AI Painting

A Survey of Multimodal Image Generation

孙熙江

2022.11.02

Introduction

Begin
2012

DALL.E
2021.02

Disco Diffusion
2022.02

MidJourney
2022.05

Noval AI
2022.10

Deep Dream
2015

GLIDE
2021.12

DALL.E2
2022.04

Stable diffusion
2022.08

CNN
2012

GAN
2014

VQ-VAE
2017

DDPM
2020

Begin

Deep Dream

GAN

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

DALL.E

作者：末知

关键词："Cyberpunk city in the night seen from below,cityscape,mist,rain,artstation,Greg Rutkowski

Disco Diffusion

a shiba inu wearing a beret and black turtleneck

DALL.E2

MidJourney

Noval AI

AI Generation

AI Painting

Contents

Preliminaries of Deep Generative Model

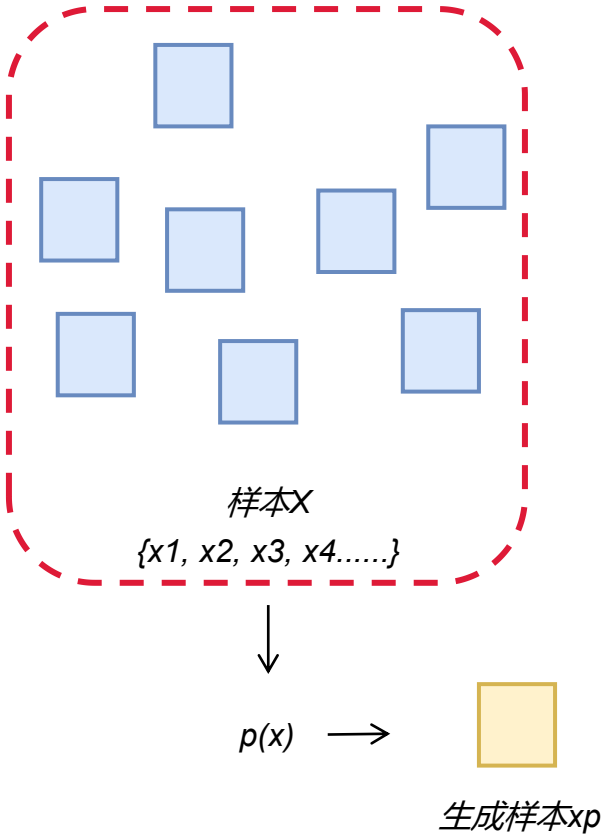
00 VAE

01 VQVAE

02 VQGAN

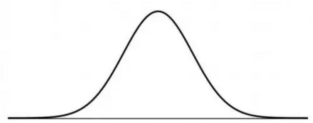
03 DDPM

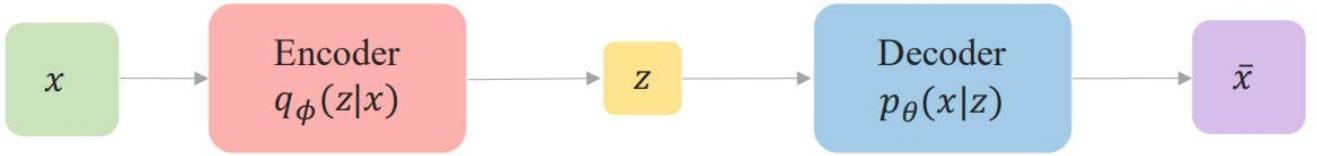
04 DDIM



VAE & VQVAE & VQGAN

p(x) = Σ p(x|z) p(z) → p(z) →

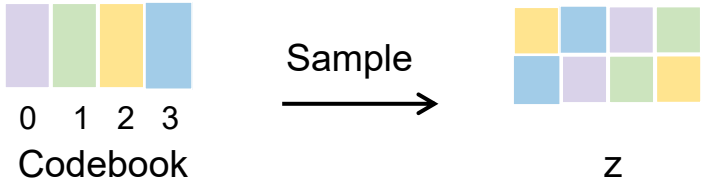


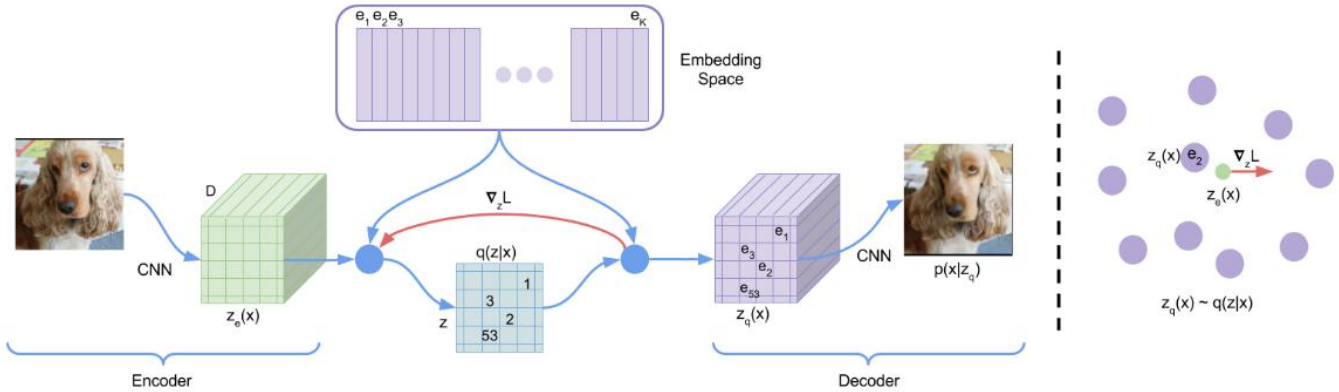


Auto-Encoding Variational Bayes

将p(z)定义为连续的均值为0方差为1的正态分布N(0, 1)

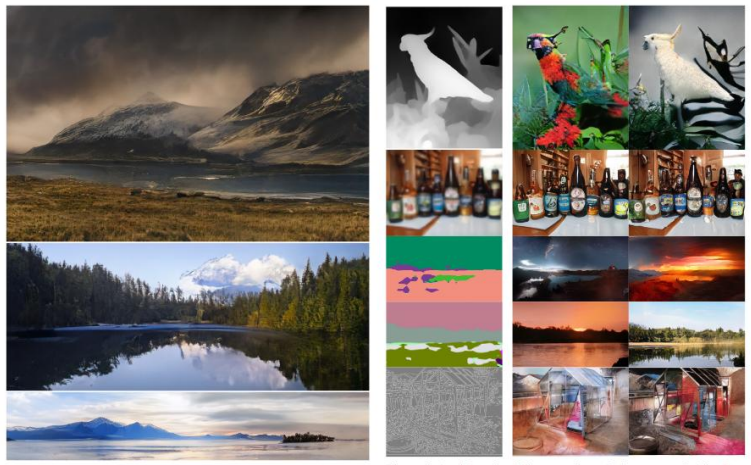
p(z) → z(cd1)z(cd2|cd1)z(cd3|cd1, cd2) .....





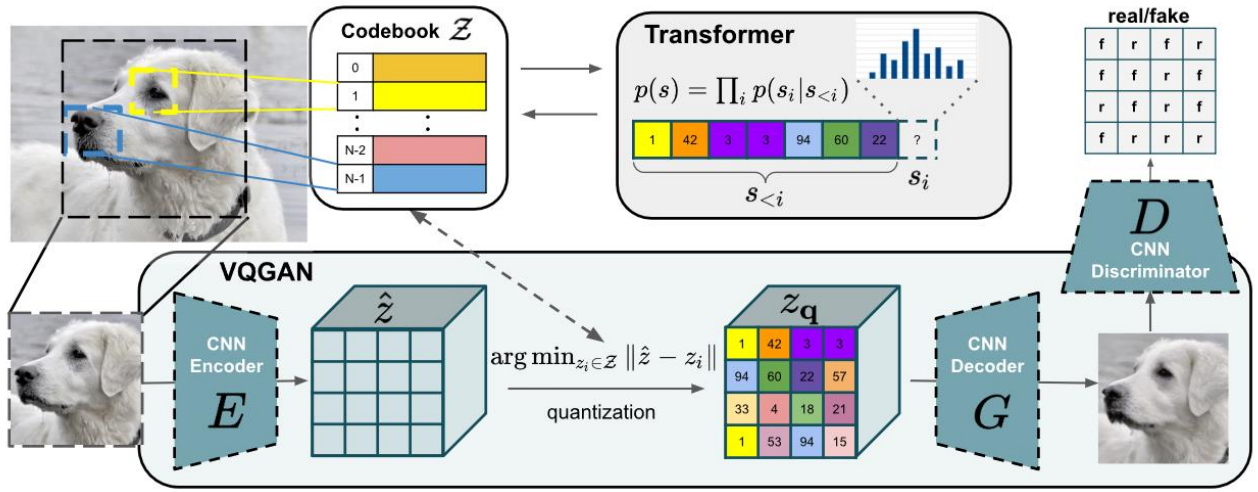
Neural Discrete Representation Learning

将p(z)定义为表示codebook中离散向量c的分布z(cb)



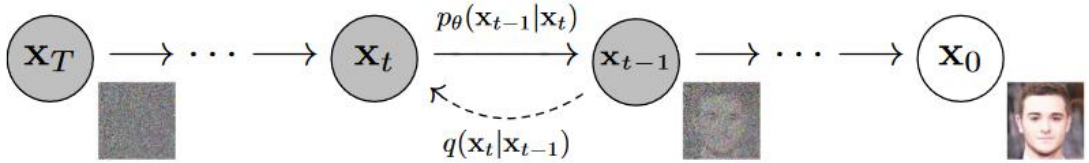
Discriminator

LPIPS



Taming Transformers for High-Resolution Image Synthesis

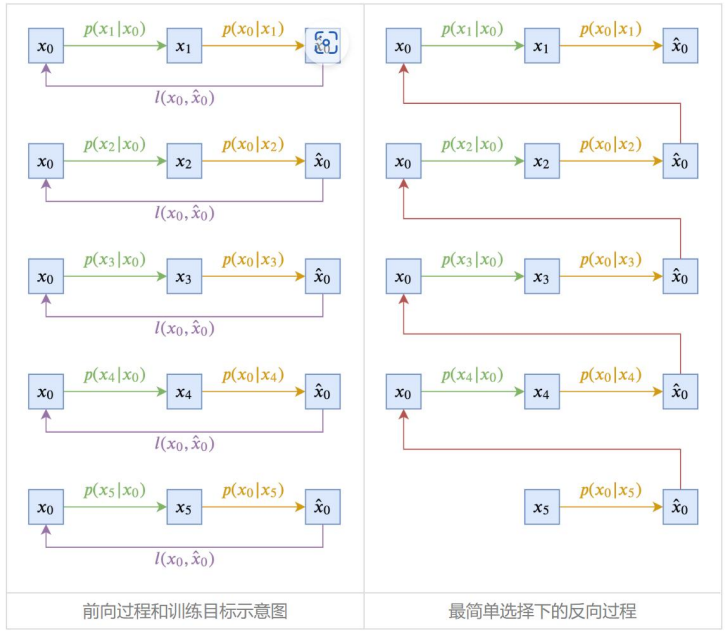
DDPM



$$Add\ Noise:\quad x_0 \;->\; x_1 \;->\; .. \;->\; x_{t-1} \;->\; x_t$$
$$x_t = a_t x_{t-1} + b_t \varepsilon_t\,, \quad \varepsilon_t \sim N(0, I) \quad {a_t}^2 + {b_t}^2 = 1$$

$$Del\ Noise:\quad \bar{x}_0 \;<-\; \bar{x}_1 \;<-\; .. \;<-\; \bar{x}_{t-1} \;<-\; x_t$$
$$x_{t-1} = \mu(x_t)$$

Denoising Diffusion Probabilistic Models



- Predict noise

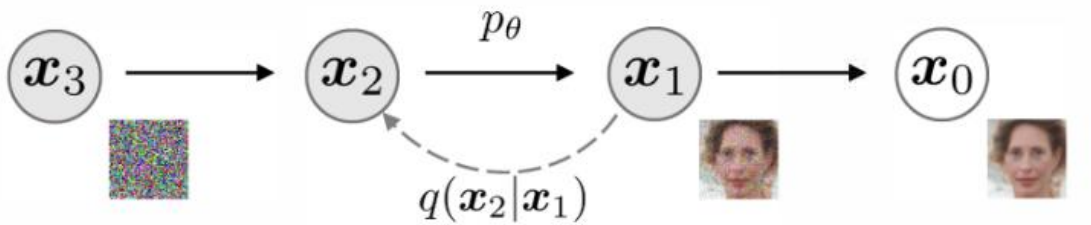
$$min||\mu(\bar{x}_t) - x_{t-1}||^2$$
$$\downarrow$$
$$min||\in_\theta(x_t, t) - \varepsilon_t||^2$$

