Text to Image Generation: Pretraining and Applications

Wangmeng Zuo

Centre on Machine Learning Research
Harbin Institute of Technology

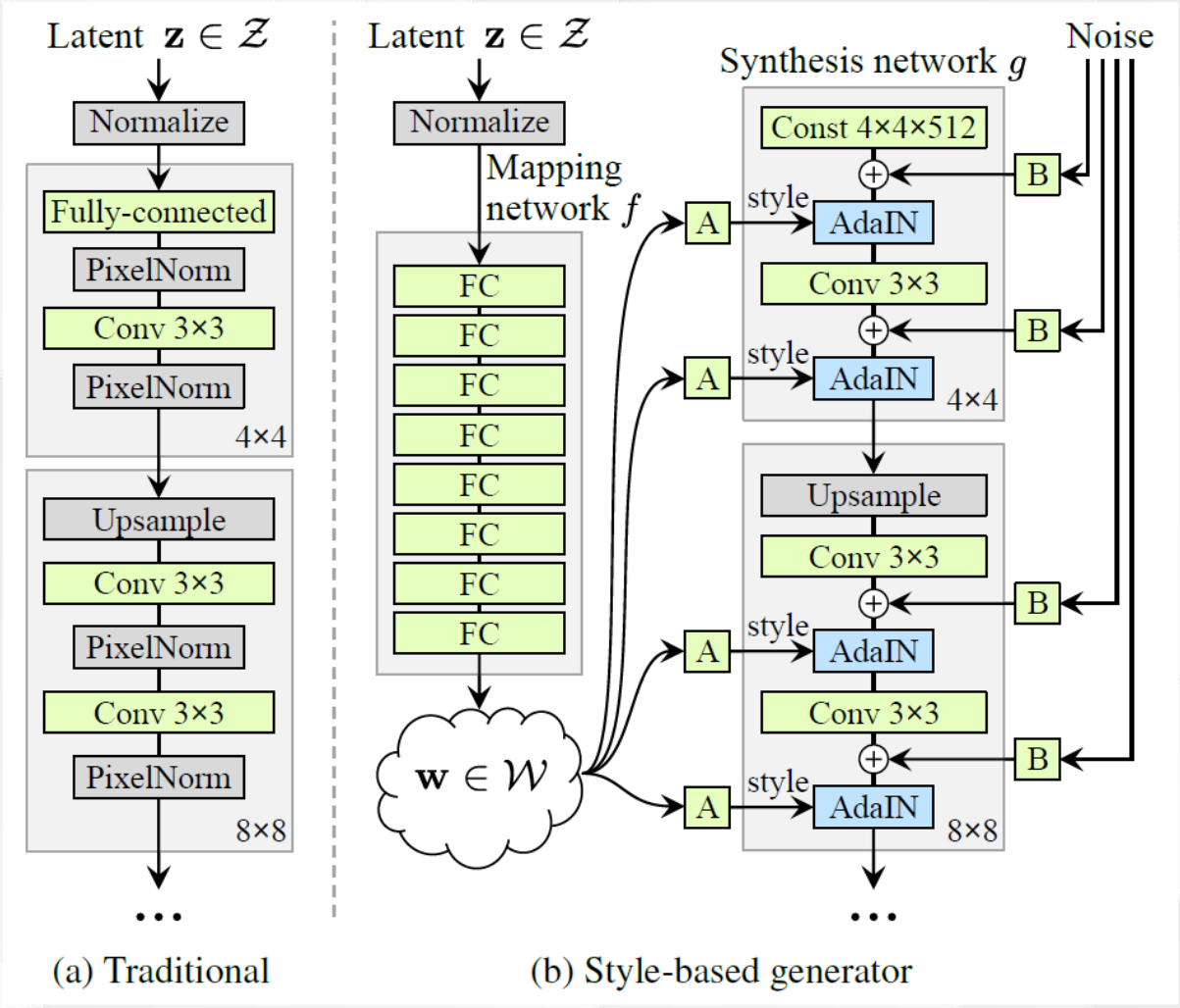
Content

- Large Scale Pretraining Models
  - GAN
  - Auto-autoregressive models
  - Diffusion models
  - Return of GAN and AR Models
  - From T2I to 3D/Video Generation

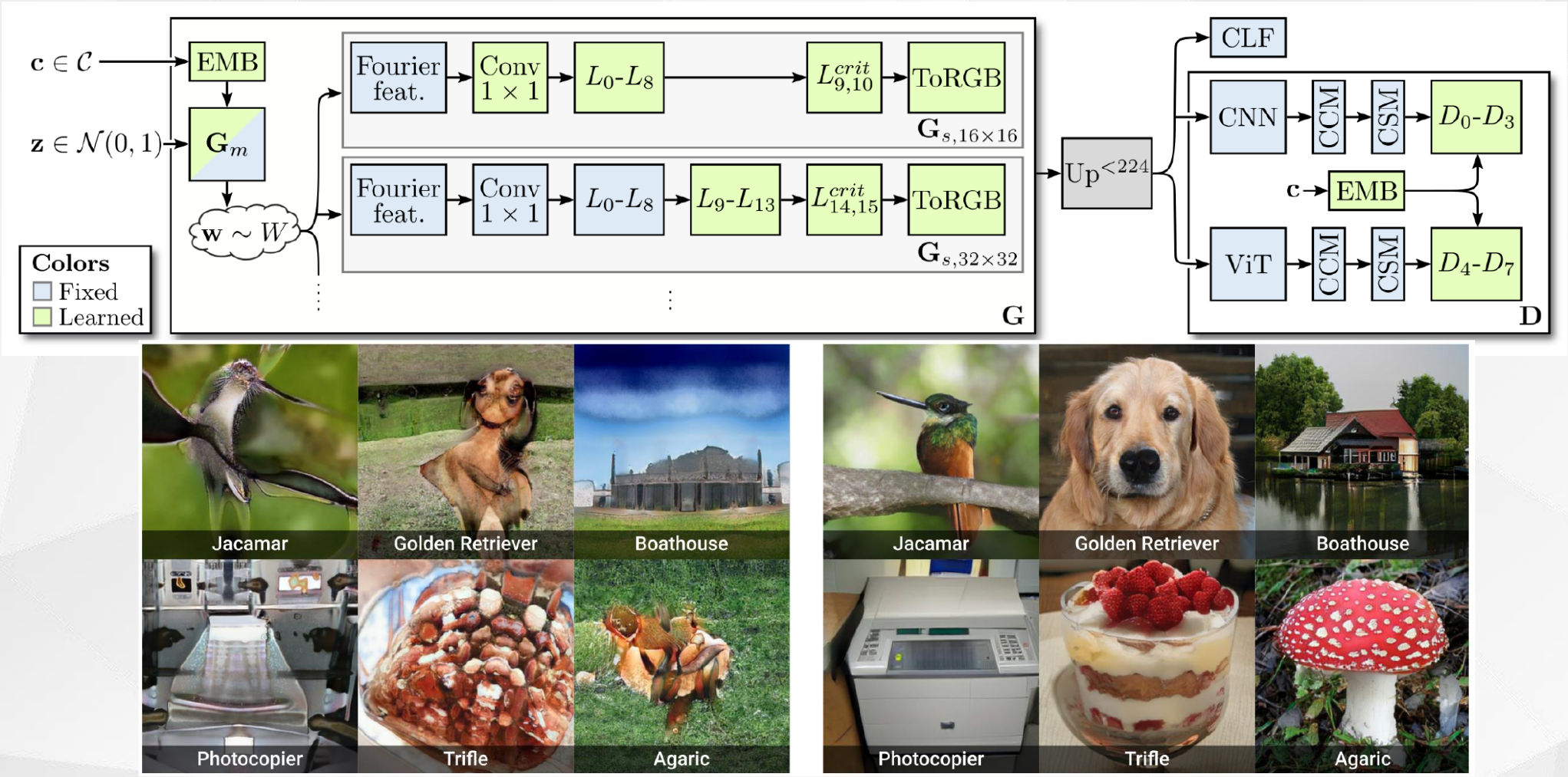
- Applications

StyleGAN 1~3 (Karras, CVPR 2019)

- Progressive GAN
- Multi-layer manipulation on style and noise
- Truncation trick



StyleGAN-XL



StyleGAN-XL: Scaling StyleGAN to Large Diverse Datasets, Siggraph 2022.

Content

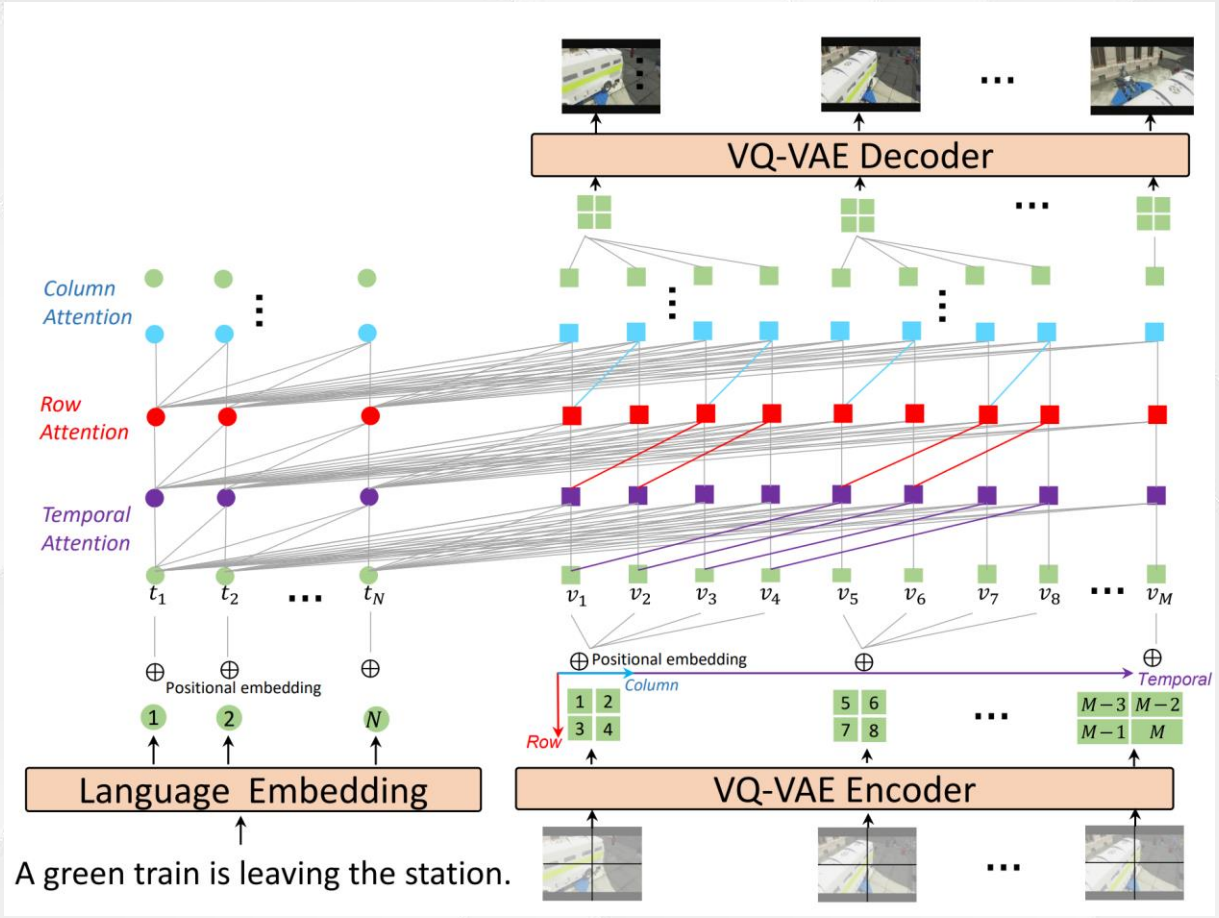
- Large Scale Pretraining Models
  - GAN
  - Auto-autoregressive models
  - Diffusion models
  - Return of GAN and AR Models
  - From T2I to 3D/Video Generation
- Applications

DALL-E: Text-to-Image Generation

- Sufficient data and scale
  - 250 million text-images pairs
  - A discrete variational autoencoder (**dVAE**) to compress each 256x256 RGB image into a 32x32 grid of image tokens.
  - An autoregressive transformer to model the joint distribution over the text and image tokens

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever, Zero-Shot Text-to-Image Generation, Arxiv 2021.

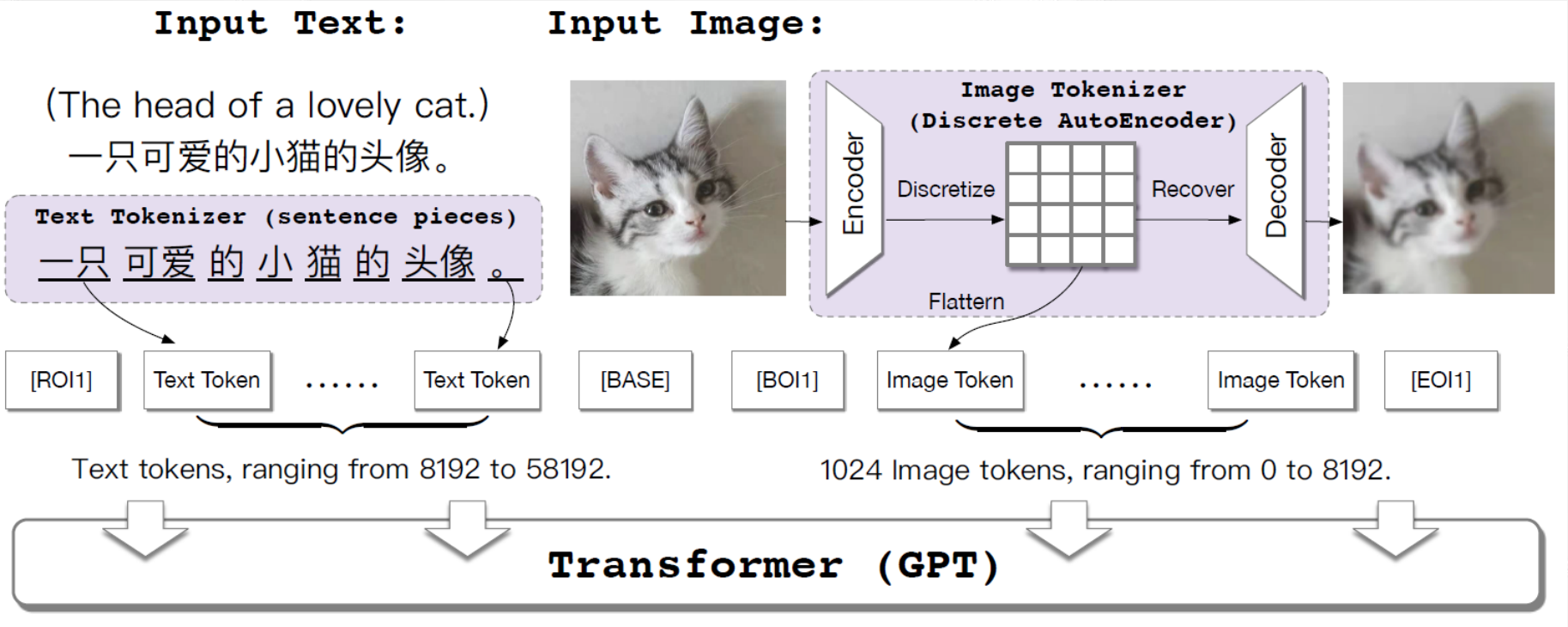
GODIVA Model



GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions, arXiv 2021

CogView: Text-to-Image Generation

- **VQ-VAE** Tokenizer

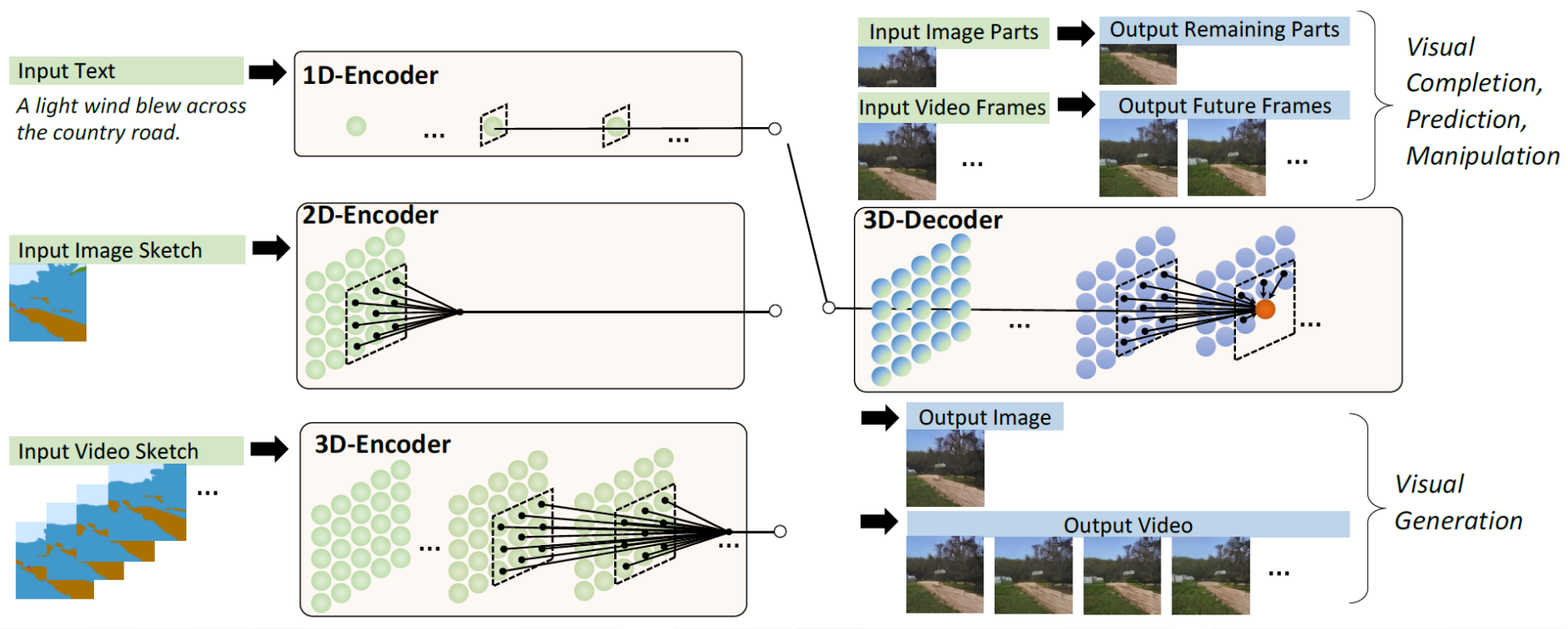


Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering Text-to-Image Generation via Transformers. Arxiv 2021

CogView: Text-to-Image Generation

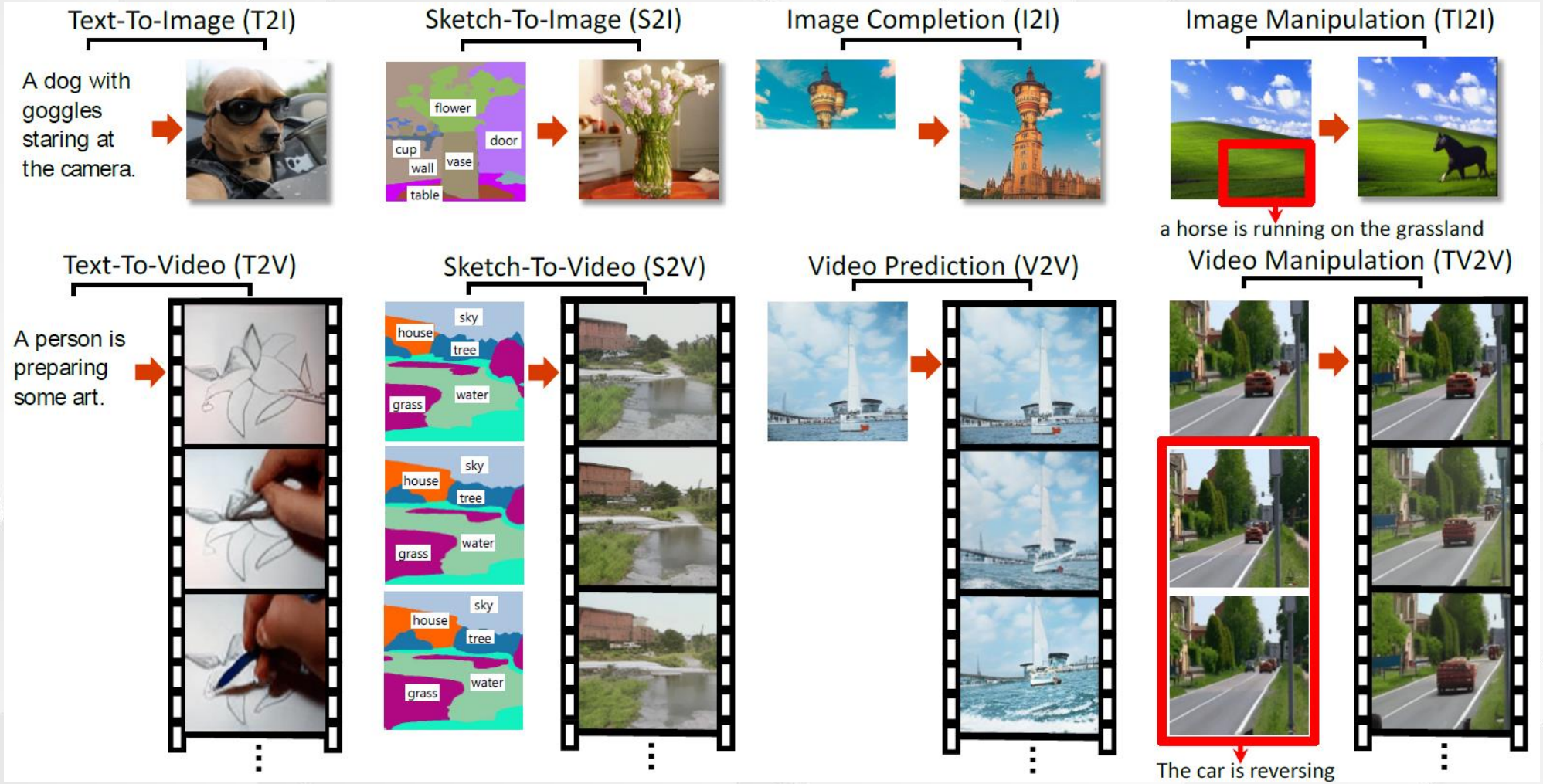
NÜWA: Neural visUalWorld creAtion

- Adaptive encoder and pre-trained decoder, many applications



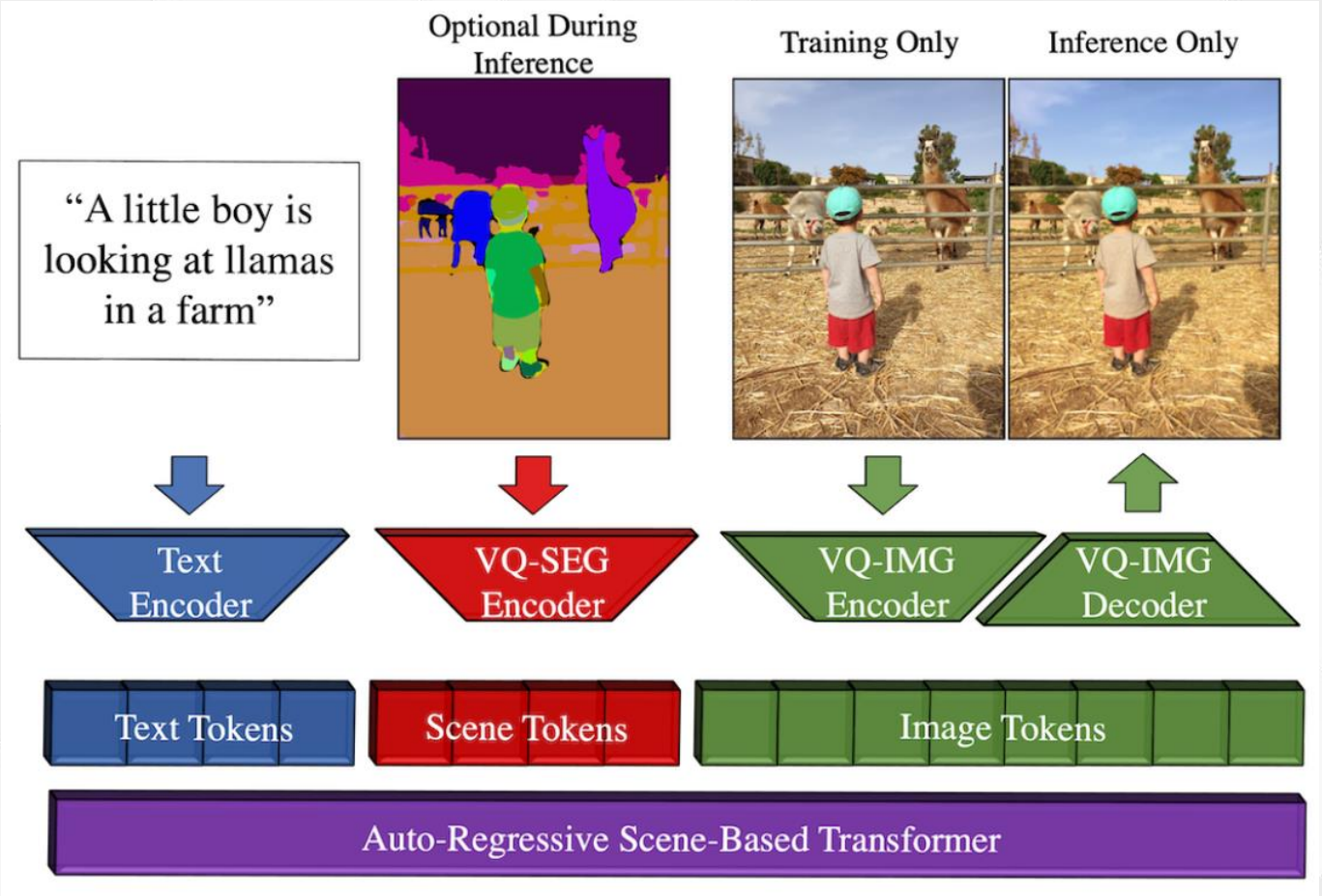
Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, Nan Duan, NÜWA: Visual Synthesis Pre-training for Neural visUalWorld creation, Arxiv 2021.

