



Text to Image Models

Stephanie Fu

Image generation *with control*

Sound
Mask
Point
Text
Layout
Image
Sketch



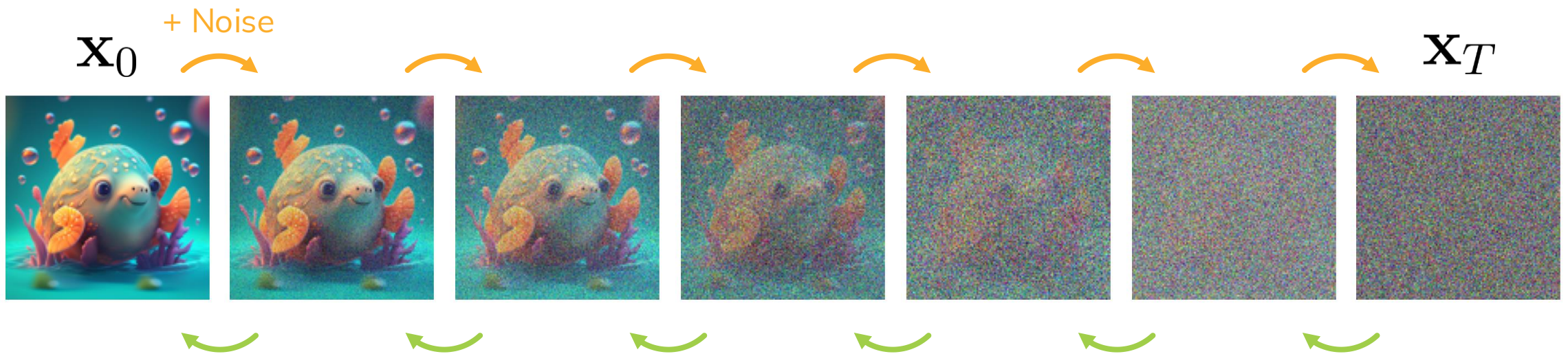
*"a hyperrealistic
peppa pig"*



Diffusion model intuition

Creating is hard... but destroying is easy!

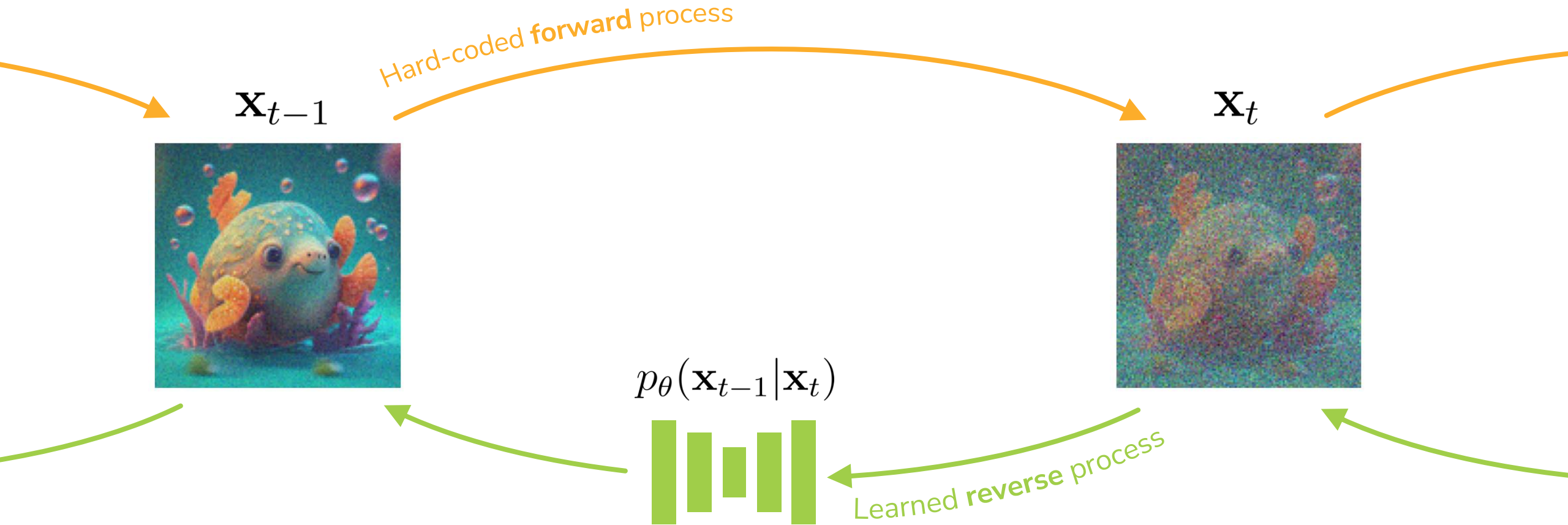
Let's **destroy** a clean image \mathbf{X}_0



and learn $p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t)$ to **reverse** the process