- Using x_t x_0 to predict x_{t-1}

$$P(x_{t-1}|x_t, x_0) = N(x_{t-1}; \kappa_t x_t + \lambda_t x_0, \sigma^2 I)$$

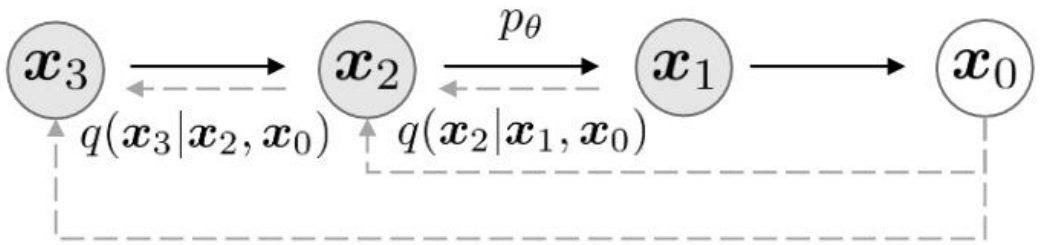
DDIM

$$Suppose:\ P(x_{t-1}|x_t, x_0) = N(x_{t-1}; \kappa_t x_t + \lambda_t x_0, \sigma^2 I)$$
$$= N(x_{t-1}; \frac{\sqrt{{\bar{b}_t}^2 - \sigma^2}}{\bar{b}_t} x_t + (\bar{a}_{t-1} - \frac{\bar{a}_t \sqrt{{\bar{b}_{t-1}}^2 - \sigma^2}}{\bar{b}_t}) x_0, \sigma^2 I)$$



Graphical models for diffusion

$$Sample:\ x_{\tau_{t-1}} = \frac{1}{a_{\tau_t}}(x_{\tau_t} - \frac{{b_{\tau_t}}^2}{{\bar{b}_{\tau_t}}^2} \in_\theta(x_t, \tau_t)) + \frac{\bar{b}_{\tau_{t-1}} {b_{\tau_t}}^2}{\bar{b}_{\tau_t}} \varepsilon$$
$$[\tau_1, \tau_2, ..., \tau_{t-1}\tau_t] \in [1,2, ..., t-1, t]$$



non-Markovain inference models

Denoising Diffusion Implicit Models

Deep Generative Model with Condition

00 Hidden Variables Guidance

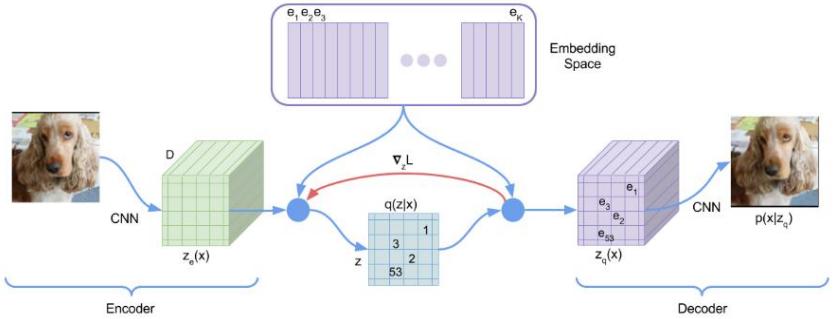
01 Classifier Diffusion Guidance

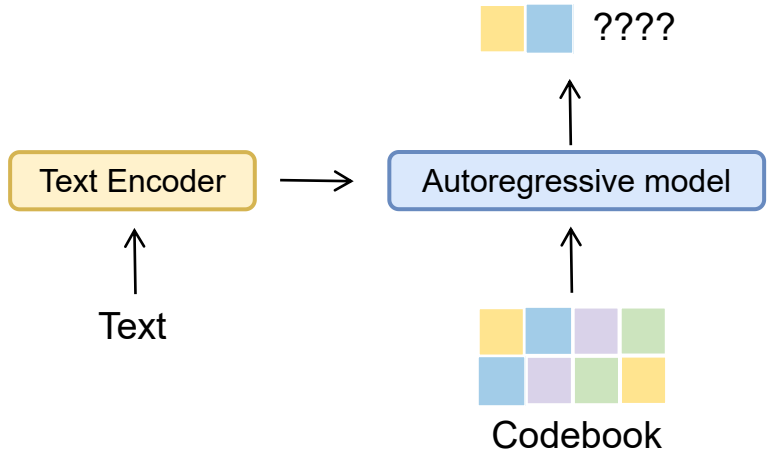
02 Classifier-Free Diffusion Guidance

Hidden Variables Guidance

VQVAE/VQGAN

Transformer-like



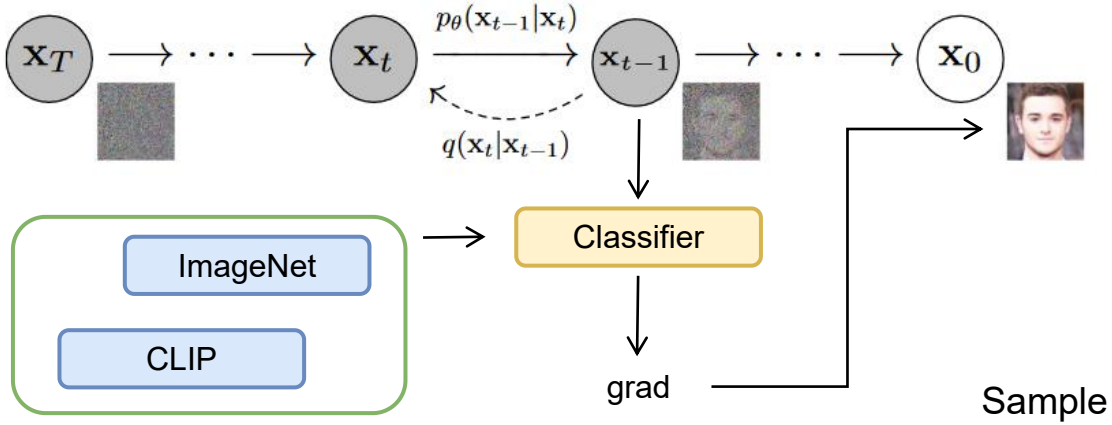


Classifier Diffusion Guidance

Diffusion Models

Sample: $\overline{\epsilon}_\theta(x_t, t) = \epsilon_\theta(x_t, t) - \overline{b}_t \omega \nabla_{x_t} log P(y|x_t)$

Diffusion Models Beat GANs on Image Synthesis

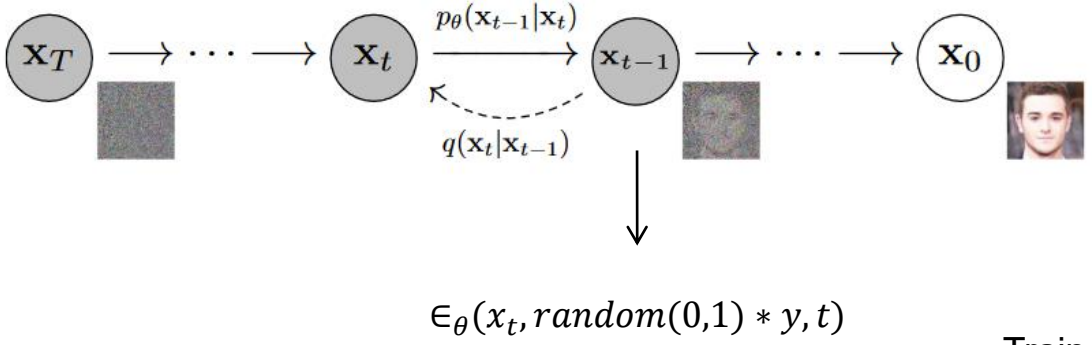


Classifier-Free Diffusion Guidance

Diffusion Models

Train: $\overline{\epsilon}_\theta(x_t, y, t) = \epsilon_\theta(x_t, y, t) + \omega(\overline{\epsilon}_\theta(x_t, y, t) - \overline{\epsilon}_\theta(x_t, t))$

Classifier-Free Diffusion Guidance



Multimodal Deep Generative Model

00 DALL.E & DALL.E mini

01 Parti

02 GLIDE & DALL.E 2

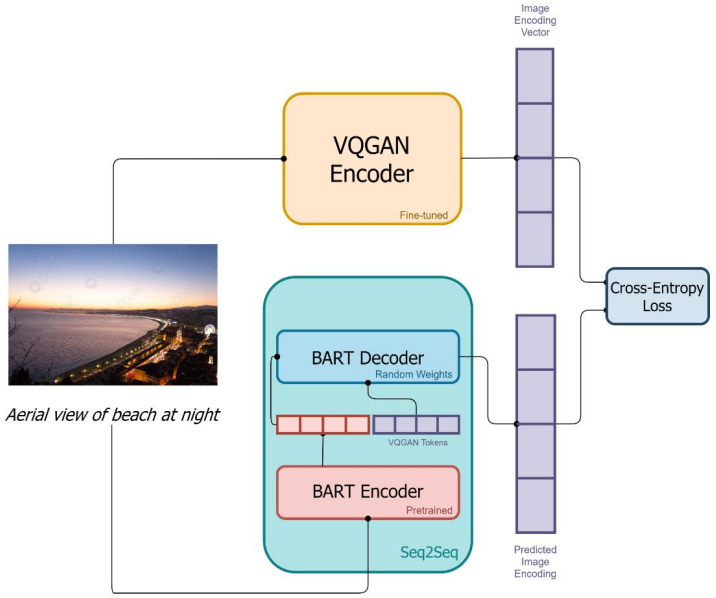
03 Imagen

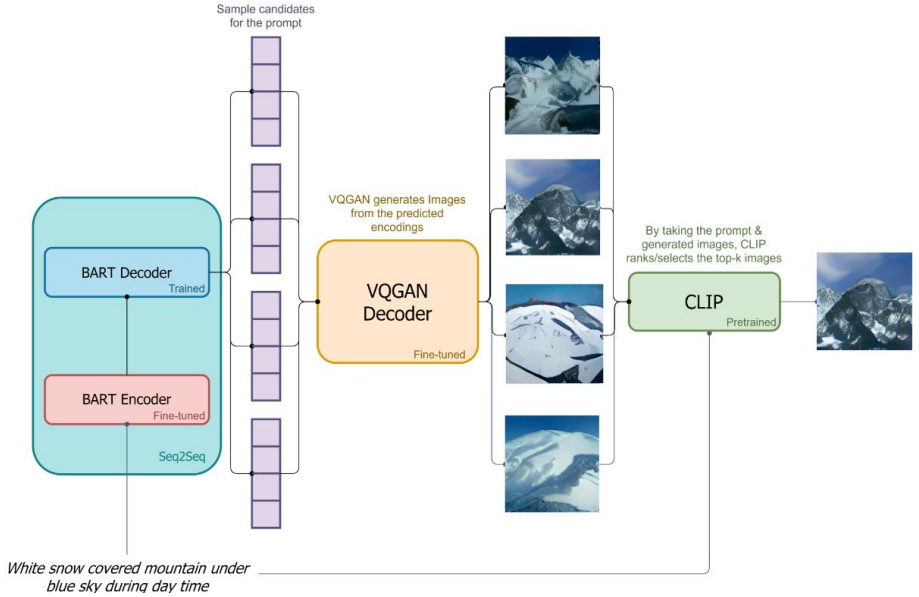
04 文心

DALL.E & DALL.E mini

(a) a tapir made of accordion. a tapir with the texture of an accordion.