NÜWA: Many Possible Applications



Make-A-Scene

- Segmentation mask: domain-specific knowledge
- Face/Object emphasis
- Classifier-free guidance



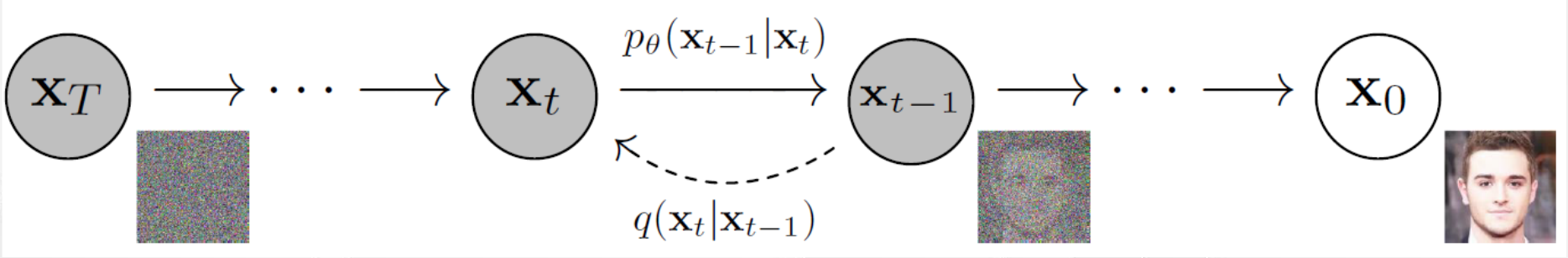
Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors, Arxiv 2022

Samples on MS-COCO Prompts

Content

- Large Scale Pretraining Models
  - GAN
  - Auto-autoregressive models
  - Diffusion models
  - Return of GAN and AR Models
  - From T2I to 3D/Video Generation
- Applications

DDPM



- Forward process or diffusion process

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$

- Reverse process

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Jonathan Ho , Ajay Jain, Pieter Abbeel, Denoising Diffusion Probabilistic Models, NeurIPS 2020

DDPM: Learning Objective

- Forward process

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t\mathbf{I}),$$
$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$$

- Reverse process

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$$

$$L_{t-1} = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)\|^2\right] + C$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}, \text{ where } \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t, \mathbf{x}_0, \boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

DDPM: Results

- Training

$$\epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)$$

- Generation

$$\epsilon_\theta(\mathbf{x}_t, t)$$

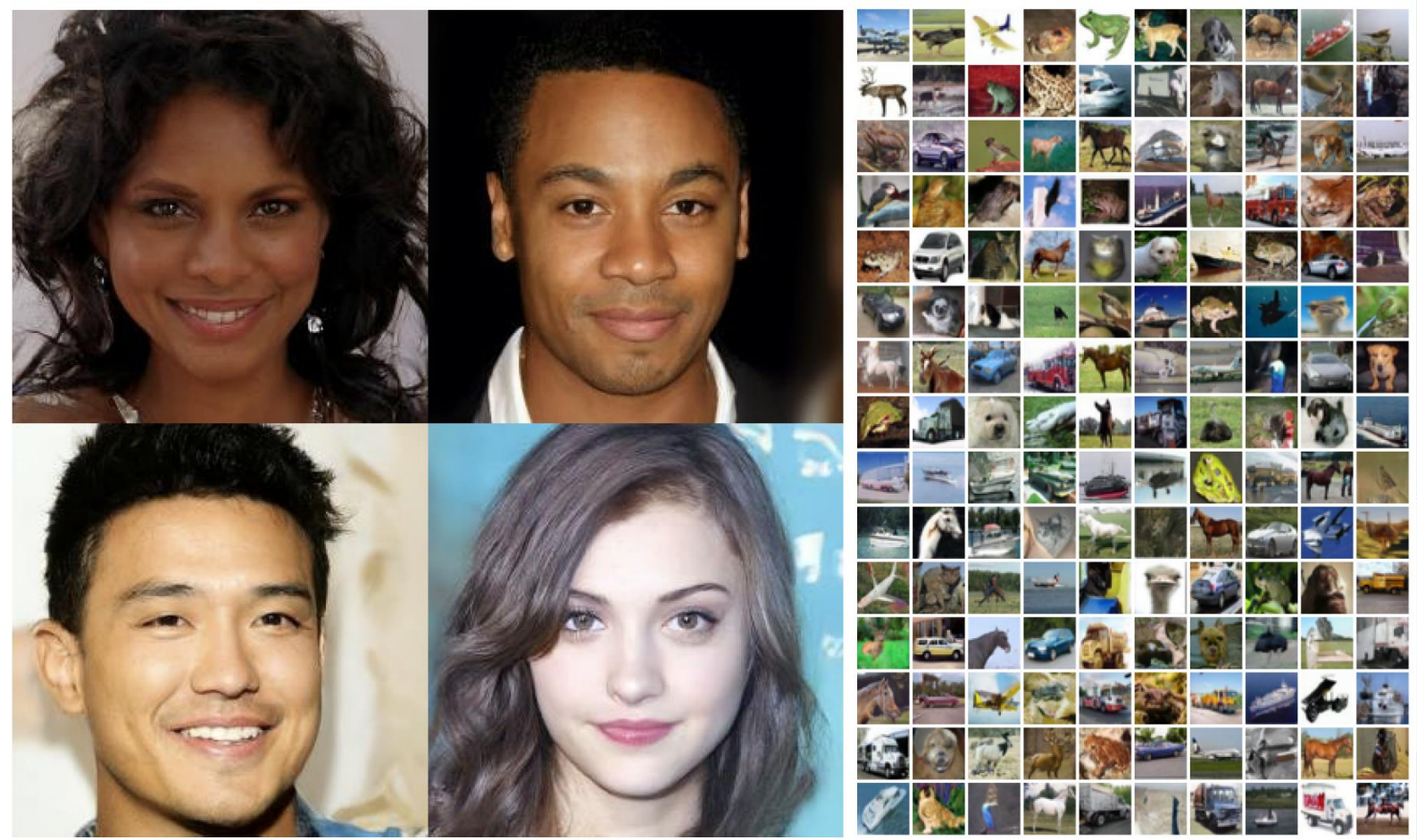
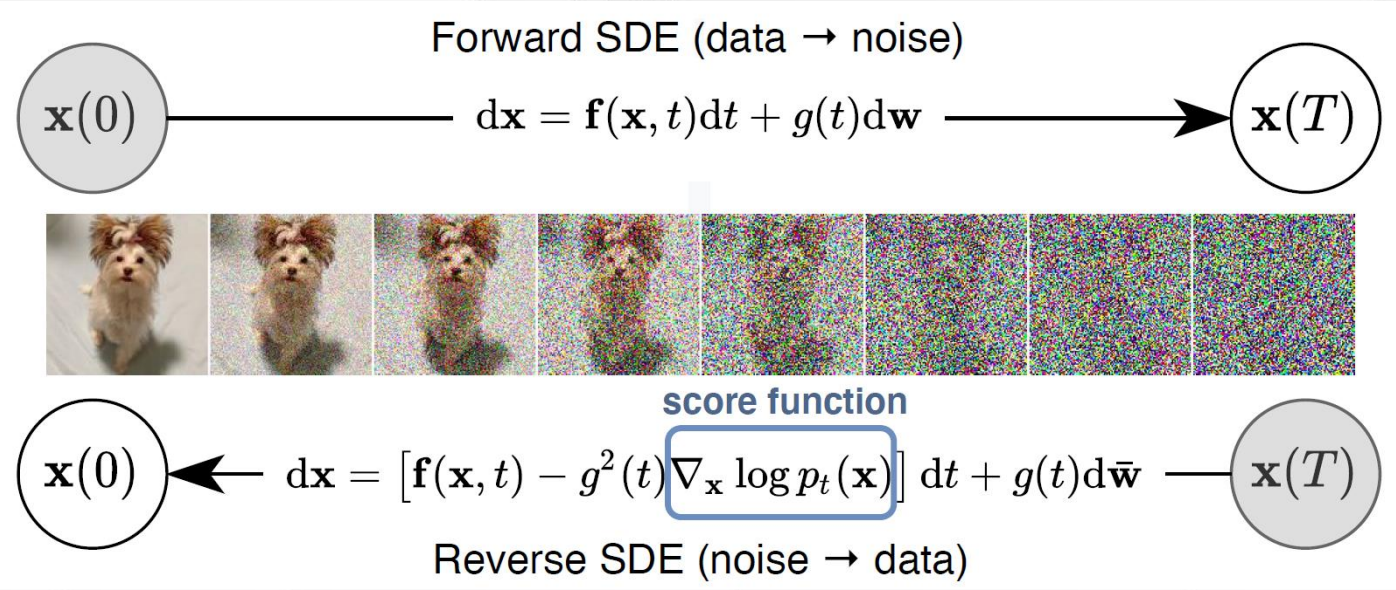


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

Score-based Generative Modeling

- Score: time-dependent gradient field
- Accurately estimate these scores with neural networks

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t$$



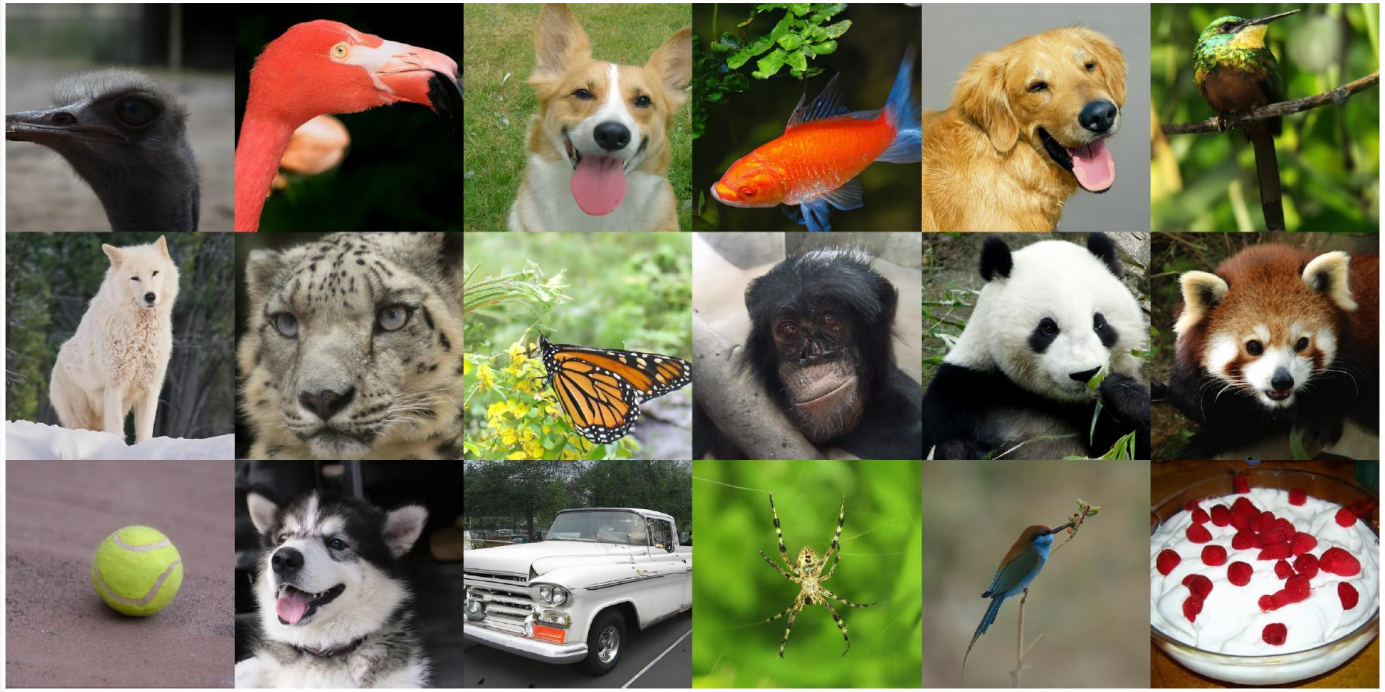
- Controllable Generation

$$\mathrm{d}\mathbf{x} = \{\mathbf{f}(\mathbf{x}, t) - g(t)^2[\nabla_{\mathbf{x}} \log p_t(\mathbf{x}) + \nabla_{\mathbf{x}} \log p_t(\mathbf{y} \mid \mathbf{x})]\}\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$$

Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021.

ADM: Diffusion Models Beat GANs

- Combining classifier guidance $p_{\theta,\phi}(x_t|x_{t+1}, y) = Z p_\theta(x_t|x_{t+1}) p_\phi(y|x_t)$
- ADM: ablated diffusion model

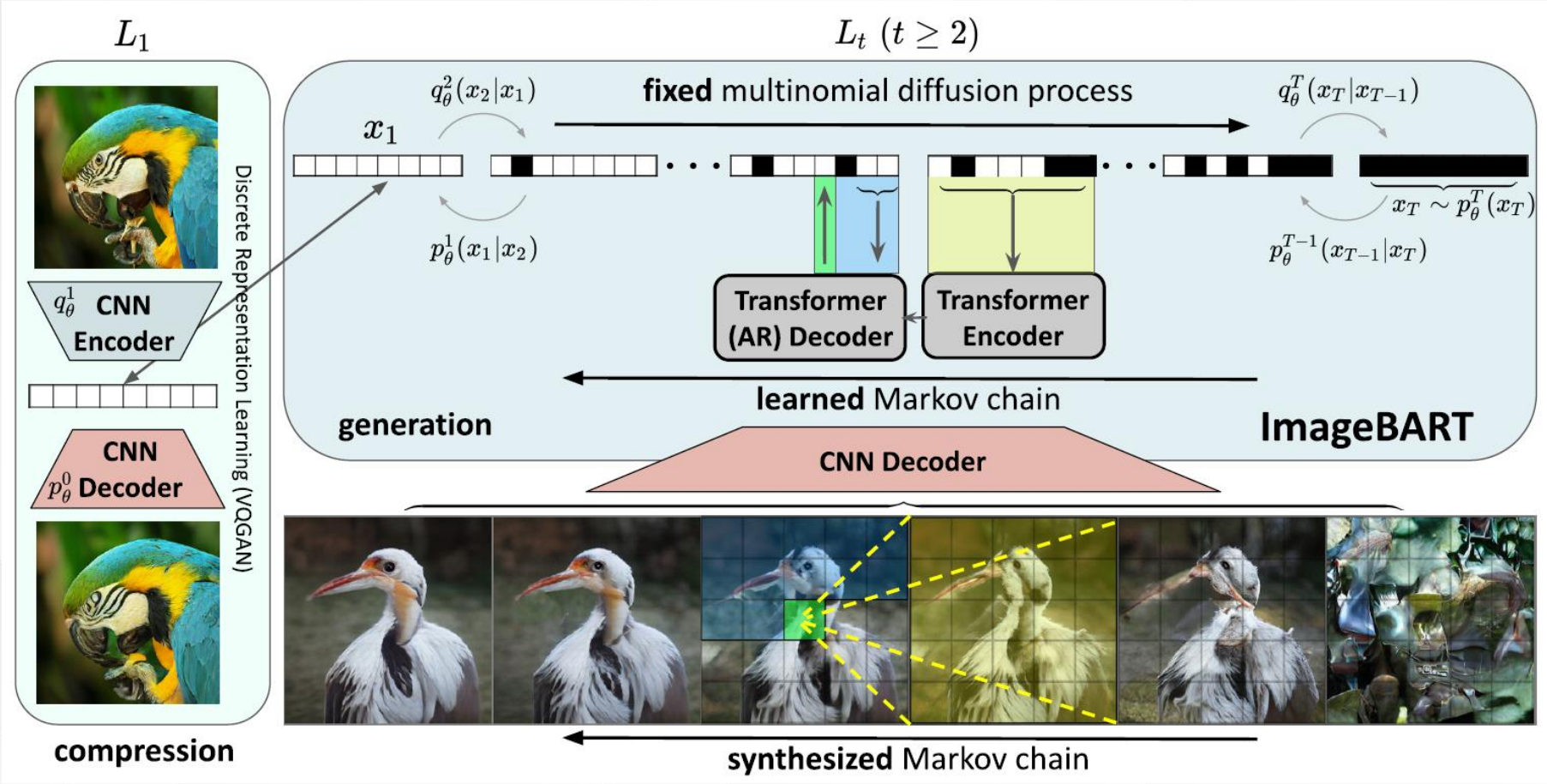


Model	FID	sFID	Prec	Rec
LSUN Bedrooms 256×256				
DCTransformer† [42]	6.40	6.66	0.44	**0.56**
DDPM [25]	4.89	9.07	0.60	0.45
IDDPM [43]	4.24	8.21	0.62	0.46
StyleGAN [27]	2.35	6.62	0.59	0.48
ADM (dropout)	**1.90**	**5.59**	**0.66**	0.51
LSUN Horses 256×256				
StyleGAN2 [28]	3.84	6.46	0.63	0.48
ADM	2.95	**5.94**	0.69	**0.55**
ADM (dropout)	**2.57**	6.81	**0.71**	**0.55**
LSUN Cats 256×256				
DDPM [25]	17.1	12.4	0.53	0.48
StyleGAN2 [28]	7.25	**6.33**	0.58	0.43
ADM (dropout)	**5.57**	6.69	**0.63**	**0.52**
ImageNet 64×64				
BigGAN-deep* [5]	4.06	3.96	**0.79**	0.48
IDDPM [43]	2.92	**3.79**	0.74	0.62
ADM	2.61	**3.77**	0.73	0.63
ADM (dropout)	**2.07**	4.29	0.74	**0.63**

Model	FID	sFID	Prec	Rec
ImageNet 128×128				
BigGAN-deep [5]	6.02	7.18	**0.86**	0.35
LOGAN† [68]	3.36			
ADM	5.91	**5.09**	0.70	**0.65**
ADM-G (25 steps)	5.98	7.04	0.78	0.51
ADM-G	**2.97**	**5.09**	0.78	0.59
ImageNet 256×256				
DCTransformer† [42]	36.51	8.24	0.36	**0.67**
VQ-VAE-2†‡ [51]	31.11	17.38	0.36	0.57
IDDPM‡ [43]	12.26	5.42	0.70	0.62
SR3†‡ [53]	11.30			
BigGAN-deep [5]	6.95	7.36	**0.87**	0.28
ADM	10.94	6.02	0.69	0.63
ADM-G (25 steps)	5.44	5.32	0.81	0.49
ADM-G	**4.59**	**5.25**	0.82	0.52
ImageNet 512×512				
BigGAN-deep [5]	8.43	8.13	**0.88**	0.29
ADM	23.24	10.19	0.73	**0.60**
ADM-G (25 steps)	8.41	9.67	0.83	0.47
ADM-G	**7.72**	**6.57**	0.87	0.42

Diffusion Models Beat GANs on Image Synthesis, Arxiv 2021.

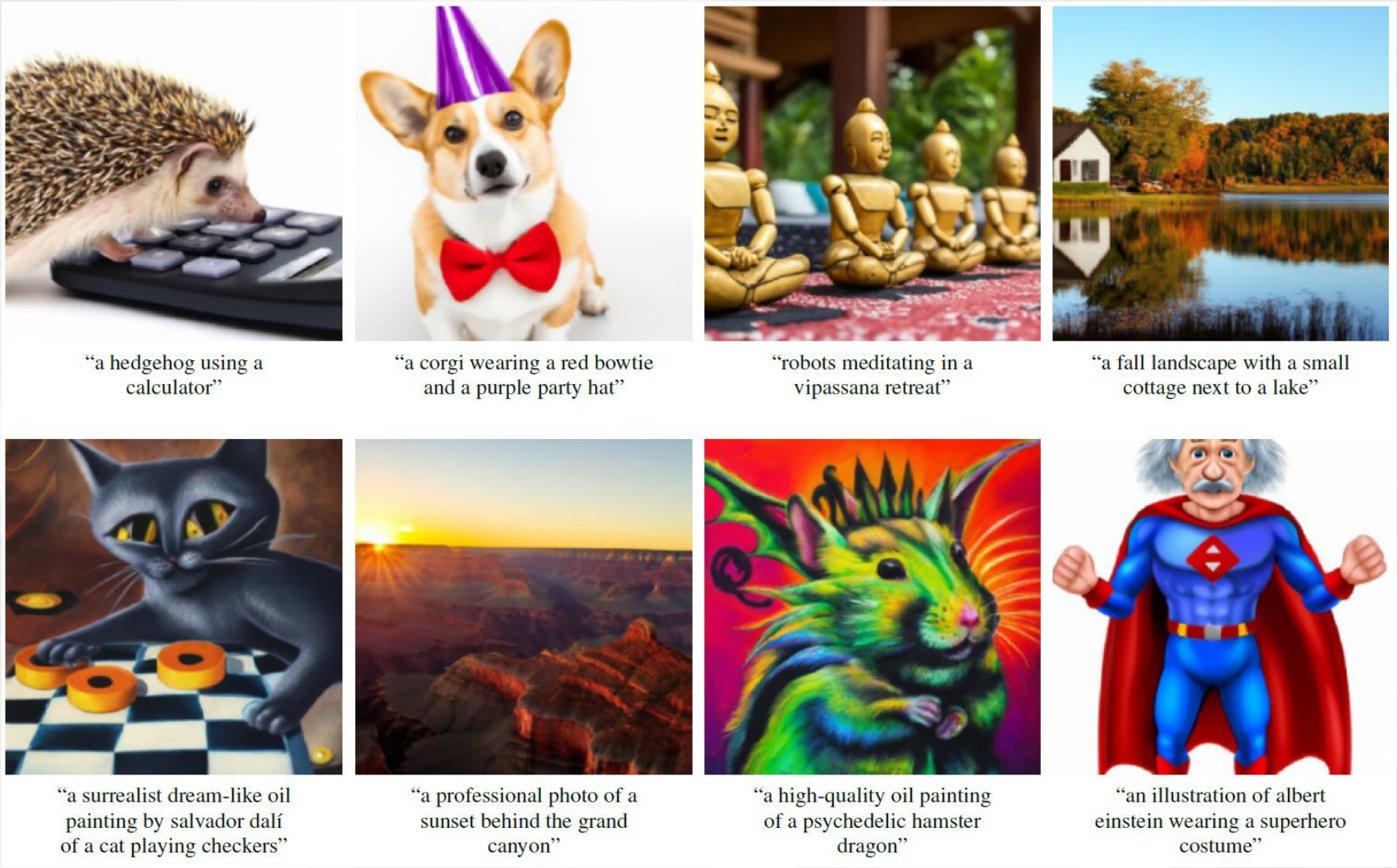
ImageBART



Patrick Esser, Robin Rombach, Andreas Blattmann, Björn Ommer, ImageBART: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Synthesis, Arxiv 2021

GLIDE

- VQVAE
- Transformer
- Diffusion model



GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, Arxiv 2021.

GLIDE: Language-guided Image Inpainting

"a man with red hair"

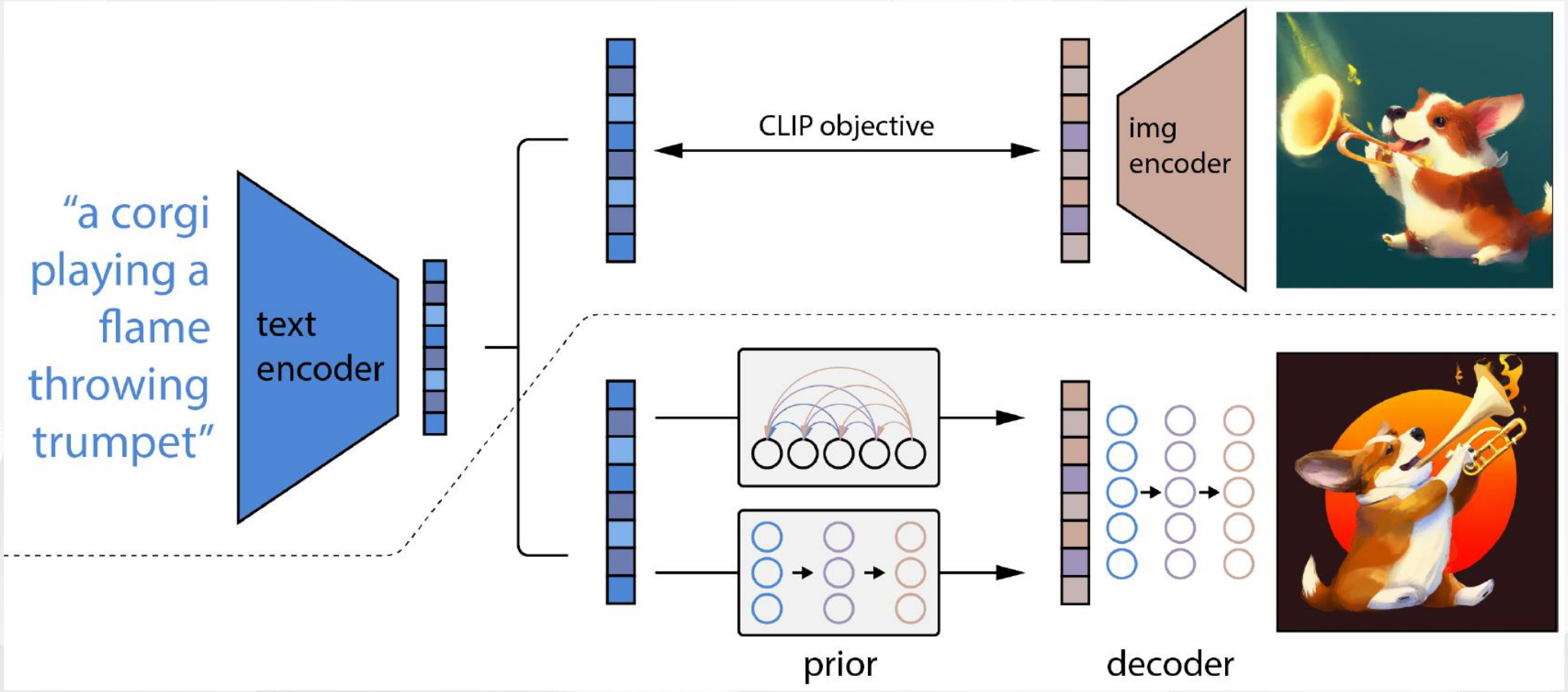
"a vase of flowers"

"an old car in a snowy forest"

"a man wearing a white hat"

