REPORT

Dragondiffusion: Enabling drag-style manipulation on diffusion models

Chong Mou[1]; Xintao Wang[2]; Jiechong Song[1]; Ying Shan[1];Jian Zhang[1]

[1]Peking University Shenzhen Graduate School, Shenzhen, China

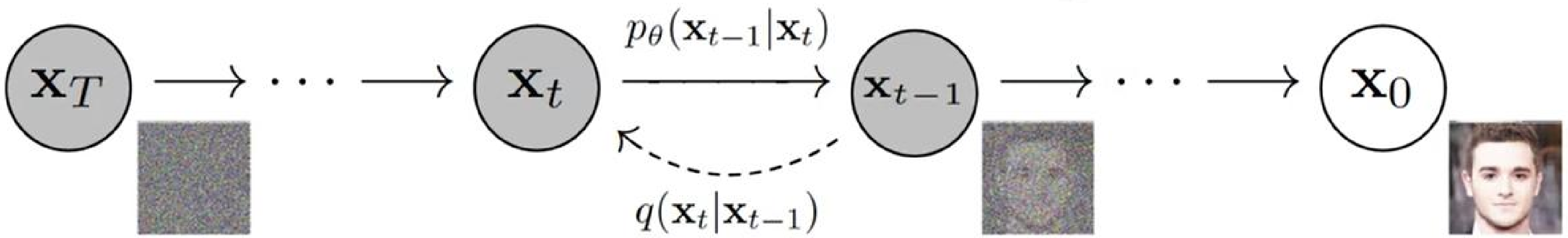
[2]ARC Lab, Tencent, PCG

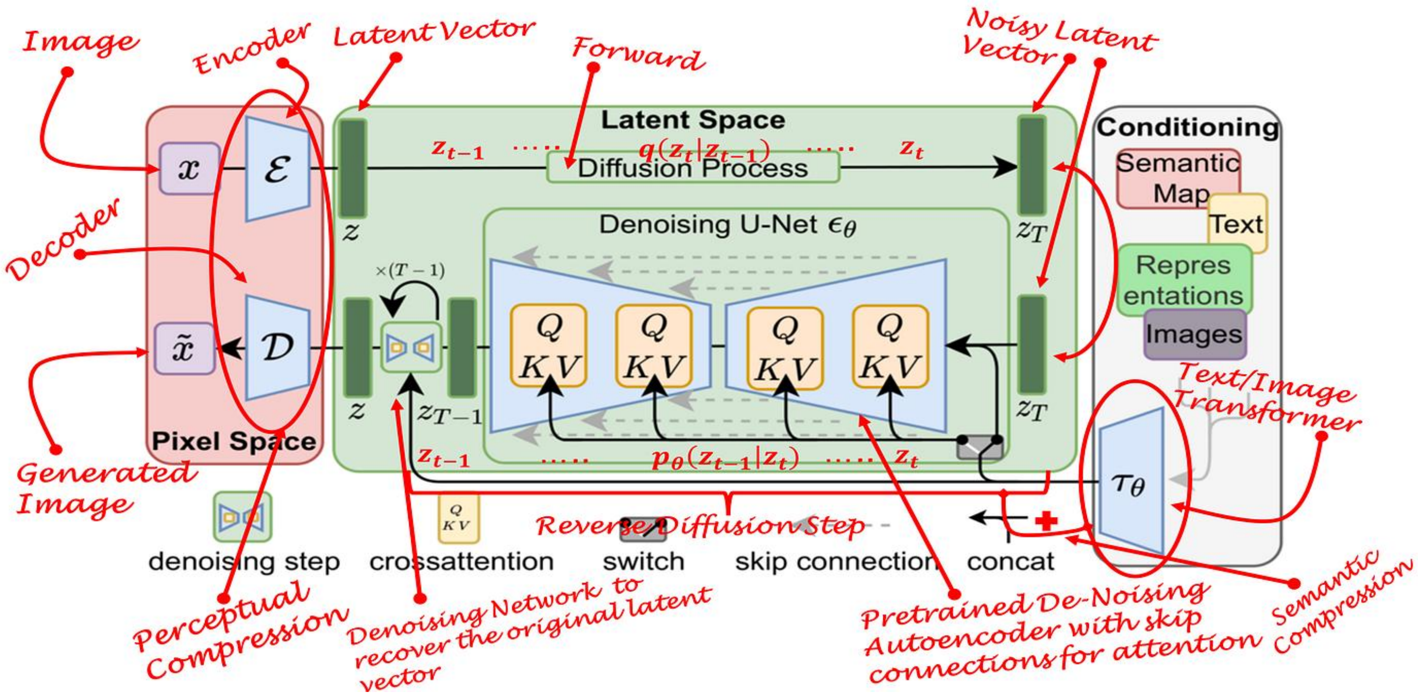
https://villa.jianzhang.tech/