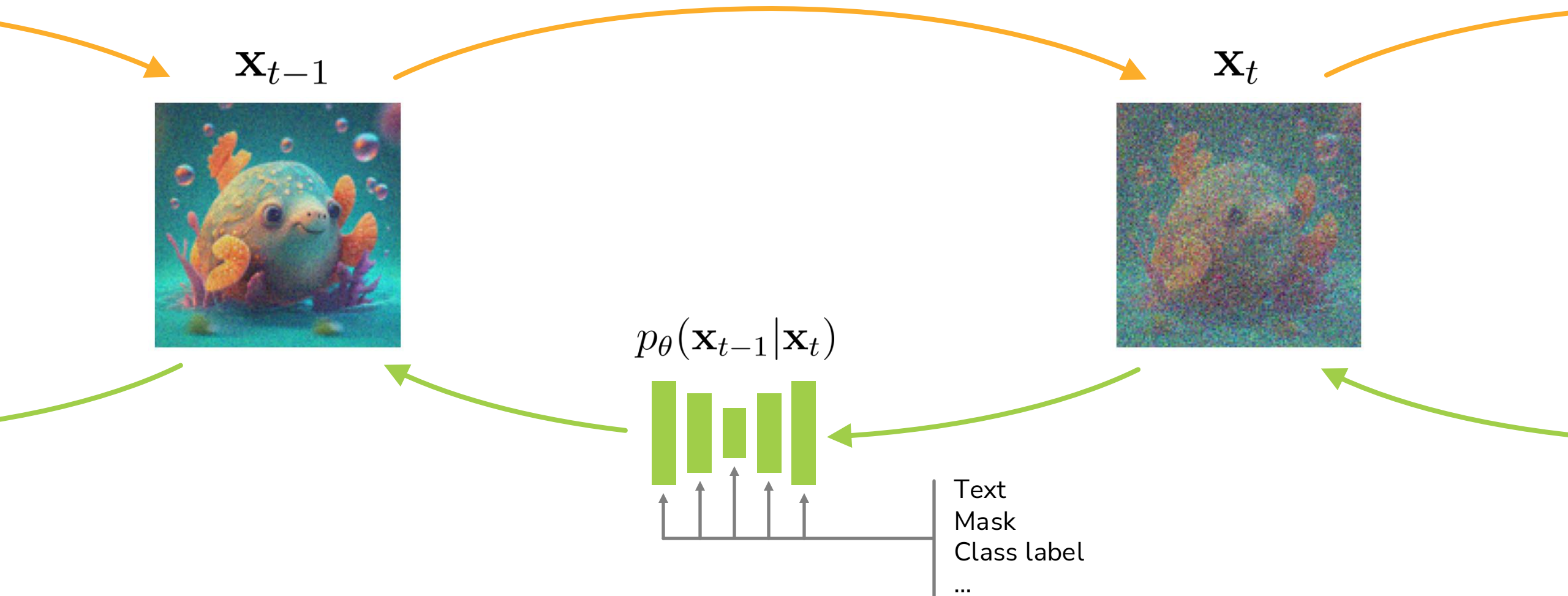
Diffusion model intuition

Predicting noise with an autoencoder



Diffusion model intuition

Adding a conditioning signal



GLIDE: An early instance of text-conditioned diffusion

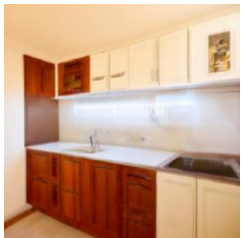
CLIP guidance



"a green train is coming down the tracks"



"a group of skiers are preparing to ski down a mountain"

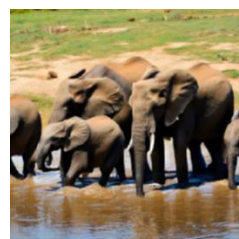
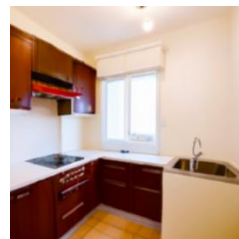
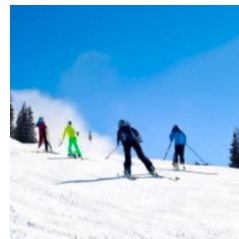
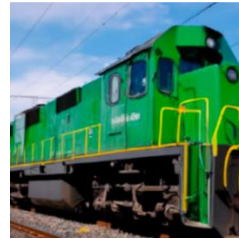