DALL.E-2: unCLIP

- High quality 1024 × 1024 iImages



Hierarchical Text-Conditional Image Generation with CLIP Latents, Arxiv 2022

DALL.E-2: unCLIP

an espresso machine that makes coffee from human souls, artstation

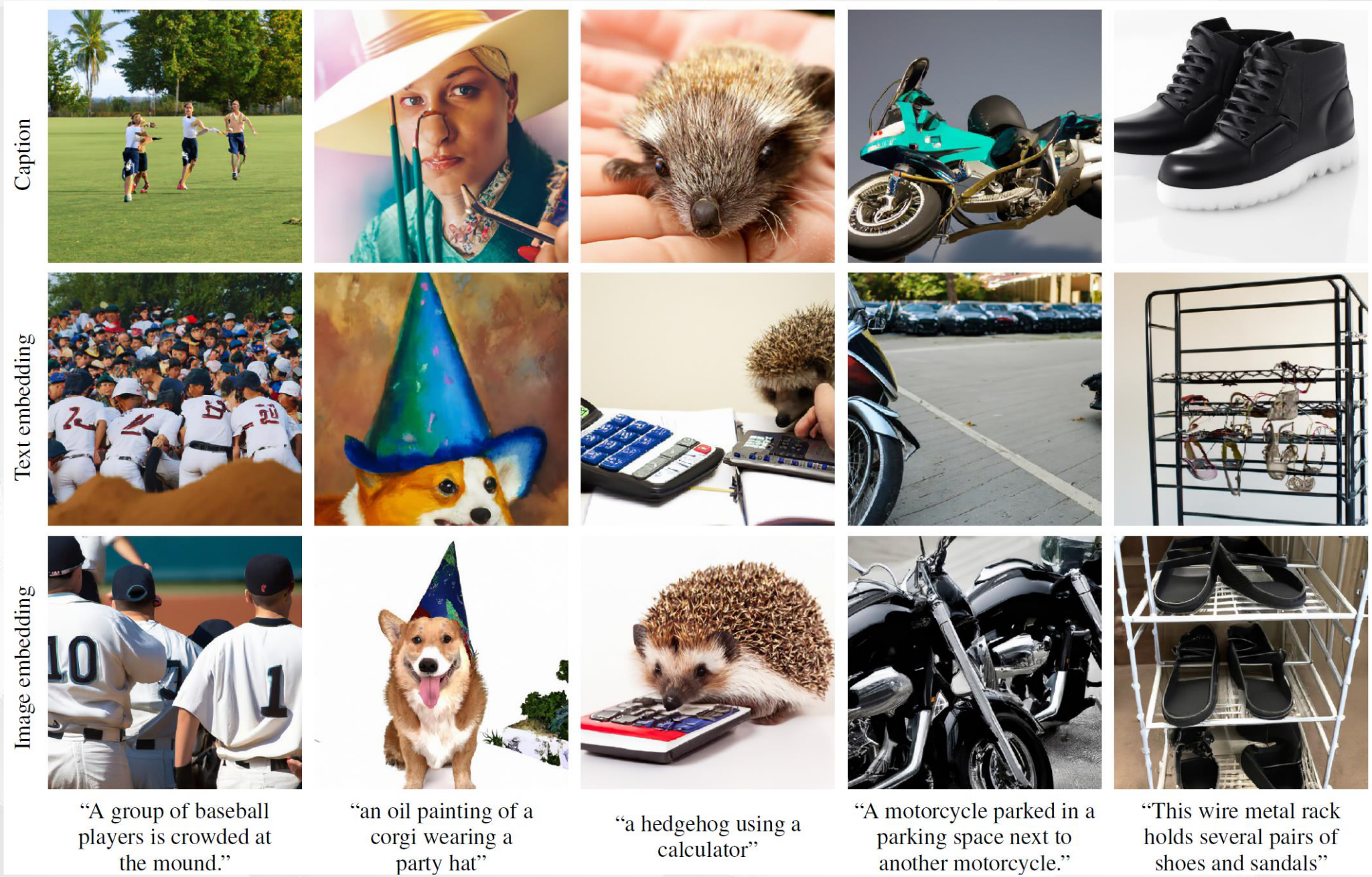
panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

a dolphin in an astronaut suit on saturn, artstation

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

a teddybear on a skateboard in times square


Caption
Text embedding
Image embedding
"A group of baseball players is crowded at the mound."
"an oil painting of a corgi wearing a party hat"
"a hedgehog using a calculator"
"A motorcycle parked in a parking space next to another motorcycle."
"This wire metal rack holds several pairs of shoes and sandals"

Samples on MS-COCO Prompts

Imagen

- Abstract
  - Generic large language model, eg. T5, pretrained on text only corpora.
  - Text-guided super-res model

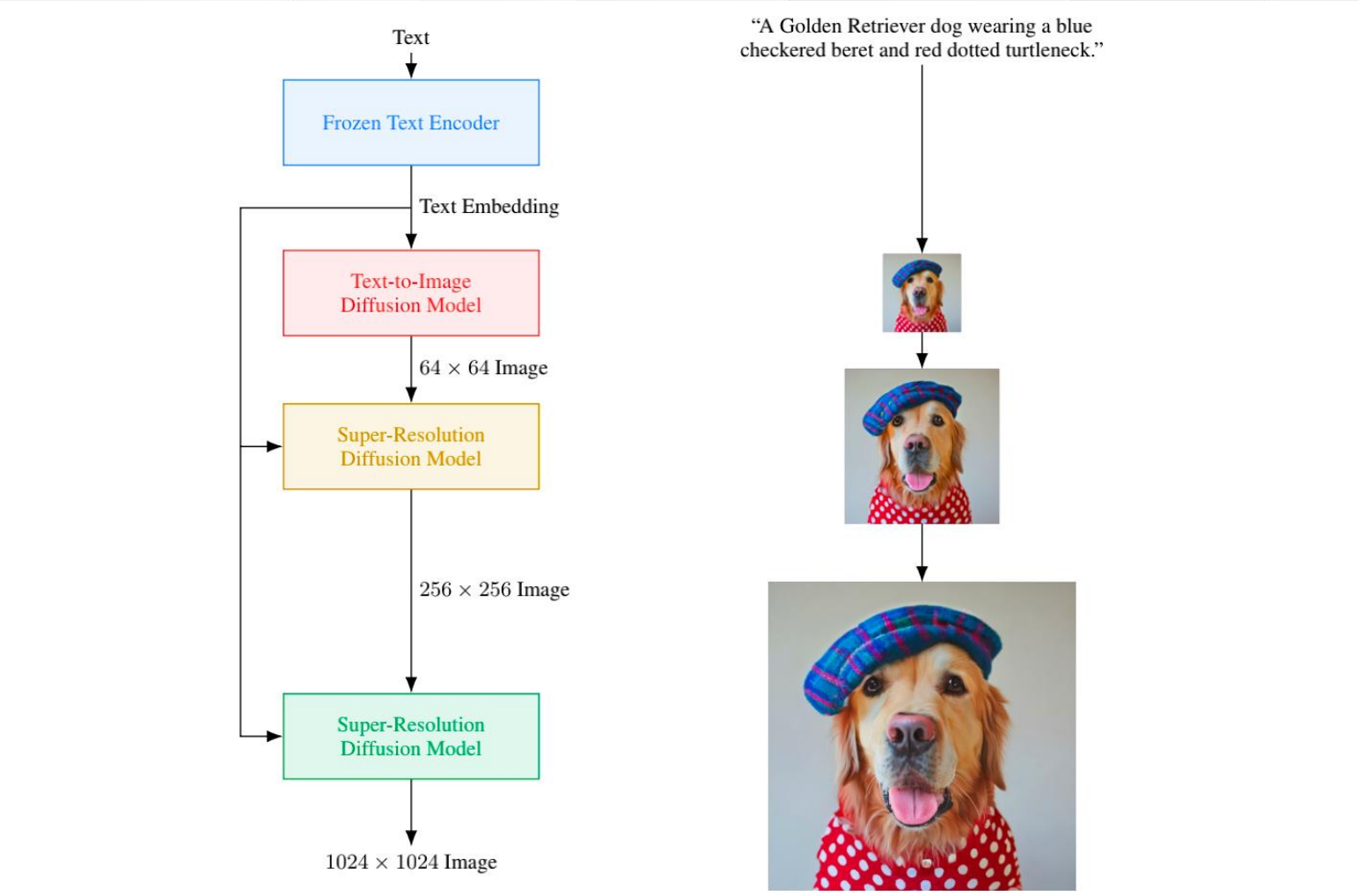
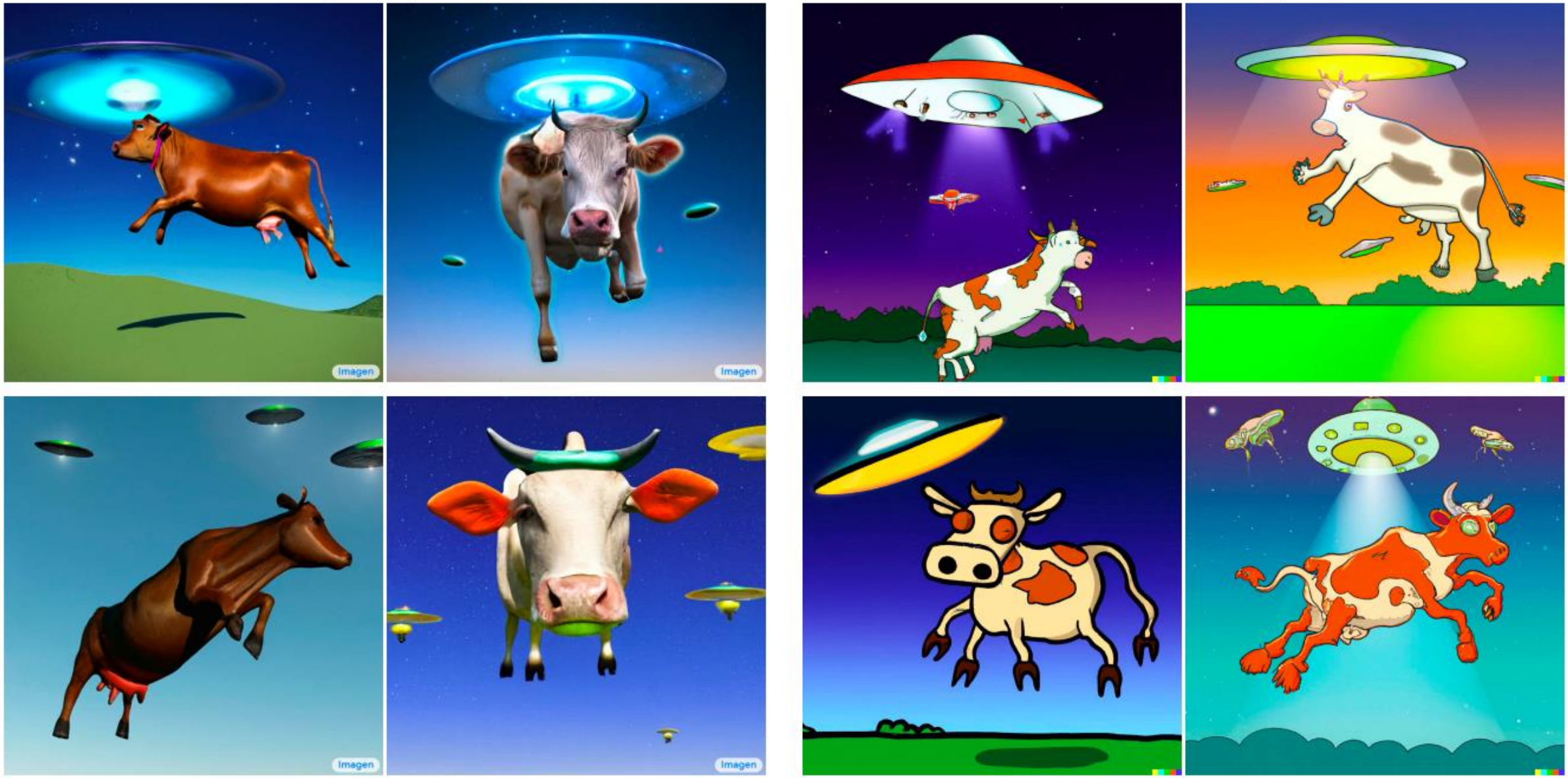


Figure A.4: Visualization of Imagen. Imagen uses a frozen text encoder to encode the input text into text embeddings. A conditional diffusion model maps the text embedding into a 64×64 image. Imagen further utilizes text-conditional super-resolution diffusion models to upsample the image, first $64 \times 64 \rightarrow 256 \times 256$, and then $256 \times 256 \rightarrow 1024 \times 1024$.

Imagen (Ours)

DALL-E 2 [54]



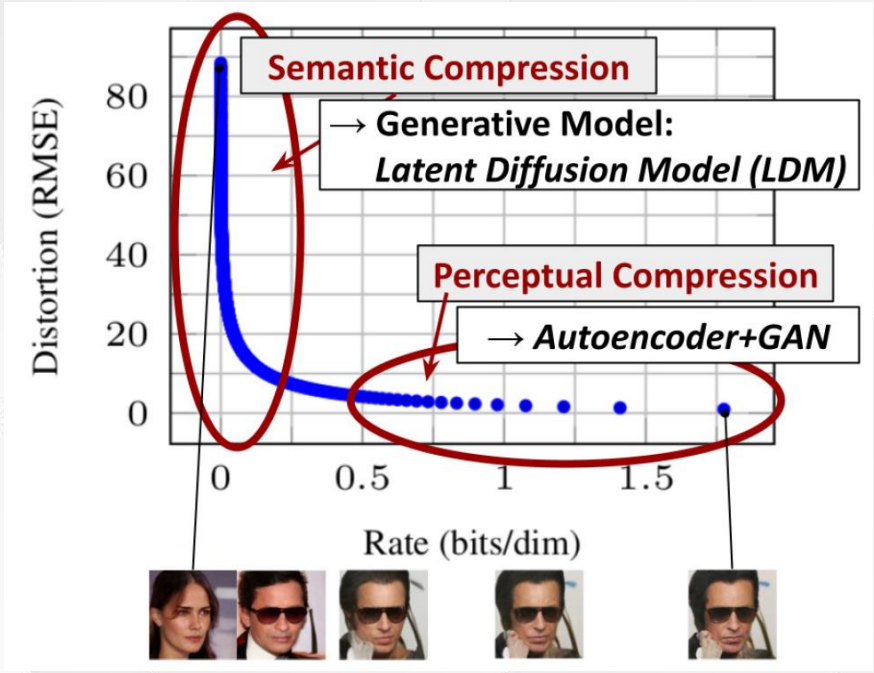
Hovering cow abducting aliens.

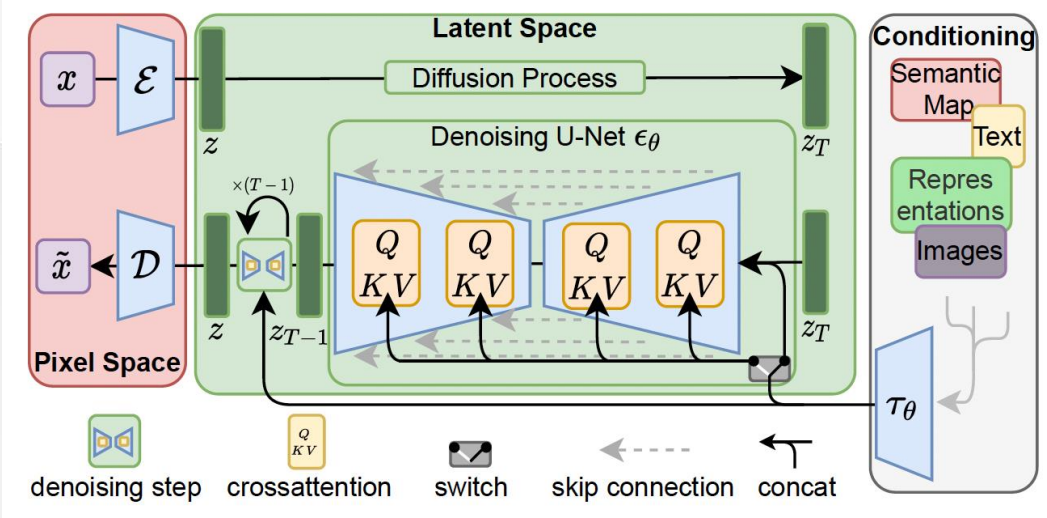
Imagen (Ours)

DALL-E 2 [54]

New York Skyline with Hello World written with fireworks on the sky.

LDM (Latent Diffusion Models)



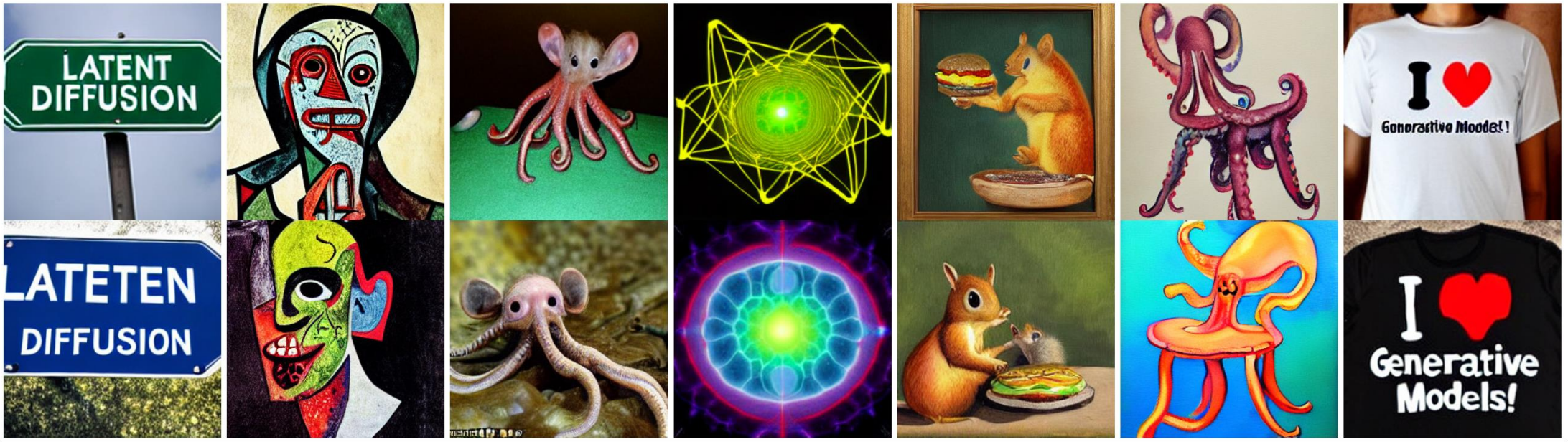


Convert an image into latent space to reduce computational resources

LDM: Results

Text-to-Image Synthesis on LAION. 1.45B Model.

'A street sign that reads "Latent Diffusion" '	*'A zombie in the style of Picasso'*	*'An image of an animal half mouse half octopus'*	*'An illustration of a slightly conscious neural network'*	*'A painting of a squirrel eating a burger'*	*'A watercolor painting of a chair that looks like an octopus'*	*'A shirt with the inscription: "I love generative models!" '*



Stable Diffusion

Stable Diffusion

Fork 4.6k | Starred 31k

Stable Diffusion was made possible thanks to a collaboration with Stability AI and Runway and builds upon our previous work:

High-Resolution Image Synthesis with Latent Diffusion Models
Robin Rombach*, Andreas Blattmann*, Dominik Lorenz, Patrick Esser, Björn Ommer
CVPR '22 Oral | GitHub | arXiv | Project page

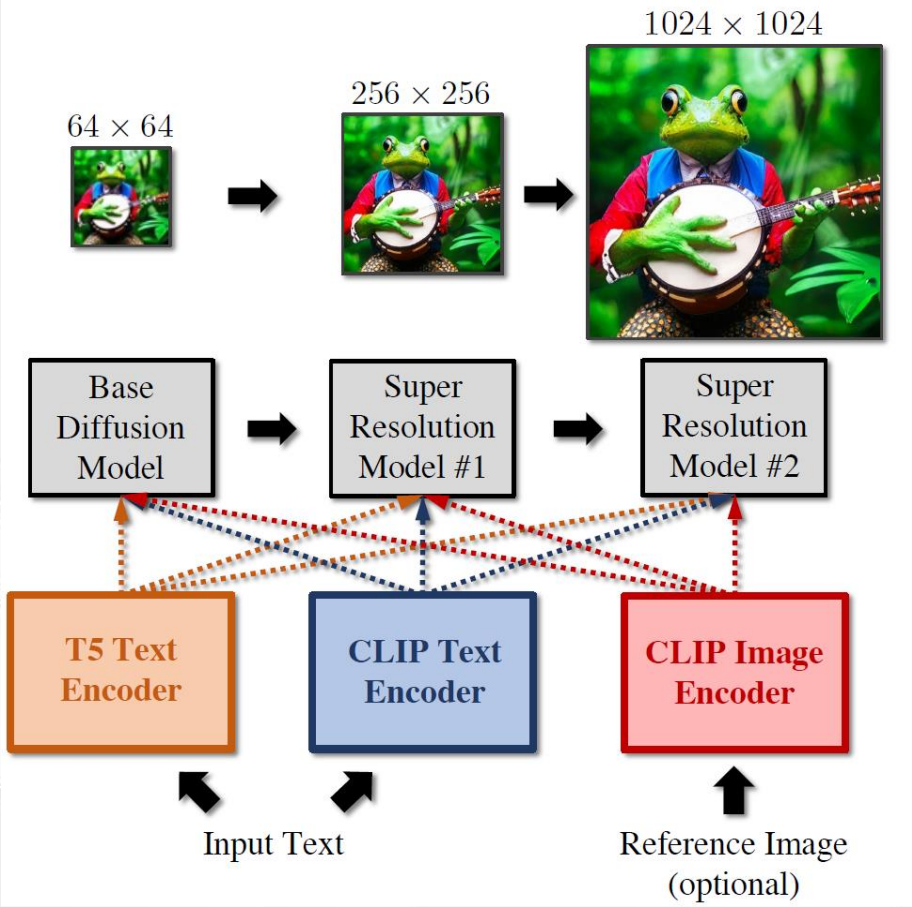
Stable Diffusion is a latent text-to-image diffusion model. Thanks to a generous compute donation from Stability AI and support from LAION, we were able to train a Latent Diffusion Model on 512x512 images from a subset of the LAION-5B database. Similar to Google's Imagen, this model uses a frozen CLIP ViT-L/14 text encoder to condition the model on text prompts. With its 860M UNet and 123M text encoder, the model is relatively lightweight and runs on a GPU with at least 10GB VRAM. See this section below and the model card.

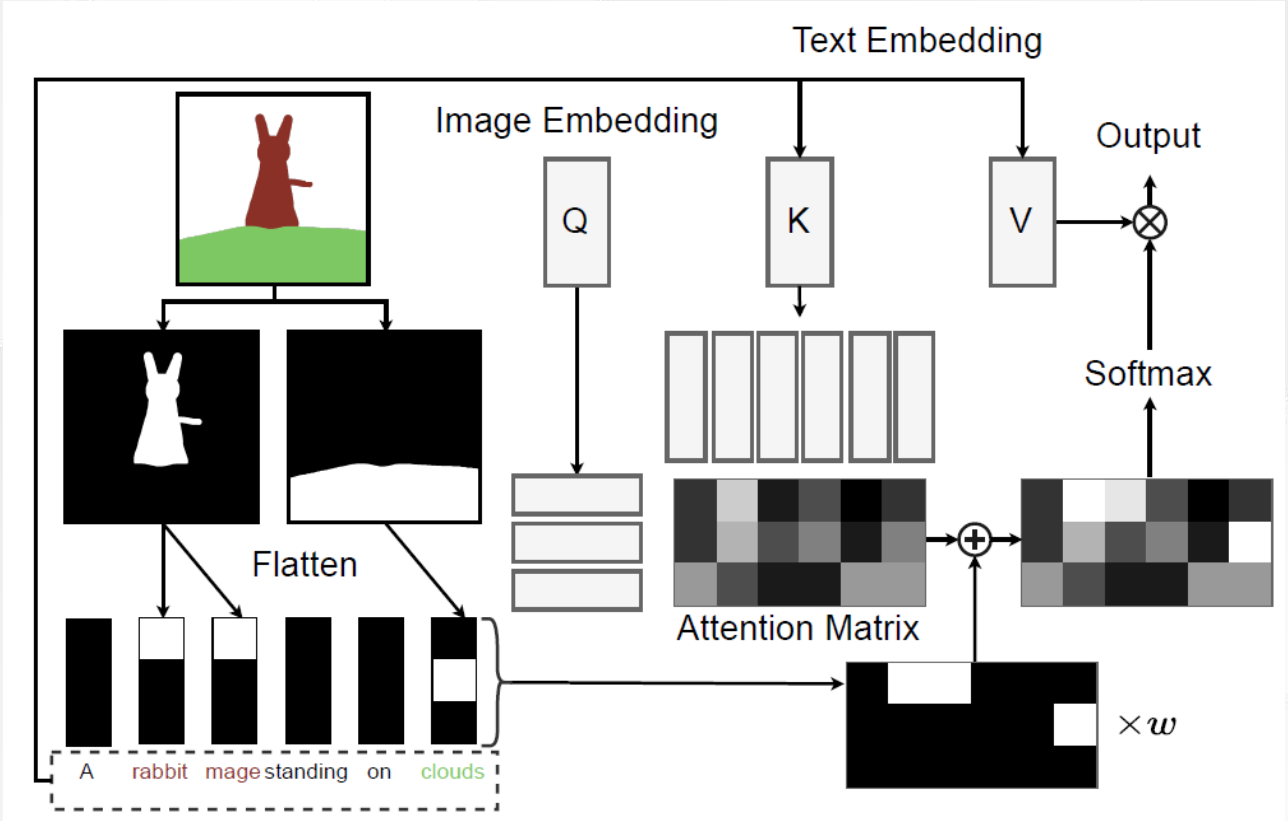
Diffusers

GitHub release v0.6.0 Contributor Covenant 2.0

集成现有Diffusion模型，直接调用

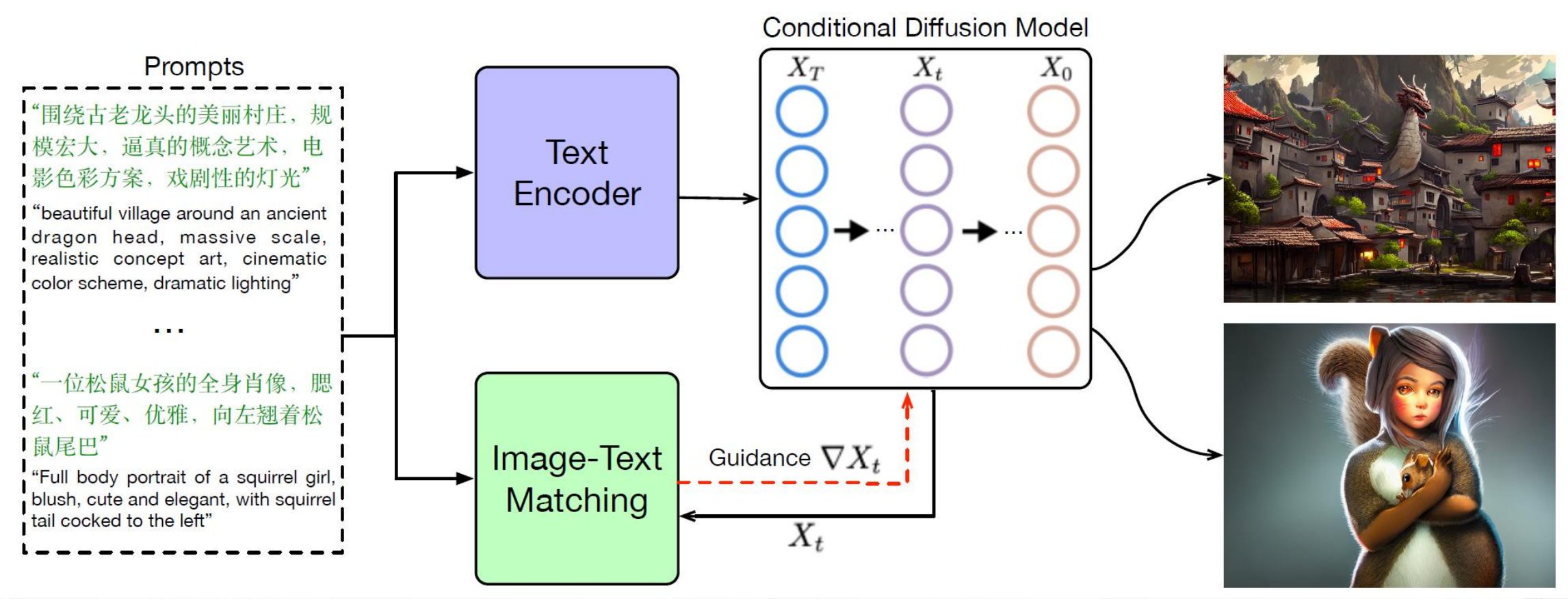
eDiffi (Nvidea)





eDiffi: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers, Arxiv 2022.