北京大學 PEKING UNIVERSITY VILLA Visual-Information Intelligent Learning LAB ARC Applied Research Center ICLR

01 Background: Text-to-Image Diffusion Model

➤ **Stable Diffusion**





$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t,t)\|_2^2\right]$$

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t,t)\|_2^2\right]$$

01

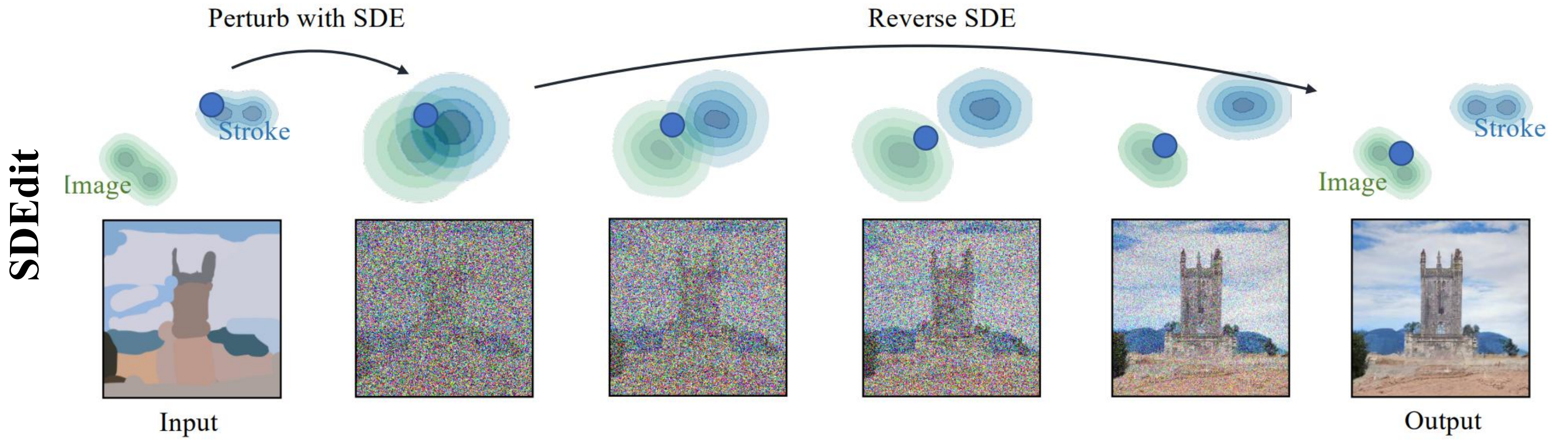
Background: Text-to-Image Diffusion Model