"a small kitchen with a low ceiling"



"a group of elephants walking in muddy water"

Classifier-free guidance



- 3.5B parameter diffusion model with text guidance
- CFG generally more realistic and preferred by human raters
- Didn't invent classifier- or classifier-free guidance! But showed CFG working at scale

Classifier-Free Diffusion Guidance

Jonathan Ho
Google Research

Tim Salimans
Google Research

GLIDE: An early instance of text-conditioned diffusion

CLIP guidance

noise prediction

$$\epsilon_{\theta}(x_t|y)$$



mean of noise prediction $\leftarrow s \cdot \text{CLIP}(x_t, y)$



Need a separate model (e.g., CLIP) trained on noisy images

Classifier-free guidance

predict noise (unconditional)

$$\epsilon_{\theta}(x_t|\emptyset)$$



predict noise (conditional)

$$\epsilon_{\theta}(x_t|y)$$



Training: apply dropout on condition

Swap out condition y for no condition (\emptyset) with probability p

Don't need separate model!

noise prediction during sampling:

$$\epsilon_{\theta}(x_t|\emptyset) + s \cdot (\underbrace{\epsilon_{\theta}(x_t|\emptyset) - \epsilon_{\theta}(x_t|y)}_{\text{step a little towards the text-conditioned image}})$$

“step a little towards the text-conditioned image”

Imagen

Model



Use a pretrained text encoder

CLIP T5 BERT



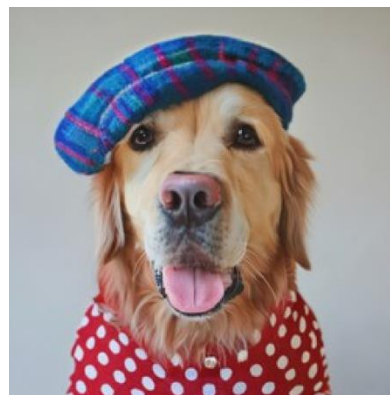
Progressively increase fidelity

Benchmark: DrawBench

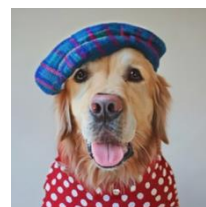
Challenging text prompts, testing:

object count
color text in scene
unusual interactions spatial relation
rare words misspelled prompts

1024 x 1024



256 x 256

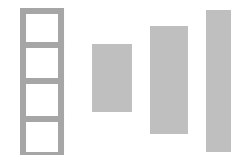


64 x 64



Text-to-Image DM

Super-resolution DM



"a Golden Retriever dog wearing a blue checkerboard beret and red dotted turtleneck"

Latent Diffusion Model (LDM)