Upainting (Baidu)



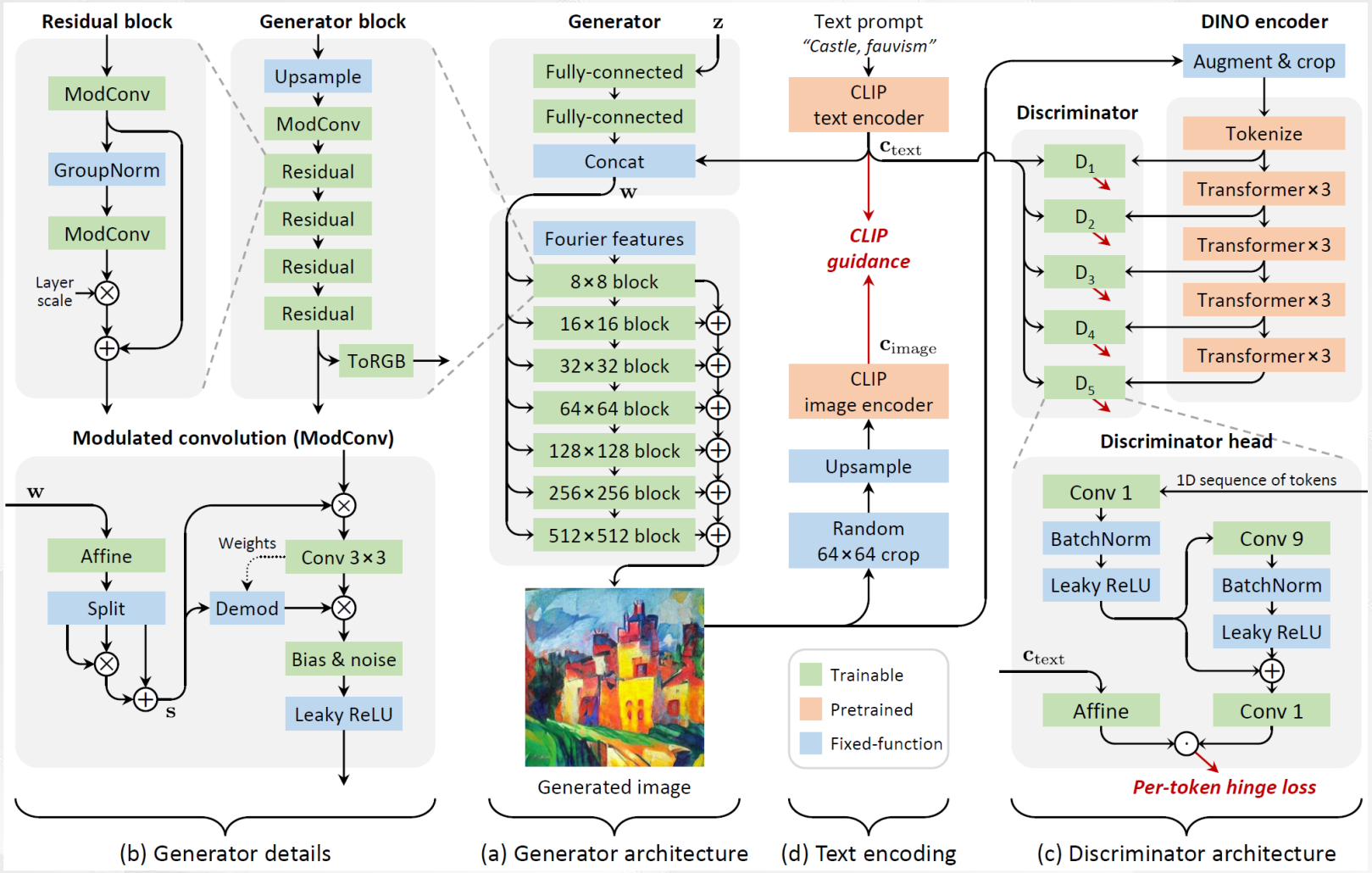
UPainting: Unified Text-to-Image Diffusion Generation with Cross-modal Guidance, Arxiv 2022.

Content

- Large Scale Pretraining Models
  - GAN
  - Auto-autoregressive models
  - Diffusion models
  - Return of GAN and AR Models
  - From T2I to 3D/Video Generation

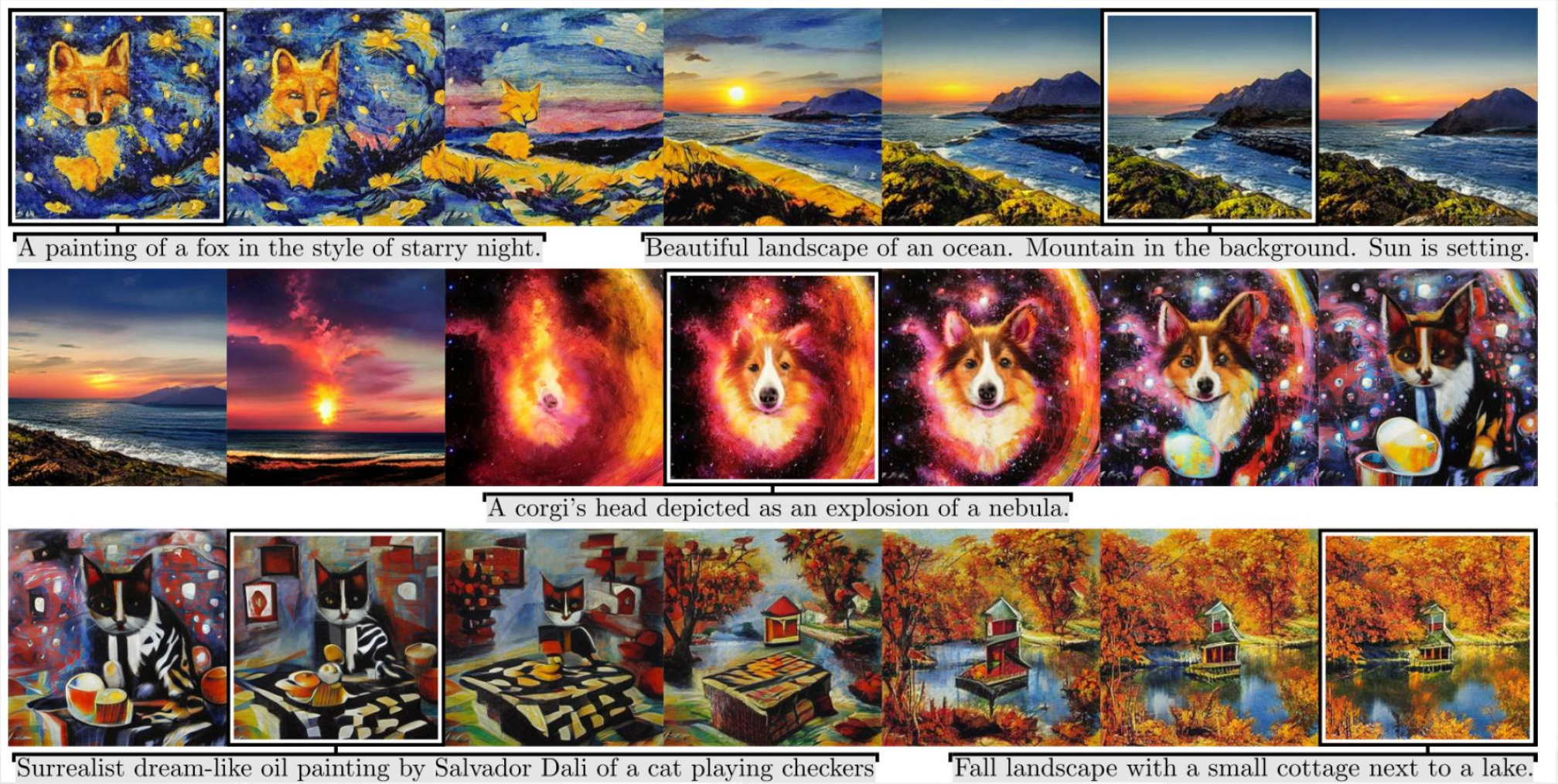
- Applications

StyleGAN-T (Arxiv 2023)



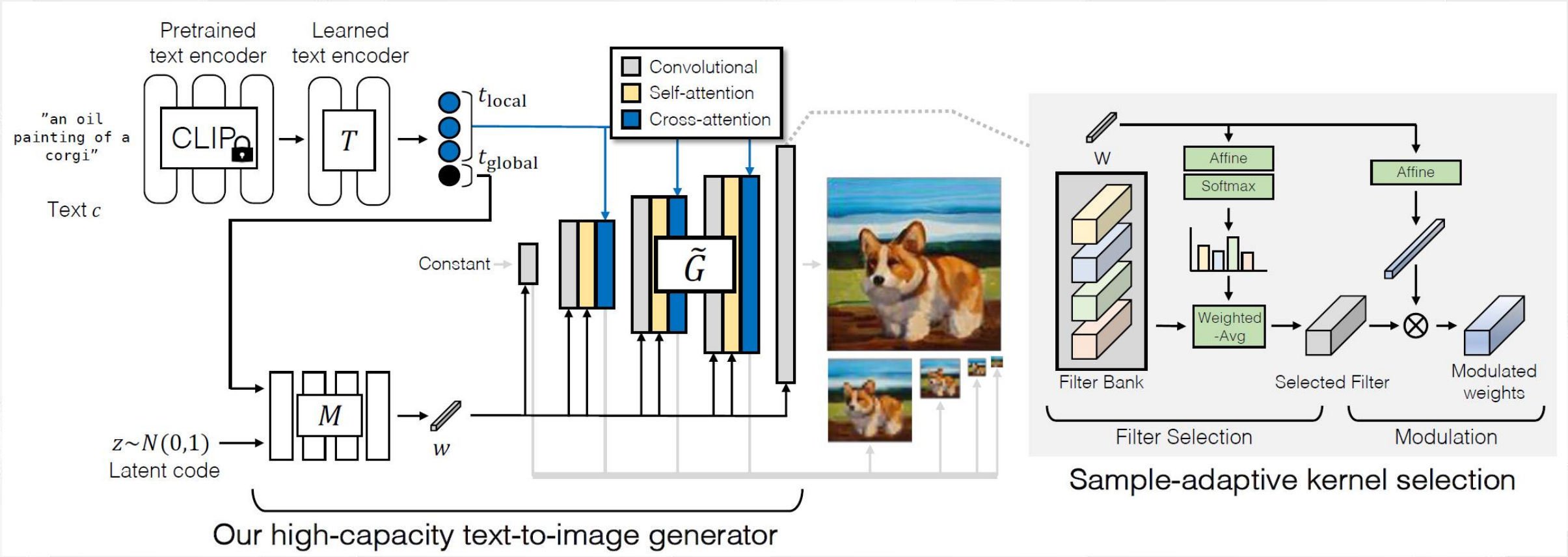
StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis, Arxiv 2023.

StyleGAN-T (Arxiv 2023)



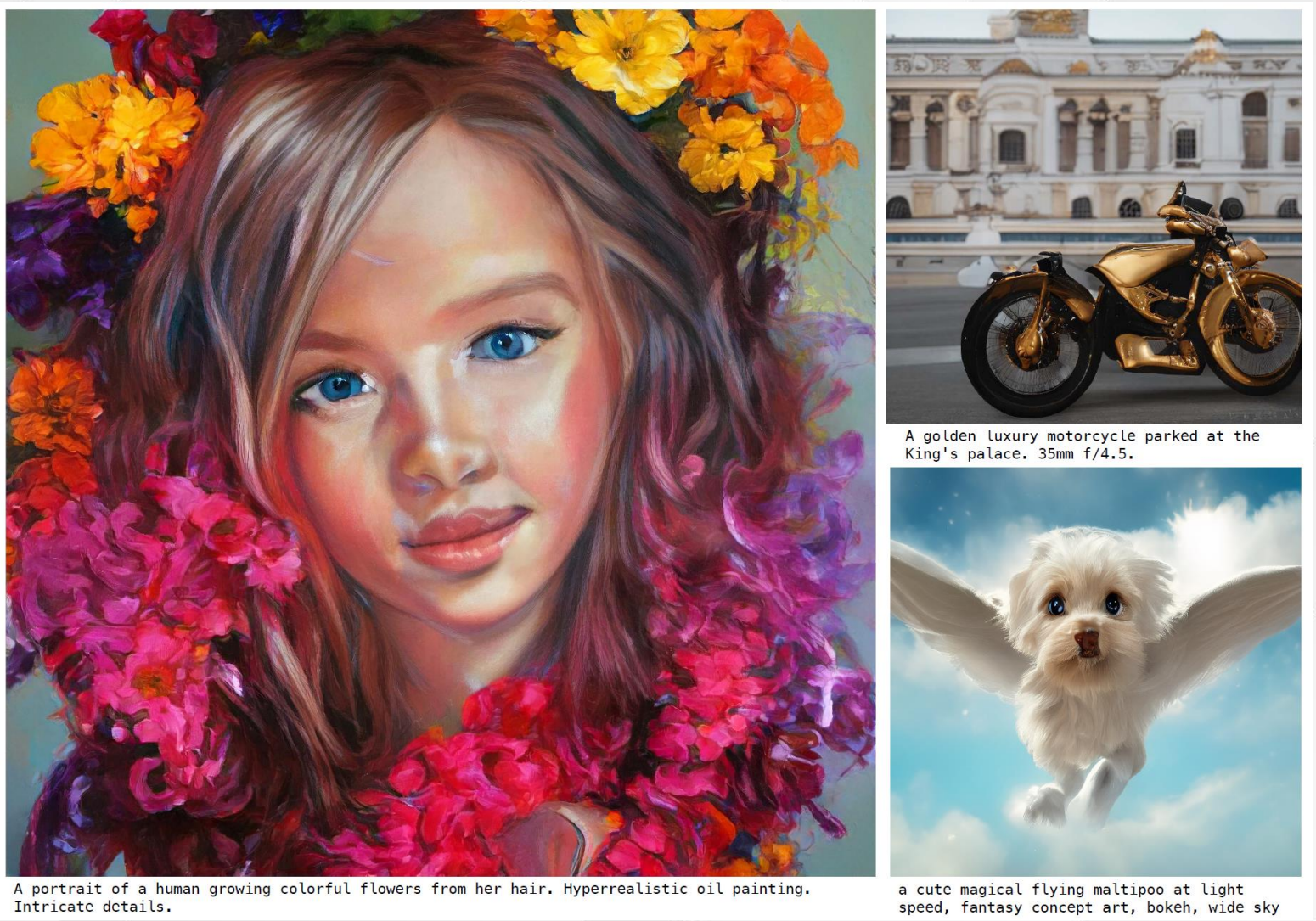
StyleGAN-T: Unlocking the Power of GANs for Fast Large-Scale Text-to-Image Synthesis, Arxiv 2023.

GigaGAN (Arxiv 2023)

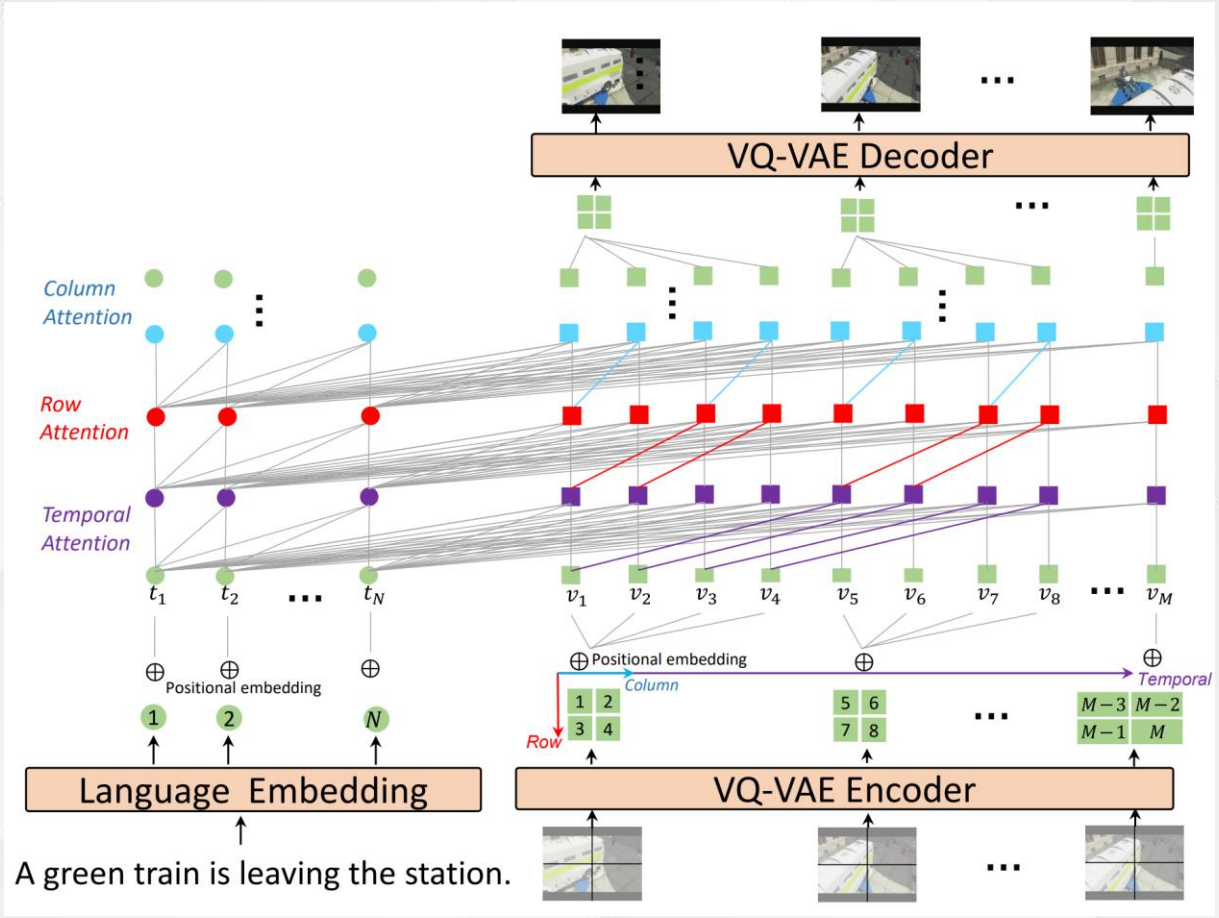


Scaling up GANs for Text-to-Image Synthesis, Arxiv 2023.

GigaGAN (Arxiv 2023)

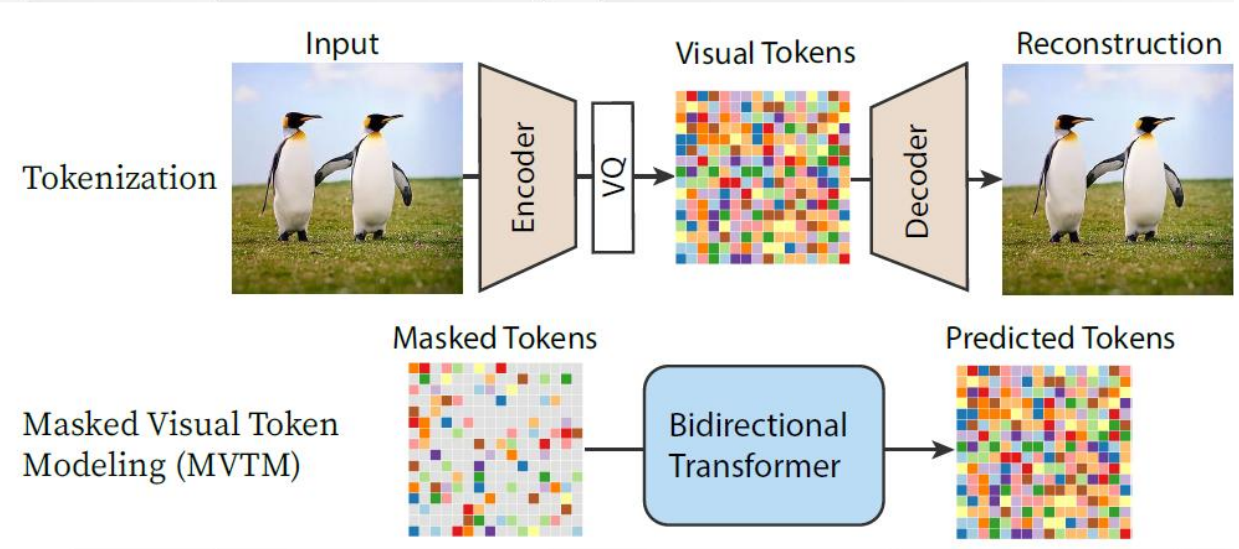


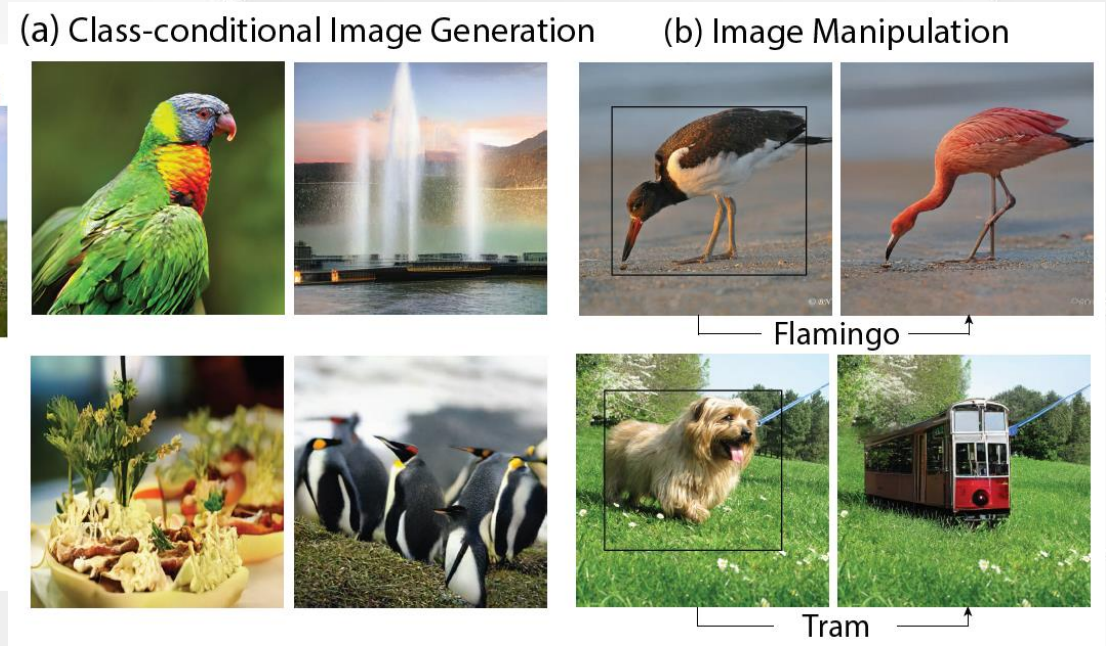
Autoregressive Model



GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions, arXiv 2021

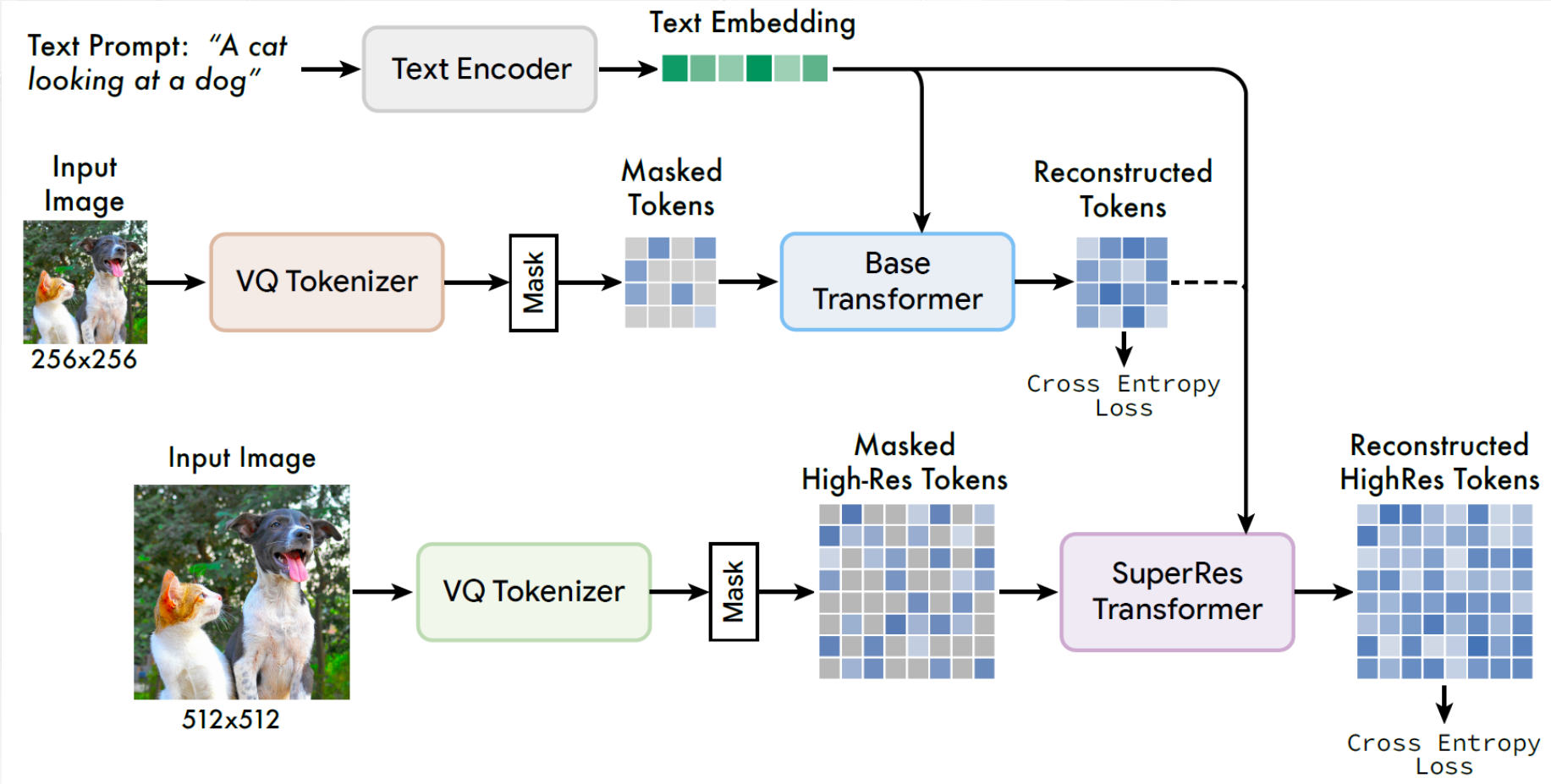
MaskGIT (CVPR 2022)





MaskGIT: Masked Generative Image Transformer, CVPR 2022.