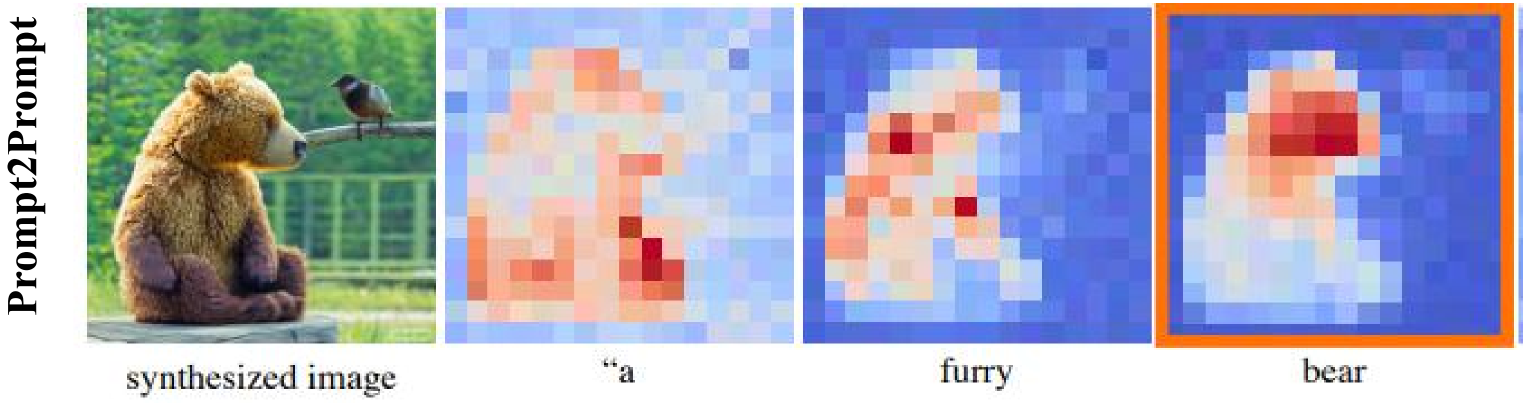
Relying on the powerful generation capabilities of SD:

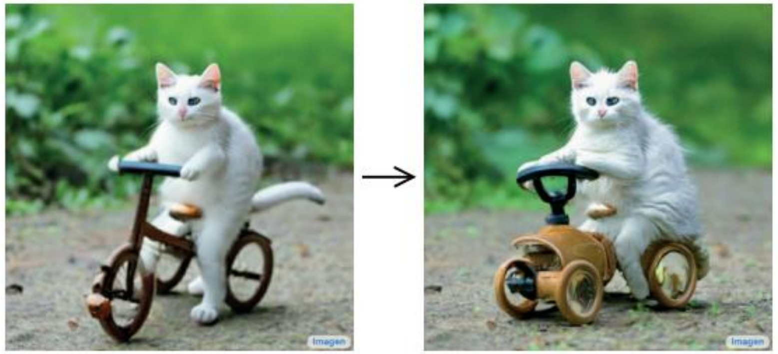
- How can we edit existing images?

  We propose DragonDiffusion.

