JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Alignments in Text-to-Image Generation

Zhengyuan Yang

Microsoft

Alignments in Text-to-Image Generation

( , a dog is running through the grass)

Q3: how to do image generation?

Image Generation

Produce visual data: aligned with human intentions

LLM for language understanding and generation

General-purpose interface

Q4: how to train multimodal LLM?
Q5: how to chain vision experts with LLM?

Image Encoder

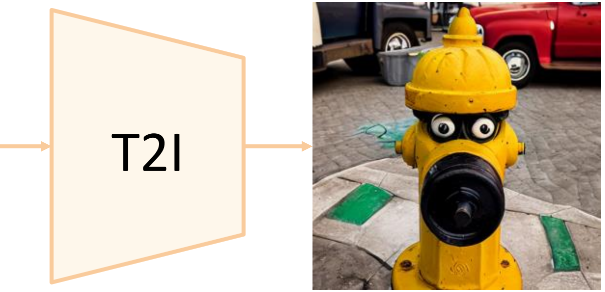
Consume visual data

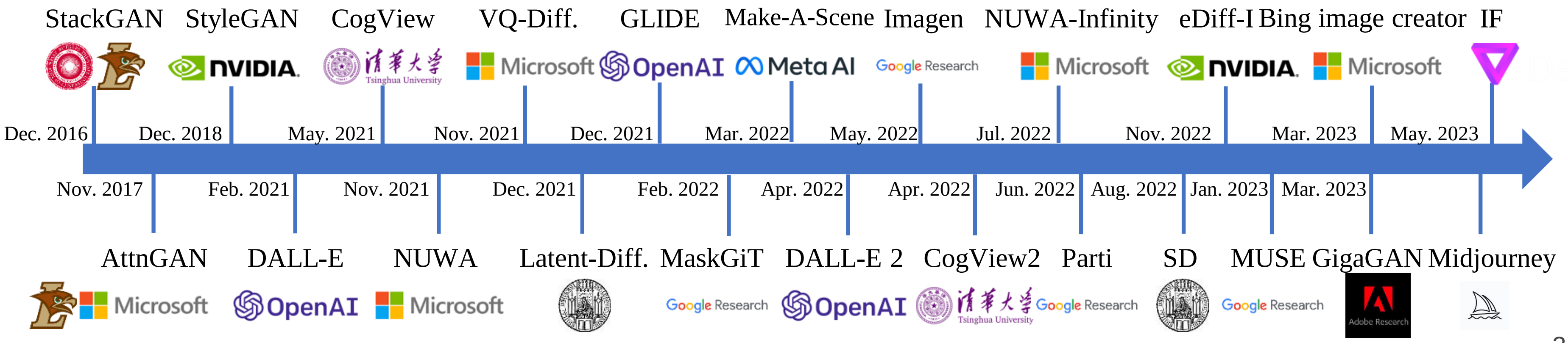
Q1: how to learn image representations?
Q2: how to extend vision models with more flexible, promptable interfaces?

Text-to-Image Generation

- Text-to-image (T2I)
- Aligning with human intentions

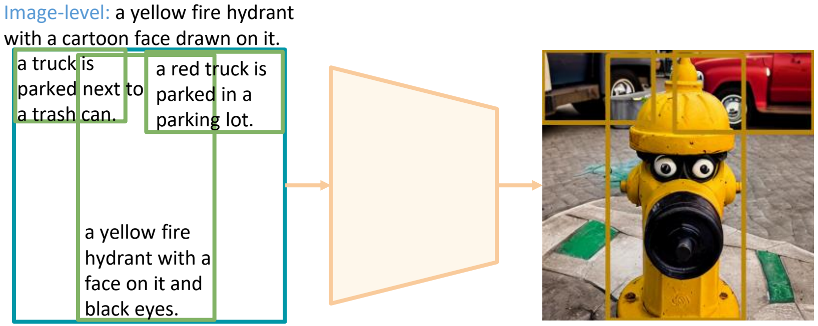
Text prompt: a yellow fire hydrant with a cartoon face drawn on it.



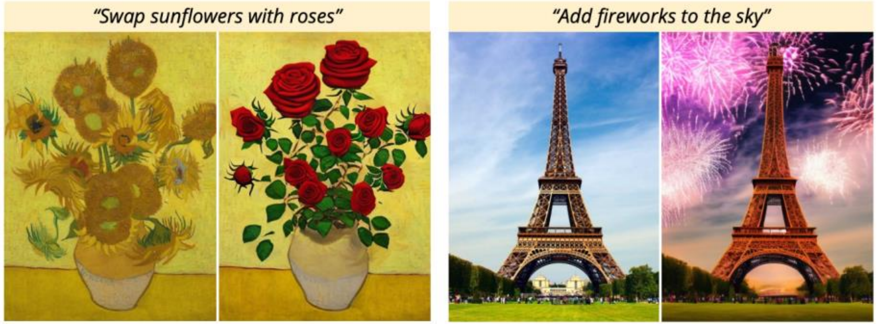


Alignments in Text-to-Image Generation

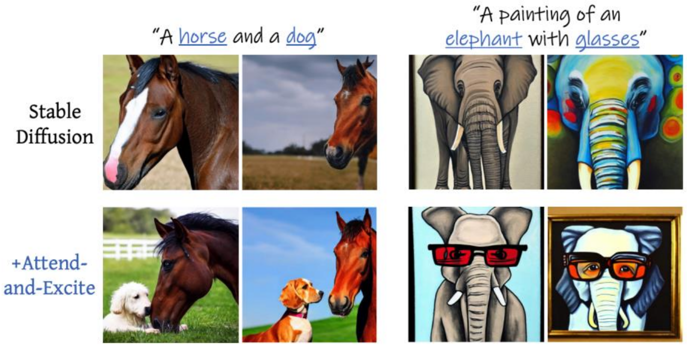
Controllable generation



Editing



Better following prompts



Concept customization

Image credit: ReCo, InstructPix2Pix, Attend-and-Excite, DreamBooth

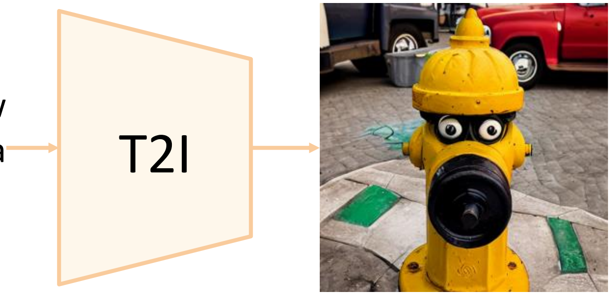
Agenda

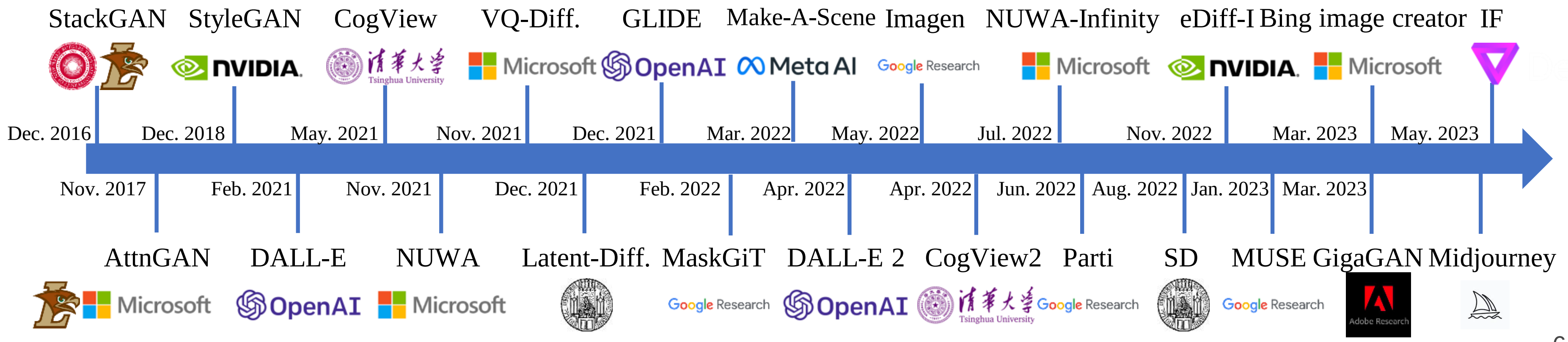
- Text-to-image (T2I) basics
- Aligning human intentions in T2I generation
  - Controllable generation
  - Editing
  - Better following prompts
  - Concept customization
- Summary and discussion

Text-to-Image Basics

- GAN
- Auto-regressive
- Non-AR Transformer
- Diffusion

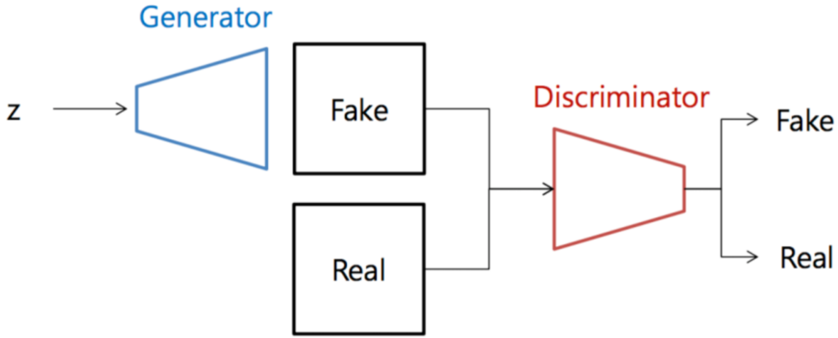
Text prompt: a yellow fire hydrant with a cartoon face drawn on it.