MUSE (Arxiv 2023)



Muse: Text-To-Image Generation via Masked Generative Transformers, Arxiv 2023.

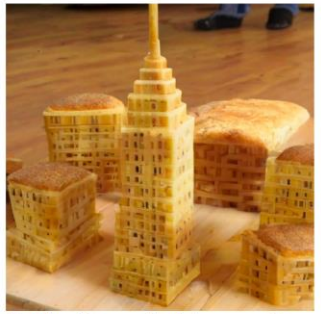
MUSE (Arxiv 2023)


A fluffy baby sloth with a knitted hat trying to figure out a laptop, close up.


A sheep in a wine glass.


A futuristic city with flying cars.

A large array of colorful cupcakes, arranged on a maple table to spell MUSE.


Manhattan skyline made of bread.


Astronauts kicking a football in front of Eiffel tower.


Two cats doing research.


3D mesh of Titanic floating on a water lily pond in the style of Monet.

Approach	Model Type	Params	FID-30K	Zero-shot FID-30K
AttnGAN (Xu et al., 2017)	GAN		35.49	-
DM-GAN (Zhu et al., 2019)	GAN		32.64	-
DF-GAN (Tao et al., 2020)	GAN		21.42	-
DM-GAN + CL (Ye et al., 2021)	GAN		20.79	-
XMC-GAN (Zhang et al., 2021)	GAN		9.33	-
LAFITE (Zhou et al., 2021)	GAN		8.12	-
Make-A-Scene (Gafni et al., 2022)	Autoregressive		7.55	-
DALL-E (Ramesh et al., 2021)	Autoregressive		-	17.89
LAFITE (Zhou et al., 2021)	GAN		-	26.94
LDM (Rombach et al., 2022)	Diffusion		-	12.63
GLIDE (Nichol et al., 2021)	Diffusion		-	12.24
DALL-E 2 (Ramesh et al., 2022)	Diffusion		-	10.39
Imagen-3.4B (Saharia et al., 2022)	Diffusion		-	7.27
Parti-3B (Yu et al., 2022)	Autoregressive		-	8.10
Parti-20B (Yu et al., 2022)	Autoregressive		3.22	7.23
Muse-3B	Non-Autoregressive		-	7.88

Muse: Text-To-Image Generation via Masked Generative Transformers, Arxiv 2023.

Content

- Large Scale Pretraining Models
  - GAN
  - Auto-autoregressive models
  - Diffusion models
  - Return of GAN and AR Models
  - From T2I to 3D/Video Generation
- Applications

Generative Process

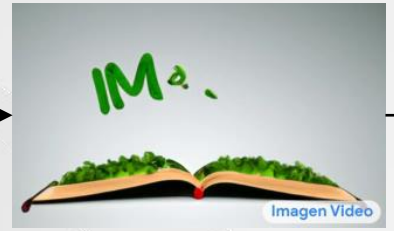
Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

Base 16×24×48 3fps

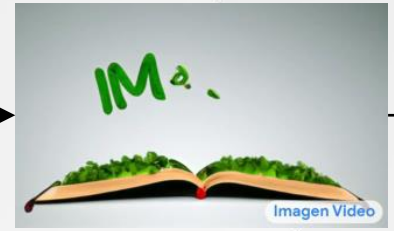
TSR 32×24×48 6fps

SSR 32×48×96 6fps

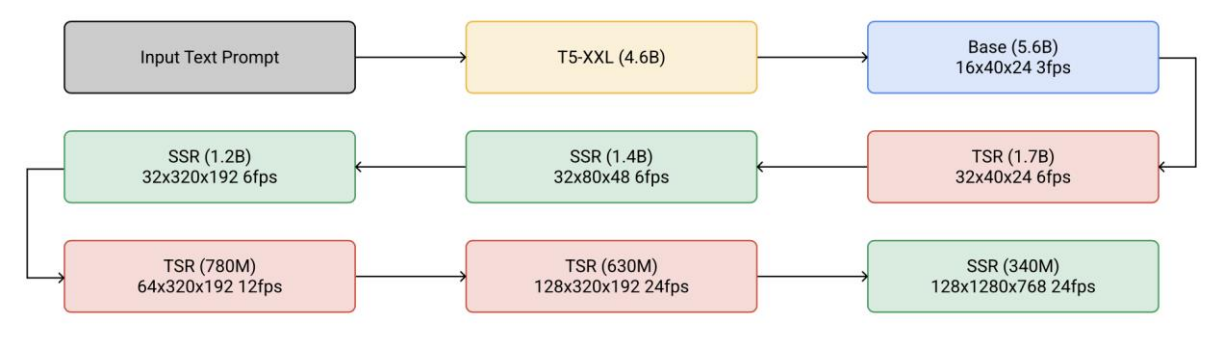
TSR 64×192×320 12fps



TSR 128×192×320 24fps



TSR 128×1280×768 24fps



SSR 128×1280×768 24fps

Imagen Video: High-Definition Video Generation with Diffusion Models, Arxiv 2022.

Imagen Video Results

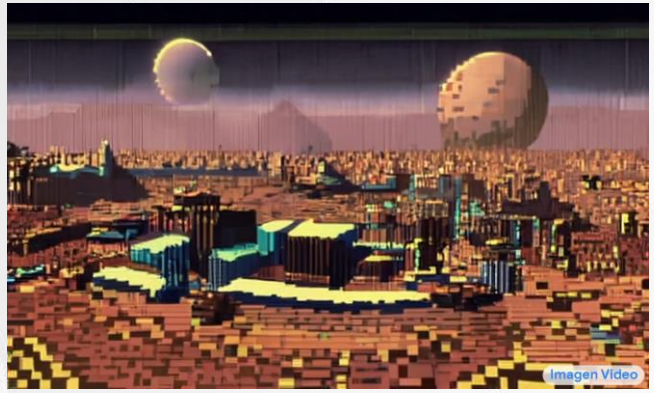
A goldendoodle playing in a park by a lake.

A giraffe underneath a microwave.

Tiny plant sprout coming out of the ground.

Flying through an intense battle between pirate ships in a stormy ocean.

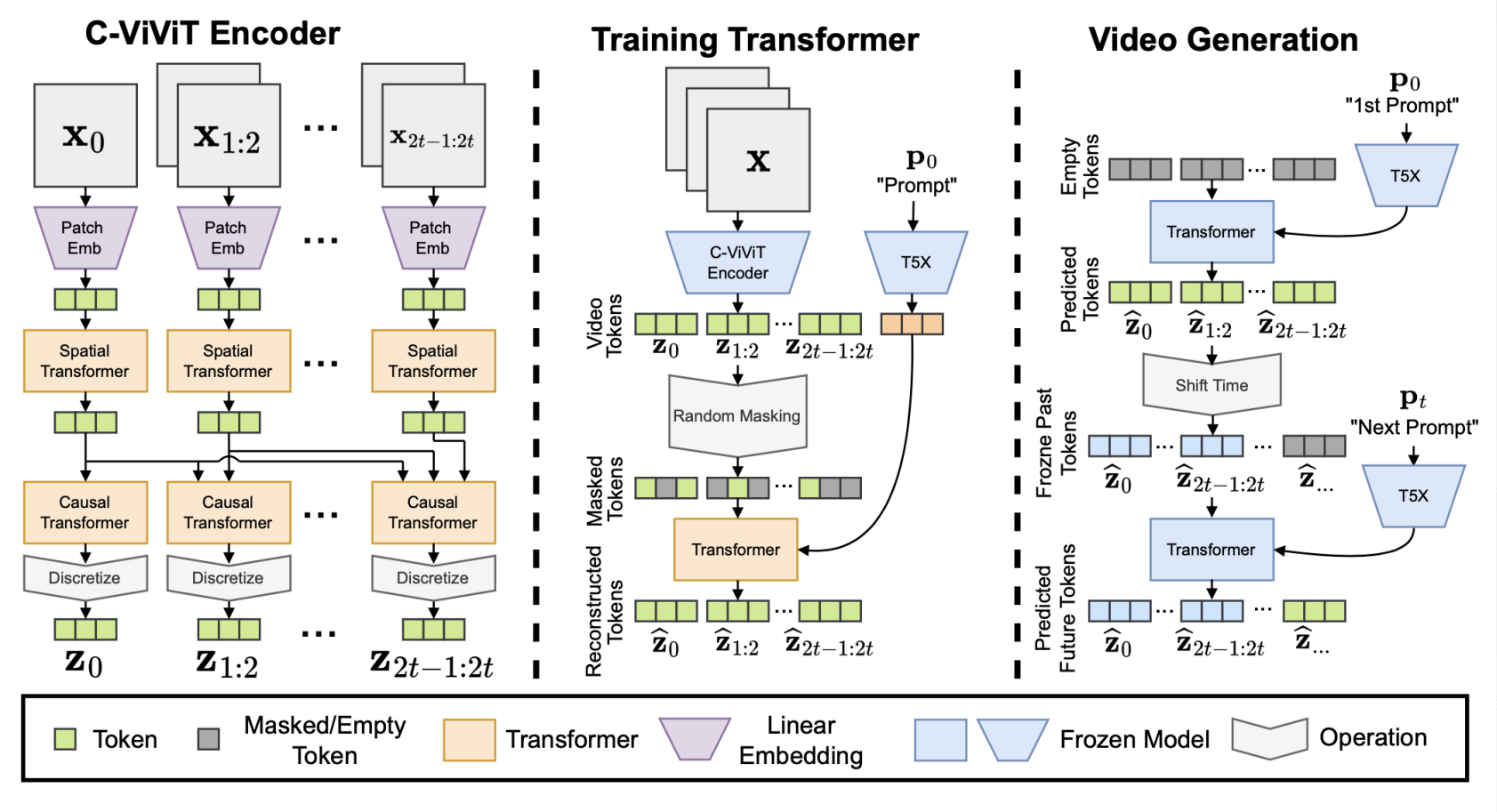
Studio shot of minimal kinetic sculpture made from thin wire shaped like a bird on white background.



Incredibly detailed science fiction scene set on an alien planet, view of a marketplace. Pixel art.

Imagen Video: High-Definition Video Generation with Diffusion Models, Arxiv 2022.

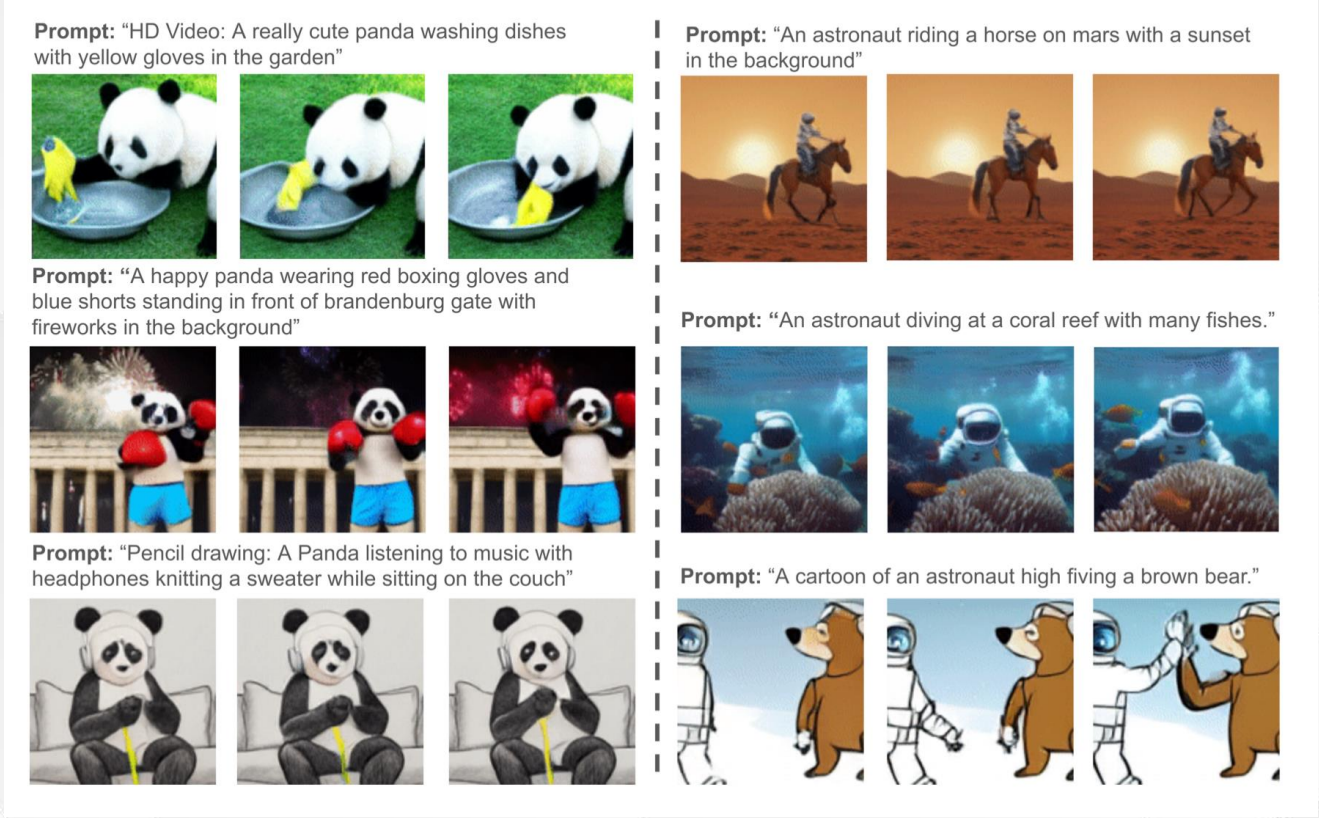
PHENAKI



PHENAKI: Variable Length Video Generation from Open Domain Textual Descriptions, Arxiv 2022.

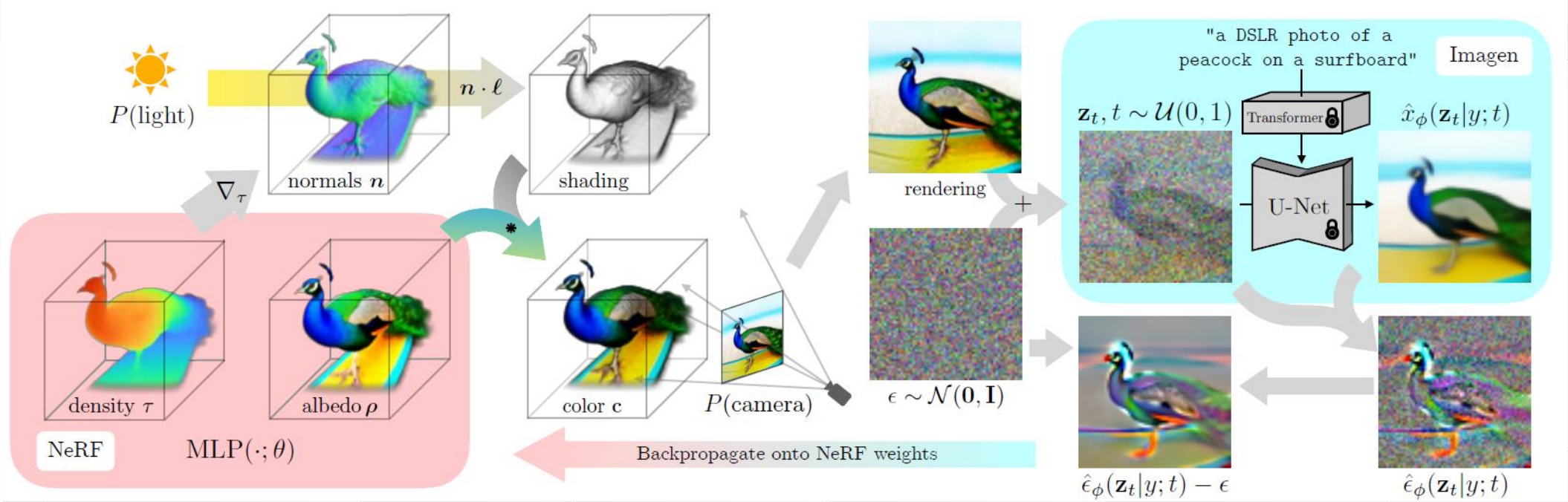
Text-to-video Results

Method	FID Image ↓	FID Video ↓
T2V [25]	82.13	14.65
SC [5]	33.51	7.34
TFGAN [5]	31.76	7.19
NUWA	28.46	7.05
Phenaki [0-Shot]	37.74	3.84



PHENAKI: Variable Length Video Generation from Open Domain Textual Descriptions, Arxiv 2022.

DreamFusion (ICLR 2023)



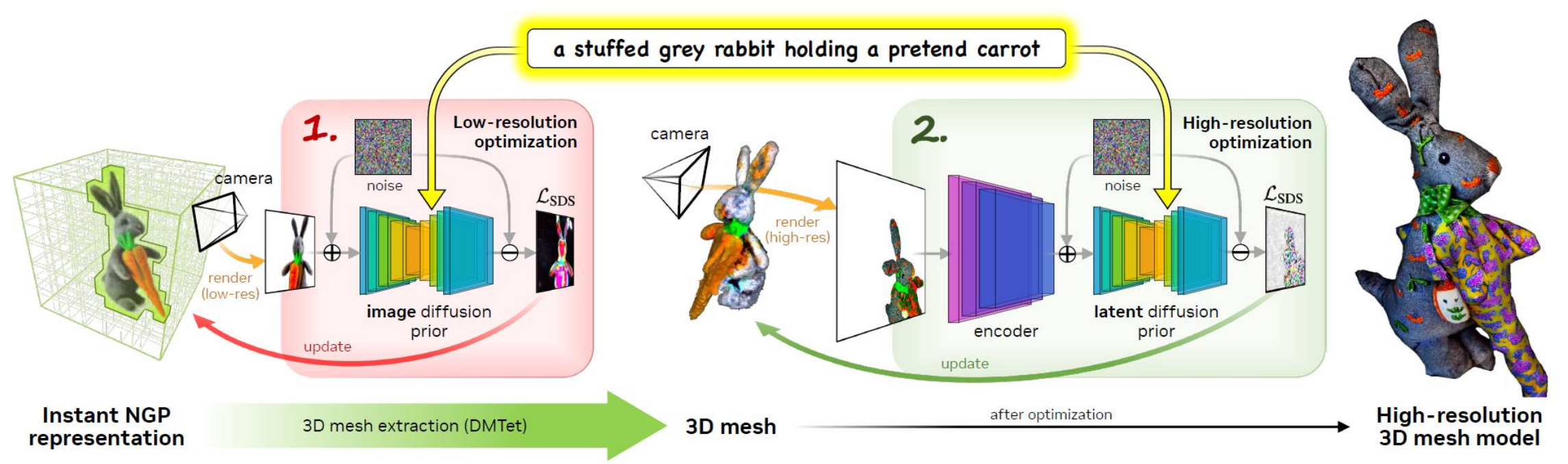
$$\nabla_\theta \mathcal{L}_{\text{SDS}}(\phi, \mathbf{x} = g(\theta)) \triangleq \mathbb{E}_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_\phi(\mathbf{z}_t; y, t) - \epsilon\right) \frac{\partial \mathbf{x}}{\partial \theta} \right]$$

Dreamfusion: Text-to-3d using 2d diffusion, ICLR 2023.

DreamFusion (ICLR 2023)

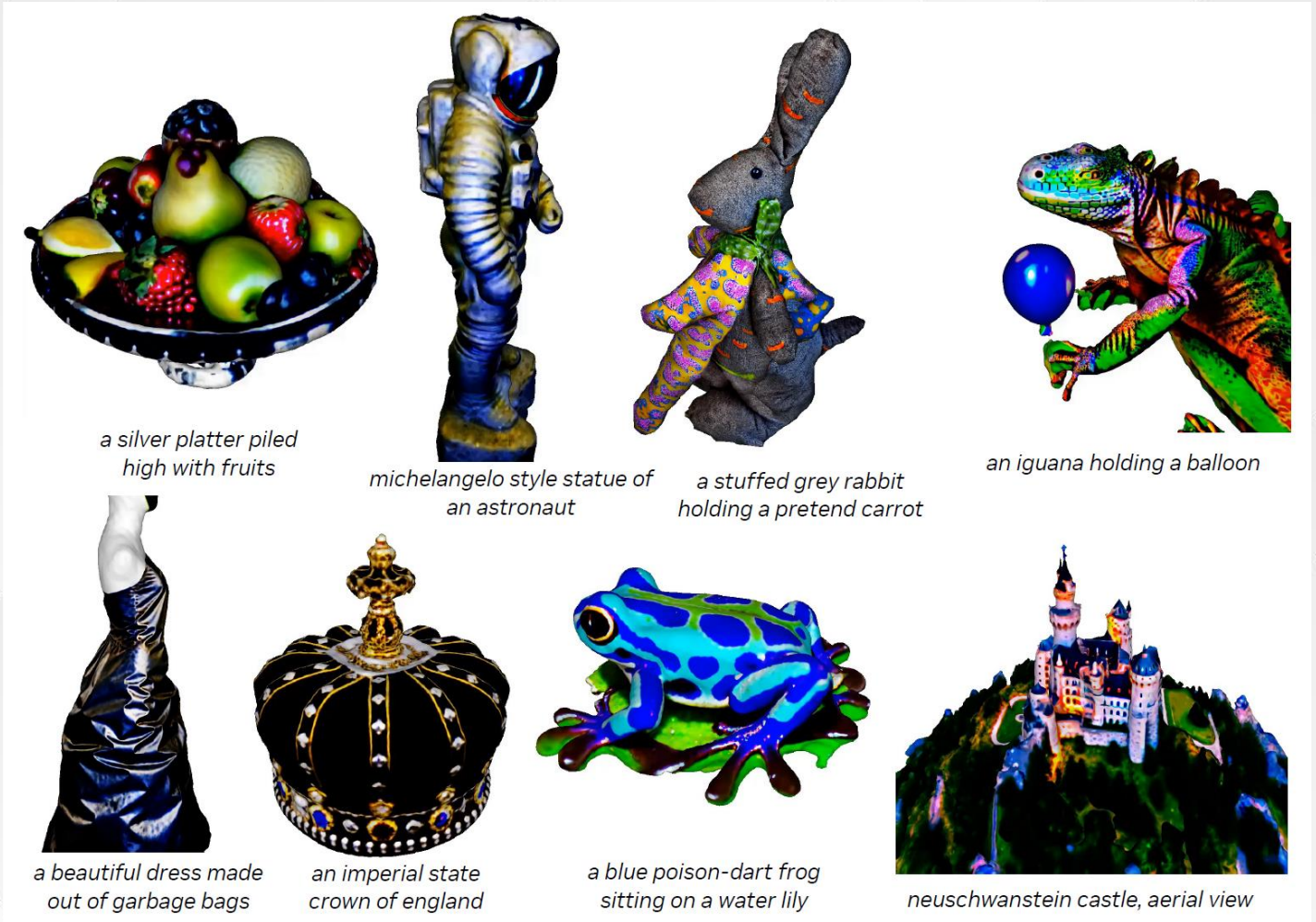
Dreamfusion: Text-to-3d using 2d diffusion, ICLR 2023.

Magic3D



Magic3D: High-Resolution Text-to-3D Content Creation, Arxiv 2022

Magic3D



Magic3D: High-Resolution Text-to-3D Content Creation, Arxiv 2022

Content

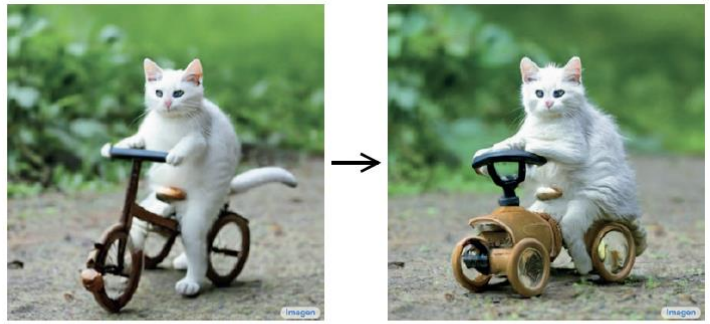
- Large Scale Pretraining Models
- Applications
  - Image Editing (Prompt-to-Prompt, Imagic, ControlNet)
  - Custom Generation
  - Latent Space

Prompt-to-Prompt Image Editing with Cross Attention Control

Amir Hertz*[1,2], Ron Mokady*[1,2], Jay Tenenbaum[1], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or*[1,2]

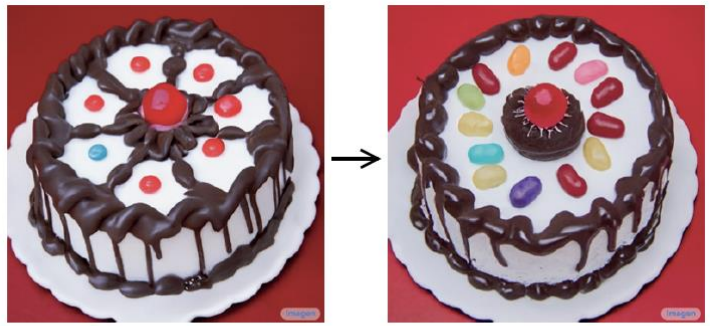
[1] *Google Research*
[2] *The Blavatnik School of Computer Science, Tel Aviv University*

"The boulevards are crowded today."