01 Background: Diffusion-based Image Editing

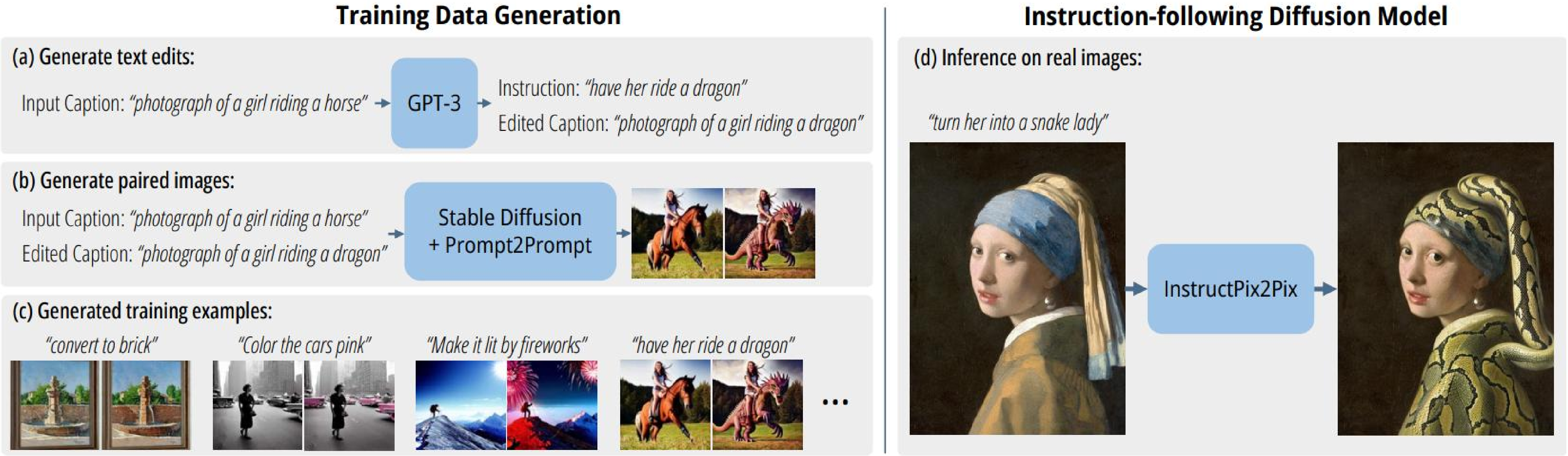






01 Background: Diffusion-based Image Editing

InstructPix2Pix:

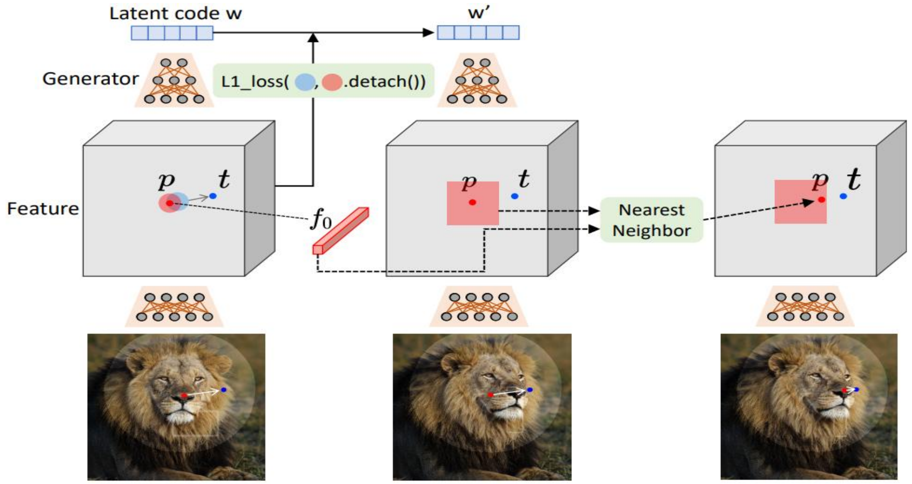


However, the correspondence between text and image features is weak, heavily relying on the design of prompts.