Diffusion in feature space

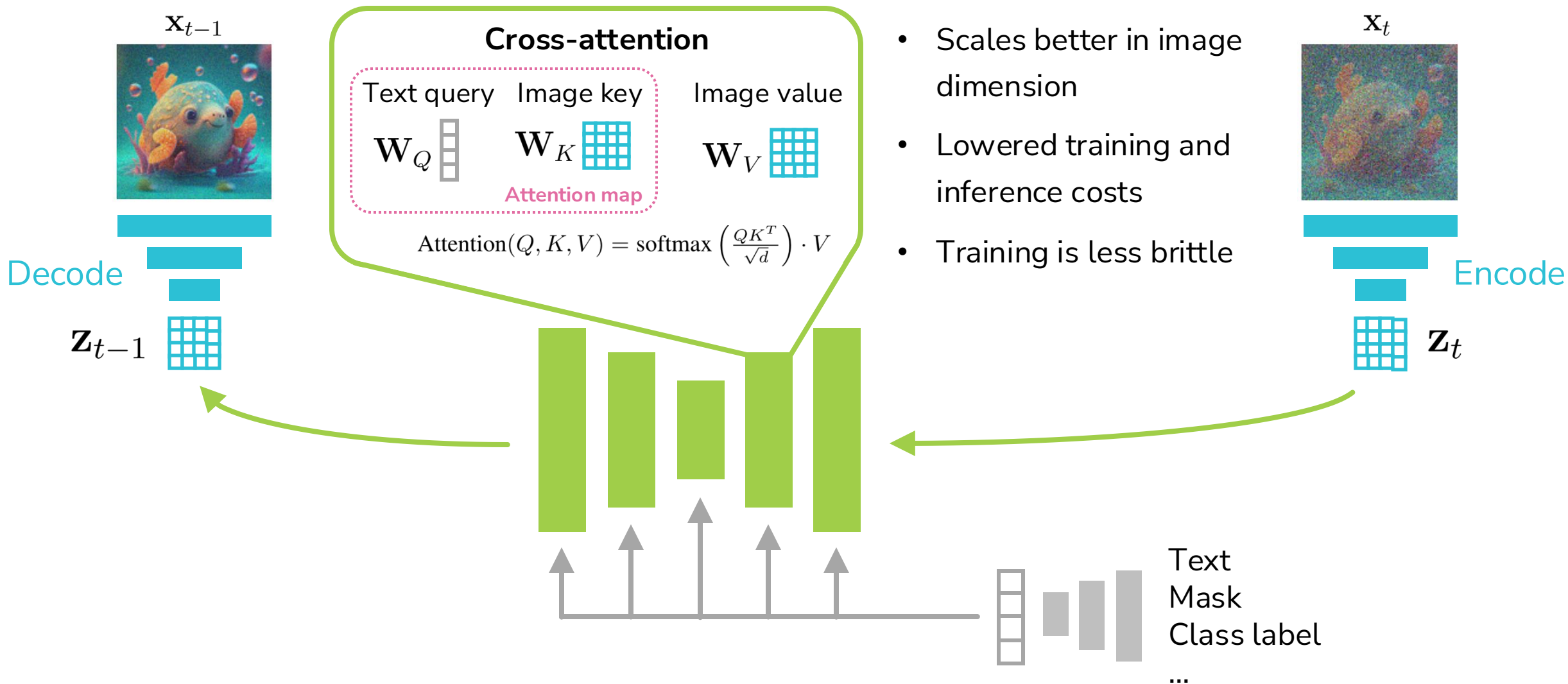


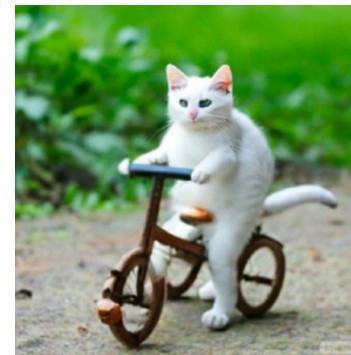
Image generation with control

Prompt-to-Prompt

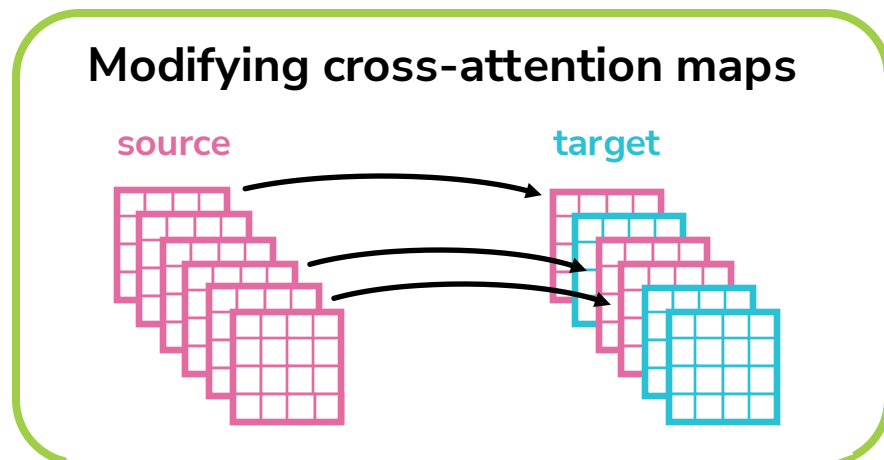
more precise

Source

Target



"photo of a cat riding a ~~bicycle~~
car"