(b) an illustration of a baby hedgehog in a christmas sweater walking a dog





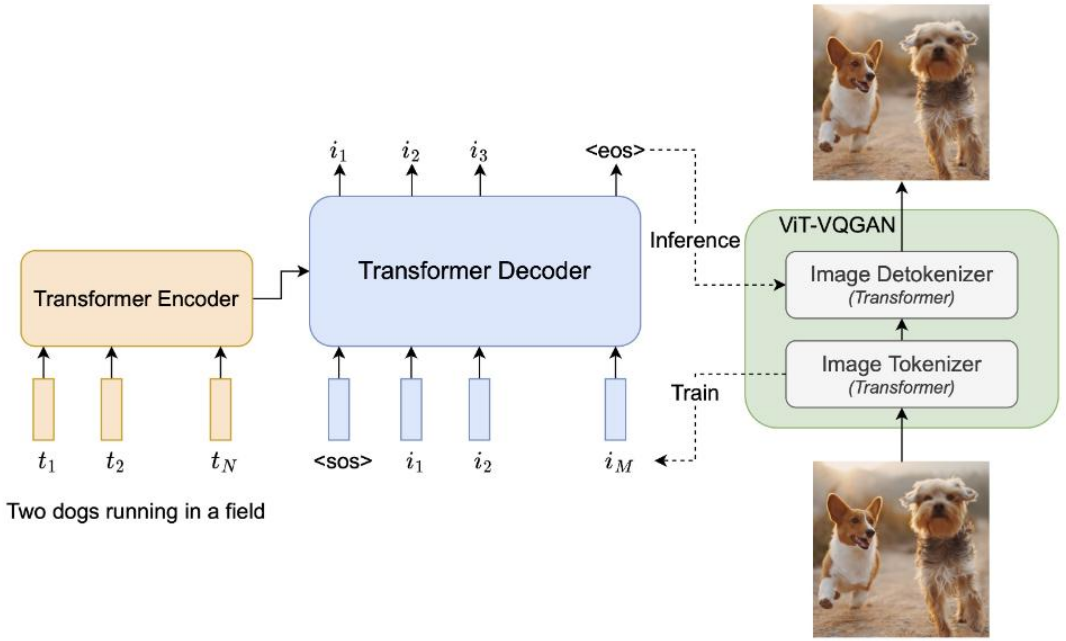
Zero-Shot Text-to-Image Generation

Parti

A. *A photo of a frog reading the newspaper named "Toaday" written on it. There is a frog printed on the newspaper too.*

B. *A portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. The city of Los Angeles is in the background. Hi-res DSLR photograph.*

C. *A high-contrast photo of a panda riding a horse. The panda is wearing a wizard hat and is reading a book. The horse is standing on a street against a gray concrete wall. Colorful flowers and the word "PEACE" are painted on the wall. Green grass grows from cracks in the street. DSLR photograph. daytime lighting.*

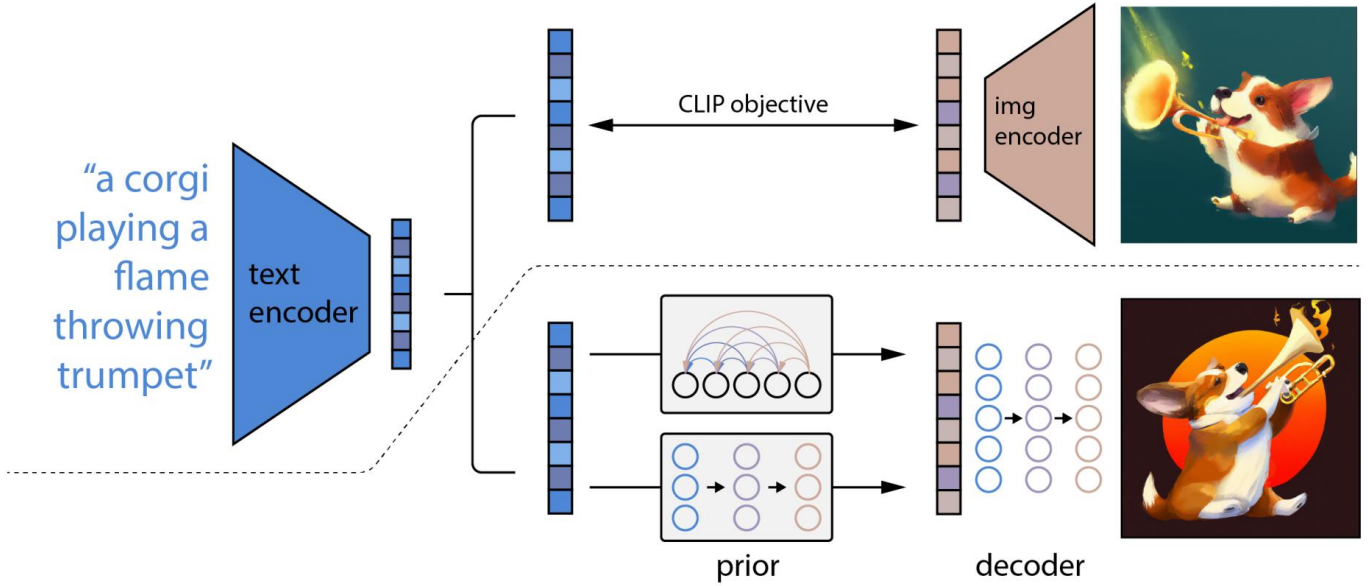


Scaling Autoregressive Models for Content-Rich Text-to-Image Generation

GLIDE & DALL.E 2

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square



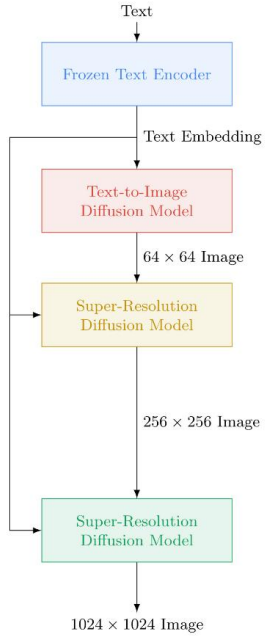
Hierarchical Text-Conditional Image Generation with CLIP Latents

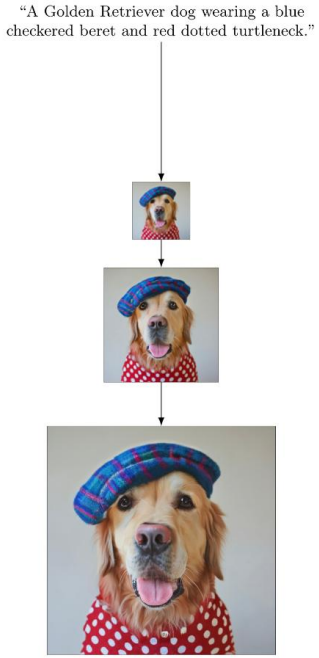
Imagen

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

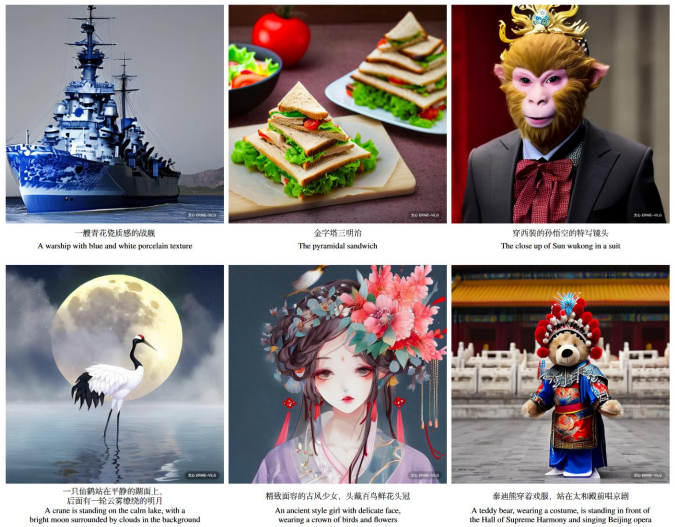
A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

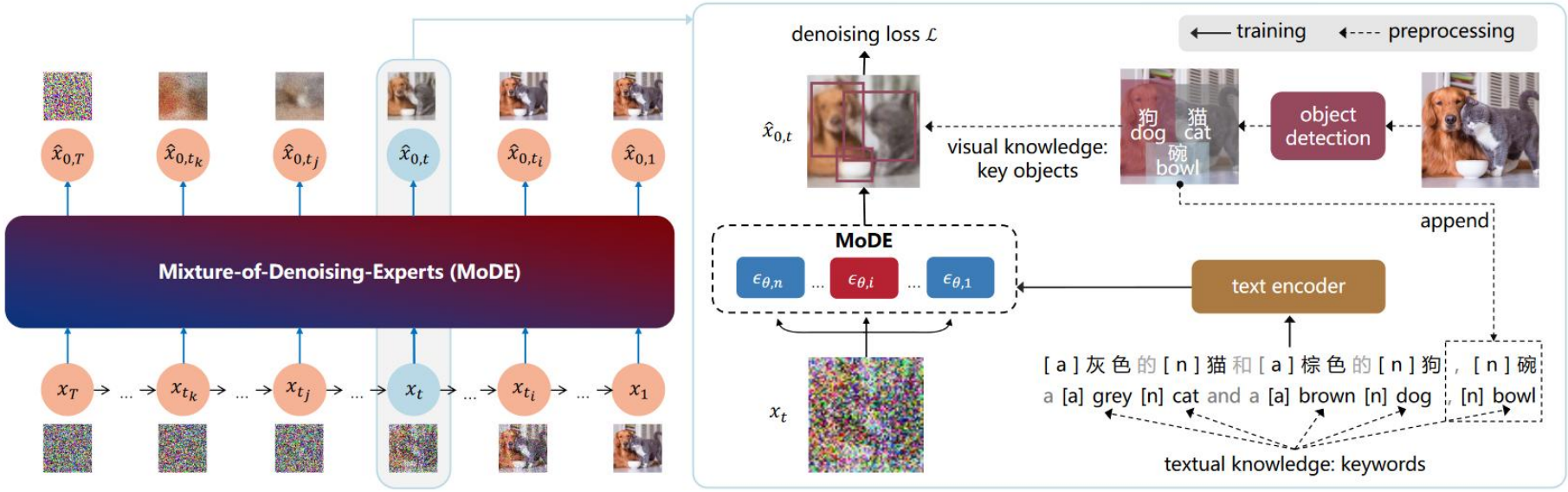




Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

文心





ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts

Knowledge-Enhanced Diffusion Model：

$$\text{Attention}(Q,K,V)' = \text{softmax}\left(\frac{W_a \cdot (QK^\top)}{\sqrt{d}}\right)V,$$

$$\mathcal{L}' = \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,I),t}\left[W_l \cdot ||\epsilon - \epsilon_\theta(z_t,t)||_2^2\right]$$

Add text prompt from image

Image captioning model

Mixture-of-Denoising-Experts：

$$\epsilon_\theta(x_t,t) = \{\epsilon_{\theta,i}(x_t,t)\}, \quad t \in S_i,$$

Knowledge-Enhanced Diffusion Model

Mixture-of-Denoising-Experts

Model	Zero-Shot FID-30k ↓
DALL-E (Ramesh et al., 2021)	27.50
CogView (Ding et al., 2021)	27.10
LAFITE (Zhou et al., 2021)	26.94
LDM (Rombach et al., 2021)	12.61
ERNIE-ViLG (Zhang et al., 2021b)	14.70
GLIDE (Nichol et al., 2022)	12.24
Make-A-Scene (Gafni et al., 2022)	11.84
DALL-E 2 (Ramesh et al., 2022)	10.39
CogView2 (Ding et al., 2022)	24.00
Imagen (Saharia et al., 2022)	7.27
Parti (Yu et al., 2022)	7.23
ERNIE-ViLG 2.0	**6.75**

Quantitative analysis method

Latent Text-to-image Deep Generative Model

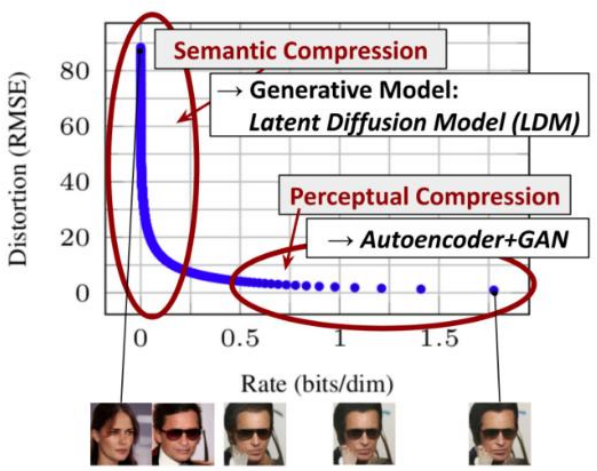
00 High-Resolution Image Synthesis with Latent Diffusion Models