01

Background: DragGAN

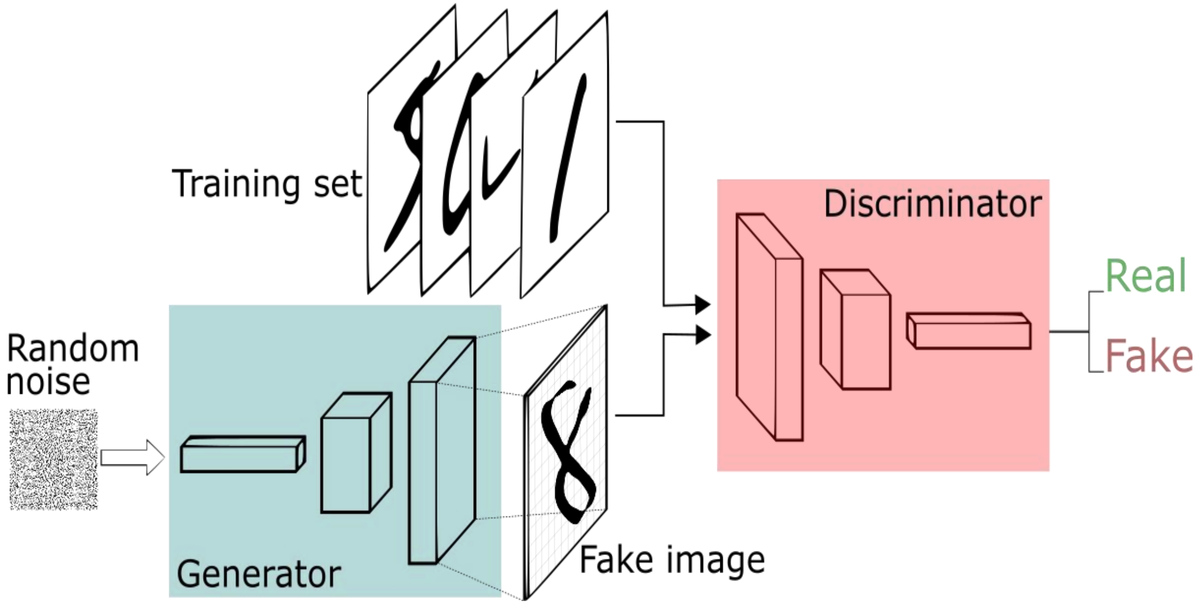
- **DragGAN**



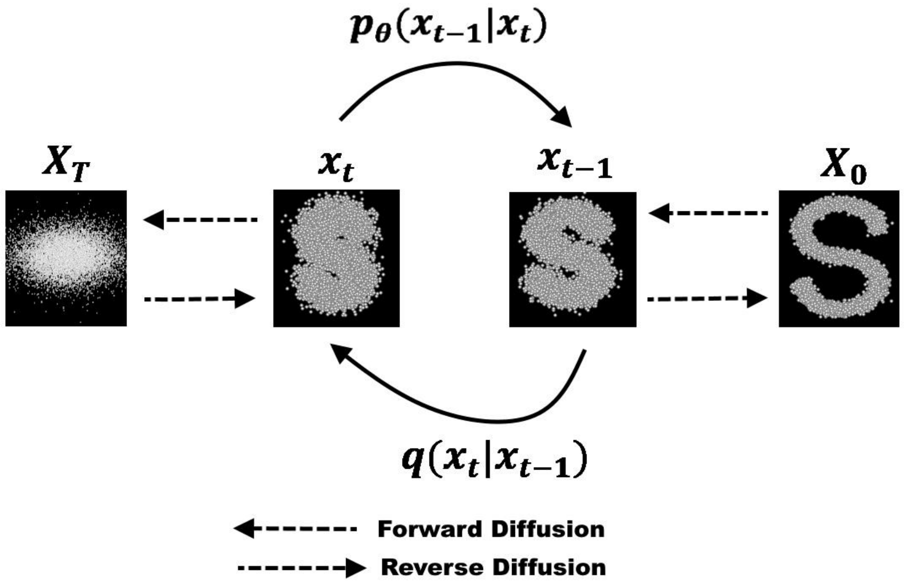
Original Image **Edit w/o alignment** **Edit w alignment**

Due to the limited capability of the GANs.