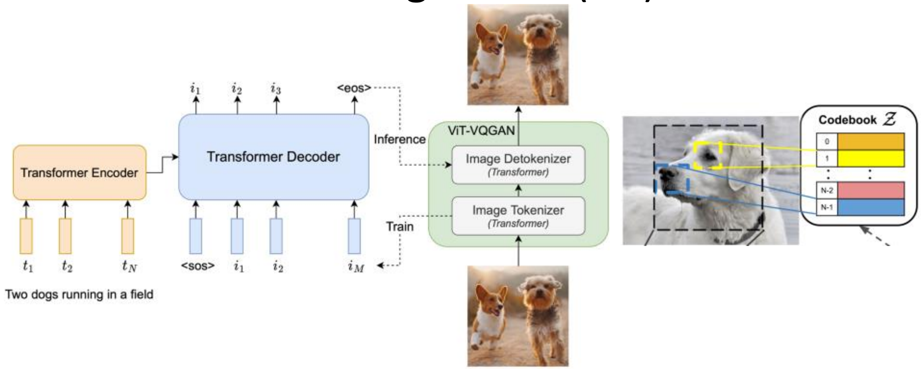


Text-to-Image Basics

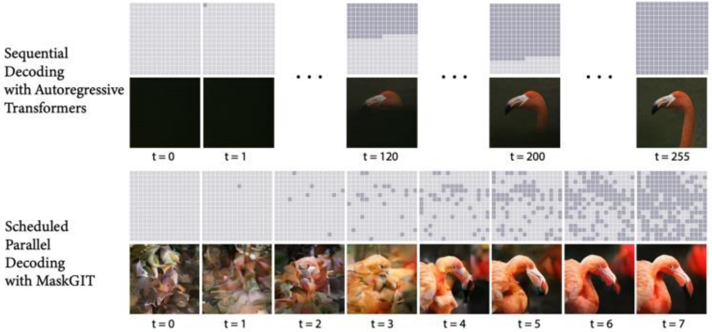
Generative Adversarial Networks (GAN)



Auto-regressive (AR)



Non-AR Transformer



Diffusion

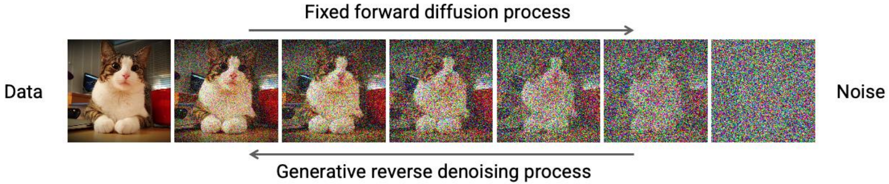
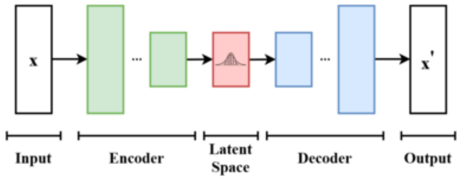


Image credit: Taming Transformer, Parti, MaskGiT, CVPR 2022 Tutorial: Denoising Diffusion-based Generative Modeling: Foundations and Applications

Stable Diffusion (SD) Basics

- SD overview
  - Variational autoencoder (VAE)
  - Condition encoder
  - Conditional denoising U-Net



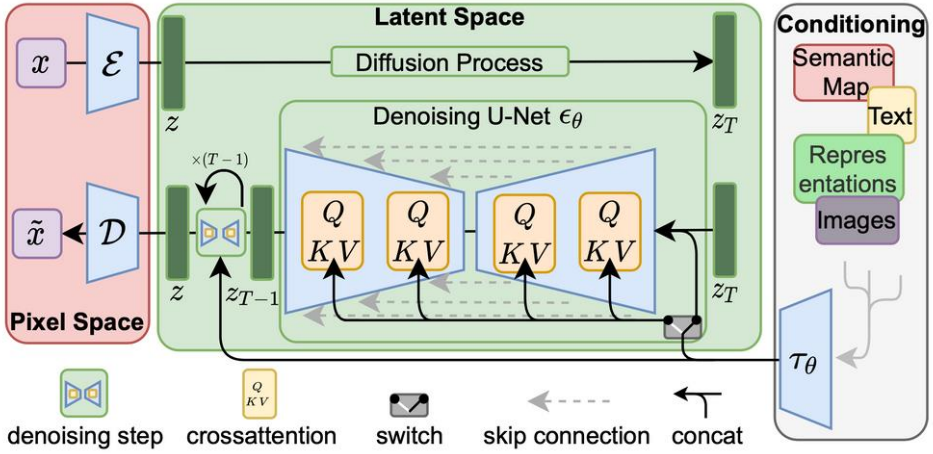
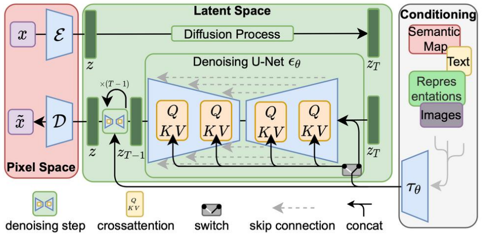


Image credit: “High-Resolution Image Synthesis with Latent Diffusion Models”, VAE-Wikipedia

Stable Diffusion (SD) Basics

- Inference flow
  - Variational autoencoder (VAE)
  - Condition encoder
  - Conditional denoising U-Net



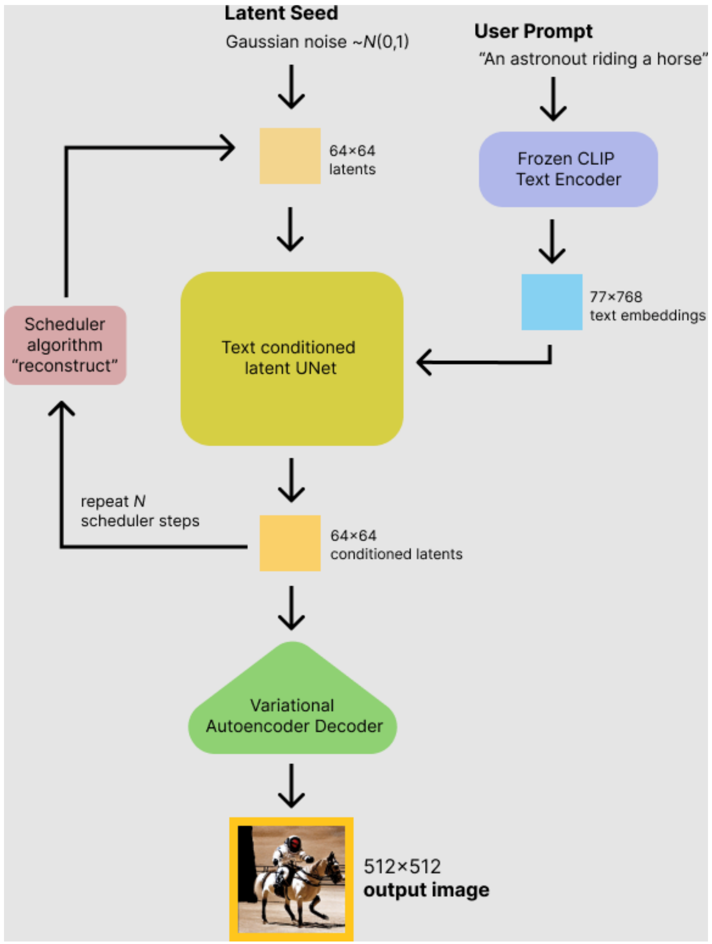
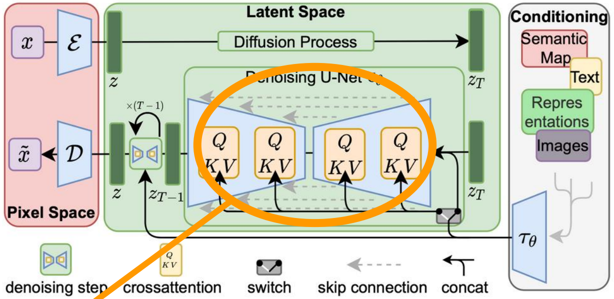
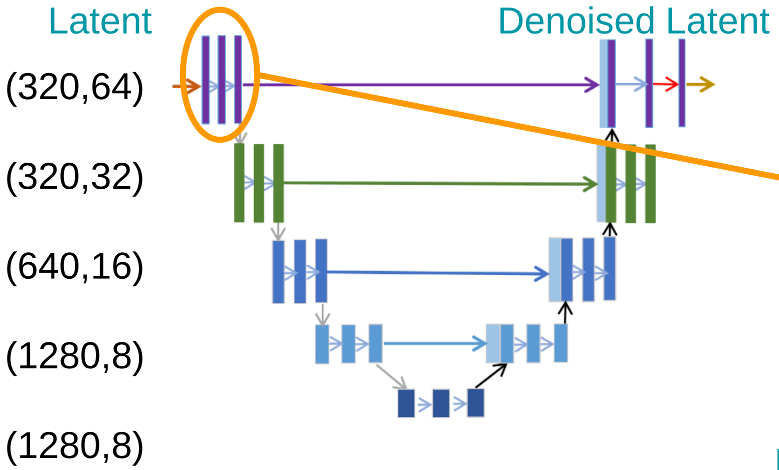


Image credit: https://huggingface.co/blog/stable_diffusion

Stable Diffusion (SD) Basics

- Zooming into conditional U-Net:
  How text condition operates on image?
  - Image-text cross attention





Latent: b*320*64*64
text: b*77*768
t: b*1280

CrossAttn
DownBlock2D

Latent': b*320*32*32

Q: latent+duplicate(linear(t))
=> b*4096*320

K, V: text => b*77*768

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

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

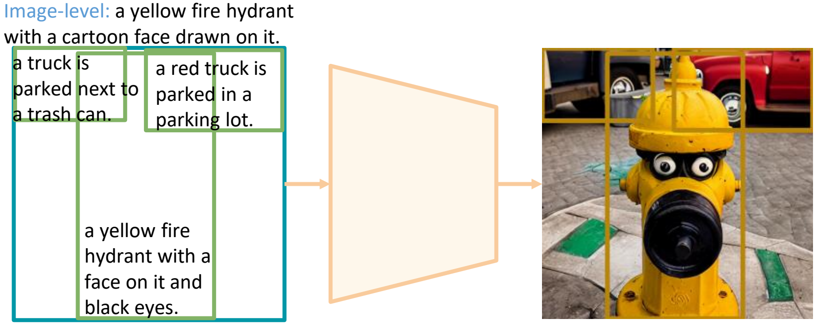
Image-text attention map of HW*77

Agenda

- Text-to-image (T2I) basics
- Aligning human intentions in T2I generation
  - Controllable generation
  - Editing
  - Better following prompts
  - Concept customization
- Summary and discussion

Controllable Generation

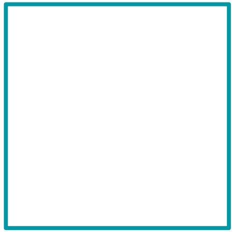
- Text+layout/box: localized description control
- Text+dense control (e.g., mask, edge, scribble, etc.)
- Inference-time guidance



[1] GLIGEN: Open-Set Grounded Text-to-Image Generation
[2] ReCo: Region-Controlled Text-to-Image Generation
[3] Diagnostic Benchmark and Iterative Inpainting for Layout-Guided Image Generation
[4] Adding Conditional Control to Text-to-Image Diffusion Models
[5] Composer: Creative and Controllable Image Synthesis with Composable Conditions
[6] SpaText: Spatio-Textual Representation for Controllable Image Generation
[7] T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models
[8] SceneComposer: Any-Level Semantic Image Synthesis
[9] Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models
[10] UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild
[11] Universal Guidance for Diffusion Models
[12] Training-Free Layout Control with Cross-Attention Guidance

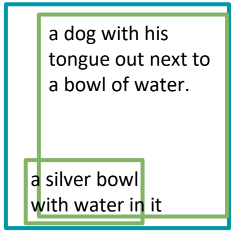
Image credit: “ReCo: Region-Controlled Text-to-Image Generation”

ReCo: Region-Controlled T2I Generation



a close up of a dog near a bowl.

Text: global image text description → Text: <u>grounded</u> global and regional descriptions (Grounded Region-Controlled texts)



a close up of a dog near a bowl.

a close up of a dog near a bowl. <145> <44> <999> <950> a dog with his tongue out next to a bowl of water. <87> <697> <613> <985> a silver bowl with water in it.

Image credit: “ReCo: Region-Controlled Text-to-Image Generation”

ReCo: Region-Controlled T2I Generation

a person standing at the plate in mid swing of a bat
<687> <204> <999> <833> baseball player … jersey.
<21> <447> <433> <840> a catcher in gray … ball.
<0> <323> <123> <827> a baseball player … jersey.