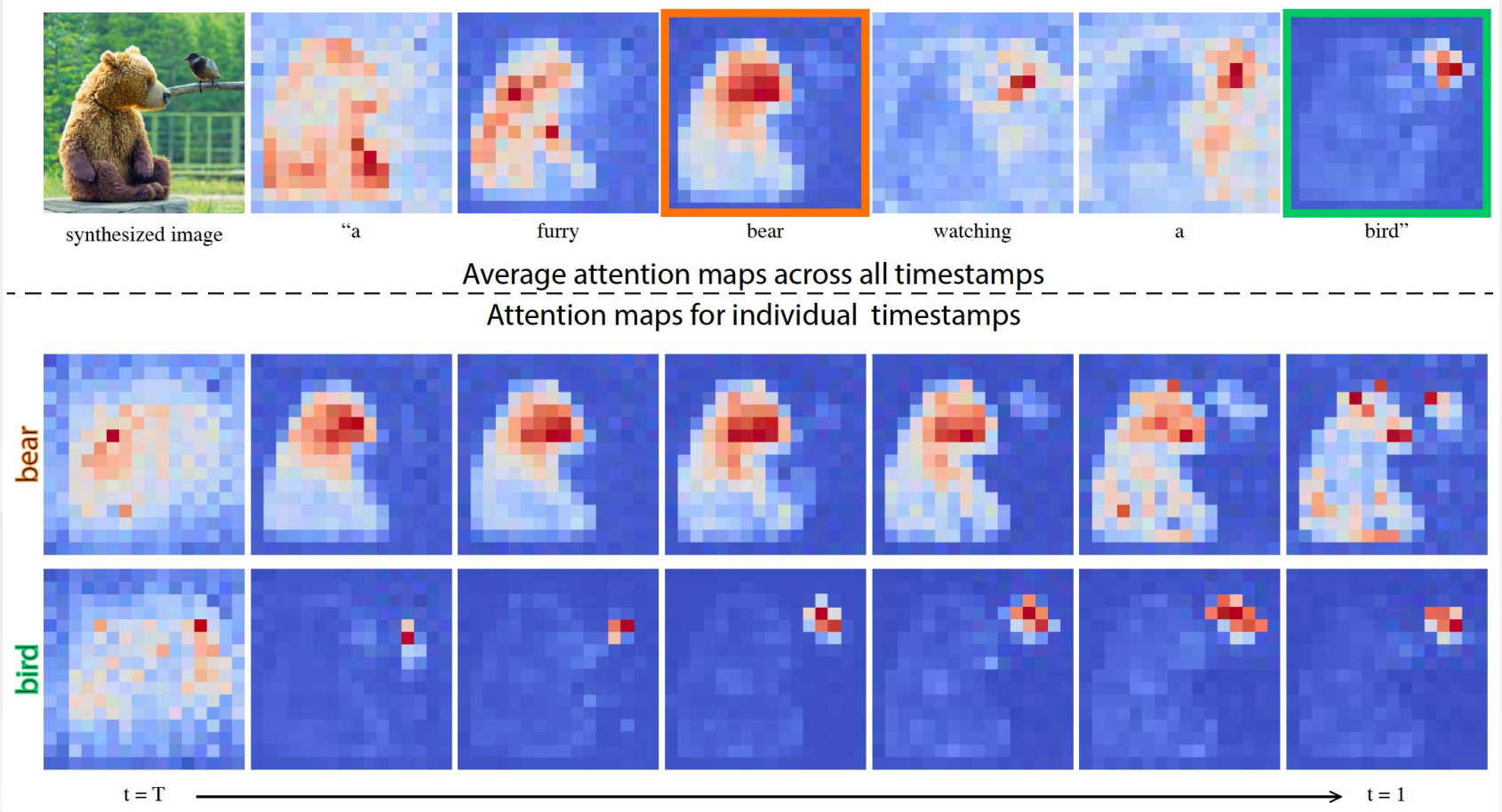
"Photo of a cat riding on a bicycle."

"Children drawing of a castle next to a river."



"a cake with decorations."

Prompt-to-Prompt: Motivation



Prompt-to-Prompt: Method

Algorithm 1: Prompt-to-Prompt image editing

Input: A source prompt $\mathcal{P}$, a target prompt $\mathcal{P}^*$, and a random seed s.
Output: A source image x_{src} and an edited image x_{dst}.
$z_T \sim N(0, I)$ a unit Gaussian random variable with random seed s;
$z_T^* \leftarrow z_T$;
for $t = T, T-1, \ldots, 1$ **do**
  $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t, s)$;
  $M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s)$;
  $\widehat{M}_t \leftarrow Edit(M_t, M_t^*, t)$;
  $z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s_t)\{M \leftarrow \widehat{M}_t\}$;
end
Return (z_0, z_0^*)

1. Word Swap

$$Edit(M_t, M_t^*, t) := \begin{cases} M_t^* & \text{if } t < \tau \\ M_t & \text{otherwise.} \end{cases}$$

2. Add a new phase

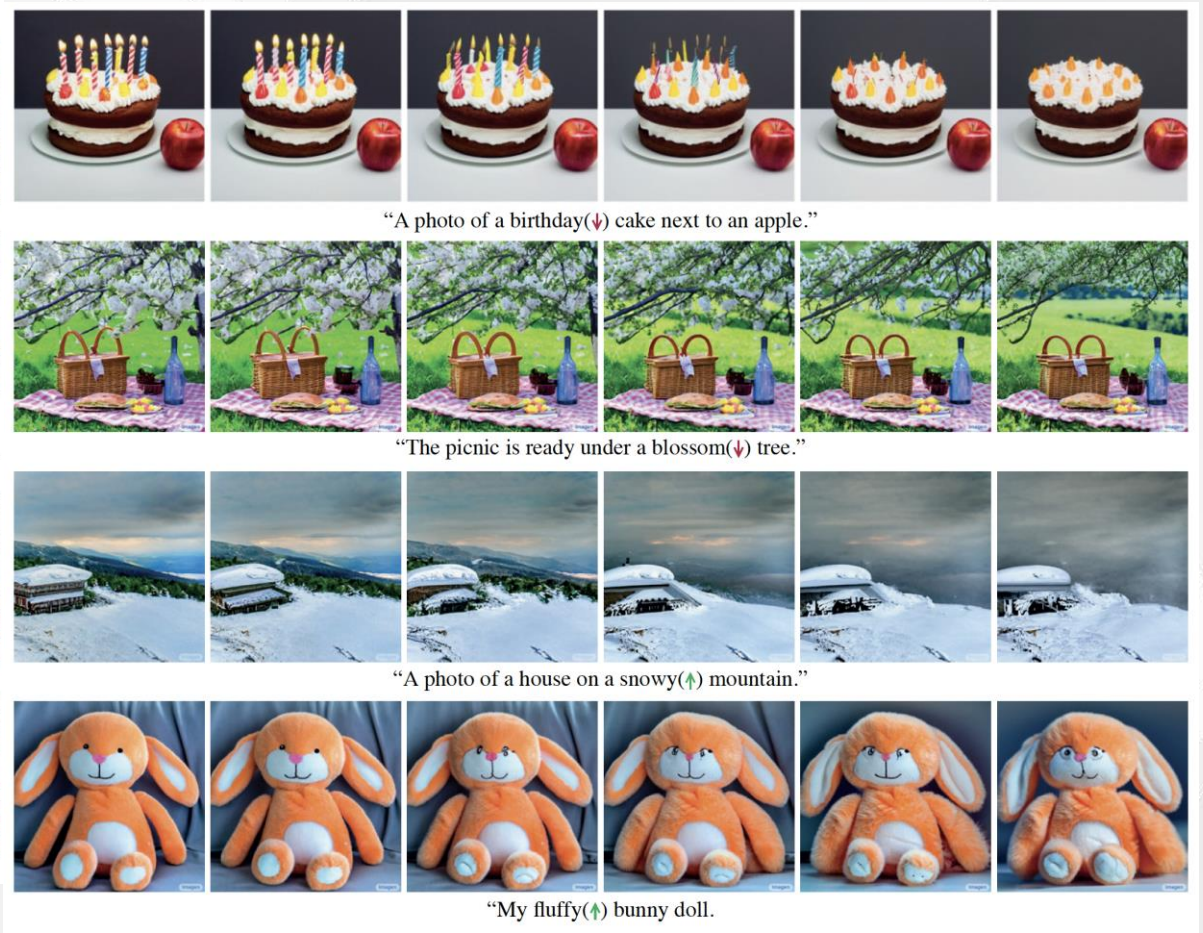
$$(Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} (M_t^*)_{i,j} & \text{if } A(j) = None \\ (M_t)_{i,A(j)} & \text{otherwise.} \end{cases}$$

3. Attention Re-weighting

$$(Edit\,(M_t, M_t^*, t))_{i,j} := \begin{cases} c \cdot (M_t)_{i,j} & \text{if } j = j^* \\ (M_t)_{i,j} & \text{otherwise.} \end{cases}$$

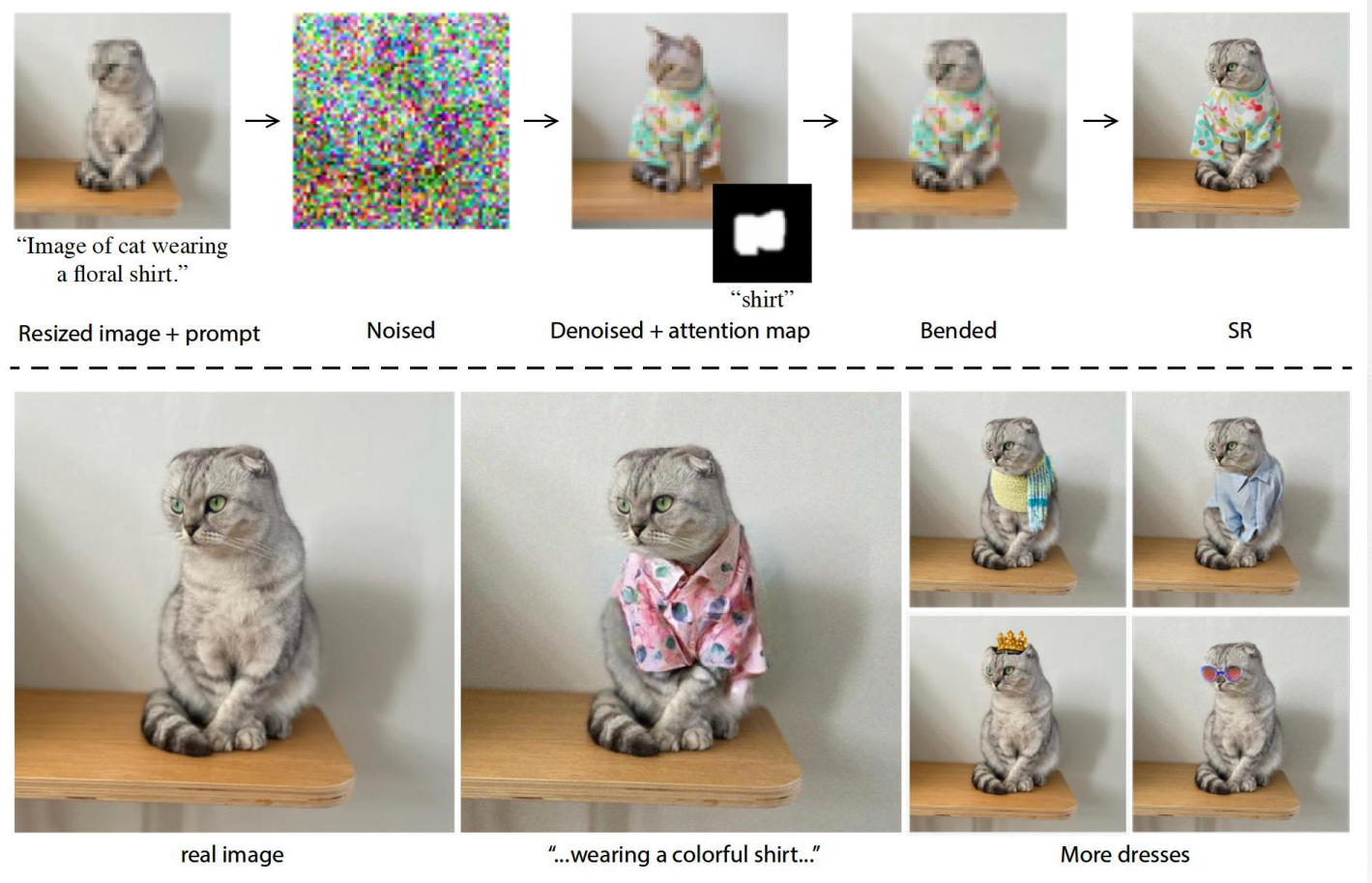
Prompt-to-Prompt: Results

Add new phase



Attention Re-weighting

Prompt-to-Prompt: Results

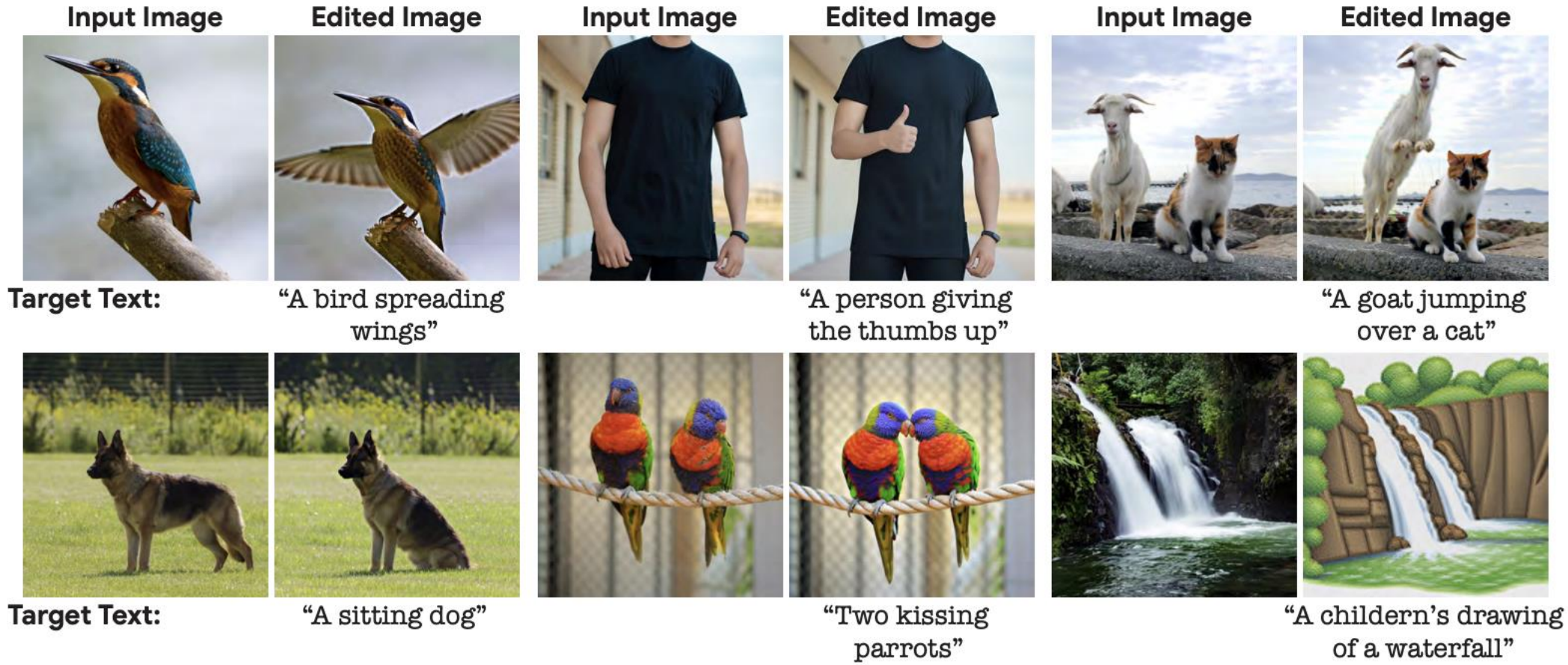


Mask-guided Editing

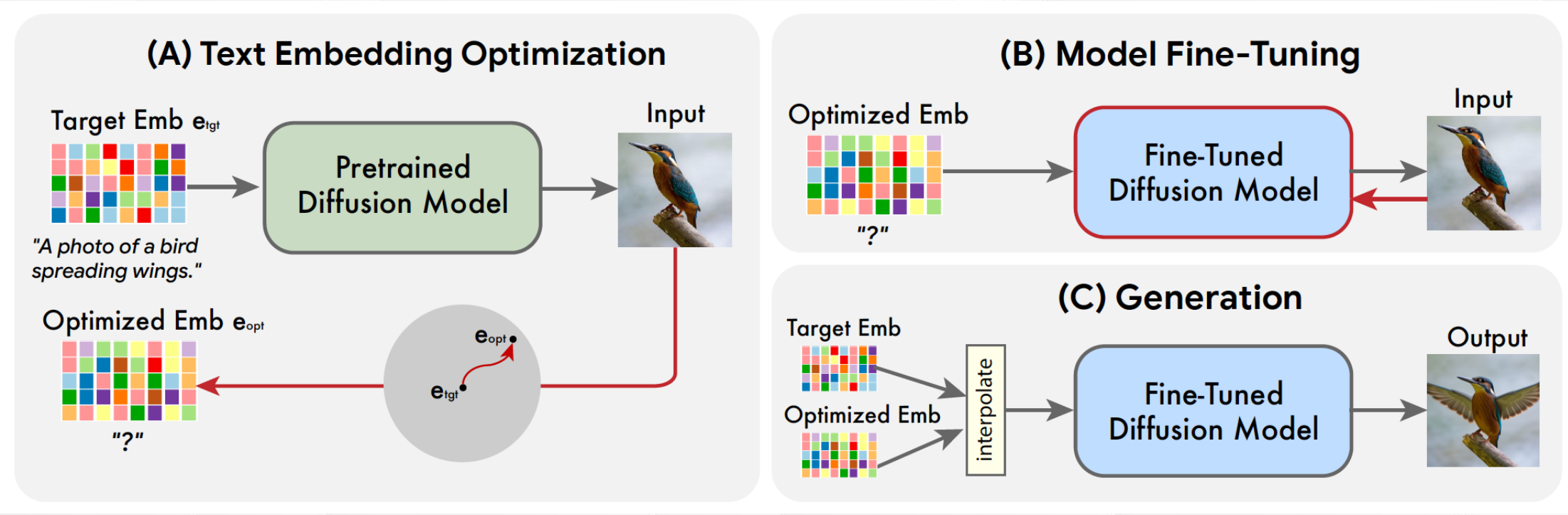
Imagic: Text-Based Real Image Editing with Diffusion Models

Bahjat Kawar* [1,2] Shiran Zada* [1] Oran Lang[1] Omer Tov[1]
Huiwen Chang[1] Tali Dekel[1,3] Inbar Mosseri[1] Michal Irani[1,3]

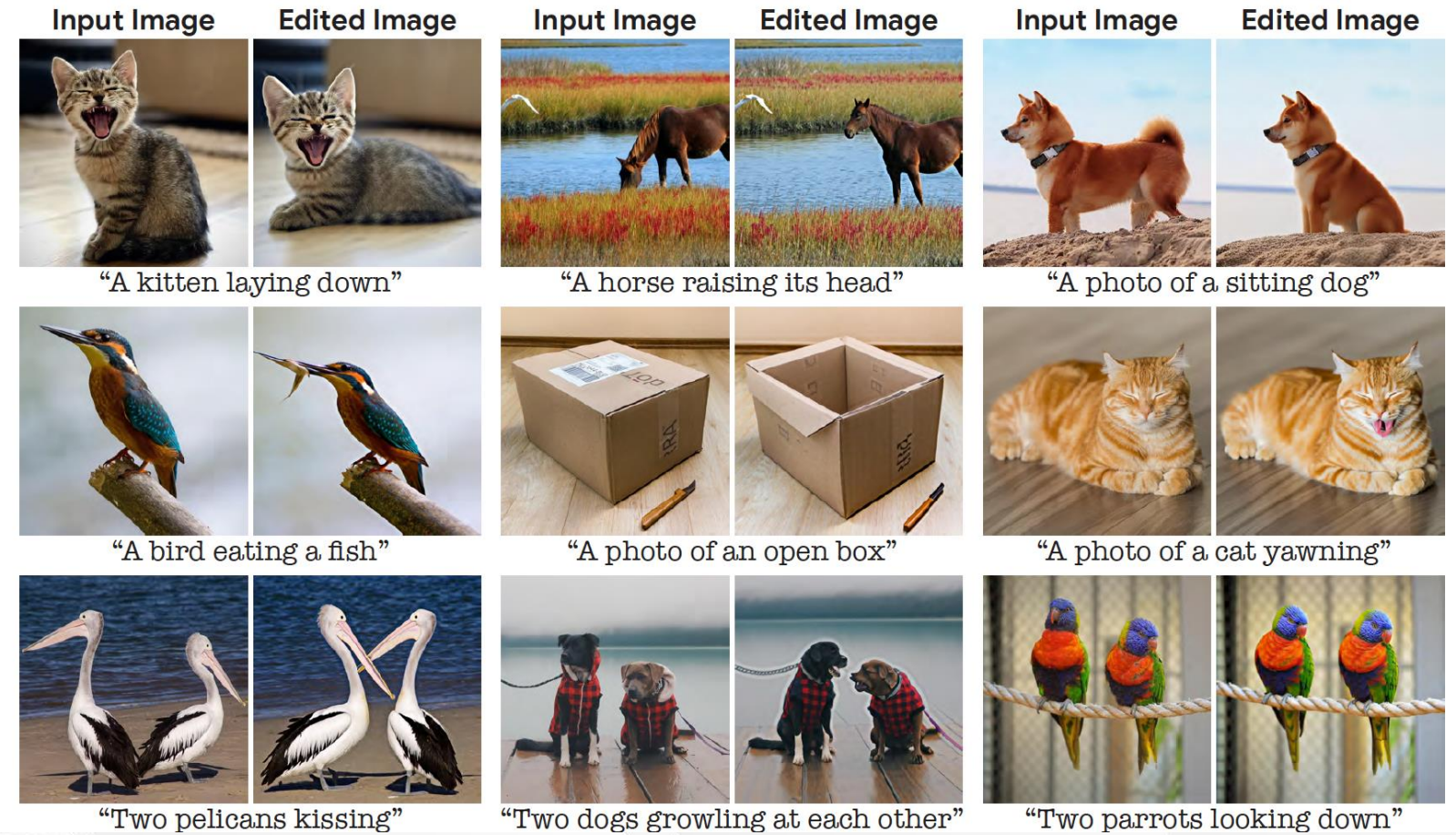
[1]Google Research [2]Technion [3]Weizmann Institute of Science



Imagic: Method



Imagic: Results



Imagic: Results

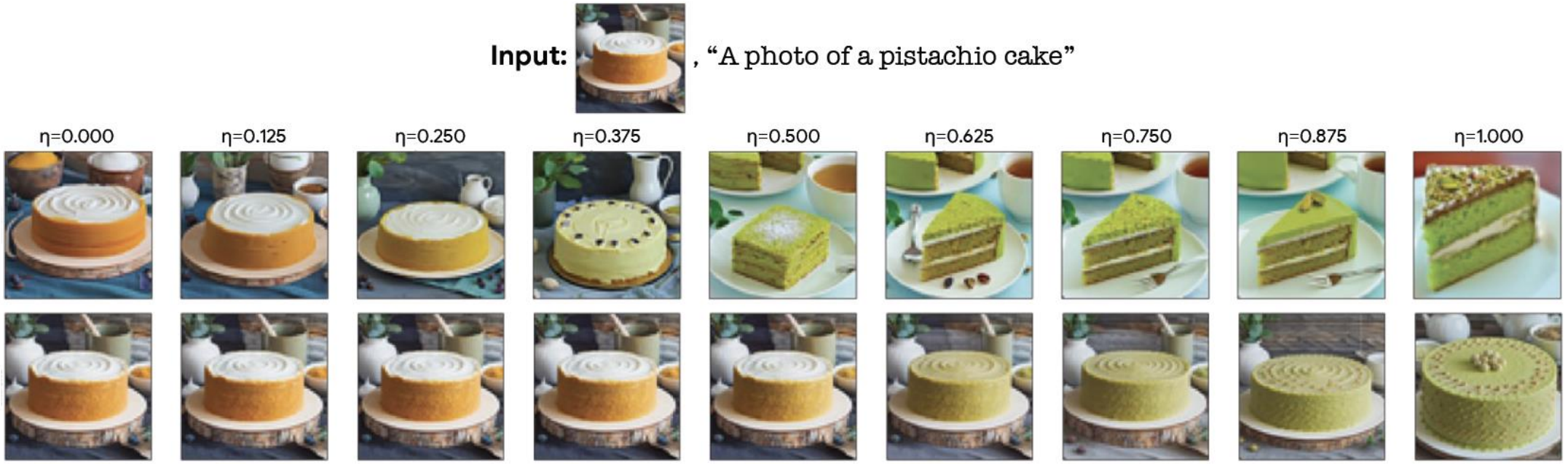
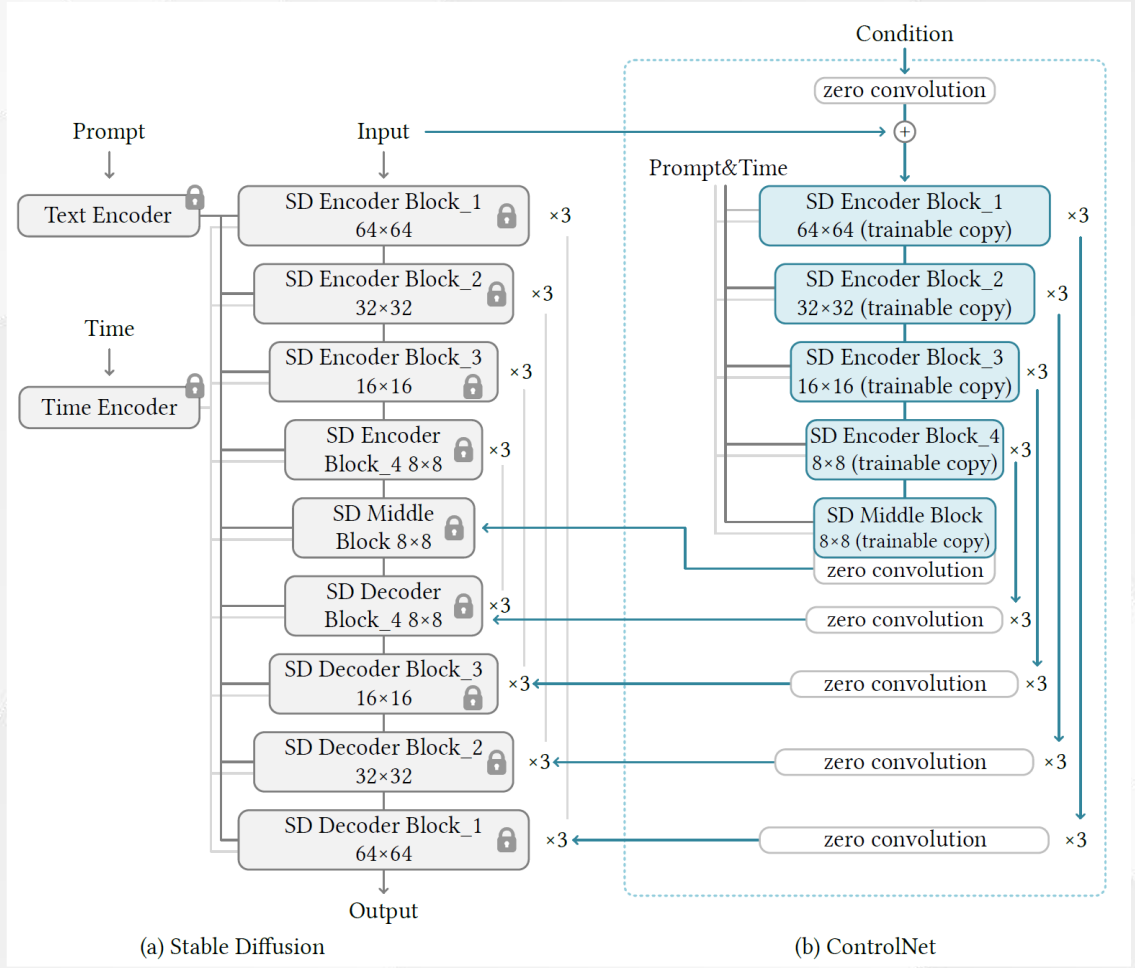