01 Background: DragGAN



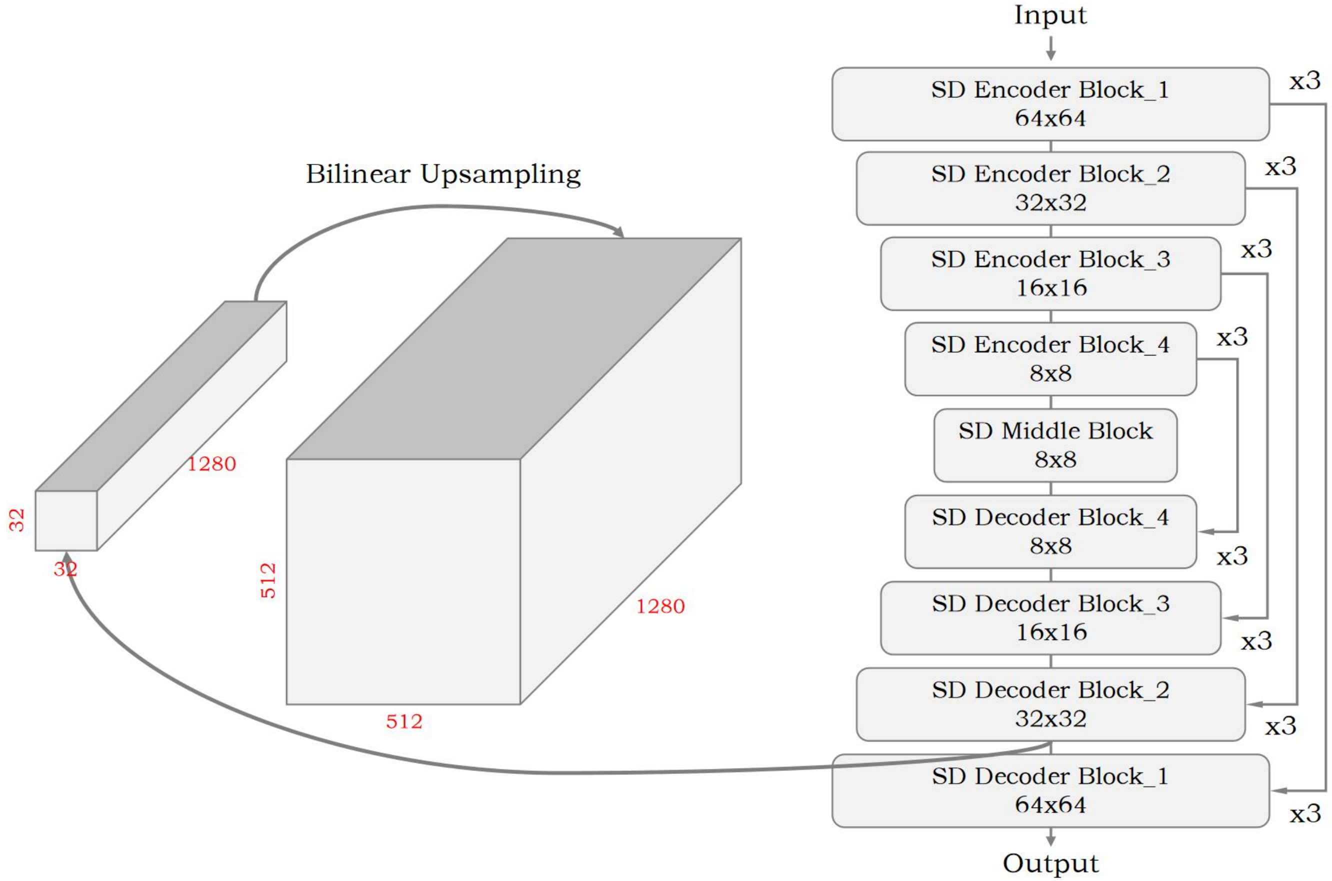
GAN model: compact and editable latent space