- Input sequence expansion: box tokens
- Grounded: box tokens operate on the text to follow
- Finetune T2I to understand box tokens

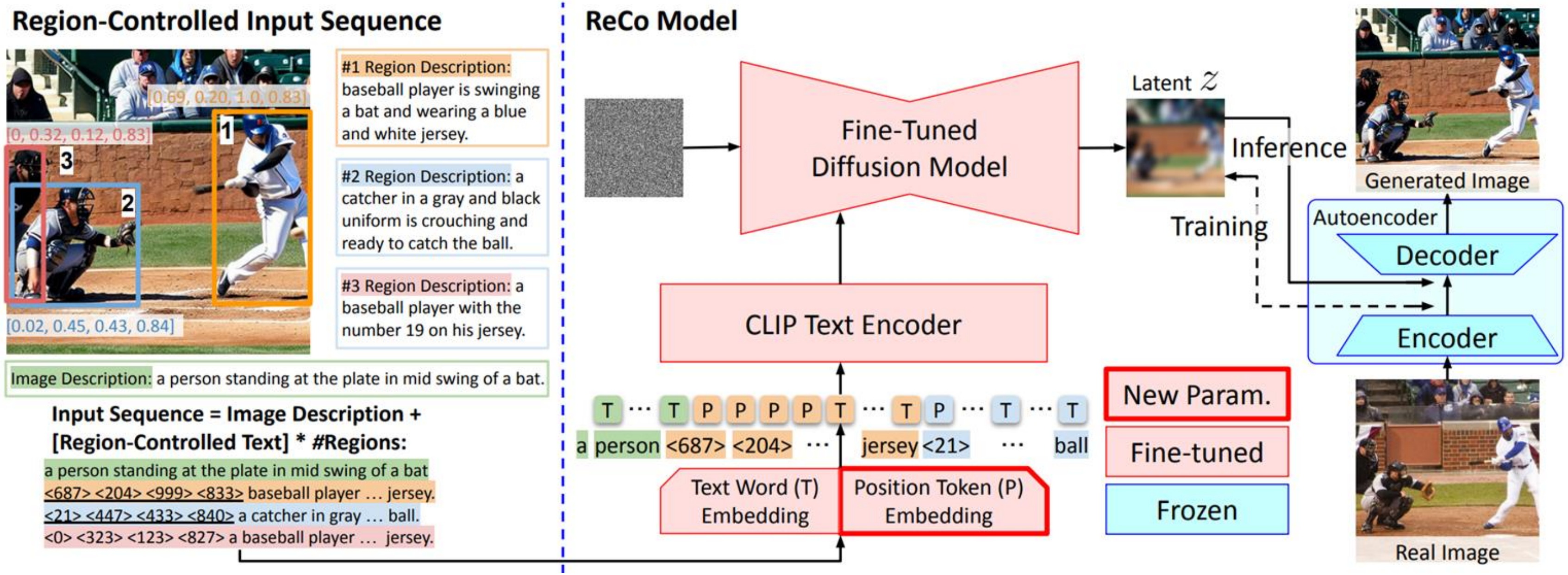
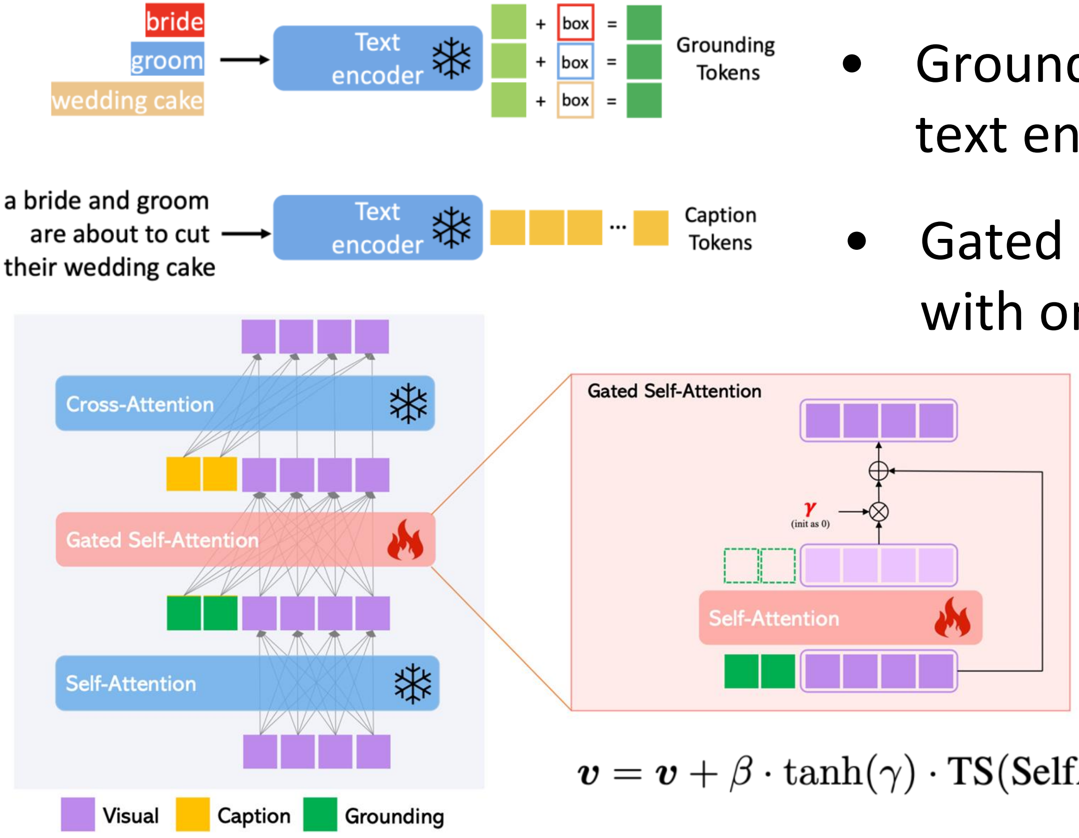


Image credit: “ReCo: Region-Controlled Text-to-Image Generation”

GLIGEN: Open-Set Grounded T2I Generation



- Grounding tokens: grounded text entity + spatial location
- Gated self-attention layer with original layers frozen

$$\boldsymbol{v} = \boldsymbol{v} + \beta \cdot \tanh(\gamma) \cdot \mathrm{TS}(\mathrm{SelfAttn}([\boldsymbol{v}, \boldsymbol{h}^e]))$$

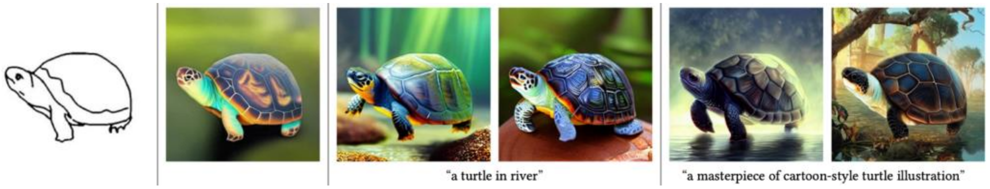
Image credit: “GLIGEN: Open-Set Grounded Text-to-Image Generation”

GLIGEN: Open-Set Grounded T2I Generation

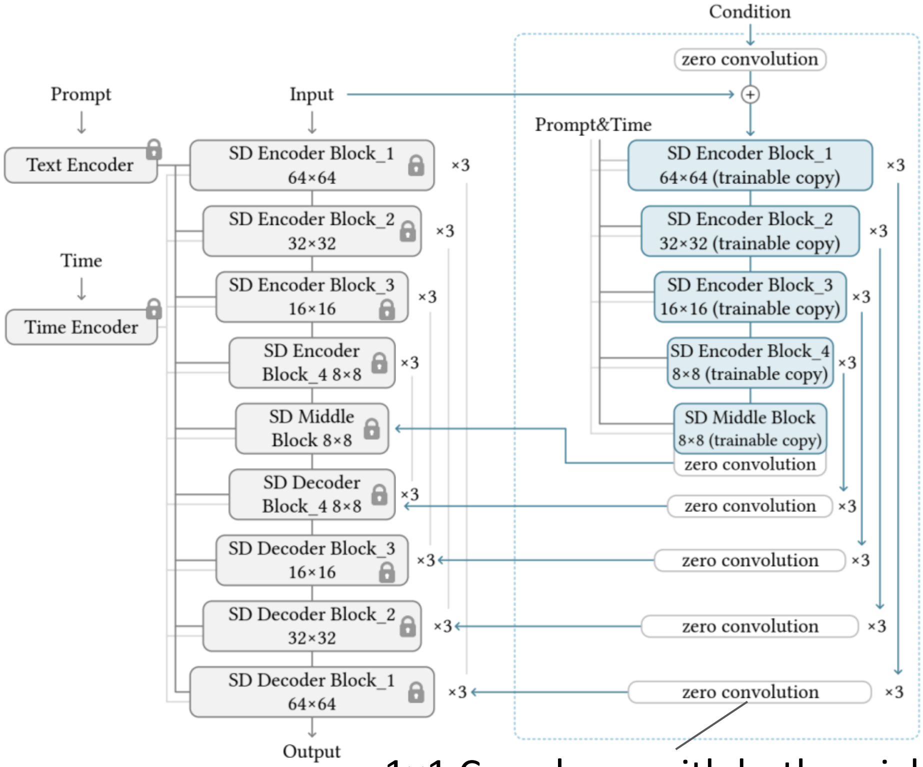
- Bounding box grounding
- Keypoint grounding
- Spatially-aligned dense conditions

Image credit: “GLIGEN: Open-Set Grounded Text-to-Image Generation”

Text+Dense Control



- Dense conditions:
  - Canny Edge
  - Hough Line
  - HED Boundary
  - User Sketching
  - Human Pose
  - Semantic Segmentation
  - Depth
  - Normal Maps
  - Cartoon Line Drawing



1×1 Conv layer with both weight and bias initialized with zeros

Image credit: “Adding Conditional Control to Text-to-Image Diffusion Models” (ControlNet)