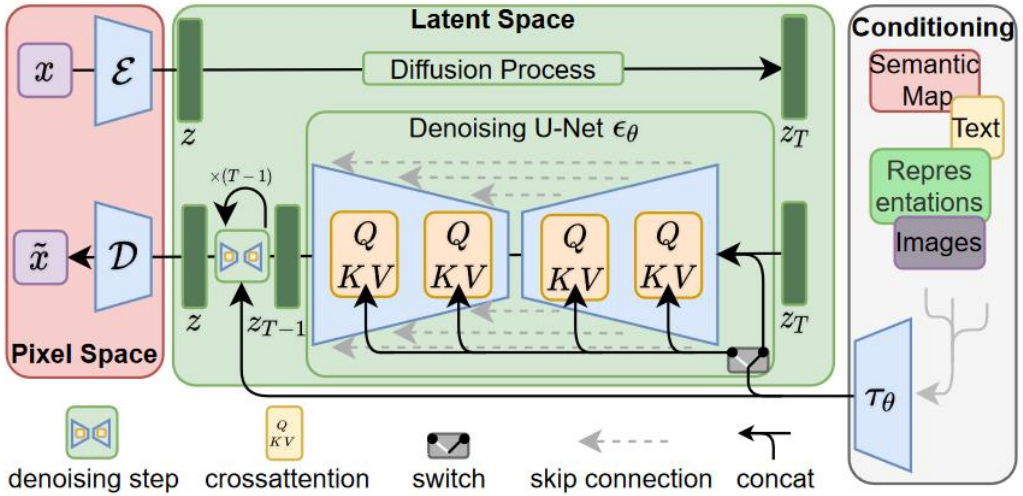
Stable-Diffusion: A latent text-to-image diffusion model

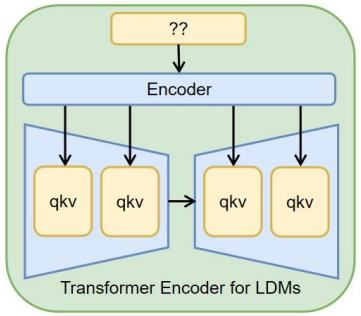
01 Frido: Feature Pyramid Diffusion for Complex Scene Image Synthesis

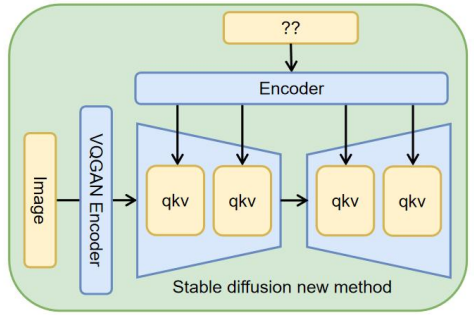
02 Pretraining is All You Need for Image-to-Image Translation

High-Resolution Image Synthesis with Latent Diffusion Models







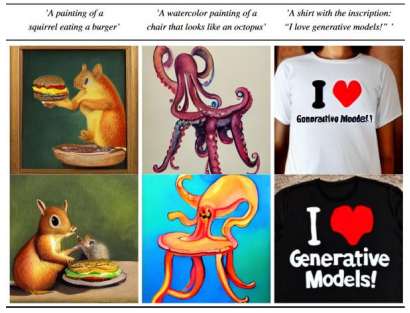


Q
KV

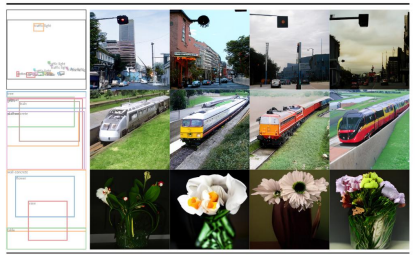
$$K = W_K^{(i)} \cdot \tau_\theta(y), \qquad V = W_V^{(i)} \cdot \tau_\theta(y).$$
$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),$$

cross-attention

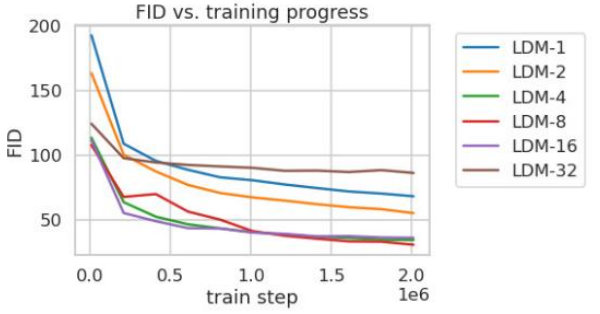
High-Resolution Image Synthesis with Latent Diffusion Models



Test to image



Layout to image

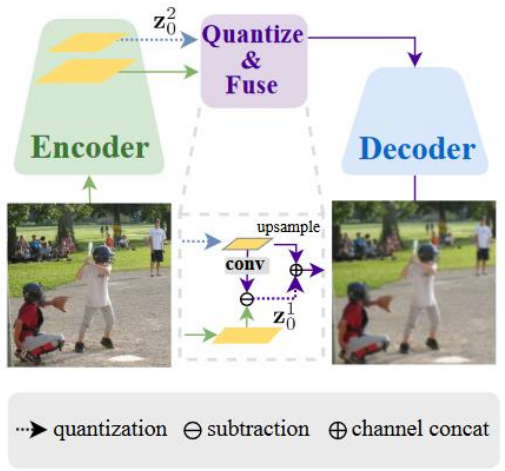


Super-Resolution

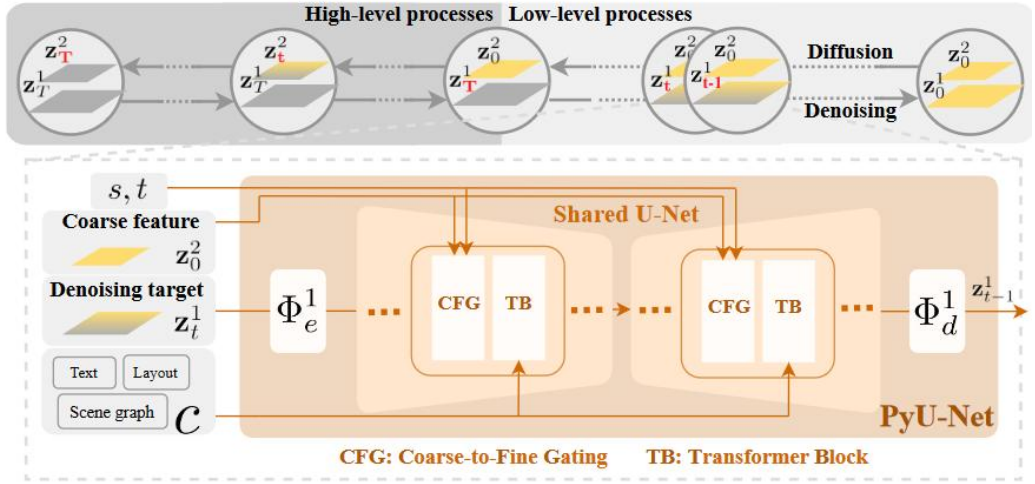
Inpainting

Method	Generator Compute	Classifier Compute	Overall Compute	Inference Throughput*	N_{params}	FID↓	IS↑	Precision↑	Recall↑
LSUN Churches 256^2									
StyleGAN2 [42]†	64	-	64	-	59M	3.86	-	-	-
LDM-8 (ours, 100 steps, 410K)	18	-	18	6.80	256M	4.02	-	0.64	0.52
LSUN Bedrooms 256^2									
ADM [15]† (1000 steps)	232	-	232	0.03	552M	1.9	-	0.66	0.51
LDM-4 (ours, 200 steps, 1.9M)	60	-	55	1.07	274M	2.95	-	0.66	0.48
CelebA-HQ 256^2									
LDM-4 (ours, 500 steps, 410K)	14.4	-	14.4	0.43	274M	5.11	-	0.72	0.49
FFHQ 256^2									
StyleGAN2 [42]	32.13‡	-	32.13†	-	59M	3.8	-	-	-
LDM-4 (ours, 200 steps, 635K)	26	-	26	1.07	274M	4.98	-	0.73	0.50
ImageNet 256^2									
VQGAN-f-4 (ours, first stage)	29	-	29	-	55M	0.58††	-	-	-
VQGAN-f-8 (ours, first stage)	66	-	66	-	68M	1.14††	-	-	-
BigGAN-deep [3]†	128-256		128-256	-	340M	6.95	$203.6_{\pm 2.6}$	0.87	0.28
ADM [15] (250 steps) †	916	-	916	0.12	554M	10.94	100.98	0.69	0.63
ADM-G [15] (25 steps) †	916	46	962	0.7	608M	5.58	-	0.81	0.49
ADM-G [15] (250 steps)†	916	46	962	0.07	608M	4.59	186.7	0.82	0.52
ADM-G,ADM-U [15] (250 steps)†	329	30	349	n/a	n/a	3.85	221.72	0.84	0.53
LDM-8-G (ours, 100, 2.9M)	79	12	91	1.93	506M	8.11	$190.4_{\pm 2.6}$	0.83	0.36
LDM-8 (ours, 200 ddim steps 2.9M, batch size 64)	79	-	79	1.9	395M	17.41	72.92	0.65	0.62
LDM-4 (ours, 250 ddim steps 178K, batch size 1200)	271	-	271	0.7	400M	10.56	$103.49_{\pm 1.24}$	0.71	0.62
LDM-4-G (ours, 250 ddim steps 178K, batch size 1200, classifier-free guidance [32] scale 1.25)	271	-	271	0.4	400M	3.95	$178.22_{\pm 2.43}$	0.81	0.55
LDM-4-G (ours, 250 ddim steps 178K, batch size 1200, classifier-free guidance [32] scale 1.5)	271	-	271	0.4	400M	3.60	$247.67_{\pm 5.59}$	0.87	0.48

Frido: Feature Pyramid Diffusion for Complex Scene Image Synthesis



(a) Architecture of MS-VQGAN.



(b) Details of the diffusion and denoising processes.

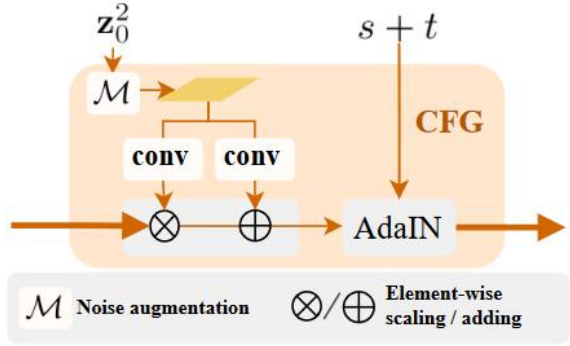
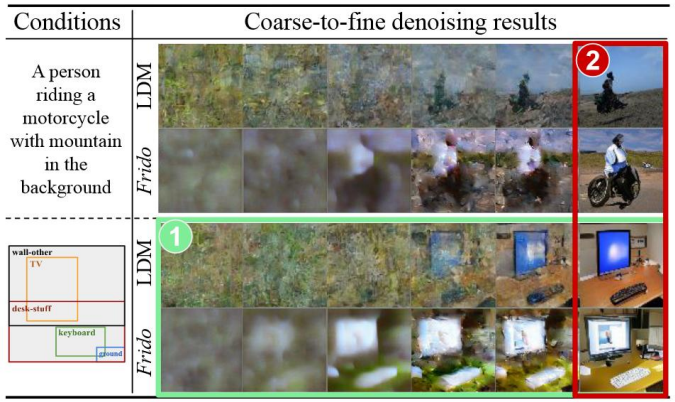


Figure 4: Framework of coarse-to-fine gating in PyU-Net.