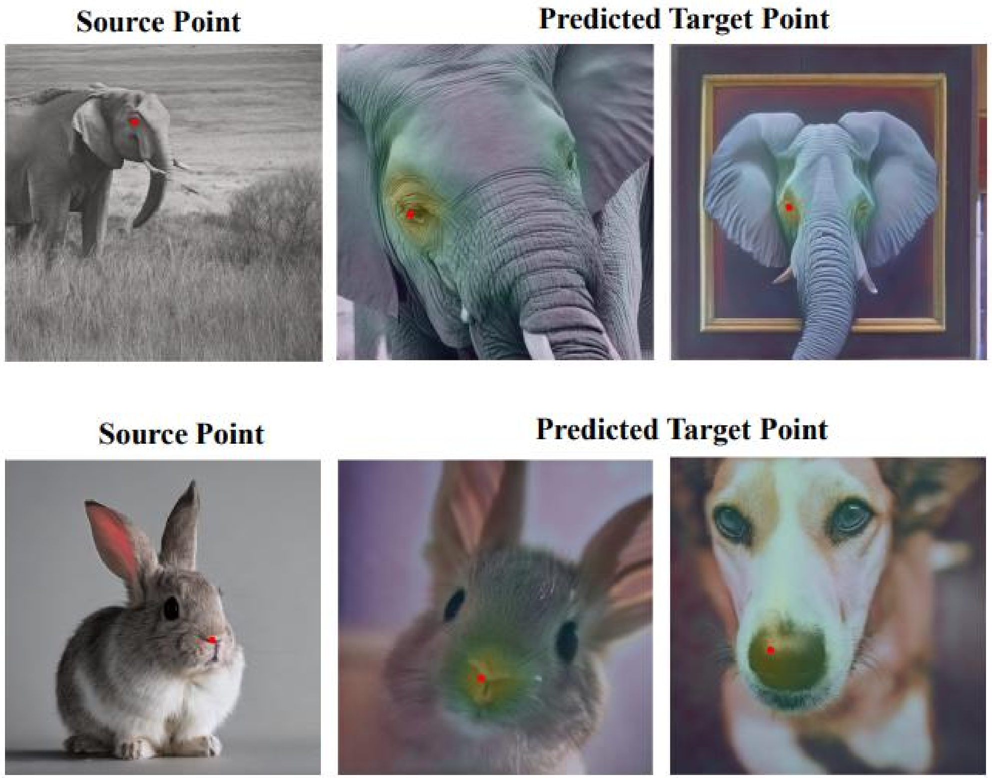


Diffusion model: Discrete latent space

Is there a way to perform fine-grained image editing based on SD?

02 Method



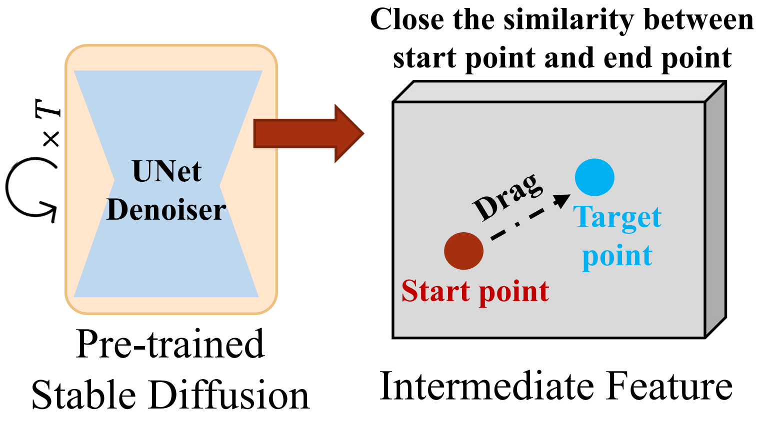


Emergent Correspondence from Image Diffusion (NeurIPS 2023)

02 Method

Score-based Guidance

- **Editing modeling:**

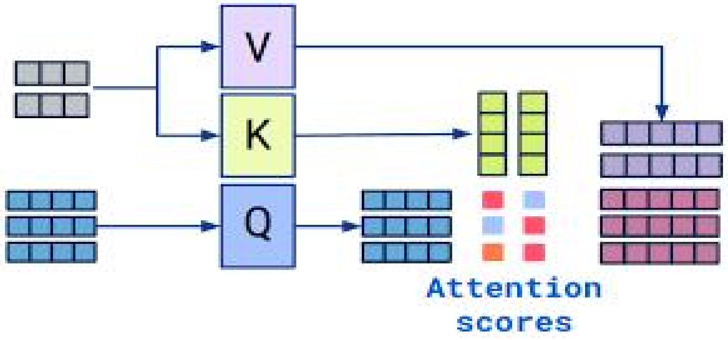


- **Design energy function for guidance:**

$$\mathcal{E} = \underbrace{w_e \cdot \mathcal{E}_{edit}}_{\text{Editing term}} + \underbrace{w_c \cdot \mathcal{E}_{content}}_{\text{Content consistency term}}$$