Uni-ControlNet, UniControl

- Unified models for different conditions
- Condition composition

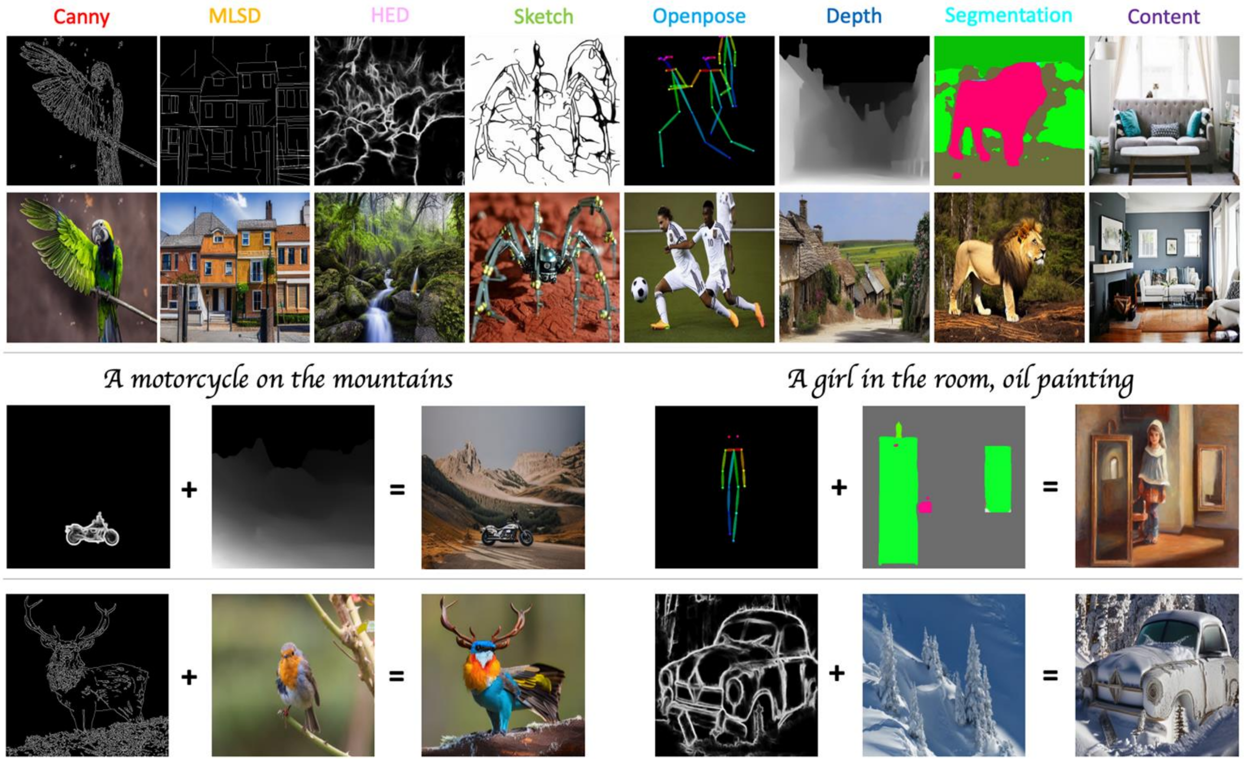


Image credit: “Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models”

Inference-time guidance

- Universal Guidance for Diffusion Models: extending classifier guidance [1] to accept any general guidance function

$$\hat{\epsilon}_\theta(z_t, t) = \epsilon_\theta(z_t, t) + s(t) \cdot \nabla_{z_t} \ell(c, f(\hat{z}_0))$$

E.g., detection:

Anchor classification, bounding box regression, and region label classification loss

Box and class labels

Faster-RCNN

Predicted “noisy” clean image

[1] Diffusion Models Beat GANs on Image Synthesis

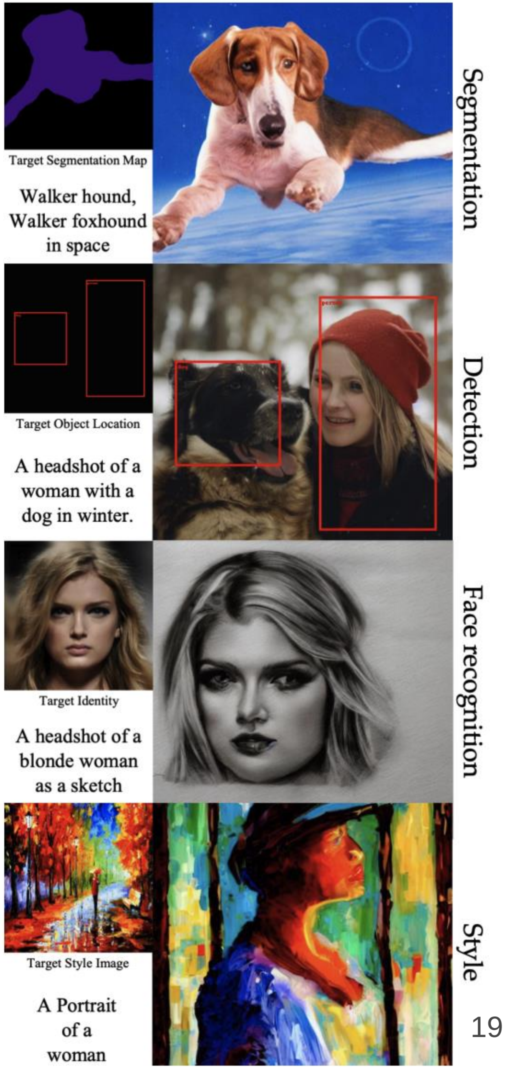
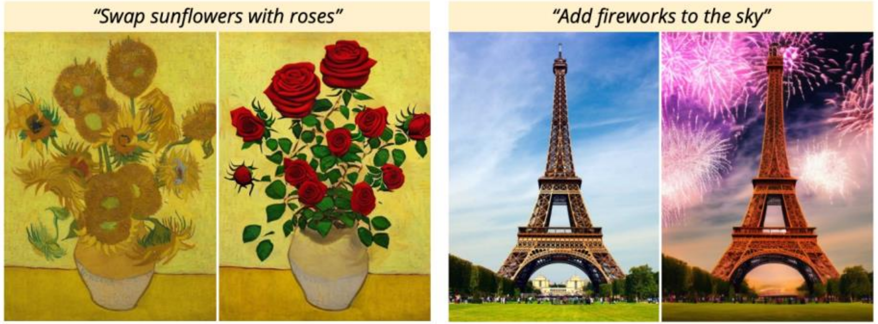


Image credit: “Universal Guidance for Diffusion Models”

Editing

- Latents spatial blend
- Image-text attention edit
- Edit instruction
- External models

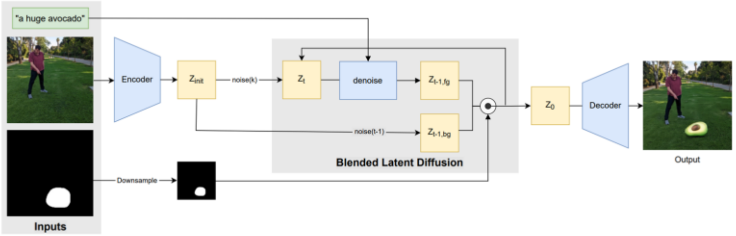


[1] Blended Diffusion for Text-driven Editing of Natural Images
[2] Blended Latent Diffusion
[3] DiffEdit: Diffusion-based semantic image editing with mask guidance
[4] eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers
[5] Region-Aware Diffusion for Zero-shot Text-driven Image Editing
[6] Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting
[7] iEdit: Localised Text-guided Image Editing with Weak Supervision
[8] EDICT: Exact Diffusion Inversion via Coupled Transformations
[9] Prompt-to-Prompt Image Editing with Cross Attention Control
[10] Imagic: Text-Based Real Image Editing with Diffusion Models
[11] SINE: SINgle Image Editing With Text-to-Image Diffusion Models
[12] InstructPix2Pix Learning to Follow Image Editing Instructions
[13] MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing
[14] Diffusion Self-Guidance for Controllable Image Generation
[15] Instruct-X-Decoder
[16] Grounded-SAM Inpainting
[17] Inpaint Anything
[18] Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models

Image credit: "InstructPix2Pix: Learning to Follow Image Editing Instructions"

Latents Spatial Blend

- Spatial editing with mask
- Image, text prompt, user input or segmented mask



$$z_t \leftarrow \underbrace{z_{fg}}_{\text{from text}} \odot m_{latent} + \underbrace{z_{bg}}_{\text{original bg image}} \odot (1 - m_{latent})$$

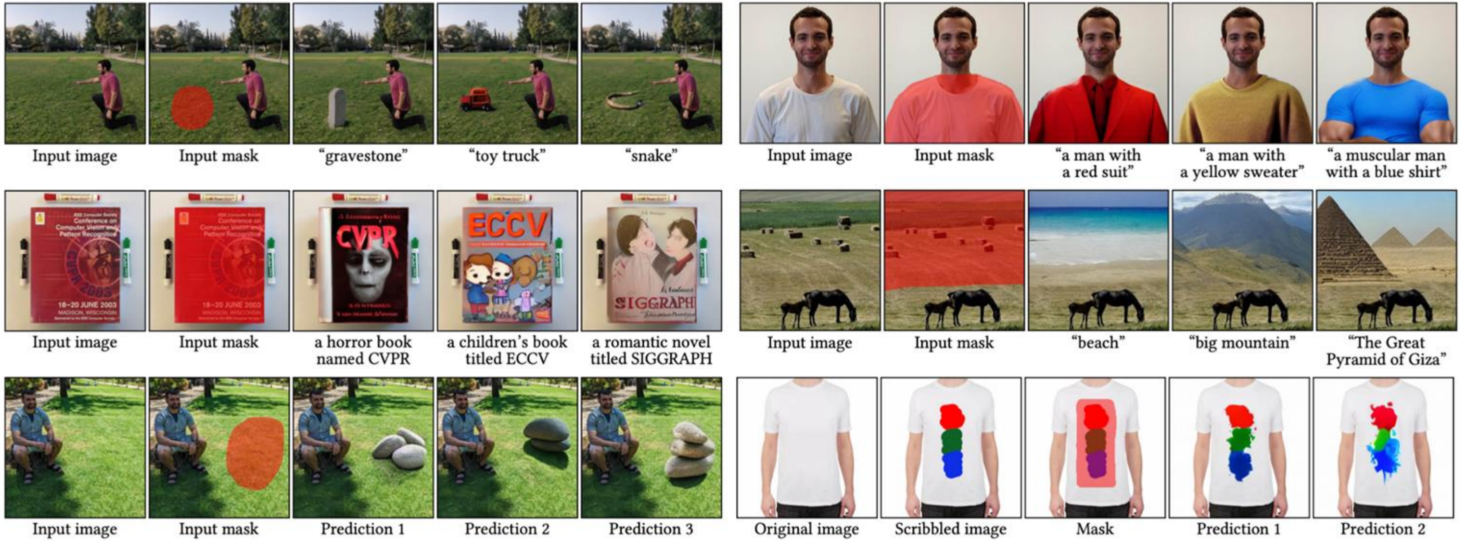


Image credit: "Blended Latent Diffusion"

Image-text Attention Edit

- Edit generated images
- Manipulate image-text cross-attention map
- Word swap, adding new phrase, attention re-weighting

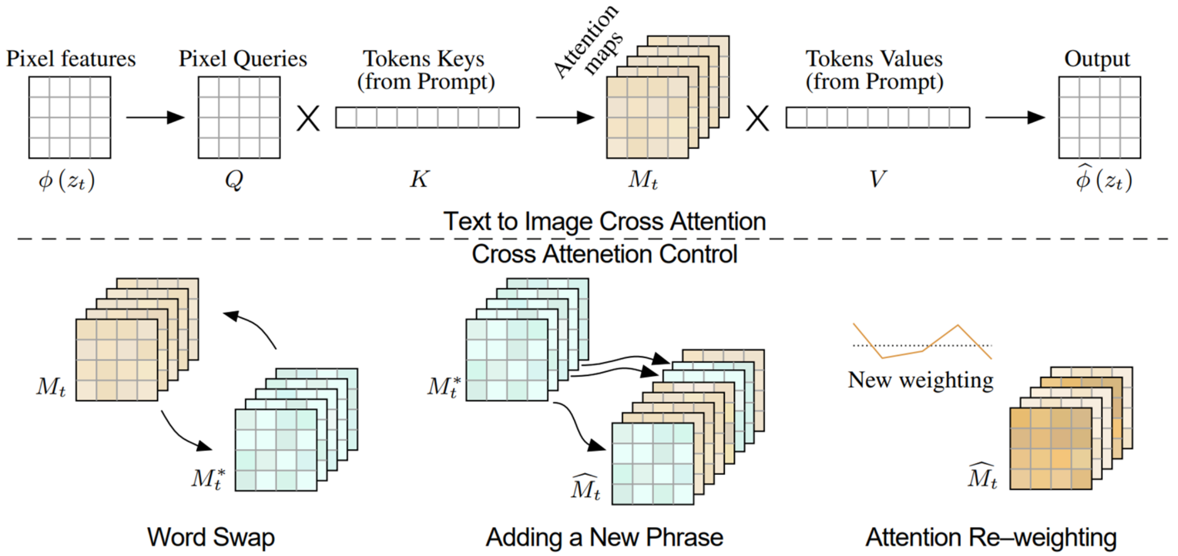
Word swap

Adding new phrase

Image credit: "Prompt-to-Prompt Image Editing with Cross Attention Control"

Image-text Attention Edit

- Maintaining two sets of cross-attention maps for edit:
  Original prompt: M_t Edited prompt: M_t^*



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

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

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

"lemon cake."