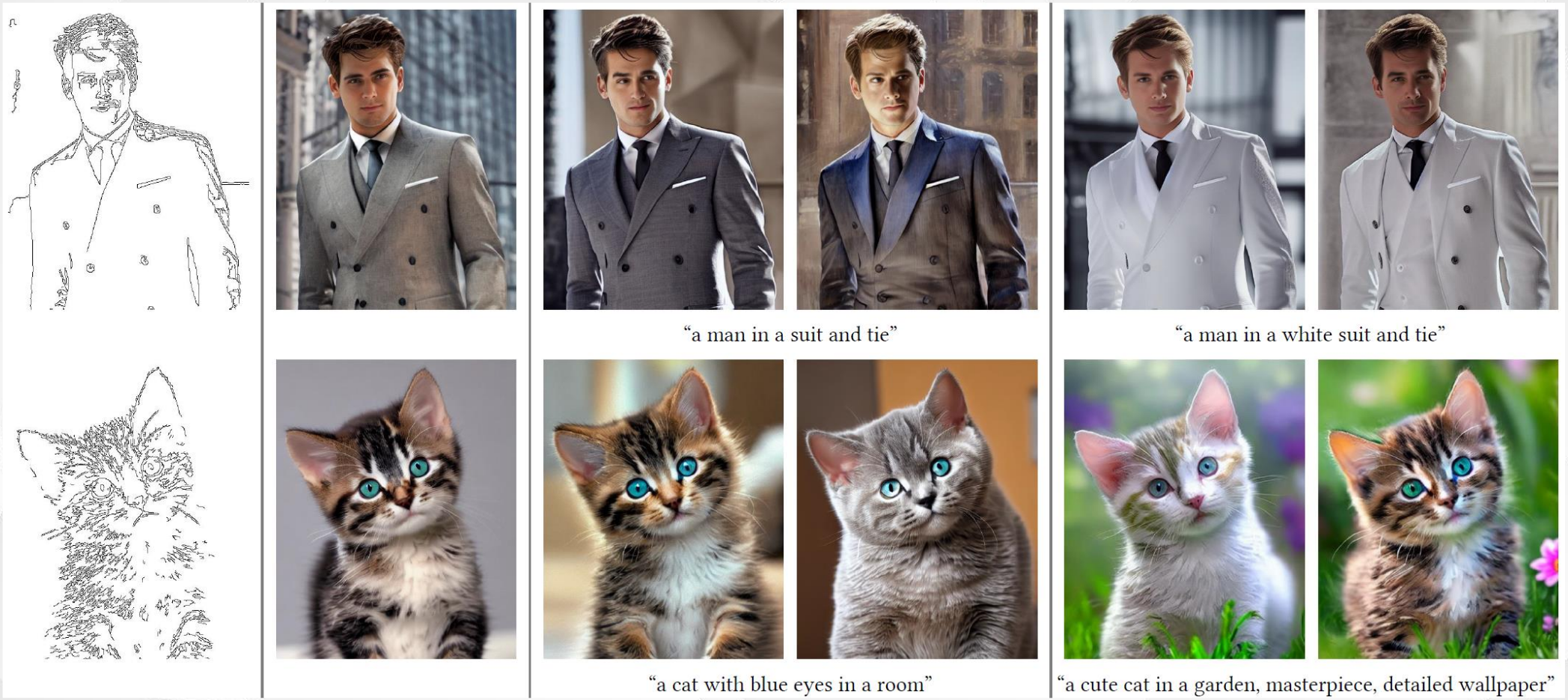
Figure 5. **Embedding interpolation.** Varying η with the same seed, using the pre-trained (top) and fine-tuned (bottom) models.

ControlNet

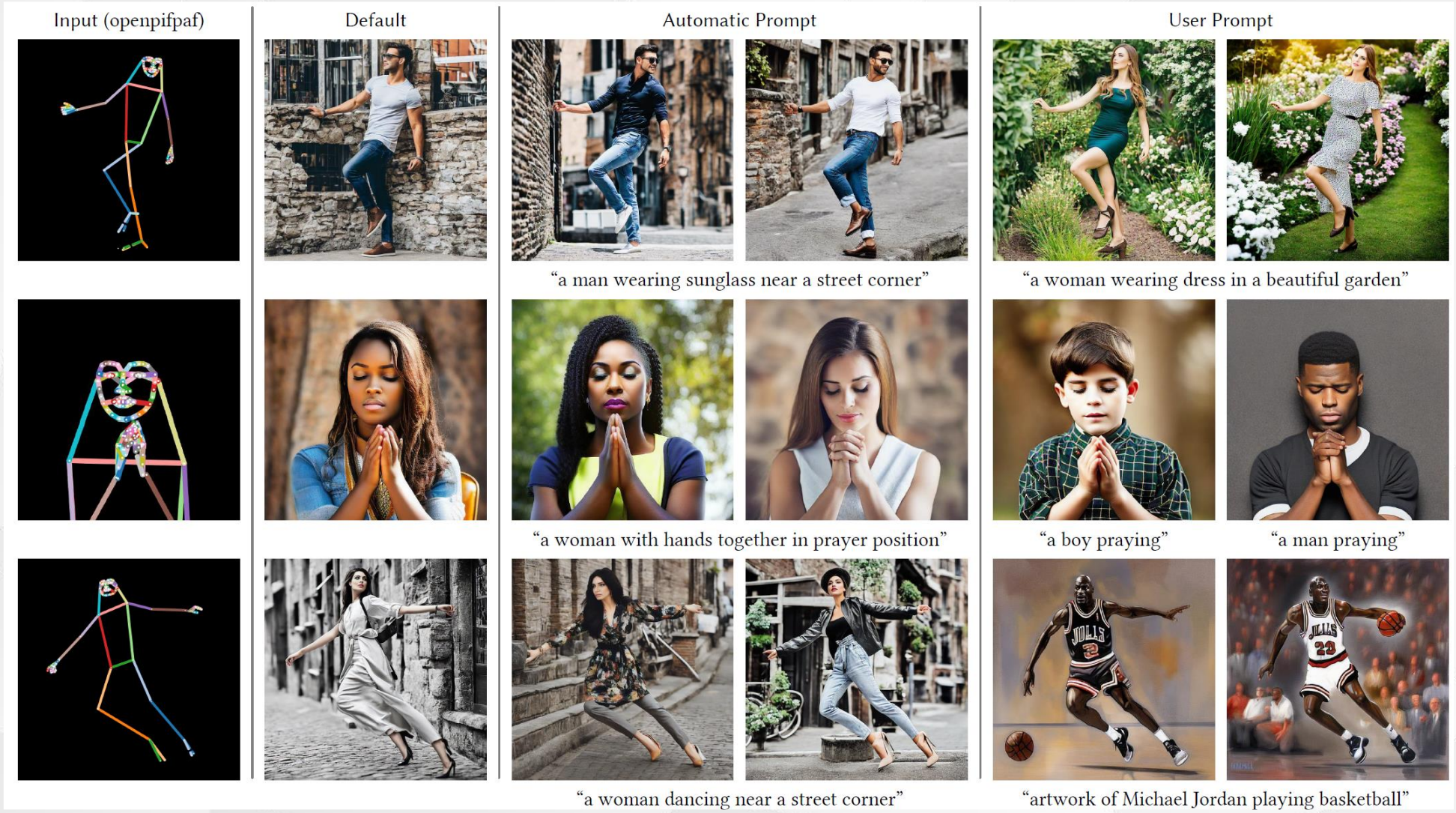


Adding Conditional Control to Text-to-Image Diffusion Models, Arxiv 2023.

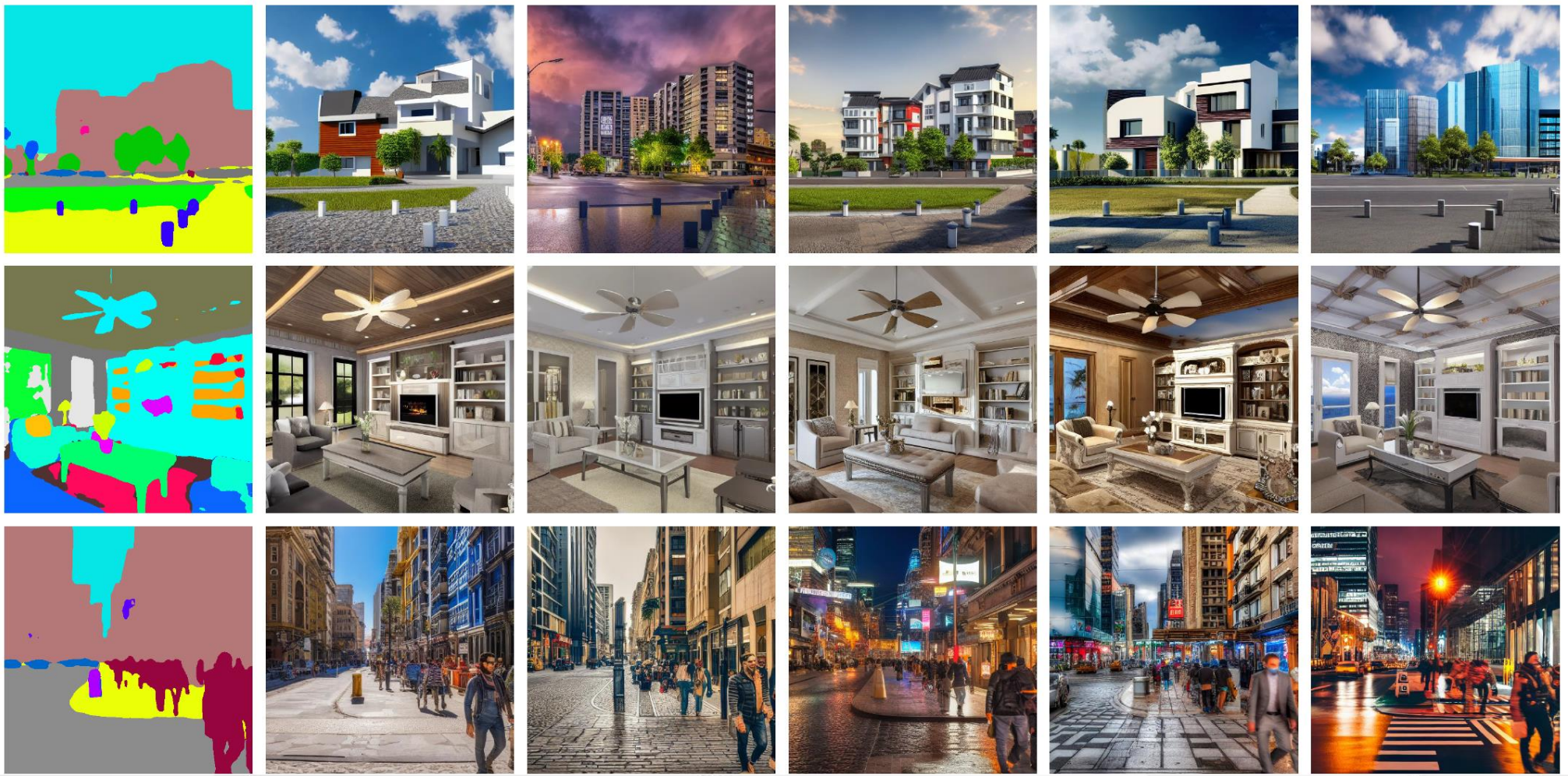
ControlNet


"a man in a suit and tie"
"a man in a white suit and tie"
"a cat with blue eyes in a room"
"a cute cat in a garden, masterpiece, detailed wallpaper"

ControlNet


Input (openpifpaf)
Default
Automatic Prompt
User Prompt
"a man wearing sunglass near a street corner"
"a woman wearing dress in a beautiful garden"
"a woman with hands together in prayer position"
"a boy praying"
"a man praying"
"a woman dancing near a street corner"
"artwork of Michael Jordan playing basketball"

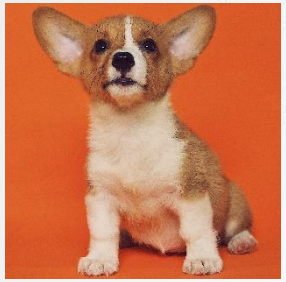
ControlNet



Content

- Large Scale Pretraining Models
- Applications
  - Image Editing
  - Custom Generation (Textual Inversion, Dreambooth, Custom Diffusion, Elite)
  - Latent Space

Customized Generation










Input —*invert*→ S*

S*

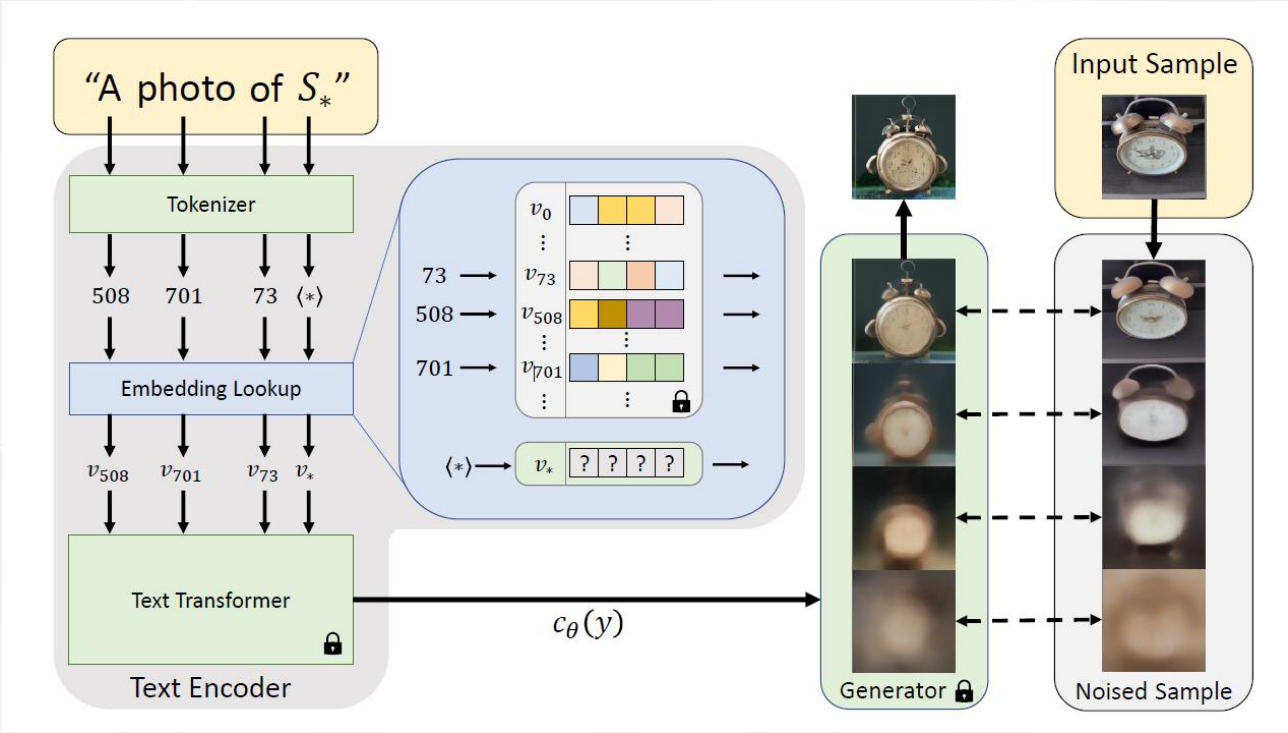
Wearing Sunglasses

In a bucket

Swimming

Textual Inversion

- Encode a visual concept as a word embedding [S*].

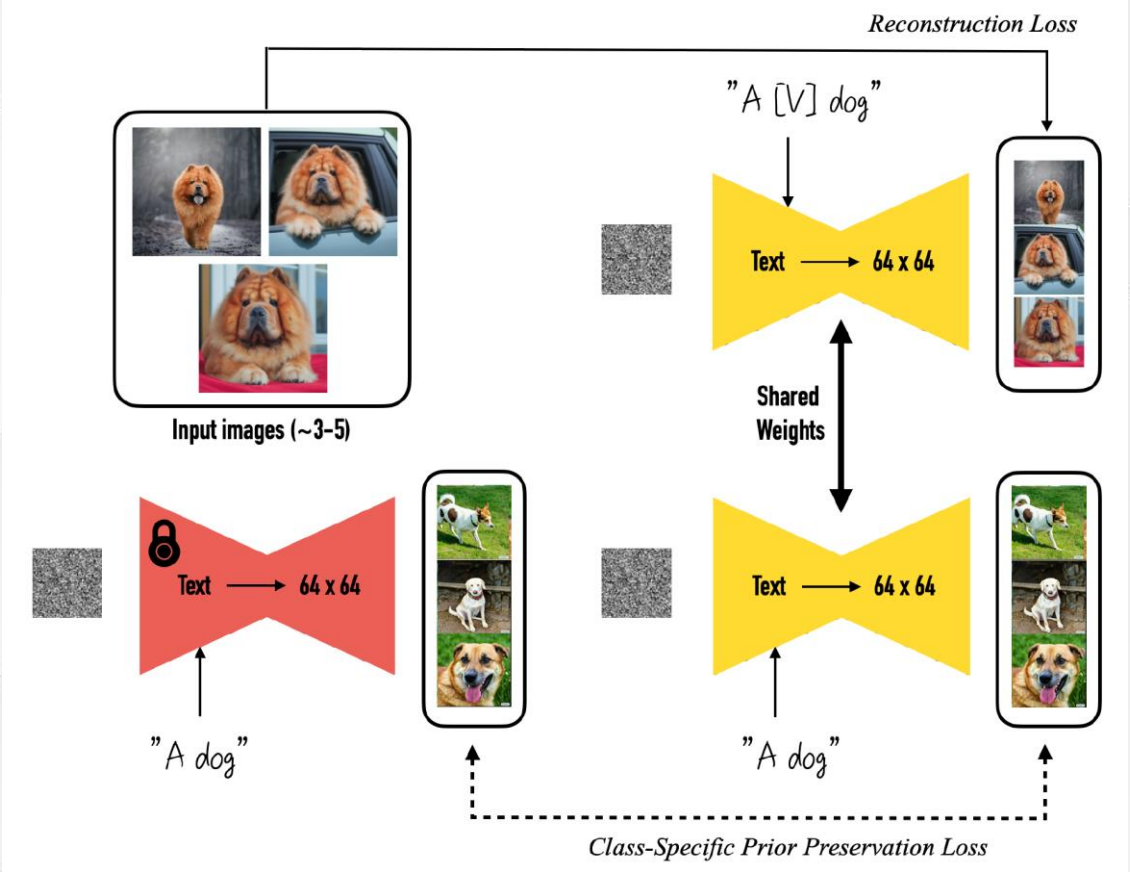


$$v_* = \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right]$$

Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." *arXiv preprint arXiv:2208.01618* (2022).

DreamBooth

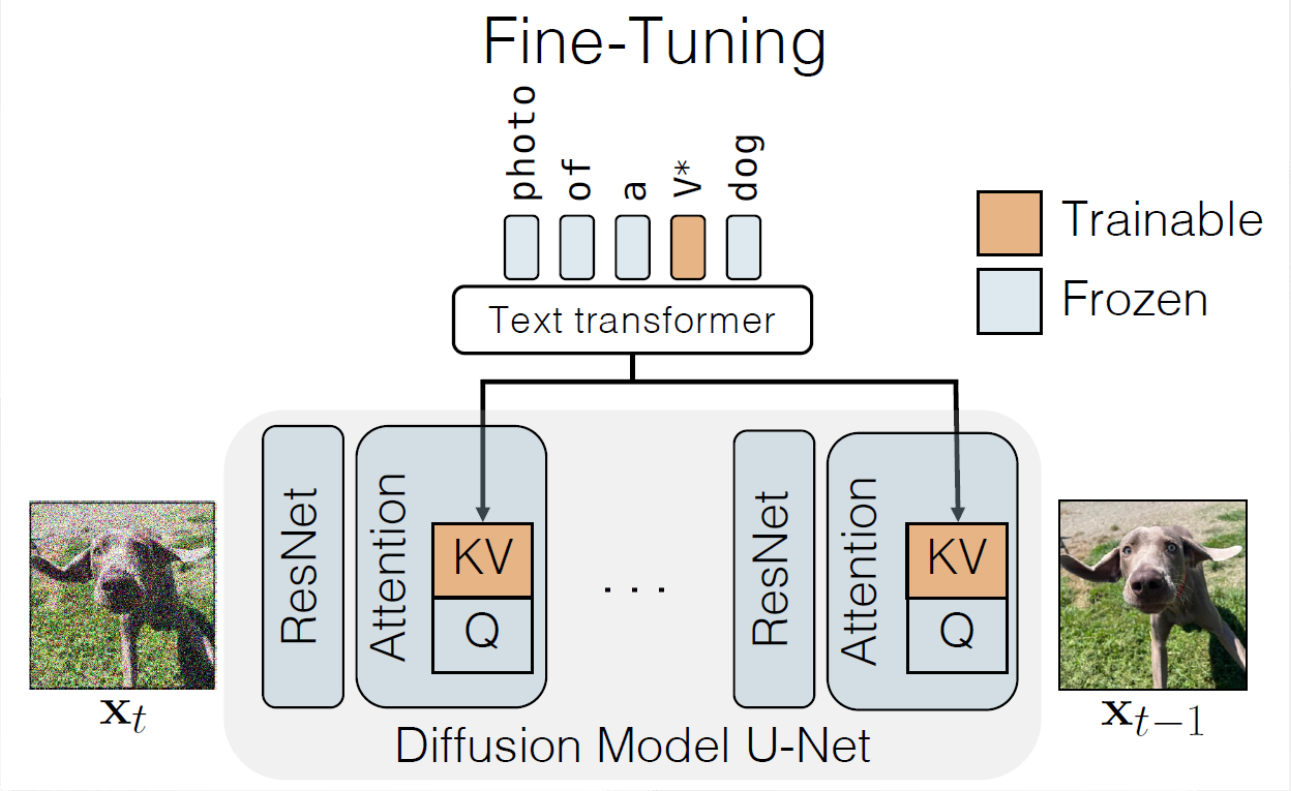
- Align a visual concept with a rare word [v] by finetuning model.



Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." *arXiv preprint arXiv:2208.12242* (2022).

Custom Diffusion

- Selectively finetune K, V mapping parameters in cross attention layers.



Kumari, Nupur, et al. "Multi-Concept Customization of Text-to-Image Diffusion." *arXiv preprint arXiv:2212.04488* (2022).

