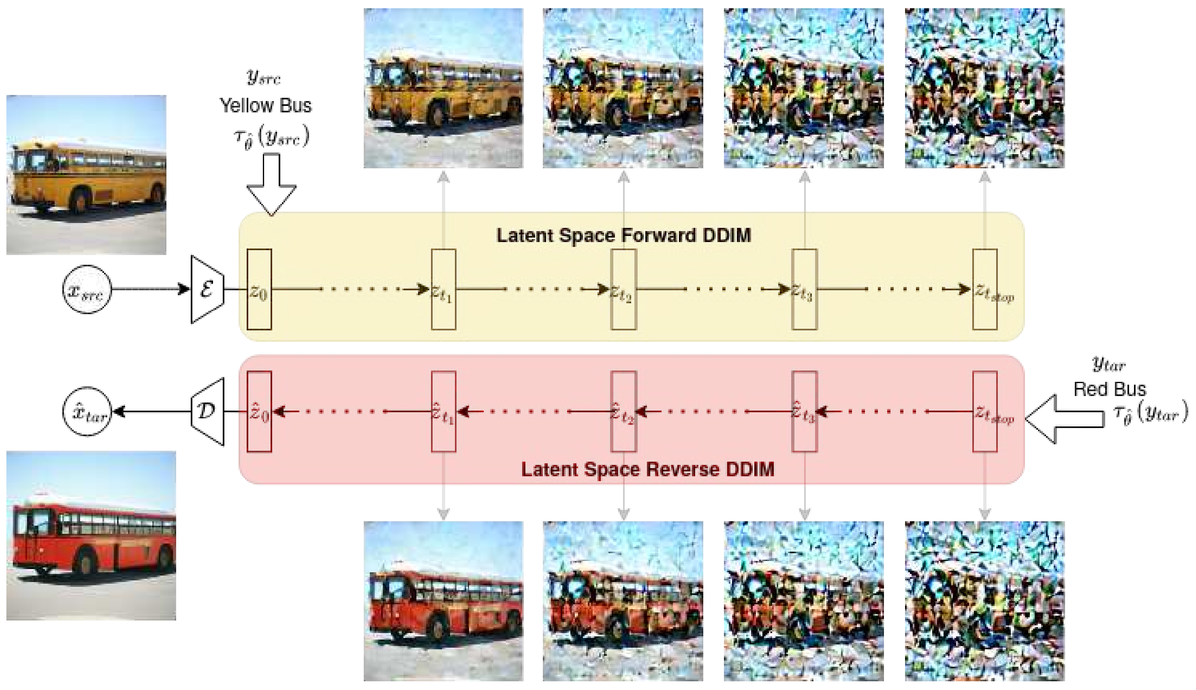
Preliminaries

Denoising Diffusion Implicit Models (DDIM) employ non-Markovian diffusion. DDIM reverse process is given as

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(x_t, t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2(\eta)}\boldsymbol{\epsilon}_\theta(x_t, t) + \sigma_t^2(\eta)\xi,$$

$\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\alpha_0 := 1$, and, α_t depends on noise variance schedule.
$\eta \in \mathbb{R}_{\geq 0}$ stochasticity hyperparameter (for fully deterministic sampling $\eta = 0$)

Overview of LDEdit



Implementation

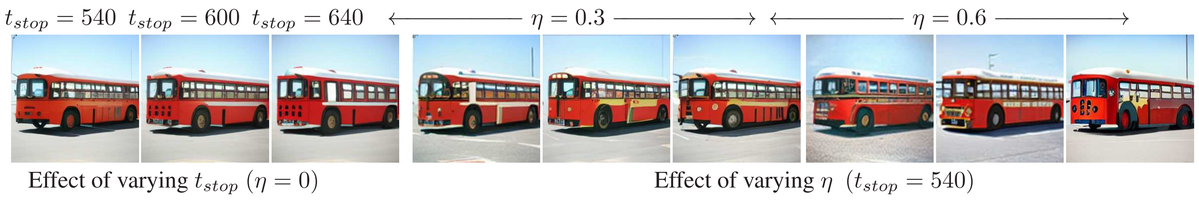
Let $\boldsymbol{f}_\theta(z_t, t, y_{cond}) = \left(\frac{z_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(z_t, t, \tau_{\tilde{\theta}}(y_{cond}))}{\sqrt{\alpha_t}} \right)$

Manipulation using LDM involves:

- Forward DDIM process conditioned on y_{src} from z_0 till $z_{t_{stop}}$ with $t_{stop} < T$
- The reverse DDIM process conditioned on y_{tar} starting from $z_{t_{stop}}$ to arrive at $\hat{z}_0$

$$z_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{f}_\theta(z_t, t, y_{tar}) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2(\eta)}\boldsymbol{\epsilon}_\theta(y_t, t, \tau_{\tilde{\theta}}(y_{tar})) + \sigma_t^2(\eta)\xi$$

- Deterministic sampling ensures a near cycle-consistency between x_{src} and $\hat{x}_{\text{tar}}$
- Use of stochasticity can produce diverse outputs by trading off consistency with x_{src}.



Effect of varying t_{stop} $(\eta = 0)$ Effect of varying η $(t_{stop} = 540)$

Visual Comparisons

Visual Comparisons

a) Reconstruction and style transfer b) Domain transfer c) Multi-attribute semantic changes

Local Editing with Mask Inputs

Simultaneous editing of multiple attributes

Left to right: (i) input (ii) girl with watermelon (iii) woman with corgi (iv) old woman with a cat-painting (v) boy with a big egg-painting (vi) man with a rabbit-painting (vii) man with a dog-painting.

Left to right: (i) Input (ii) girl with a dog (iii) woman with a dog (iv) old woman with a dog (v) boy with a basketball (vi) man with a basketball (vii) old man with a basketball

Artistic style transfer from text prompts

References

[1] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *ECCV*, 2022.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[3] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. In *CVPR*, 2022.

[4] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *CVPR*, 2022.

[5] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.