Goal

"monster cake."

M_t

"lemon cake."

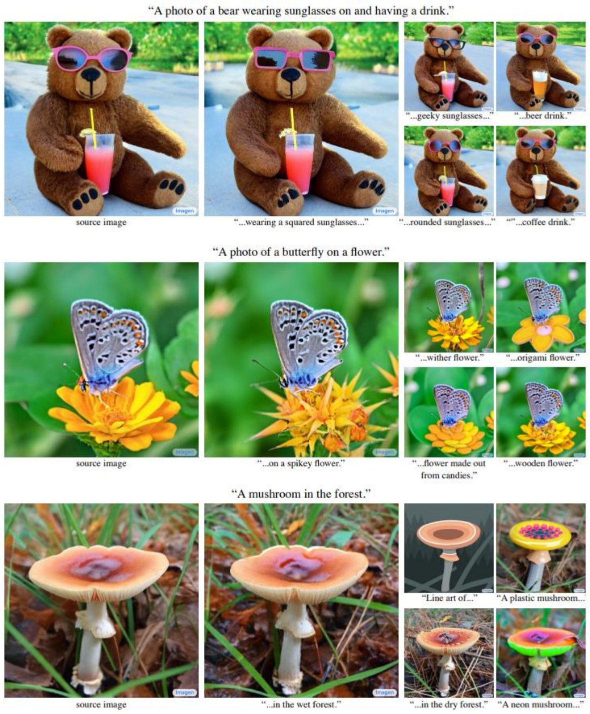
M_t^*

"monster cake."

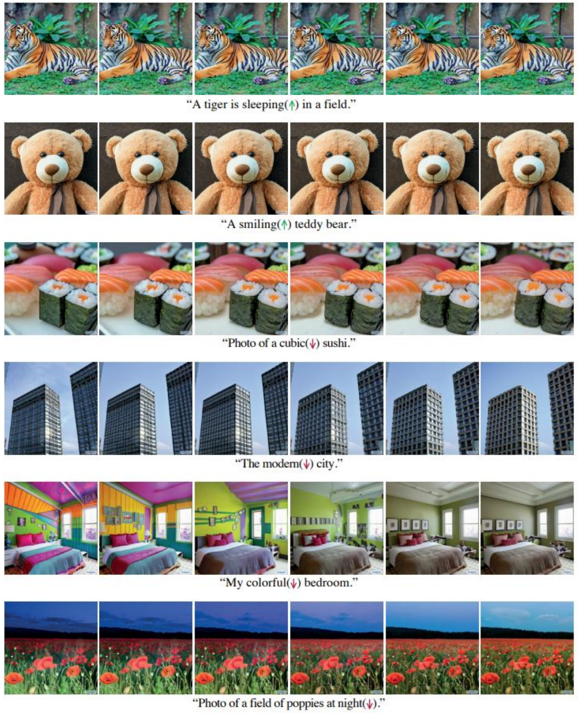
Image credit: "Prompt-to-Prompt Image Editing with Cross Attention Control"

Image-text Attention Edit

Word Swap



Adding a New Phrase



Attention Re–weighting

Image credit: "Prompt-to-Prompt Image Editing with Cross Attention Control"

Imagic

- Generated => natural image edits
- E.g., different dogs

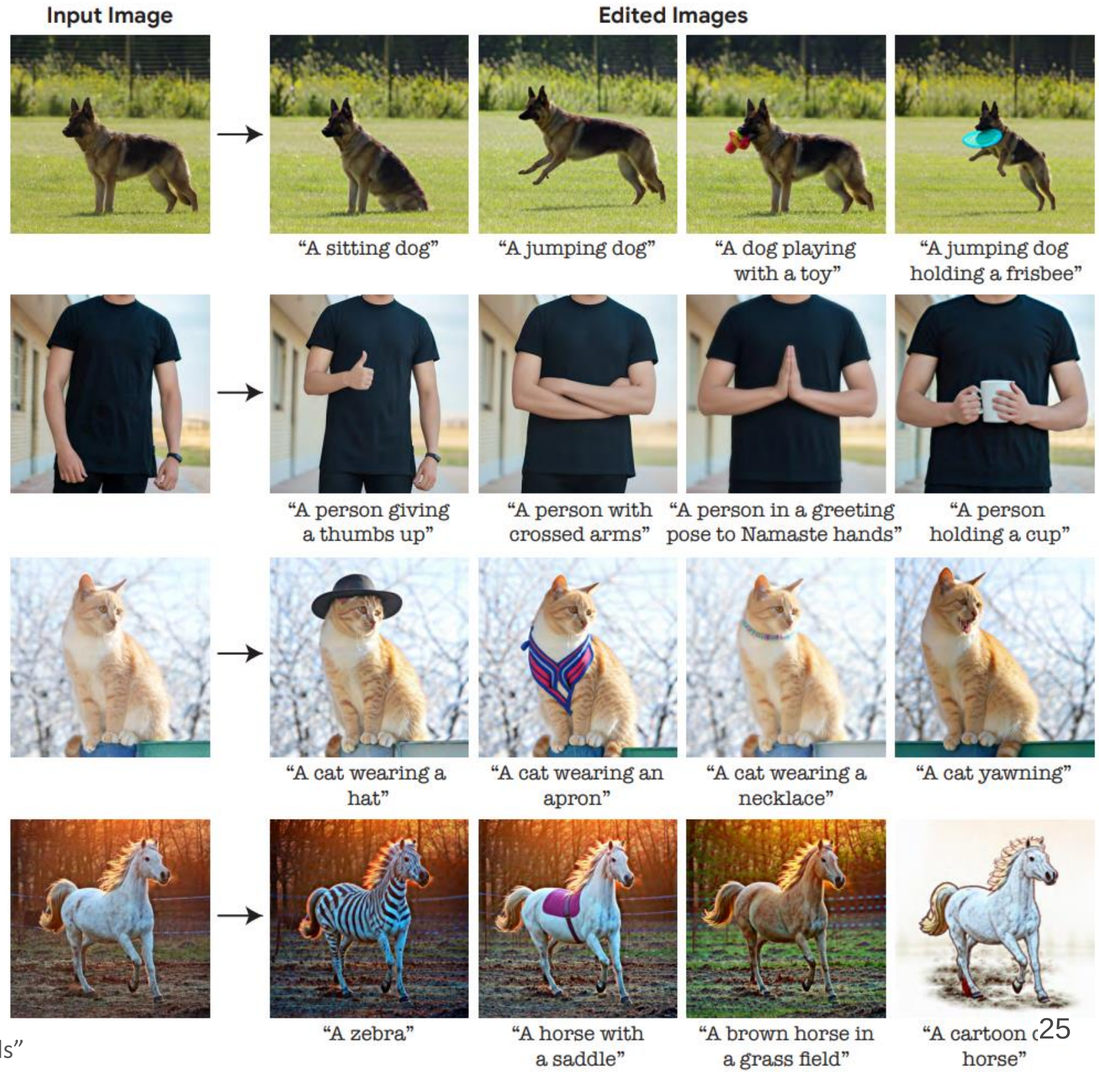
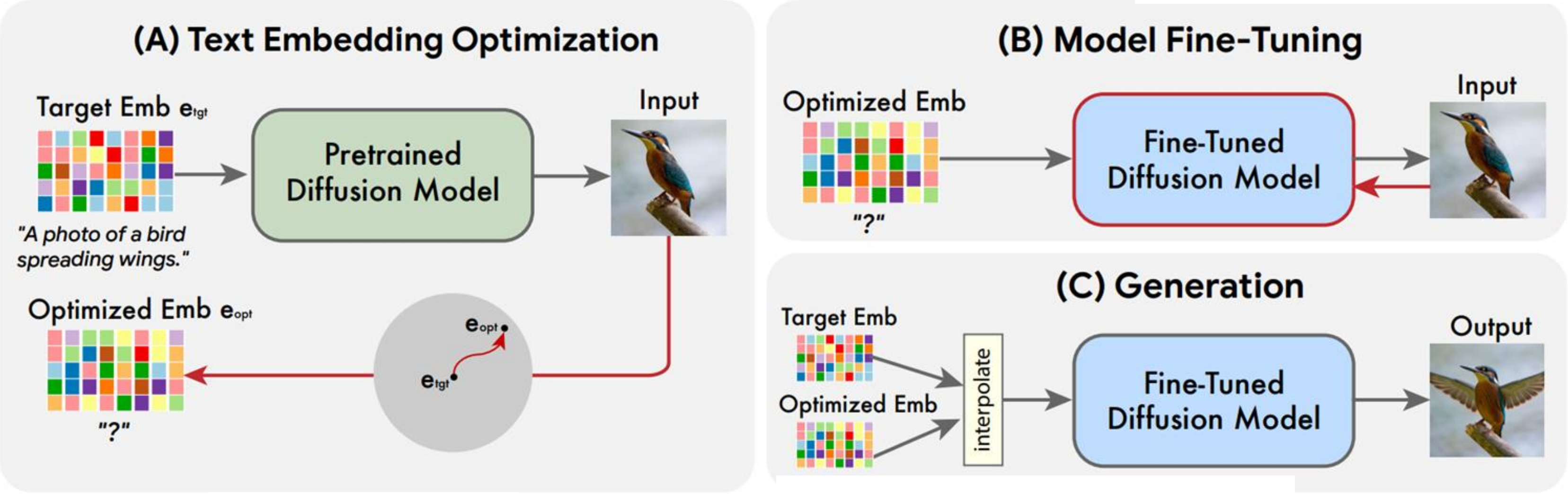


Image credit: “Imagic: Text-Based Real Image Editing with Diffusion Models”

Imagic

- Obtain original text

$$\mathcal{L}(\mathbf{x}, \mathbf{e}, \theta) = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - f_\theta(\mathbf{x}_t, t, \mathbf{e})\|_2^2\right]$$



$$\bar{\mathbf{e}} = \eta \cdot \mathbf{e}_{tgt} + (1 - \eta) \cdot \mathbf{e}_{opt}$$

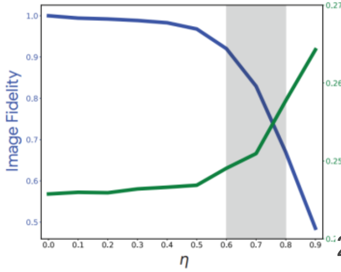


Image credit: "Imagic: Text-Based Real Image Editing with Diffusion Models"