Frido: Feature Pyramid Diffusion for Complex Scene Image Synthesis



Problem:

Most of existing DMs might not be able to capture image semantics or compositions in real-world complex scenes

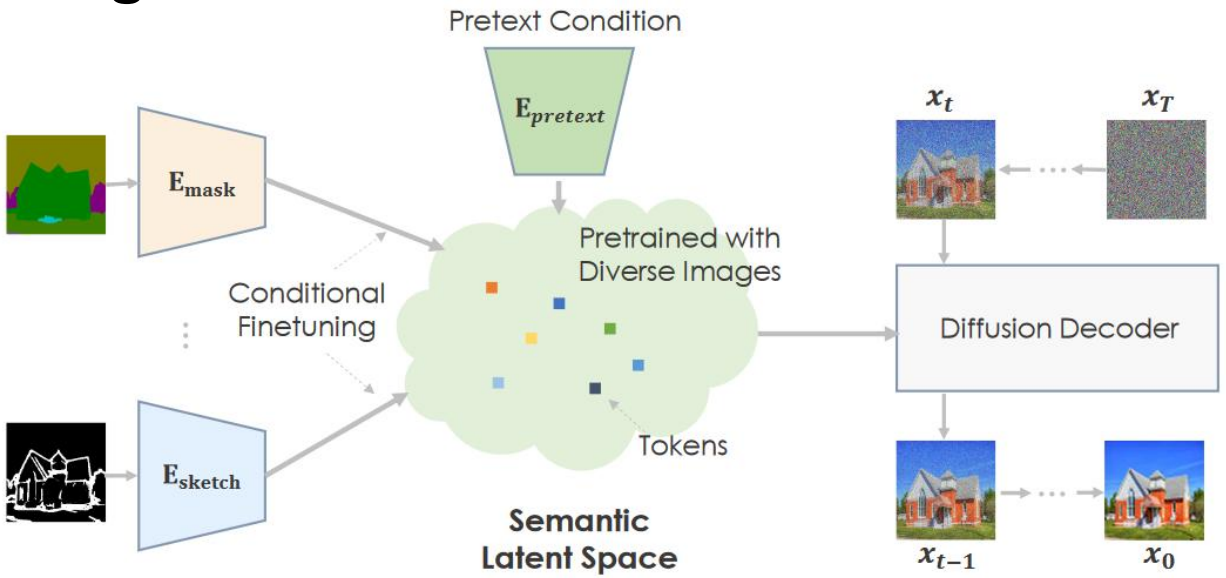
Methods	FID↓	IS↑	CLIP↑
Methods under standard T2I setting			
AttnGAN (Xu et al. 2018)	33.10	23.61	-
Obj-GAN (Li et al. 2019)	36.52	24.09	-
DM-GAN (Zhu et al. 2019)	27.34	**32.32**	-
DF-GAN (Tao et al. 2022)	21.42	-	-
LDM-8† (Rombach et al. 2022)	17.61	19.34	0.6500
VQ-diffusion‡ (Gu et al. 2022)	14.06	21.85	0.6770
LDM-8-G†	12.27	27.86	0.6927
Frido-f16f8	15.38	19.32	0.6607
Frido-f16f8-G	**11.24**	26.82	**0.7046**
Methods with external pre-trained CLIP			
LAFITE-CLIP‡ (Zhou et al. 2022)	**8.12**	**32.24**	0.7915
Frido-f16f8-G-CLIPr	8.97	27.43	**0.7991**

Pretraining is All You Need for Image-to-Image Translation

Old Method -> train the decoder with encoder fixed

First Step -> train the encoder with decoder fixed
Second Step -> finetune encoder and decoder jointly

GT | Condition | SPADE | OASIS | Ours

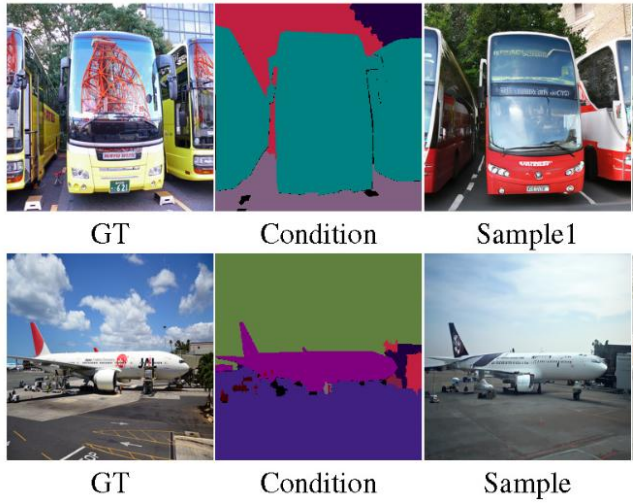


Pretraining is All You Need for Image-to-Image Translation

Input | Fixed decoder | Full model | Input | Fixed decoder | Full model

(a) Finetune strategy.

Finetune strategy	FID
Fixed decoder	12.6
One-stage finetune	13.3
Two-stage finetune	8.9



GT | Condition | Sample1

GT | Condition | Sample

Fixed encoder -> DMs don't know the semanteme which get from image encoder

Fixed decoder -> DMs don't konw the semanteme which represents spatial properties

One-stage finetune -> Change the DMs' semanteme and image encoder's semanteme

Two-stage finetune -> First, image encoder learn the semanteme of DMs
Second, change the DMs' semanteme and image encoder's semanteme of spatial properties

Other problems

Downstream task of Deep Generative Model

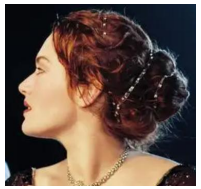
00 Subject-Driven Generation

01 Image Edition

Subject-Driven Generation

[identifier]

"...draw me like one of your French girls."



what Rose looks like

a subset of Jack's prior

1、describe

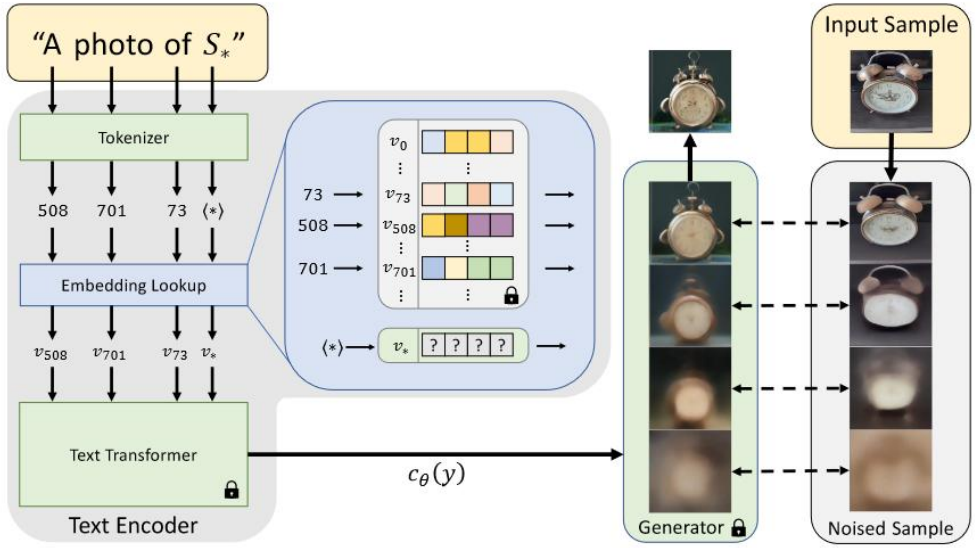
Blonde hair, Blue eyes.....

2、fine-tuning

Deep Generation Model

Embedding

for computer



An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

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

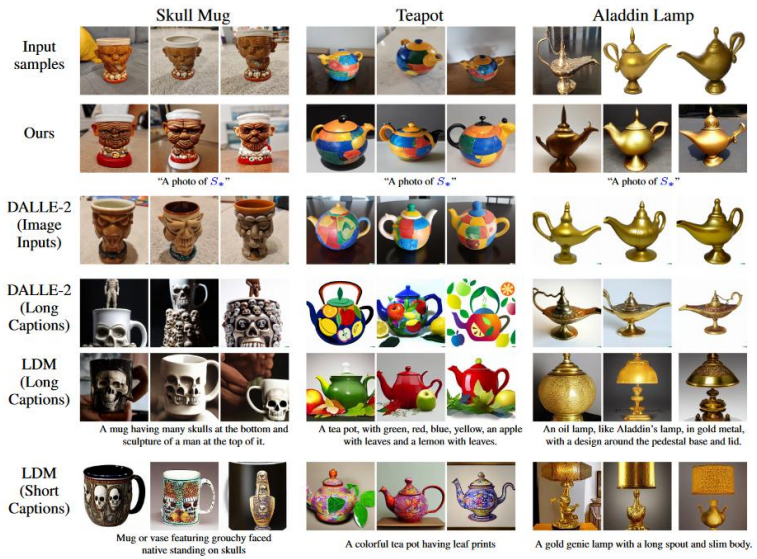
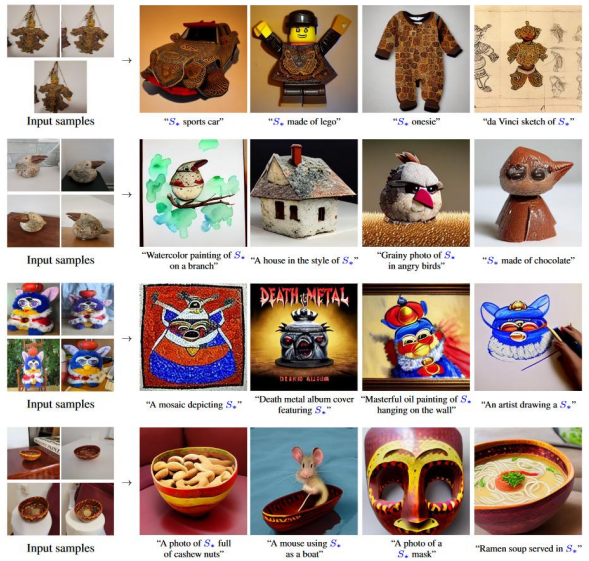


Image Variations



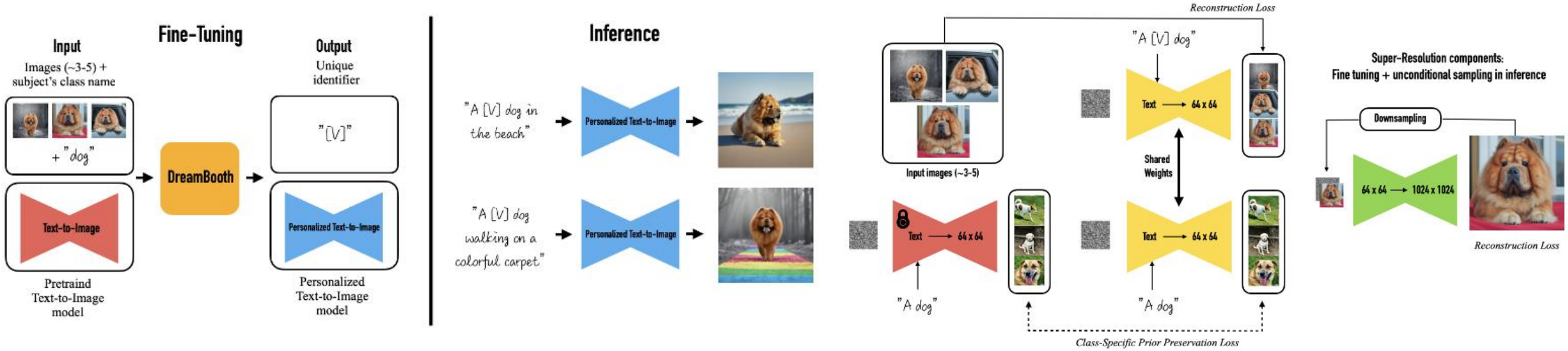
Text-guided synthesis

Style transfer

Concept compositons

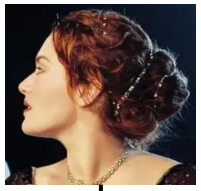
Subject-Driven Generation

[identifier][class noun]



DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

1、describe



Blonde hair

Blue eyes

necklace

T2I domain

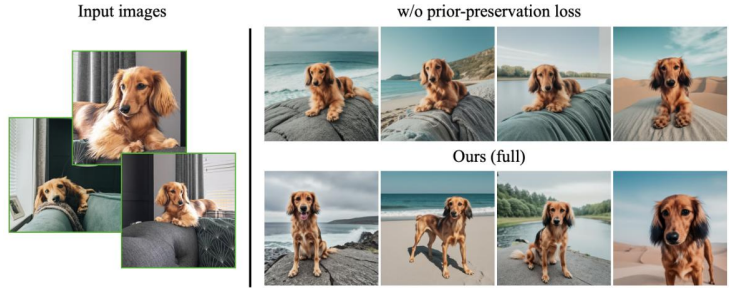
2、fine-tuning

girl celled Rose

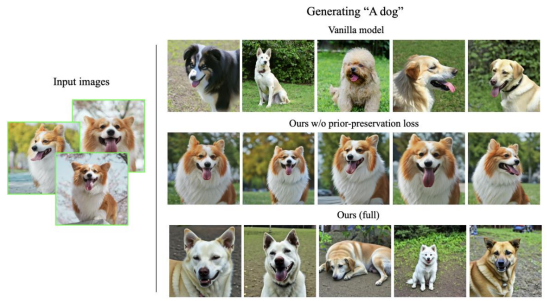
girl with Blonde hair

girl with Blue eyes

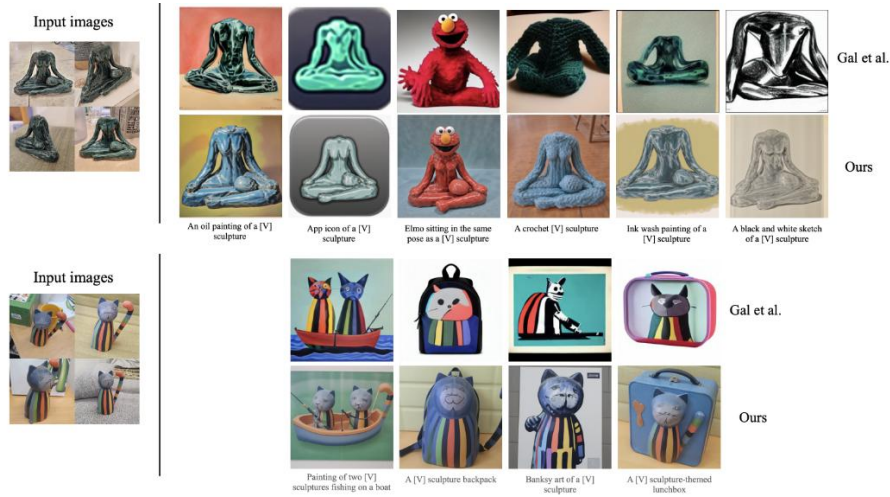
T2I domain



Overfitting



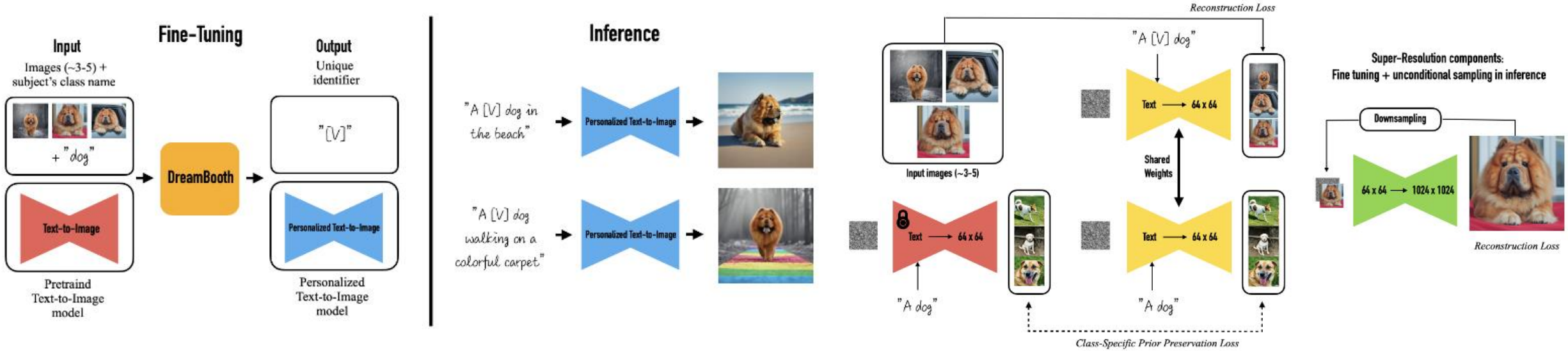
Language-drift



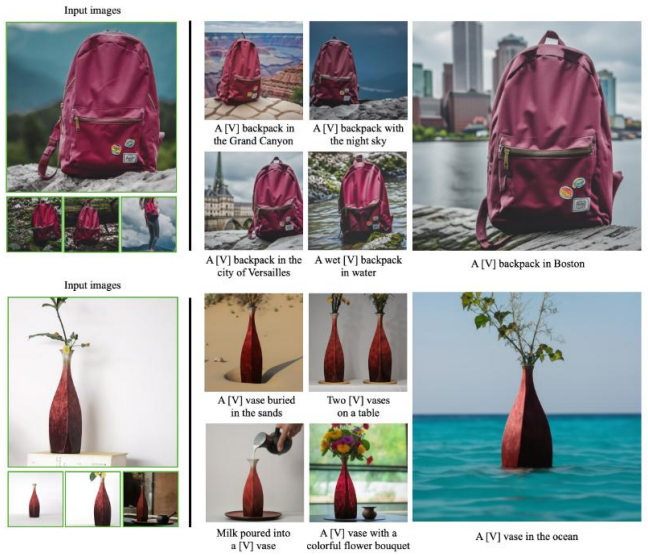
Comparison with Text Inversion

Subject-Driven Generation

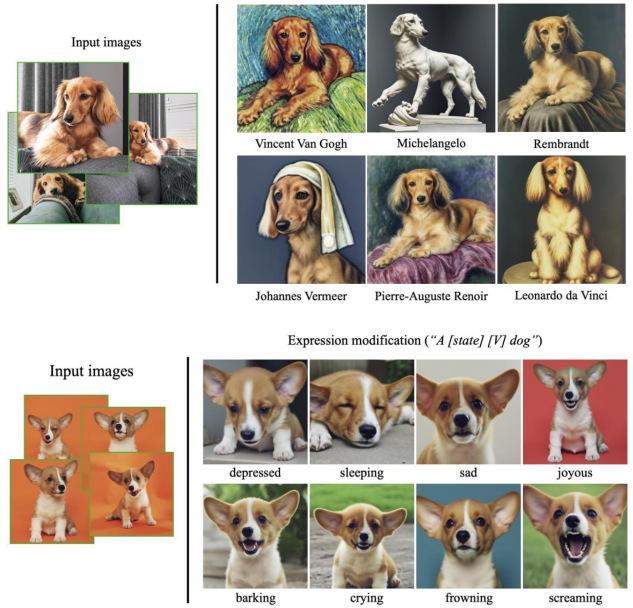
[identifier][class noun]

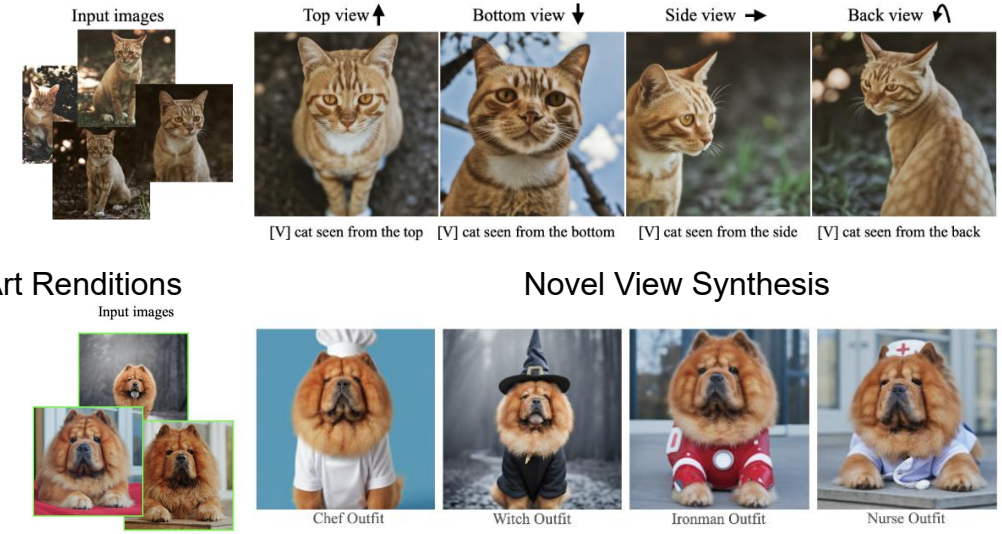


DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

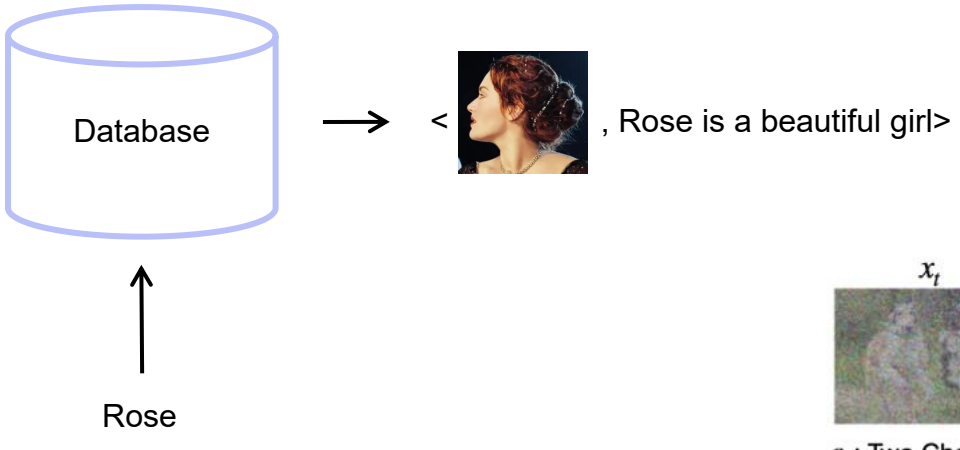


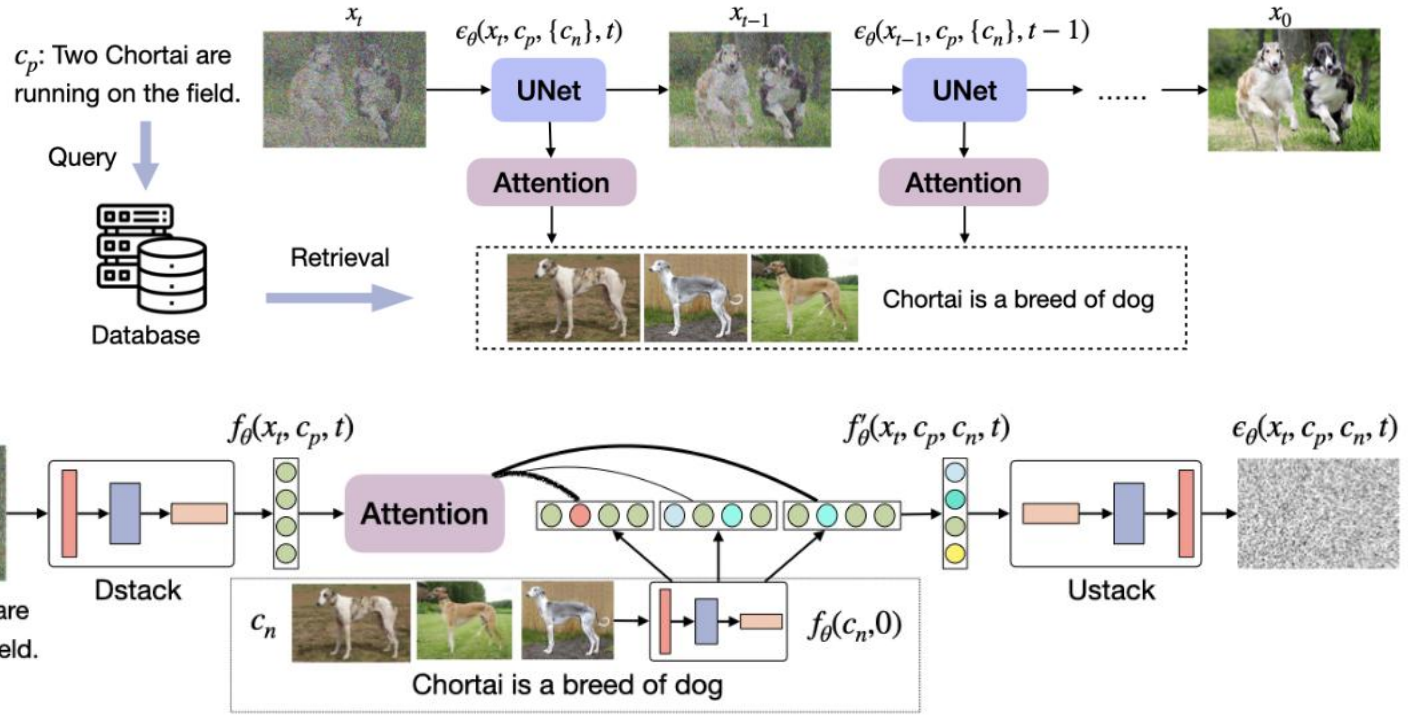
Recontextualization





Retrieval-Augmented Generation





Re-imagen: retrieval-augmented test-to-image generator

$$\hat{\epsilon}_p = w_p \cdot \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, \boldsymbol{c}_n, t) - (w_p - 1) \cdot \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}_n, t)$$
$$\hat{\epsilon}_n = w_n \cdot \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, \boldsymbol{c}_n, t) - (w_n - 1) \cdot \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}_p, t)$$