Comparison

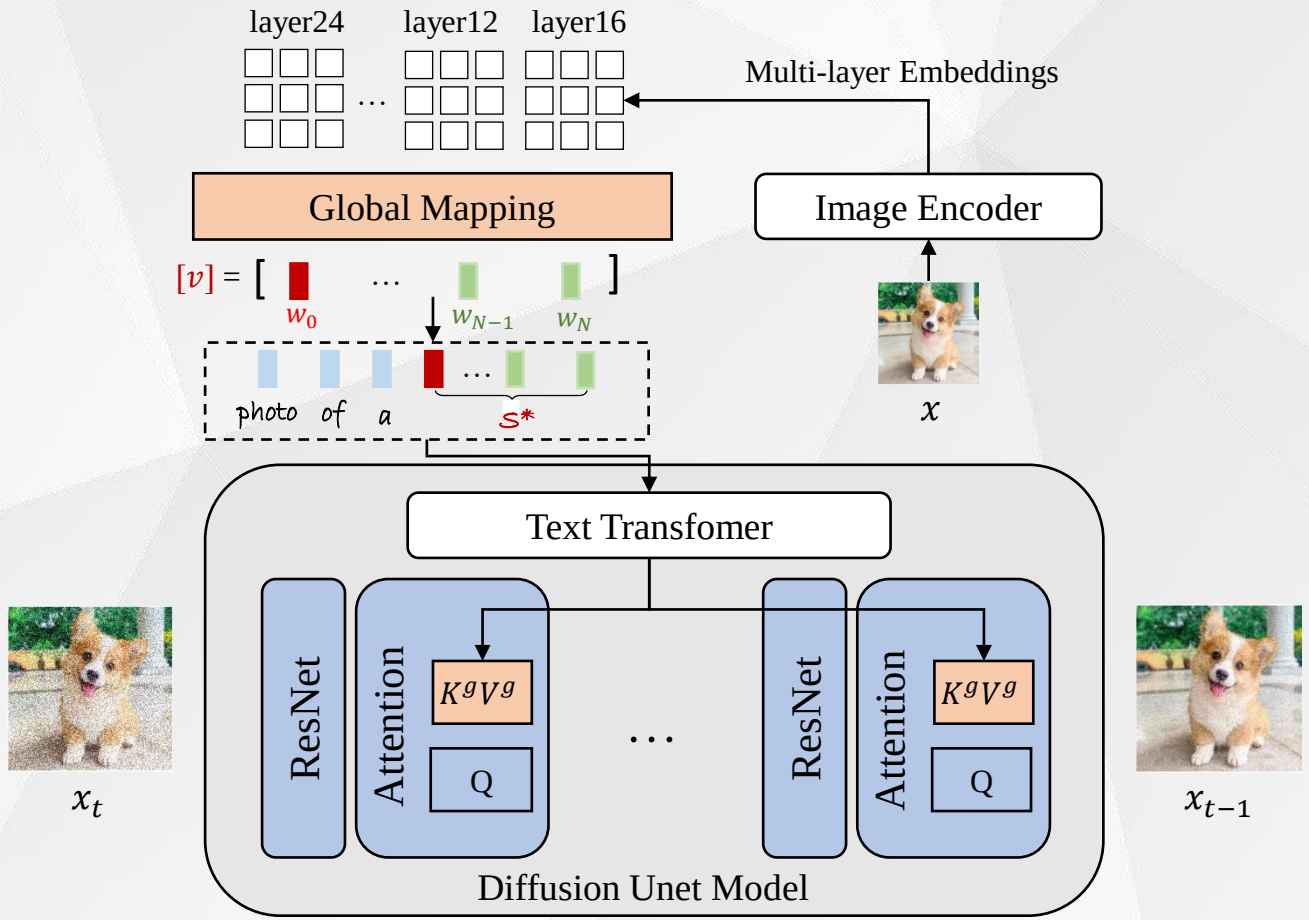
S* wearing sunglasses

Target Images | Custom Diffusion | DreamBooth | Textual Inversion

Method	Text-alignment (↑)	Image-alignment (↑)	KID (×10, ↓)	Time (↓)
Textual Inversion	0.670	0.827	22.27	50 min
DreamBooth	0.781	0.776	32.53	30 min
Custom Diffusion	0.795	0.775	20.96	6 min

Kumari, Nupur, et al. "Multi-Concept Customization of Text-to-Image Diffusion." *arXiv preprint arXiv:2212.04488* (2022).

ELITE: Global



$$L_{global} = L_{LDM} + \lambda_{global}\|v\|_1$$

Generation and Editability

Input

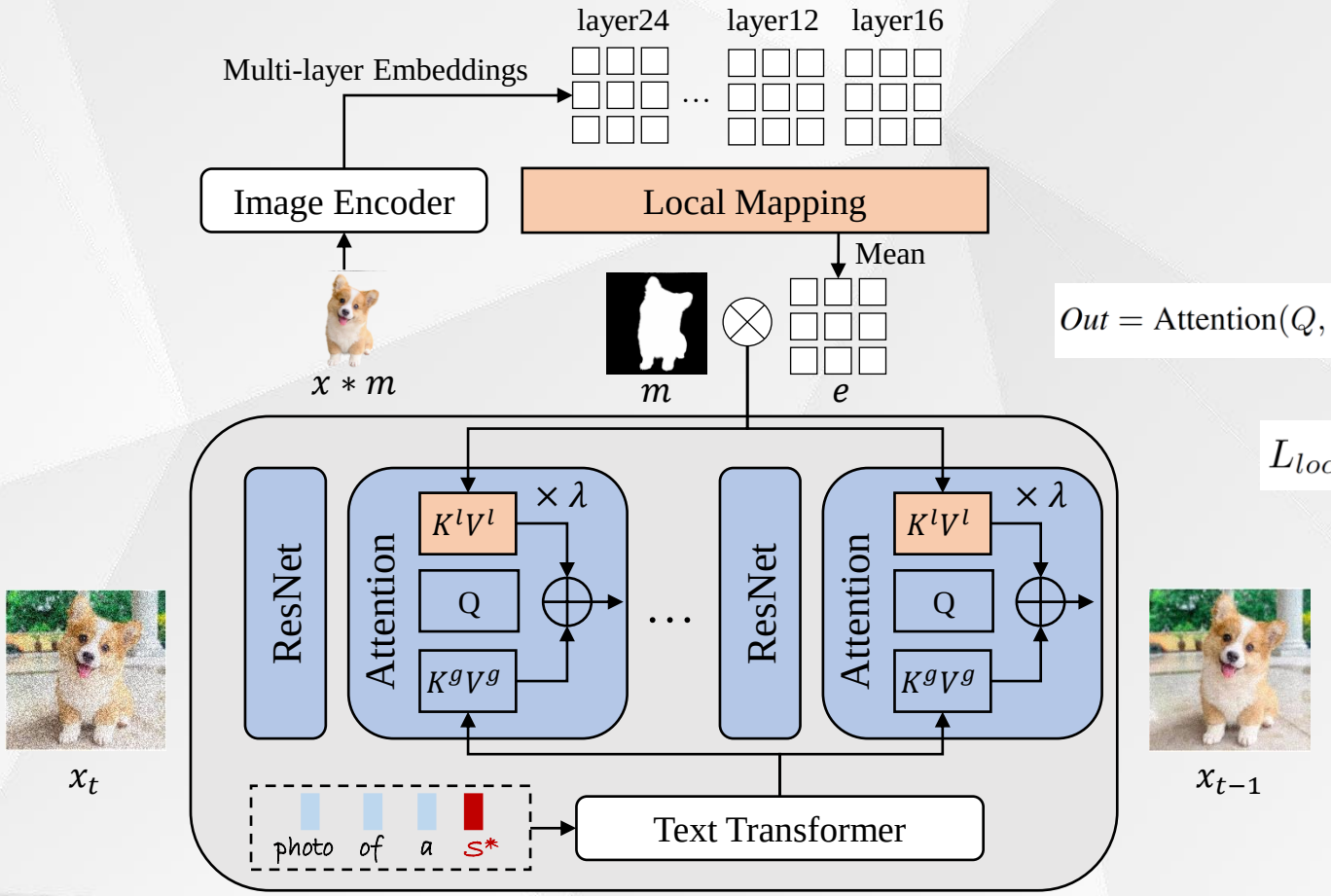
A photo of a S*

A photo of a S* in a bucket

[v]

[w0]

ELITE: Local Mapping



$$Out = \text{Attention}(Q, K^g, V^g) + \lambda \text{Attention}(Q, K^l, V^l)$$

$$L_{local} = L_{LDM} + \lambda_{local} \|V^l\|_1$$

Effect of Local Mapping

Input

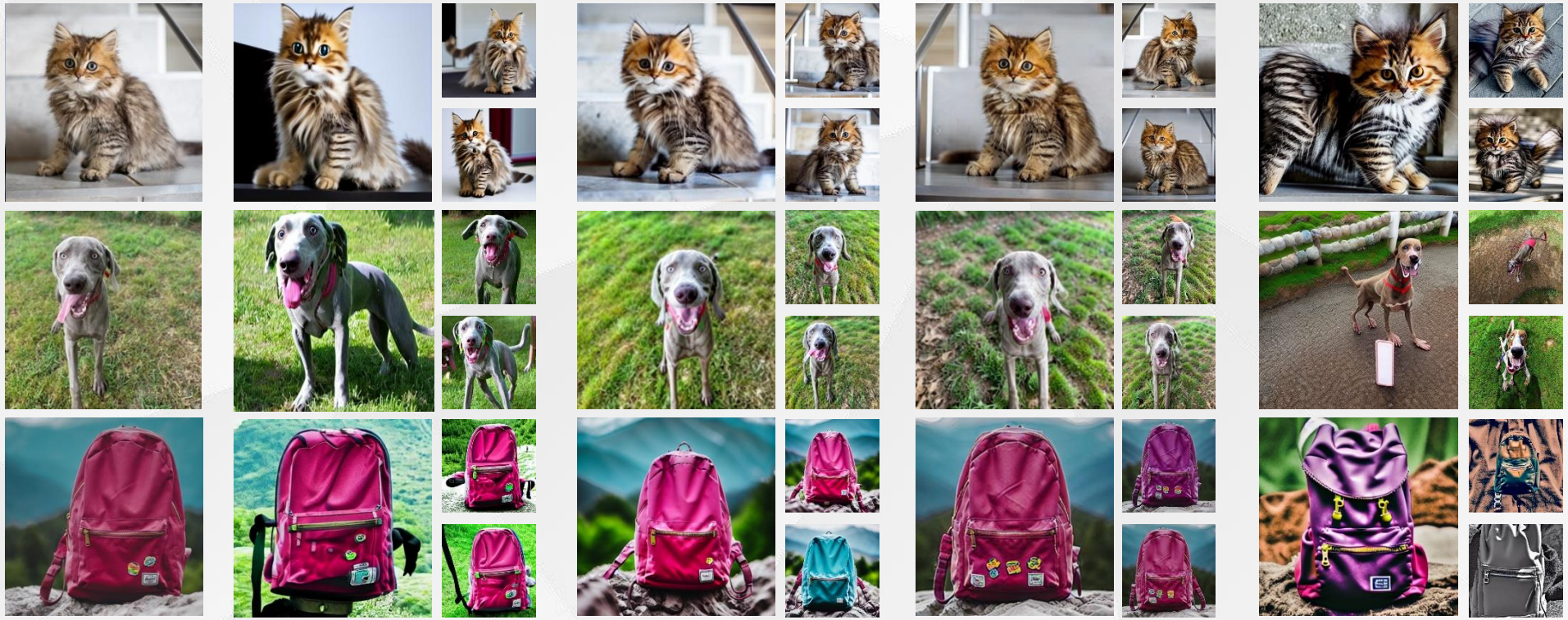
A photo of a S*

A photo of a S* in a bucket

Ours w/o Local

Ours

Inversion Comparison



Input | Ours | Custom Diffusion | Dreambooth | Textual Inversion

Editing Comparison

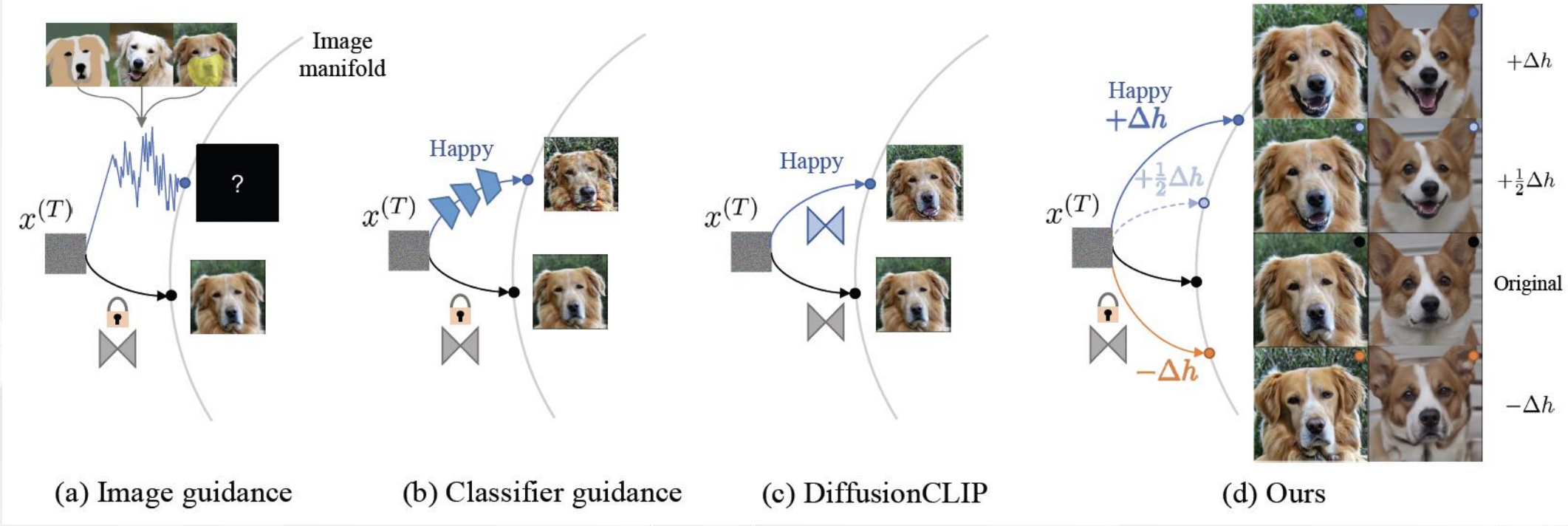
Content

- Large Scale Pretraining Models

- Applications
  - Image Editing
  - Custom Generation
  - Latent Space

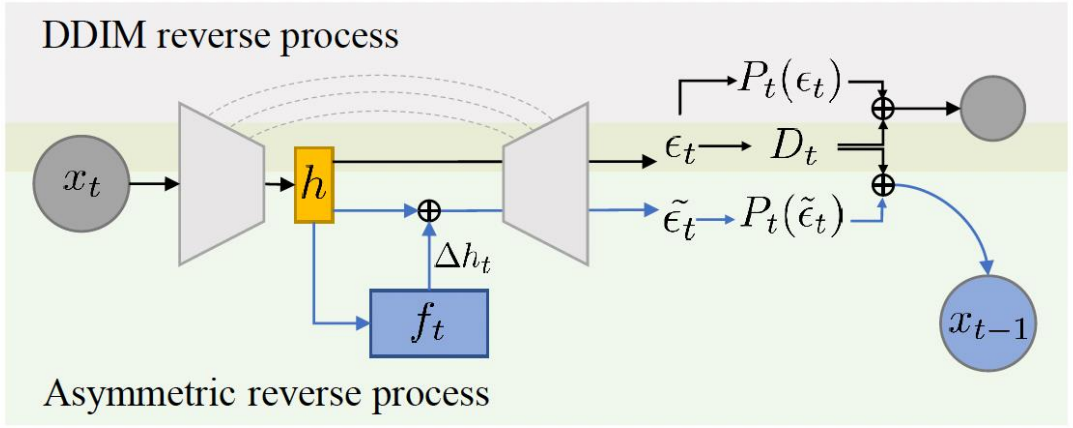
DIFFUSION MODELS ALREADY HAVE A SEMANTIC LATENT SPACE

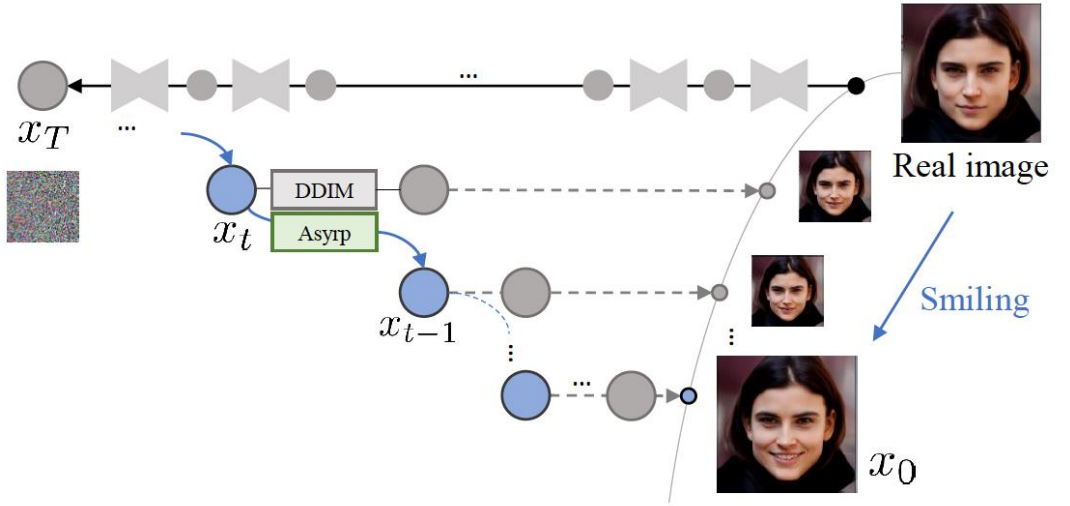
Mingi Kwon, Jaeseok Jeong, Youngjung Uh
Department of Artificial Intelligence
Yonsei University
Seoul, Republic of Korea



(a) Image guidance (b) Classifier guidance (c) DiffusionCLIP (d) Ours

Latent Space: Method





DDIM Sampling:

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\,\mathbf{P}_t(\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t)) + \mathbf{D}_t(\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t)) + \sigma_t \boldsymbol{z}_t,$$

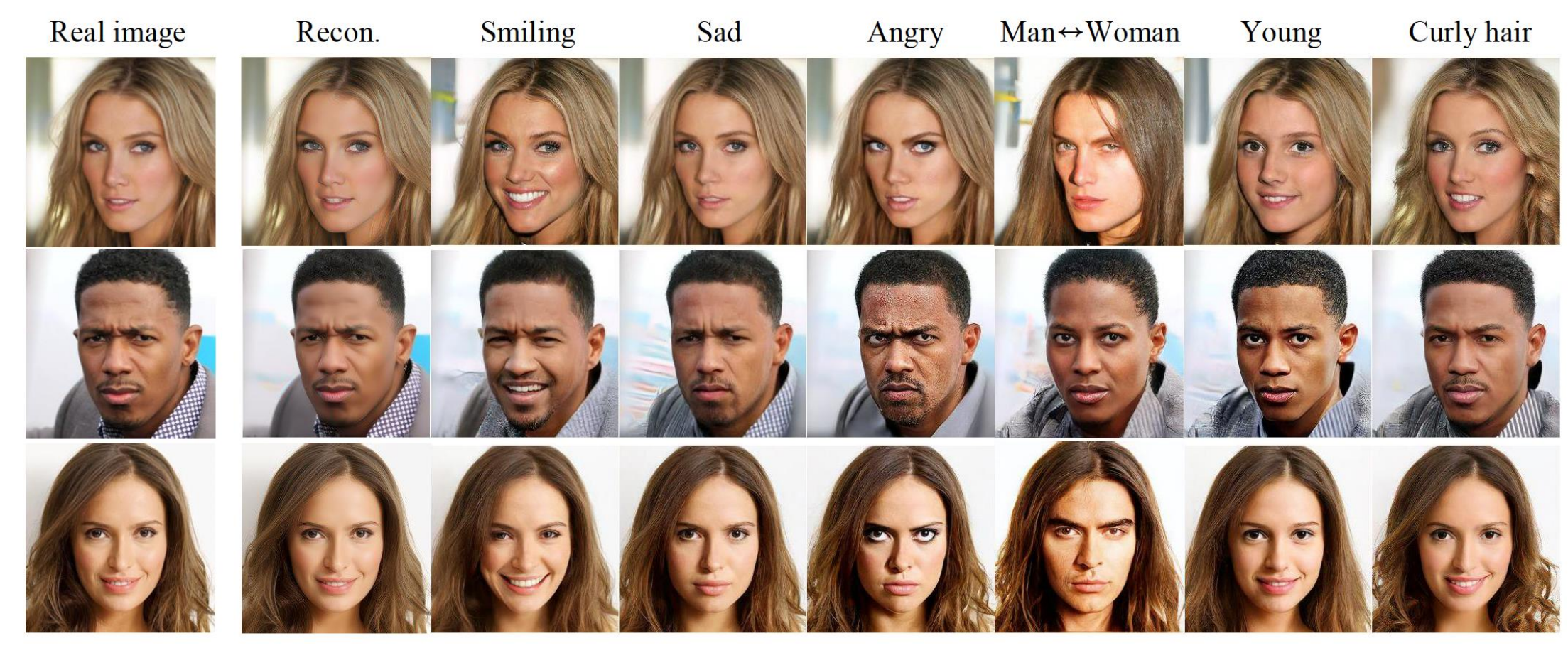
Latent Space Editing:

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\,\mathbf{P}_t(\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t|\Delta\boldsymbol{h}_t)) + \mathbf{D}_t(\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t)) + \sigma_t \boldsymbol{z}_t,$$

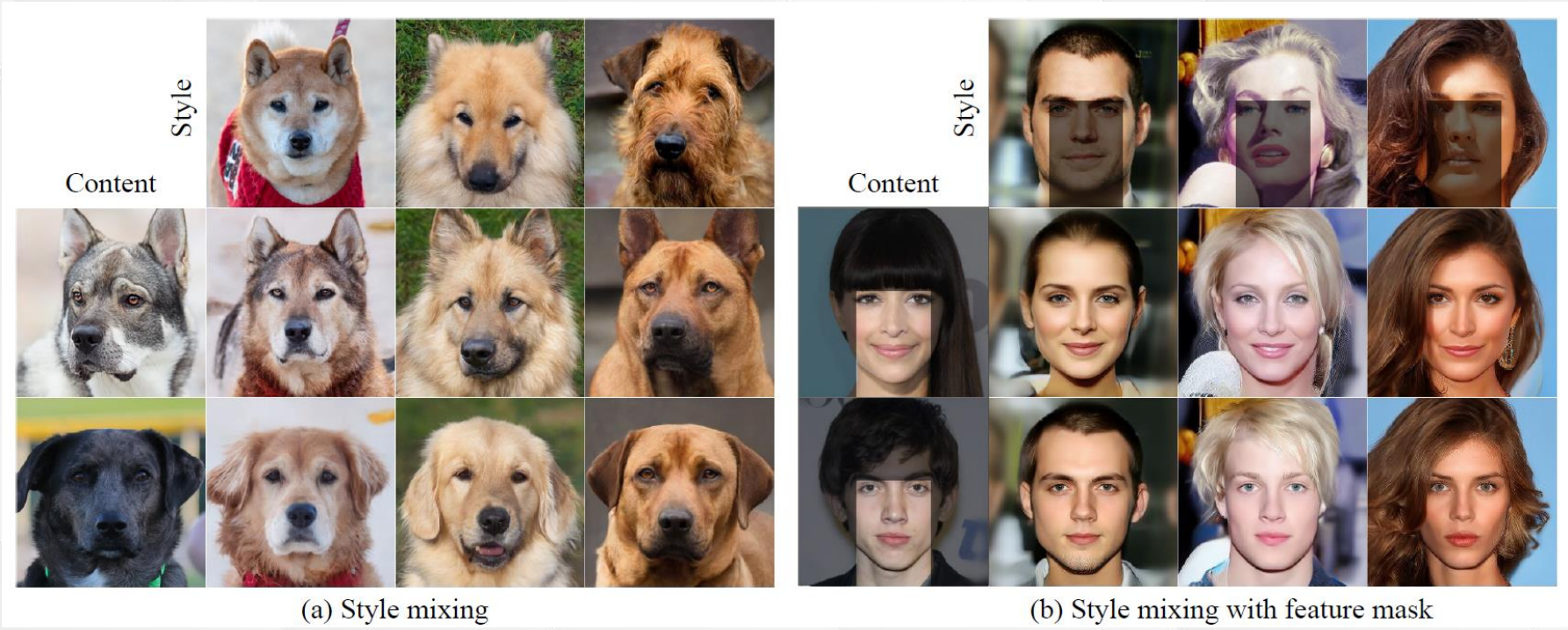
Modified Sampling:

$$p_\theta^{(t)}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \begin{cases} \mathcal{N}\left(\sqrt{\alpha_{t-1}}\,\mathbf{P}_t(\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t|\boldsymbol{f}_t)) + \mathbf{D}_t, \sigma_t^2 \boldsymbol{I}\right), \ \eta = 0 & \text{if } T \geq t \geq t_{\text{edit}} \\ \mathcal{N}\left(\sqrt{\alpha_{t-1}}\,\mathbf{P}_t(\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t)) + \mathbf{D}_t, \sigma_t^2 \boldsymbol{I}\right), \ \eta = 0 & \text{if } t_{\text{edit}} > t \geq t_{\text{noise}} \\ \mathcal{N}\left(\sqrt{\alpha_{t-1}}\,\mathbf{P}_t(\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t)) + \mathbf{D}_t, \sigma_t^2 \boldsymbol{I}\right), \ \eta = 1 & \text{if } t_{\text{noise}} > t \end{cases}$$

Latent Space: Results



Training-free Style Transfer



(a) Style mixing (b) Style mixing with feature mask

$$\tilde{\boldsymbol{h}}_t \leftarrow f((m \otimes \boldsymbol{h}_t), (m \otimes \boldsymbol{h}_t^{content}), \gamma) \oplus (1 - m) \otimes \boldsymbol{h}_t$$

Training-free Style Transfer Emerges from h-space in Diffusion models, Arxiv 2023

Discovering Interpretable Directions

$$\mathbf{J}_t^{\mathrm{T}} \mathbf{J}_t \mathbf{v} = \frac{\partial}{\partial \mathbf{h}_t} \left\langle \boldsymbol{\epsilon}_t^{\theta}(\mathbf{x}_t, \mathbf{h}_t), \mathbf{J}_t \mathbf{v} \right\rangle$$

Algorithm 1 Jacobian subspace iteration

Input: $\mathbf{f} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$, $\mathbf{h} \in \mathbb{R}^{d_{\text{in}}}$ and $\mathbf{V} \in \mathbb{R}^{d_{\text{in}} \times k}$
Output: $(\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^{\mathrm{T}})$ – k top singular values and vectors of the Jacobian $\partial \mathbf{f} / \partial \mathbf{h}$

```
y ← f(h)
if V is empty then
    V ← i.i.d. standard Gaussian samples
end if
Q, R ← QR(V)                          ▷ Reduced QR decomposition
V ← Q                                 ▷ Ensures V^T V = I
while stopping criteria do
    U ← ∂f(h + aV)/∂a at a = 0        ▷ Batch forward
    V̂ ← ∂(U^T y)/∂h
    V, Σ², R ← SVD(V̂)                 ▷ Reduced SVD
end while
Orthonormalize U
```

Discovering Interpretable Directions in the Semantic Latent Space of Diffusion Models, Arxiv 2023.

Summary

- Pretraining: From text-to-image to text-to-X
  - 3D
  - Video
- Downstream tasks
  - Generation
  - Editing
  - Understanding
  - Image -> Video
- New Paradigm: Pretraining -> Downstream tasks
  - Network architectures, tasks
  - Self-supervised vision learning
  - Pretraining -> Downstream: VL, GAN, T2I