InstructPix2Pix

- Obtain original text => Instruction-style text
  “a bird standing”, “a bird spreading wings” => “have wings spread”

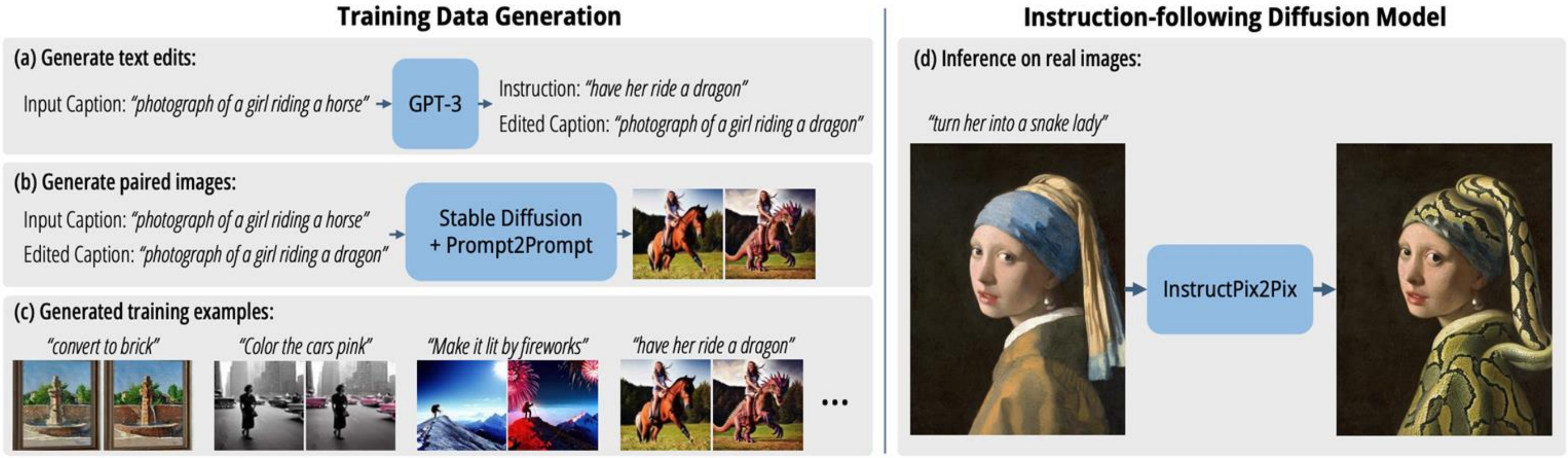
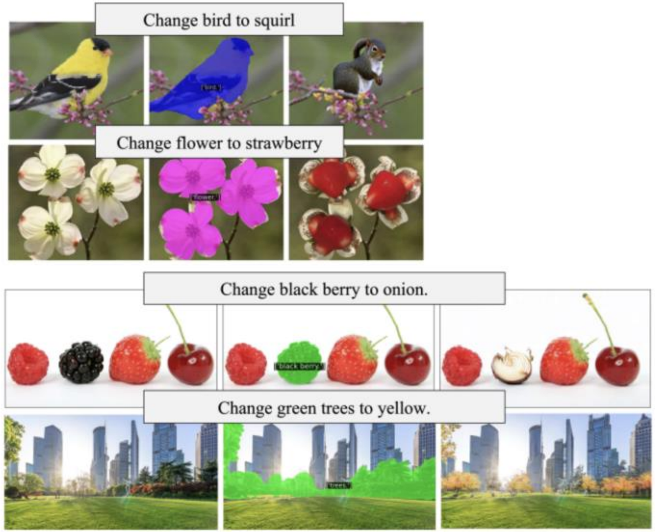


Image credit: “InstructPix2Pix: Learning to Follow Image Editing Instructions”

Editing Systems with External Models

- Segmentation



- LLM

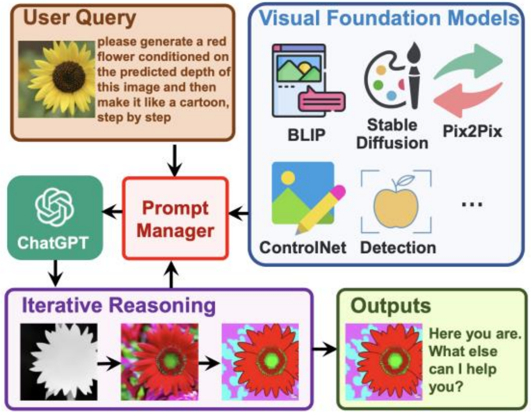
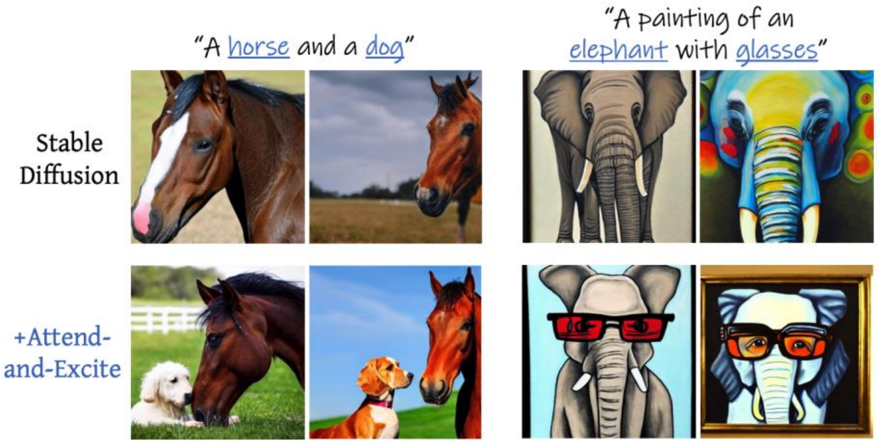


Image credit: “Instruct-X-Decoder”, “Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models ”

Better Following Prompts

- Test-time latents
- Test-time attention
- Alignment tuning



[1] Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis
[2] Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models
[3] Grounded Text-to-Image Synthesis with Attention Refocusing
[4] Compositional Visual Generation with Composable Diffusion Models
[5] Aligning Text-to-Image Models using Human Feedback
[6] ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation
[7] Training Diffusion Models with Reinforcement Learning
[8] DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models
[9] Better Aligning Text-to-Image Models with Human Preference

Image credit: "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models"

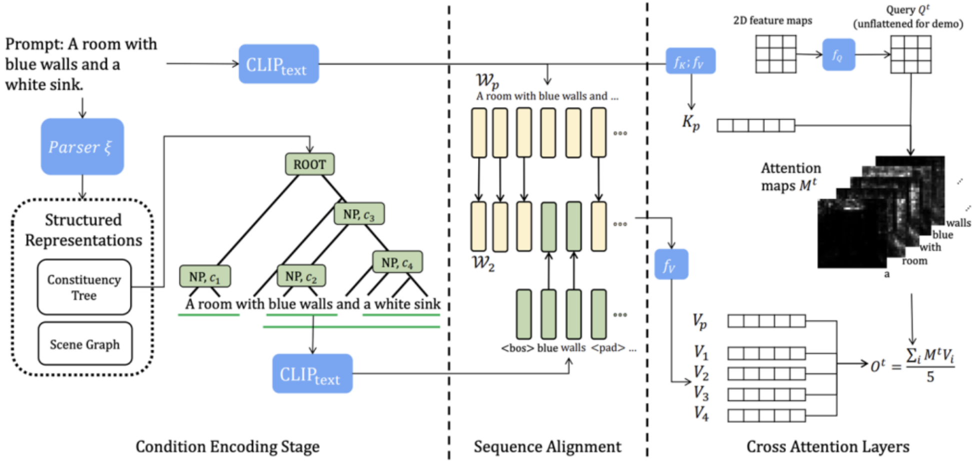
StructureDiffusion

A red car and a white sheep.	*A brown bench sits in front of an old white building*	*A blue backpack and a brown elephant*
Attribute leakage	Interchanged attributes	Missing objects

Image credit: “Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis”

StructureDiffusion

- Manipulating values in cross-attention based on linguistic parsing tree to enforce language structure
- Look at all noun phrases



Q: latent+duplicate(linear(t))
=> b*4096*320

K, V: text => b*77*768

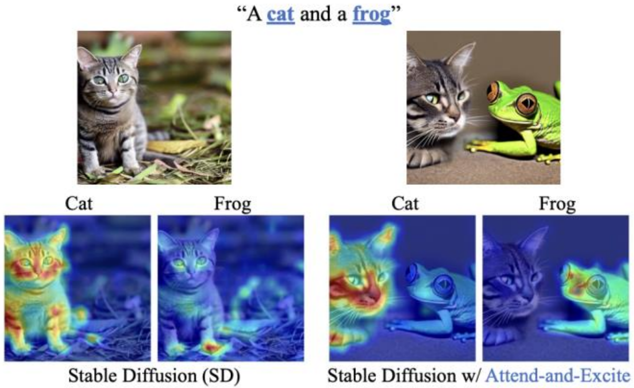
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V, \text{ with}$$

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

Image credit: “Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis”

Attend-and-Excite

- Enhance the maximal attention for the most neglected subject token
- Updates the latent with attention loss



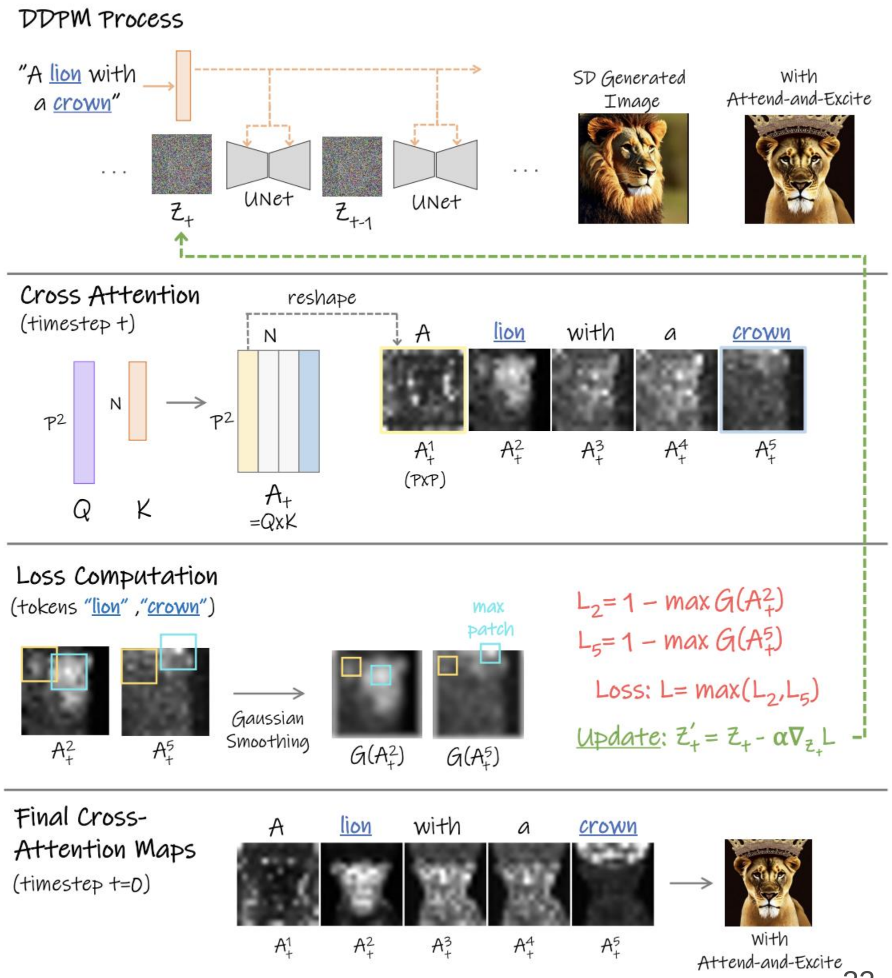


Image credit: "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models"