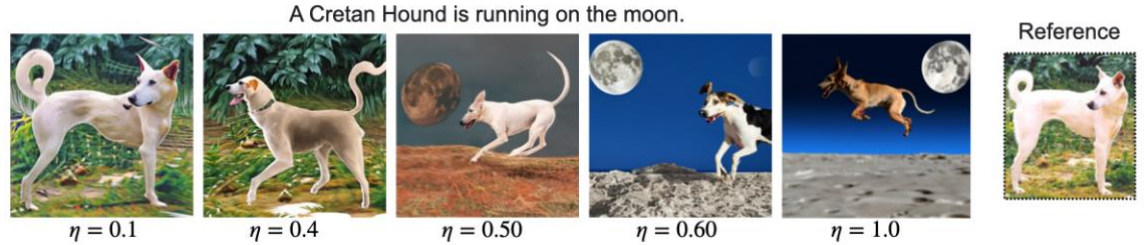


Image Edition

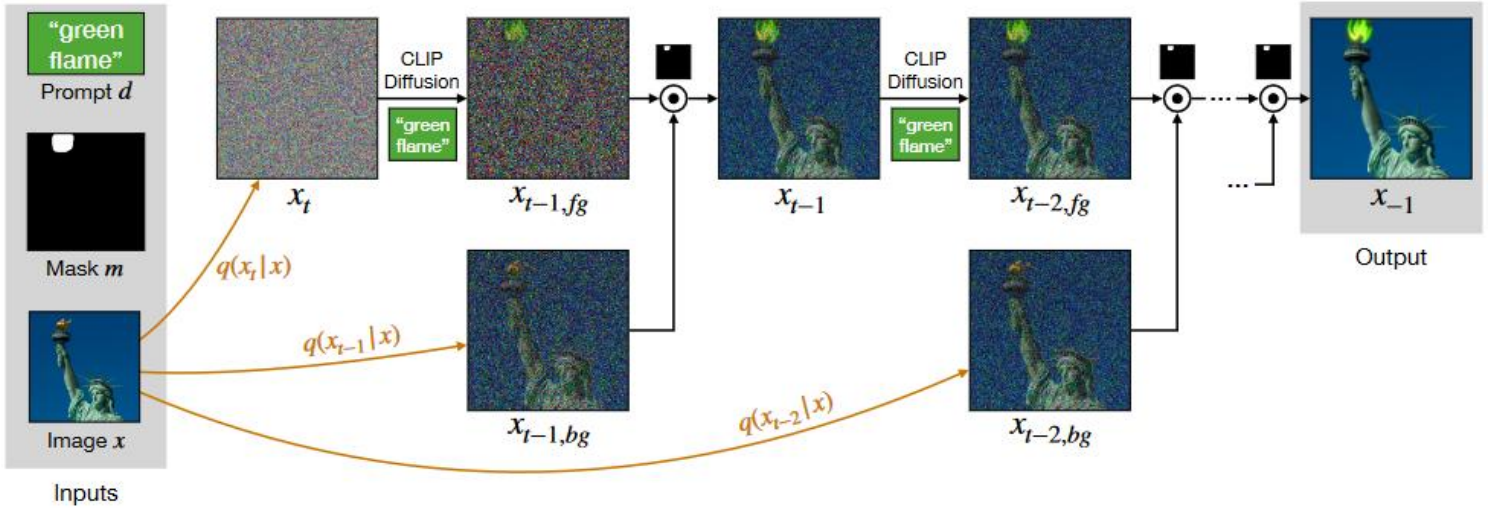
foreground preservation loss:

$$\mathcal{D}_{CLIP}(x, d, m) = D_c(CLIP_{img}(x \odot m), CLIP_{txt}(d))$$

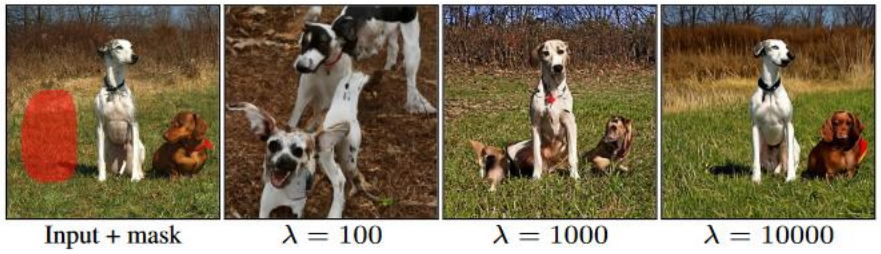
background preservation loss:

$$\mathcal{D}_{bg}(x_1, x_2, m) = d(x_1 \odot (1 - m), x_2 \odot (1 - m))$$

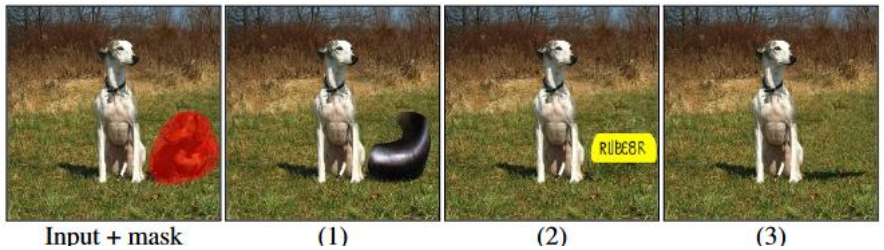
$$d(x_1, x_2) = \frac{1}{2}(MSE(x_1, x_2) + LPIPS(x_1, x_2))$$



Blended Diffusion for Text-driven Editing of Natural Images



problem: Local CLIP-guided diffusion



problem: Text-driven blended diffusion

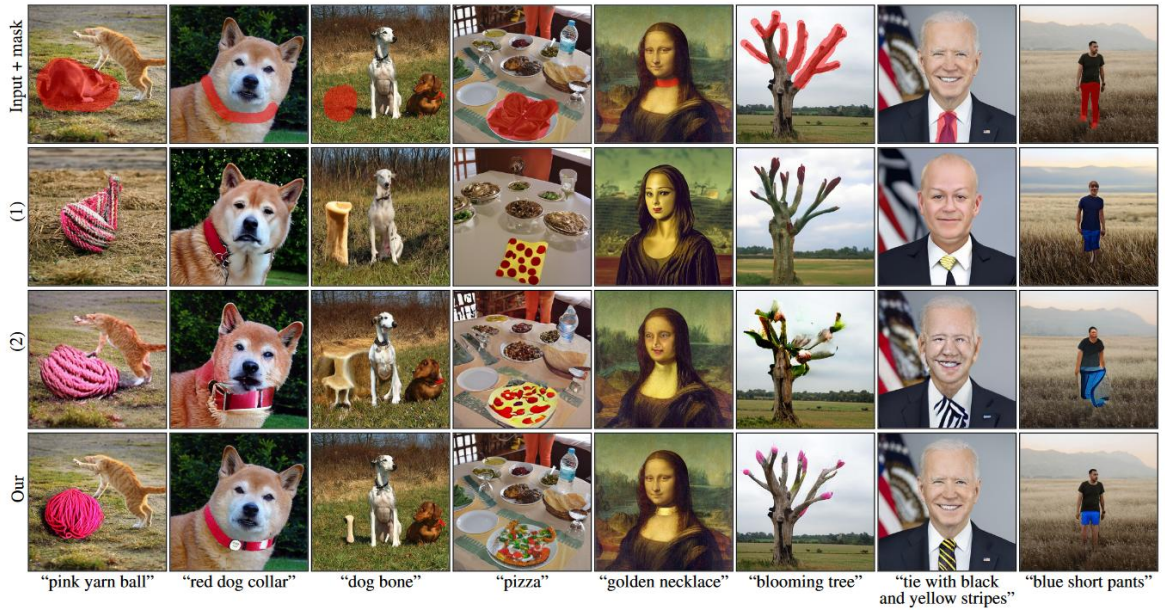
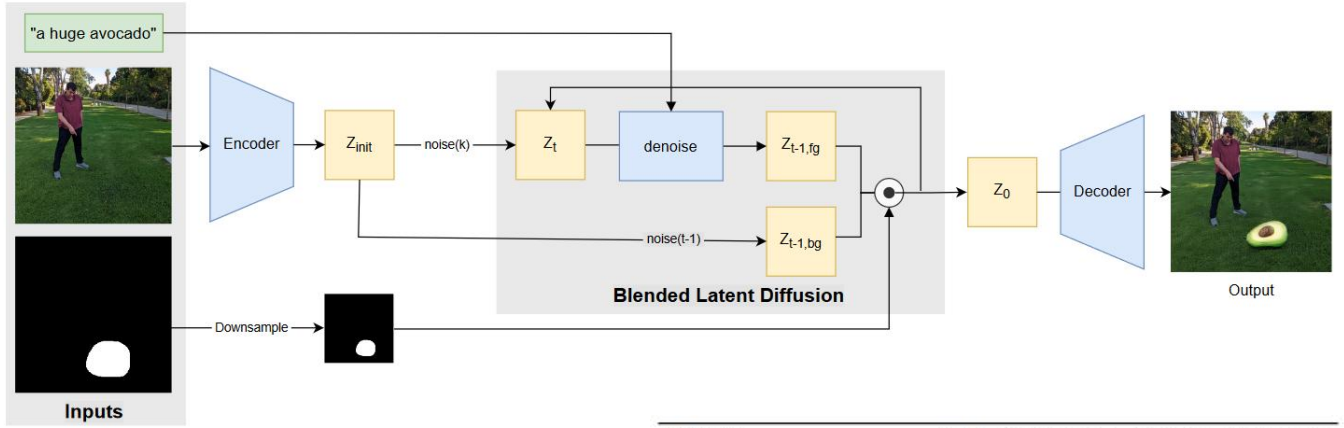


Image Edition

Blended Diffusion VS Blended Latent Diffusion

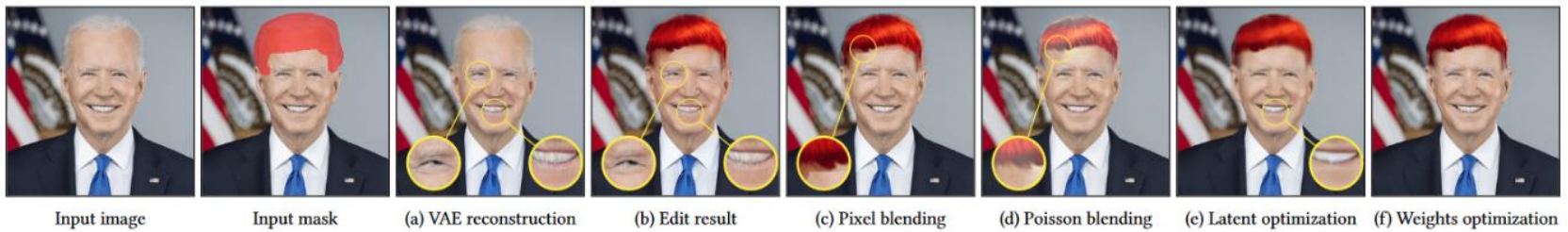
Other problems

Blended Latent Diffusion



Method	Batch Precision ↑	Batch Diversity ↑	Best Result Precision ↑
Blended Diffusion	10.4%	0.106	36%
Local CLIP-guided diffusion	10.49%	**0.419**	38%
PaintByWord++	-	-	0%
GLIDE-filtered	1.87%	0.114	4%
Ours	**28.66%**	0.115	**54%**

Method	Batch Size	Single Image (sec) ↓	Full Batch (sec) ↓	Per Image in Batch (sec) ↓
Blended Diffusion	64	27	1472	23
Blended Diffusion	24*	27	552	23
Local CLIP-guided diffusion	64	27	1472	23
Local CLIP-guided diffusion	24*	27	552	23
PaintByWord++	-	78	-	-
GLIDE-filtered	24	7	89	3.7
Ours (without background opt.)	24	6	53	2.2
Ours (with background opt.)	24	25	72	3



(a) VAE重建效果； (b) LDM编辑效果； (c) LDM编辑后与图像重叠； (d) 泊松混合后； (e) 隐变量优化后； (f) 微调decoder后

Problem one: Imperfect reconstruction

Method A: latent optimization $z^* = \underset{z}{\operatorname{argmin}} \|D(z)\odot m, \hat{x}\odot m\| + \lambda \|D(z)\odot(1-m), x\odot(1-m)\|$

Method B: fine-tuning the decoder's weights $\theta^* = \underset{\theta}{\operatorname{argmin}} \|D_\theta(z_0)\odot m, \hat{x}\odot m\| + \lambda \|D_\theta(z_0)\odot(1-m), x\odot(1-m)\|,$

Problem two: Thin mask

Image Edition

CLIP loss:

Global target loss:

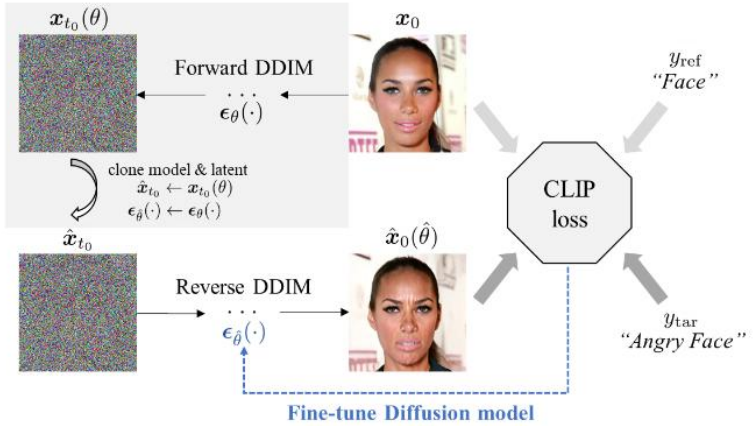
$$\mathcal{L}_{\text{global}}(\boldsymbol{x}_{\text{gen}}, y_{\text{tar}}) = D_{\text{CLIP}}(\boldsymbol{x}_{\text{gen}}, y_{\text{tar}}),$$