X_{t-1}

X_t

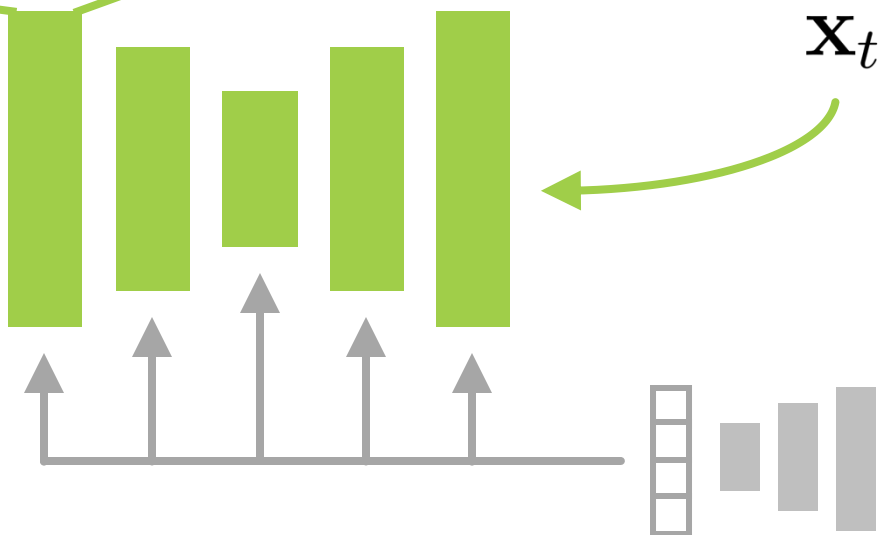
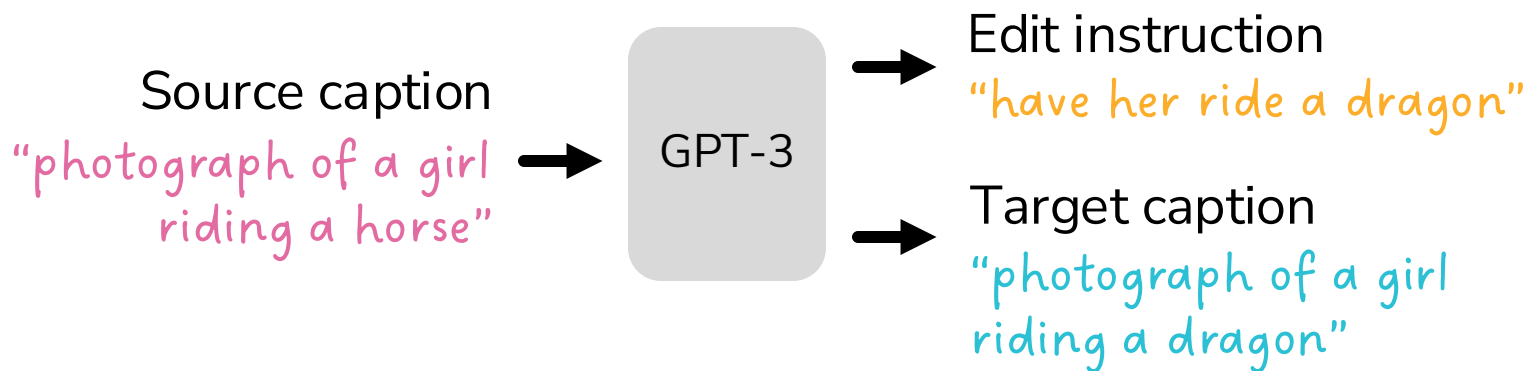