Attend-and-Excite

Image credit: "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models"

DDPO

- RL for optimizing diffusion models on different downstream objectives

Training process

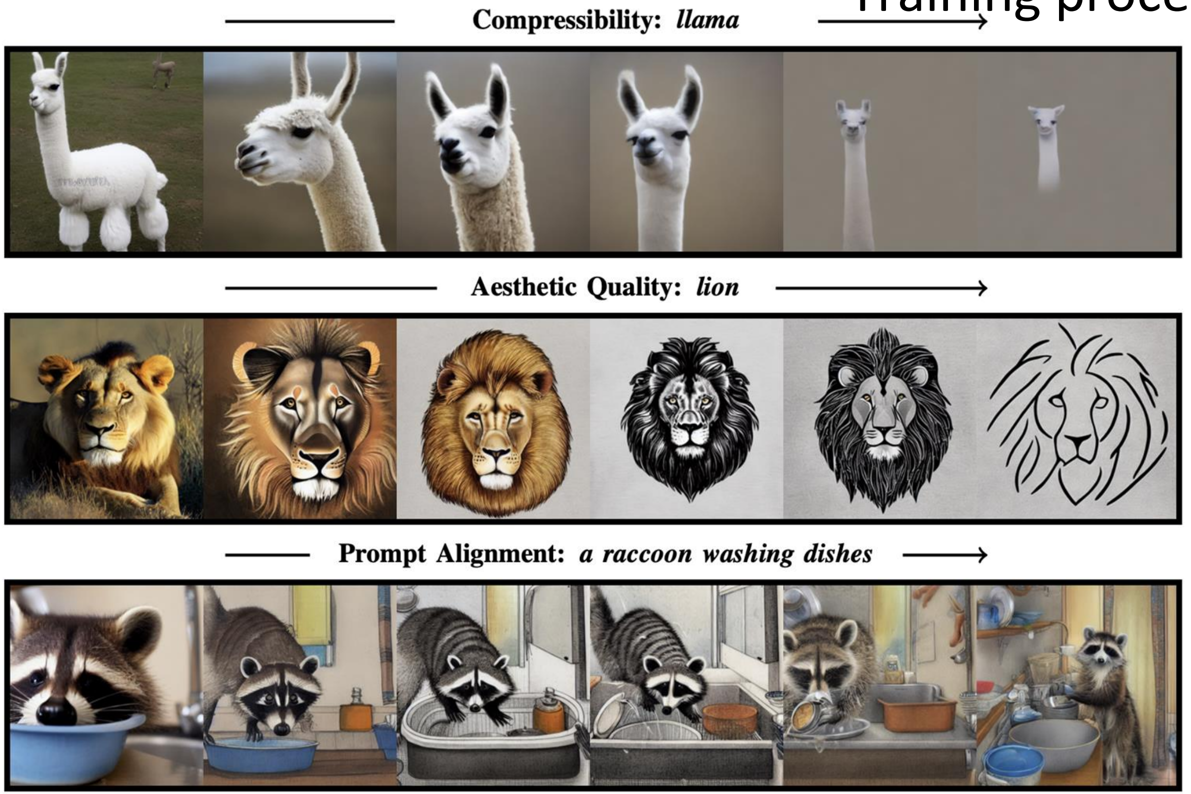
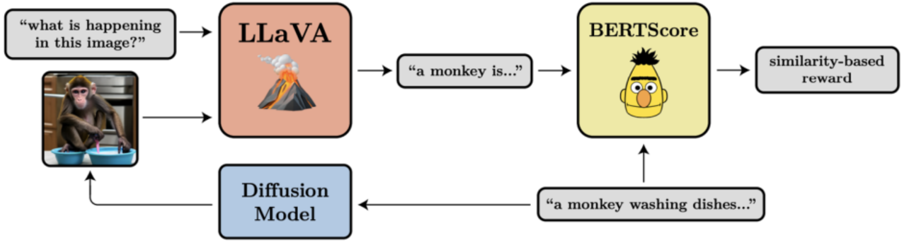


Image credit: “Training Diffusion Models with Reinforcement Learning”

DDPO

- VLM similarity reward to improve image-prompt alignment



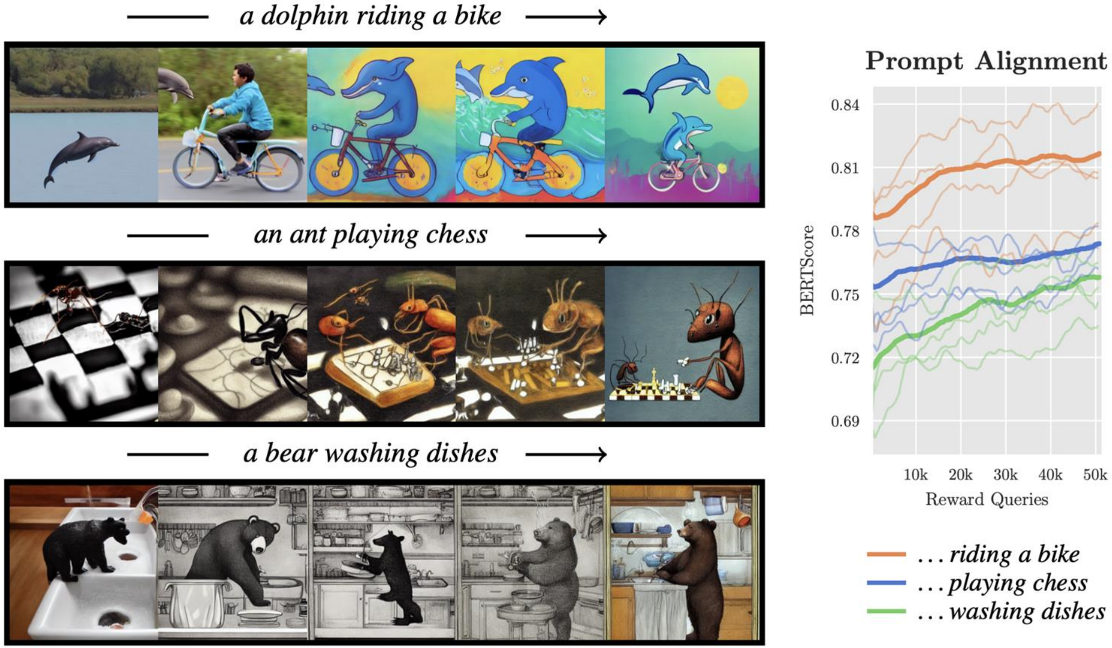


Image credit: "Training Diffusion Models with Reinforcement Learning"

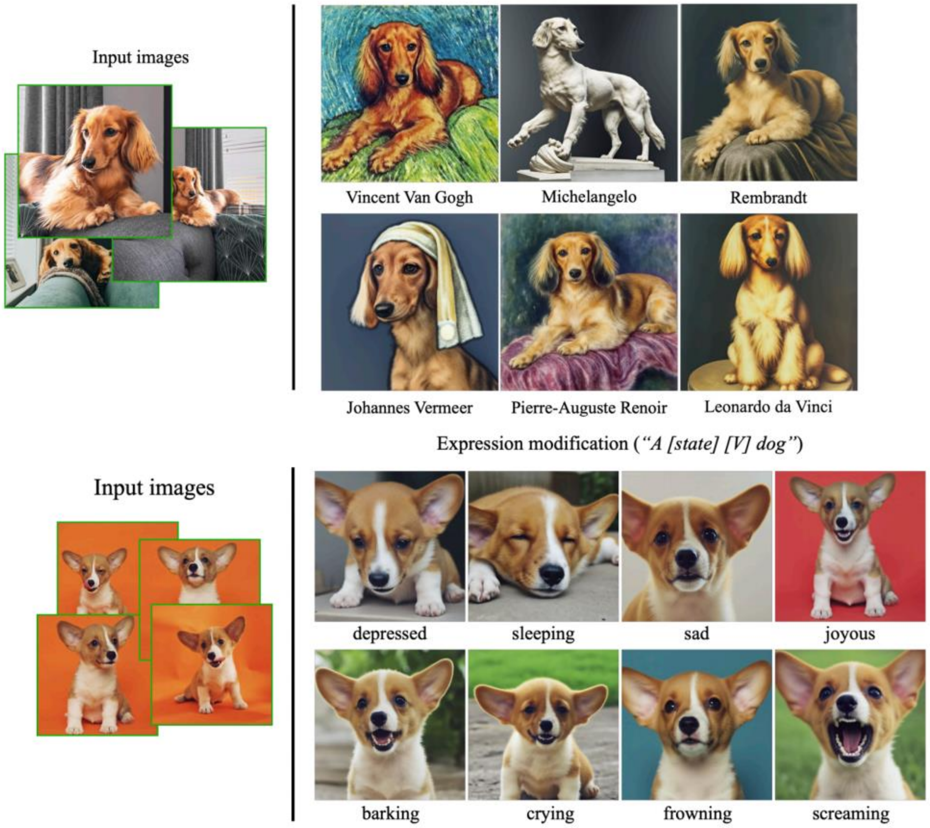
Concept Customization

- Single-concept customization
- Multi-concept customization
- Without test-time finetuning

[1] An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion
[2] DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation
[3] Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Models
[4] ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation
[5] Multi-Concept Customization of Text-to-Image Diffusion
[6] Break-A-Scene: Extracting Multiple Concepts from a Single Image
[7] Paint by Example: Exemplar-based Image Editing with Diffusion Models
[8] BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing
[9] Face0: Instantaneously Conditioning a Text-to-Image Model on a Face
[10] FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention
[11] Unified Multi-Modal Latent Diffusion for Joint Subject and Text Conditional Image Generation
[12] Re-Imagen: Retrieval-Augmented Text-to-Image Generator
[13] InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning
[14] Subject-driven text-to-image generation via apprenticeship learning