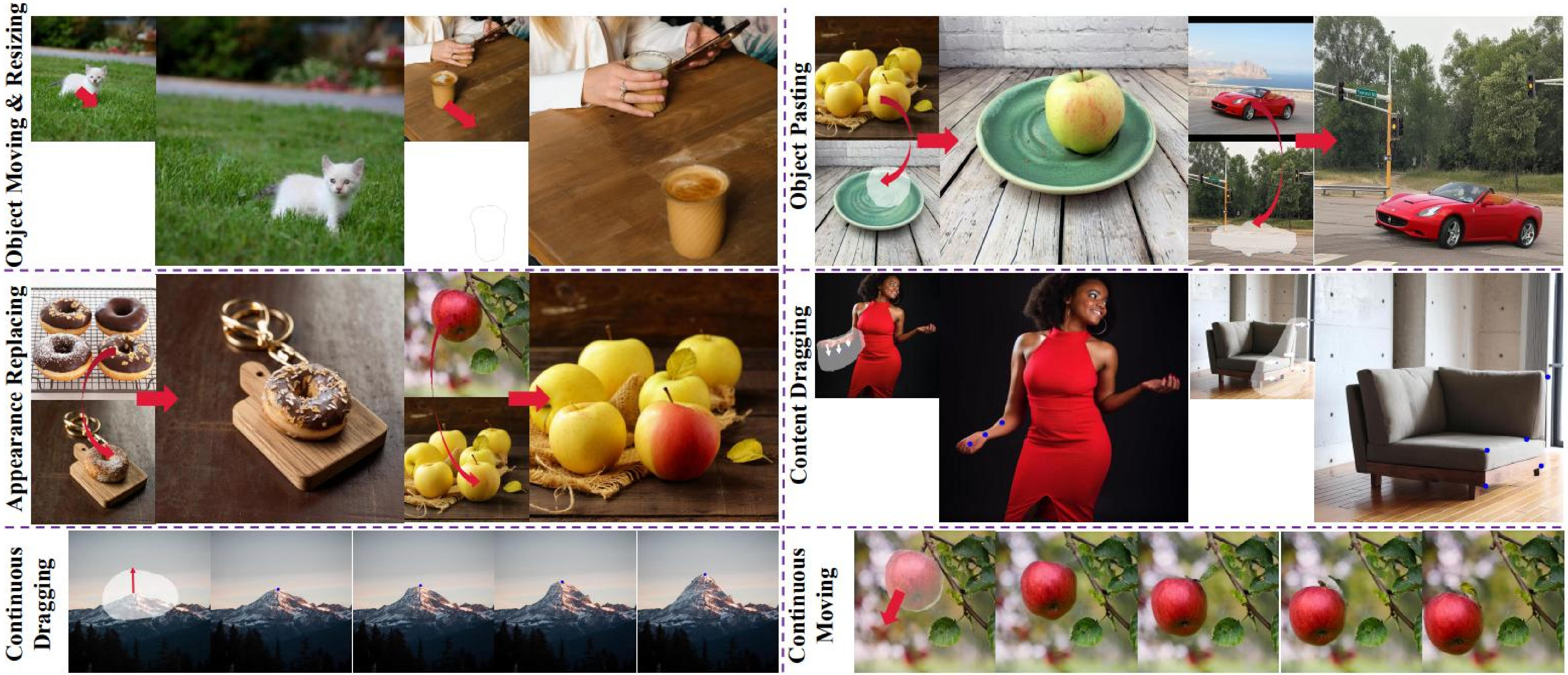
$$\begin{cases} \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t|\mathbf{y}) = \nabla_{\mathbf{x}_t} \log \left(\dfrac{q(\mathbf{y}|\mathbf{x}_t)q(\mathbf{x}_t)}{q(\mathbf{y})} \right) \\ \qquad \propto \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log q(\mathbf{y}|\mathbf{x}_t), \\ \tilde{\epsilon}_\theta^t(\mathbf{x}_t) = \epsilon_\theta^t(\mathbf{x}_t) + \eta \cdot \nabla_{\mathbf{x}_t} \mathcal{E}(\mathbf{x}_t, \mathbf{y}), \end{cases}$$

Content consistency via visual cross-attention



Key and Value are the diffusion feature from the reference image.

03 Results



Object Moving & Resizing
Object Pasting
Appearance Replacing
Content Dragging
Continuous Dragging
Continuous Moving

04 Demo

Thanks!

Chong Mou[1]; Xintao Wang[2]; Jiechong Song[1]; Ying Shan[1];Jian Zhang[1]

[1]Peking University Shenzhen Graduate School, Shenzhen, China

[2]ARC Lab, Tencent, PCG