Local directional loss:

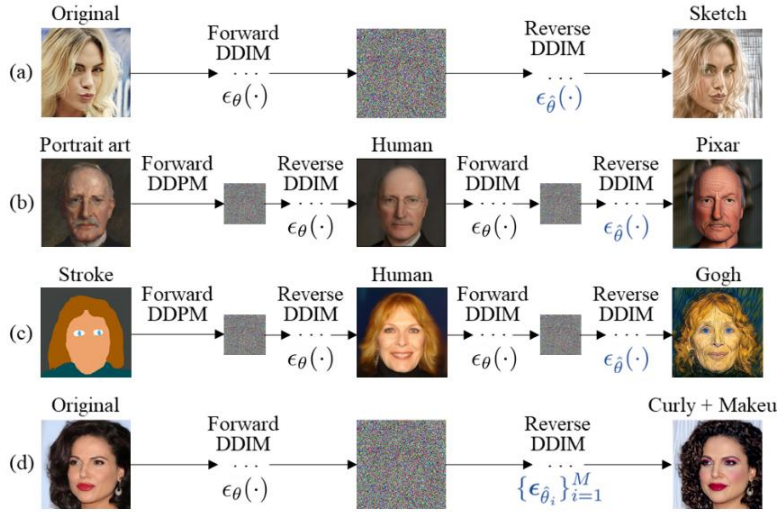
$$\mathcal{L}_{\text{direction}}(\boldsymbol{x}_{\text{gen}}, y_{\text{tar}}; \boldsymbol{x}_{\text{ref}}, y_{\text{ref}}) := 1 - \frac{\langle \Delta I, \Delta T \rangle}{\|\Delta I\| \|\Delta T\|}, \qquad (9)$$

where

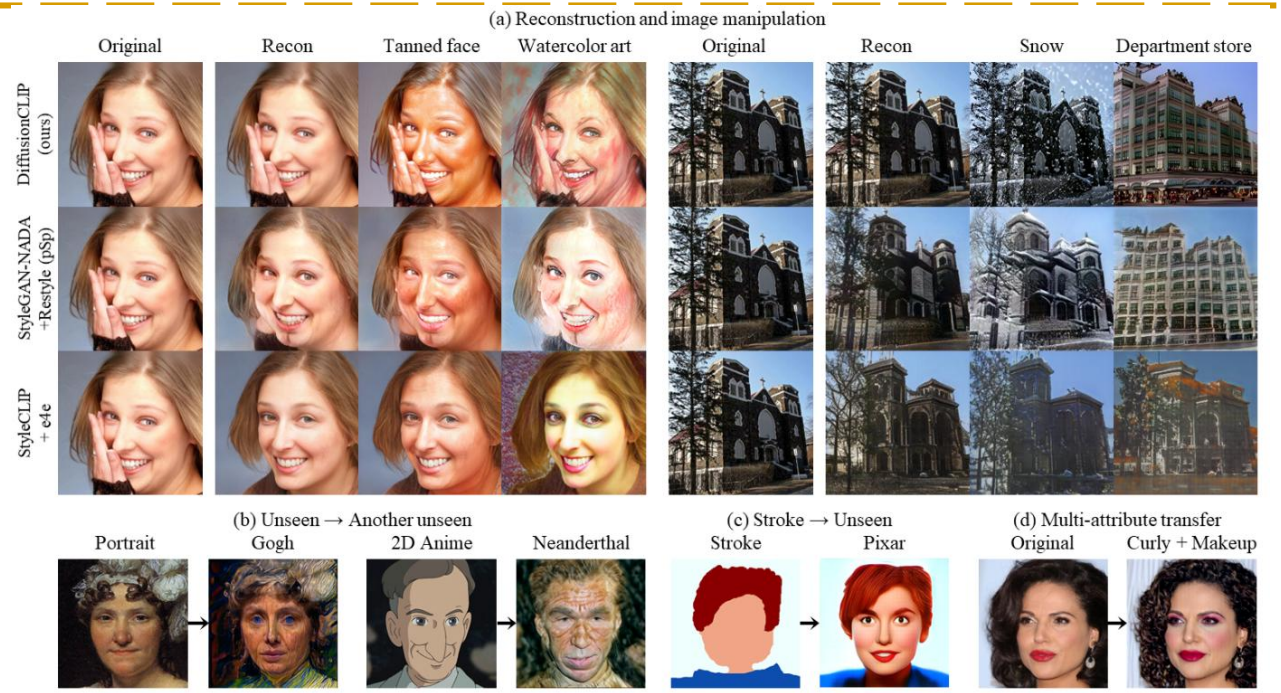
$$\Delta T = E_T(y_{\text{tar}}) - E_T(y_{\text{ref}}),\ \Delta I = E_I(\boldsymbol{x}_{\text{gen}}) - E_I(\boldsymbol{x}_{\text{ref}}).$$

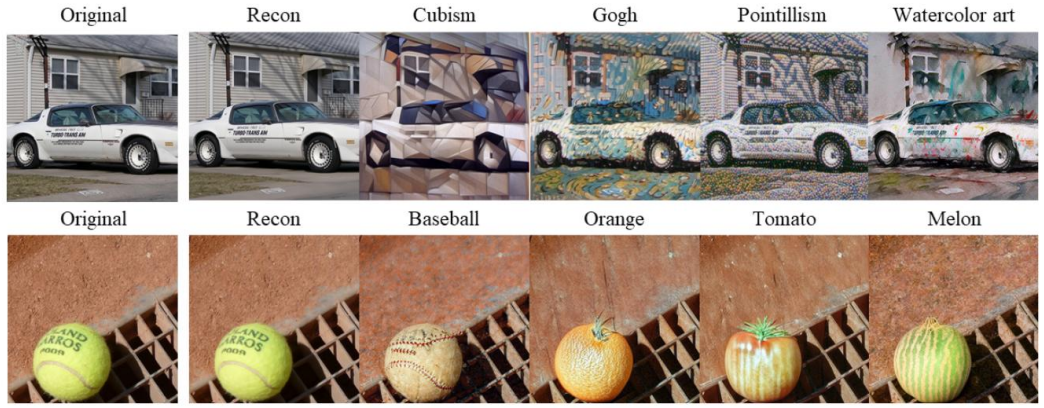


$$\mathcal{L}_{\text{direction}}\left(\hat{\boldsymbol{x}}_0(\hat{\theta}), y_{\text{tar}}; \boldsymbol{x}_0, y_{\text{ref}}\right) + \mathcal{L}_{\text{id}}(\hat{\boldsymbol{x}}_0(\hat{\theta}), \boldsymbol{x}_0),$$



DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation





	CelebA-HQ			LSUN-Church	
	$\mathcal{S}_{\text{dir}}\uparrow$	SC↑	ID↑	$\mathcal{S}_{\text{dir}}\uparrow$	SC↑
StyleCLIP	0.13	86.8%	0.35	0.13	67.9%
StyleGAN-NADA	0.16	89.4%	0.42	0.15	73.2%
DiffusionCLIP (Ours)	**0.17**	**93.7%**	**0.70**	**0.20**	**78.1%**

Image Edition

qkv

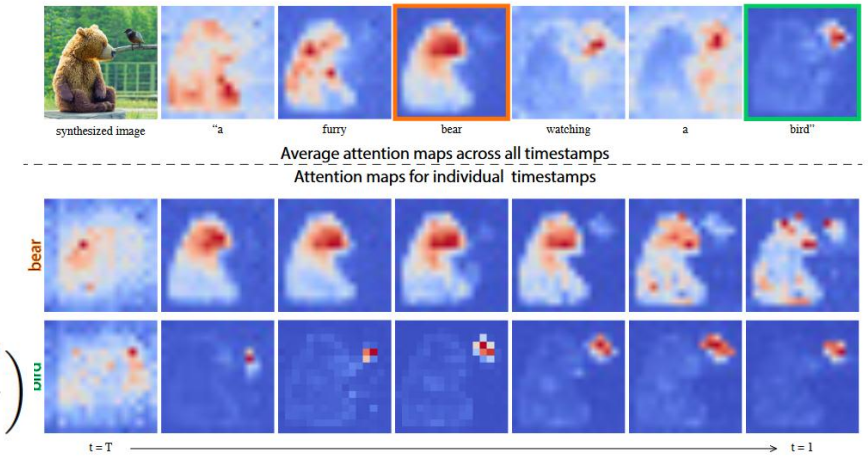
$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),$$

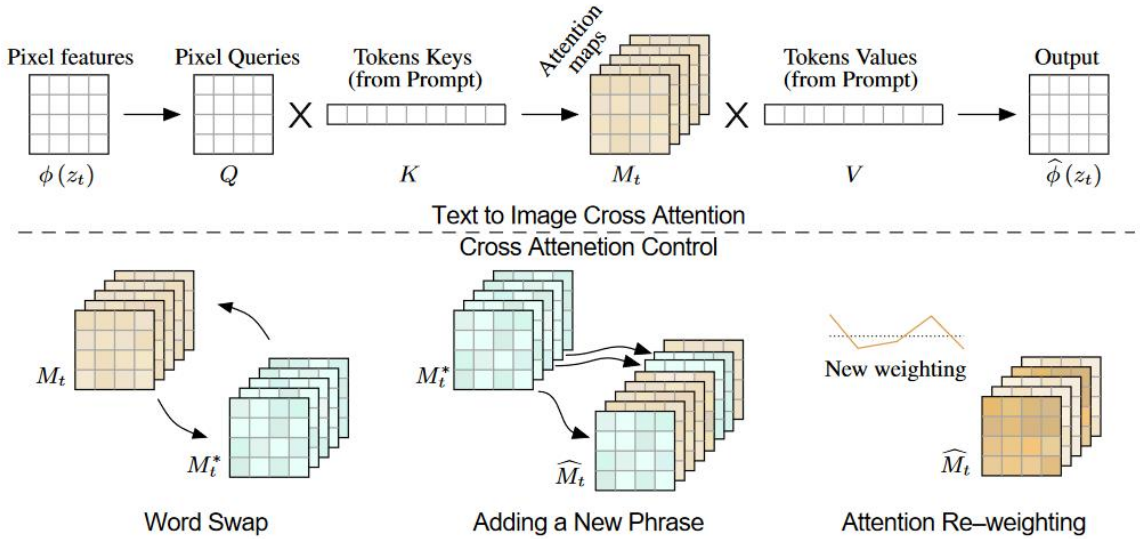
$$K = W_K^{(i)} \cdot \tau_\theta(y),$$

$$V = W_V^{(i)} \cdot \tau_\theta(y).$$

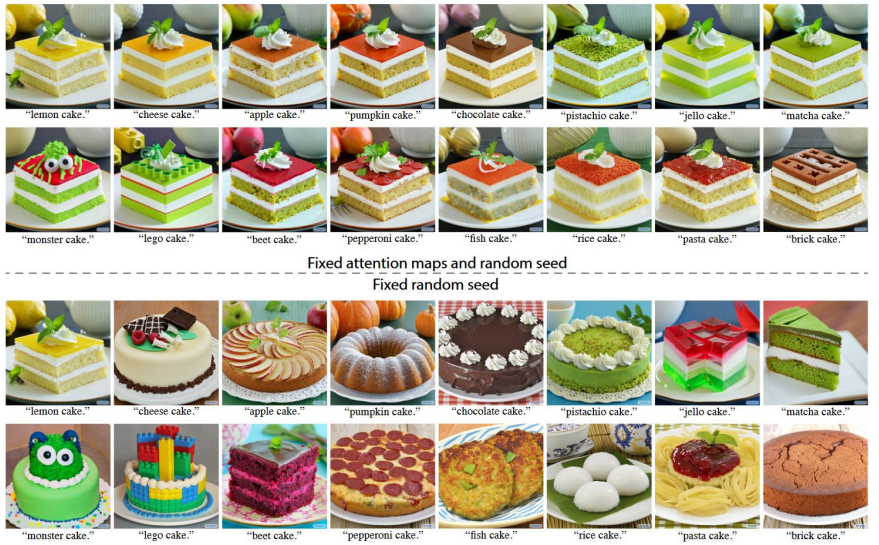
$$M = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

cross-attention





Prompt-to-Prompt Image Editing with Cross Attention Control



- Edit objects in an image
- Globally edit an image
- Amplify or attenuate the semantic effect of a word

Thanks