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

Single-Concept Customization



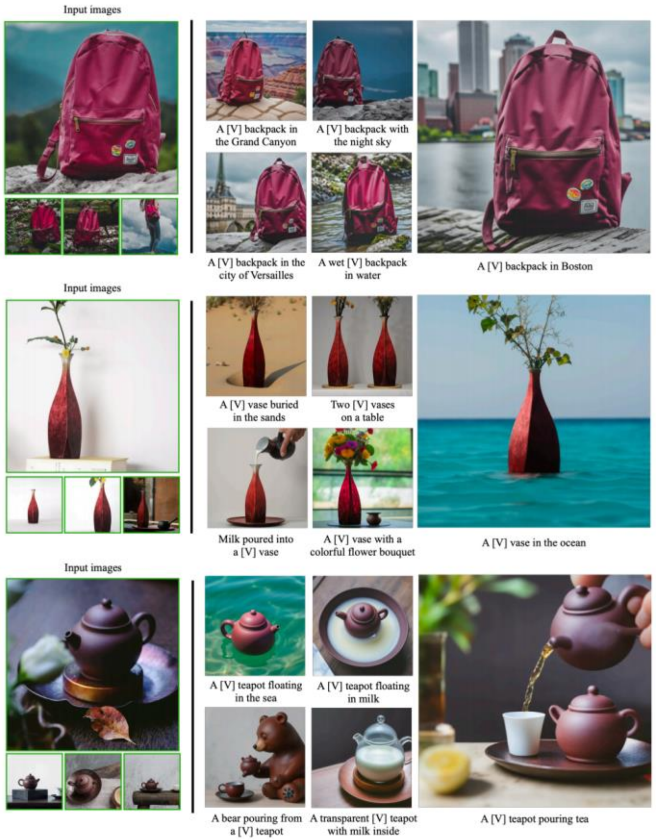
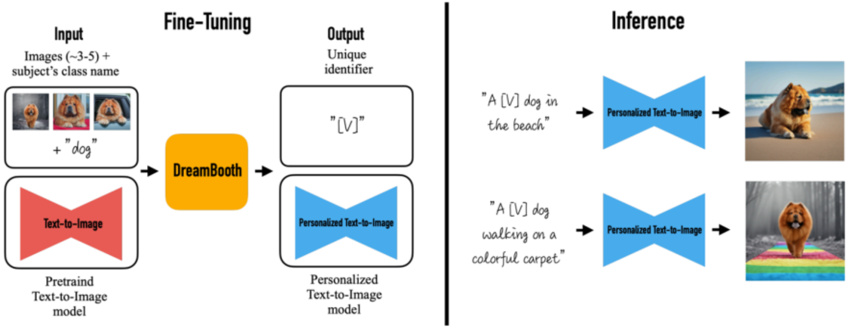


Image credit: “DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation”

Single-Concept Customization

- Tuning unique identifier [V] for customized subject
- Originally generated samples to alleviate forgetting



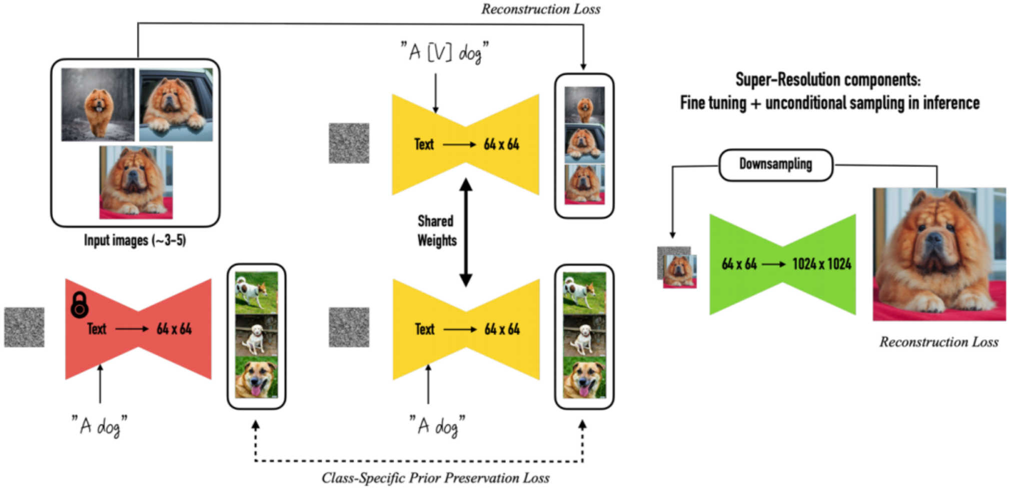
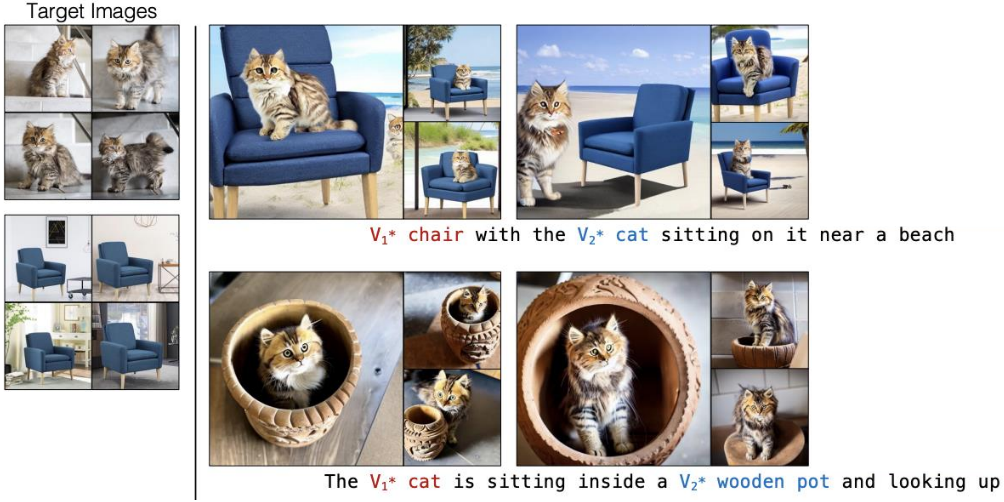


Image credit: "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation"

Multi-Concept Customization

- Multi-concept customization [V1], [V2], …
  from single image or multiple images



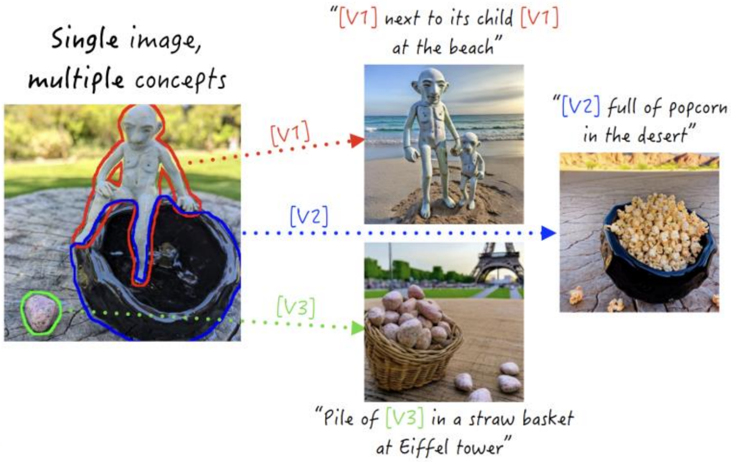
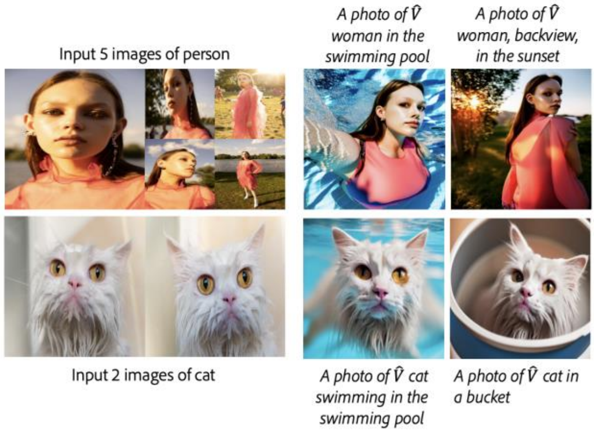


Image credit: "Multi-Concept Customization of Text-to-Image Diffusion"
"Break-A-Scene: Extracting Multiple Concepts from a Single Image"

Without Test-Time Finetuning

- Retrieve-augmented/ In-context generation
- Similar customization, but w/o test-time finetuning



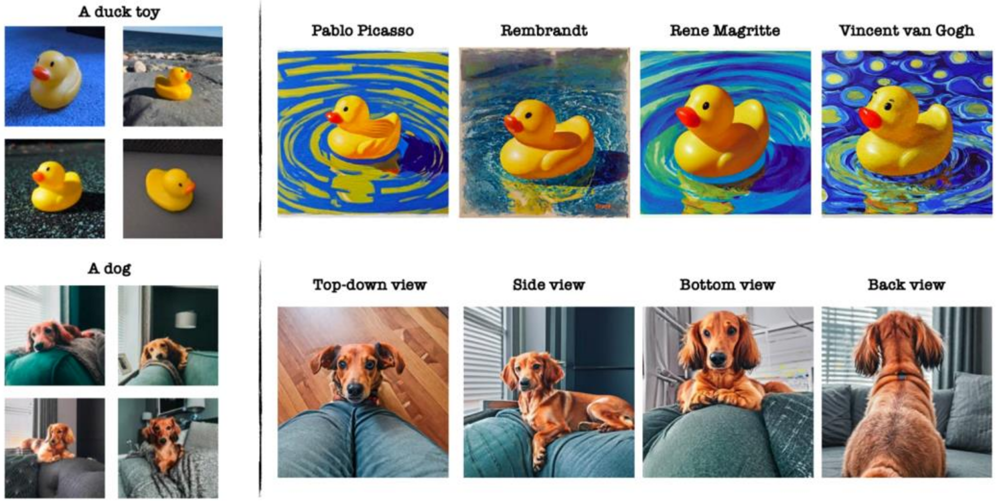


Image credit: "Subject-driven text-to-image generation via apprenticeship learning"
"InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning"

Agenda

- Text-to-image (T2I) basics
- Aligning human intentions in T2I generation
  - Controllable generation
  - Editing
  - Better following prompts
  - Concept customization
- Summary and discussion

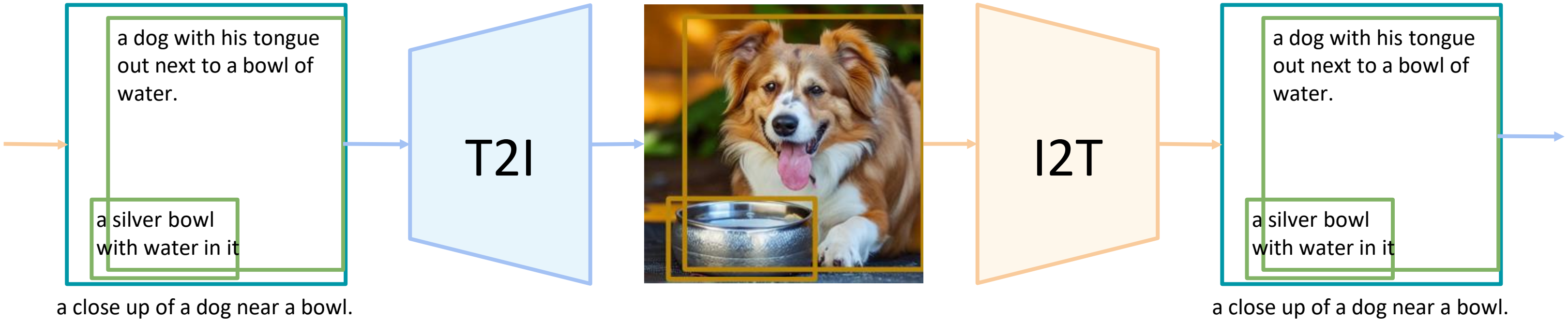
Discussion

Open-source v.s. Closed-source

Stable Diffusion
DeepFloyd AI

Parti
Pathways Autoregressive Text-to-Image Model
Imagen
DALL·E 2
runway
Adobe Firefly

Consuming and producing visual data: Understanding (I2T) and generation (T2I) loop



Thank you!