Image generation with control

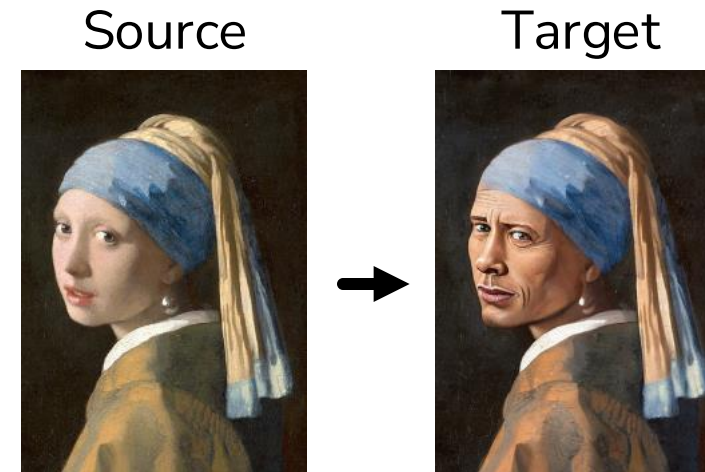
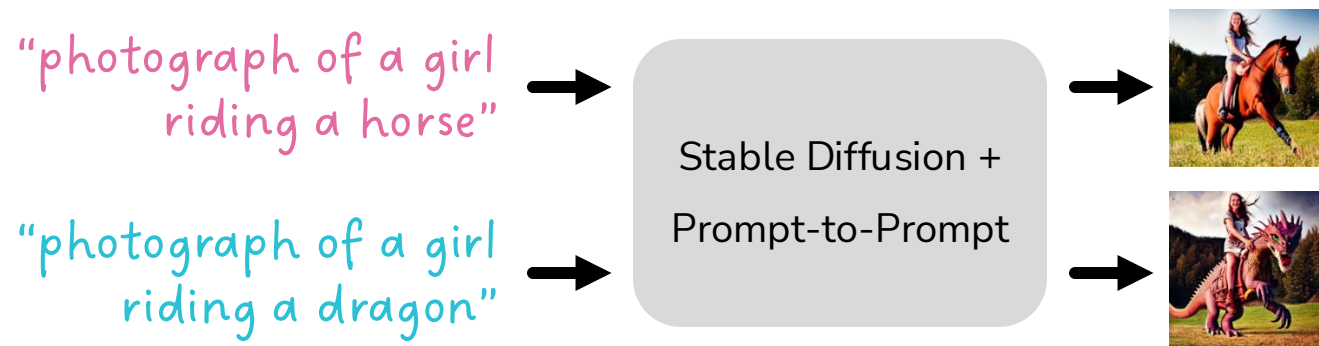
InstructPix2Pix

more precise

Step 1: Generate text edits



Step 2: Generate paired images



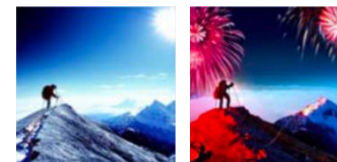
"turn her into Dwayne the Rock Johnson"

Resulting training dataset

"have her ride a dragon"



"make it lit by fireworks"



"color the cars pink"

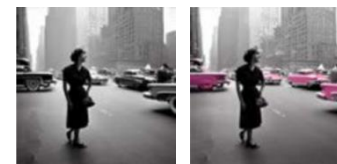


Image generation with control

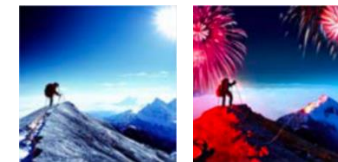
InstructPix2Pix

more precise

"have her ride a dragon"



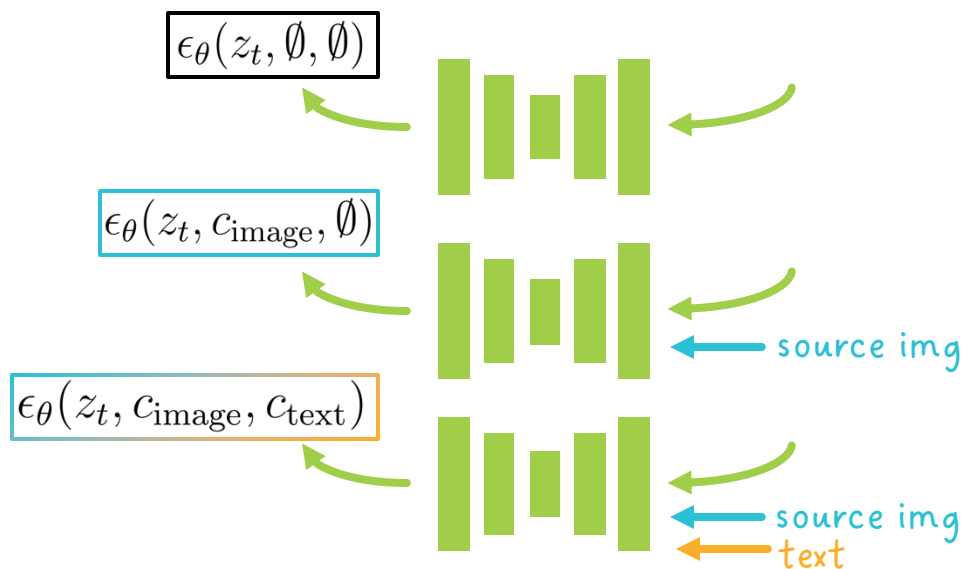
"make it lit by fireworks"



Step 3: Train latent diffusion model

Training dataset


Classifier-free guidance for both image and edit instruction




$$\begin{aligned}
 \epsilon_{\theta}(z_t, C_{image}, C_{text}) = & \epsilon_{\theta}(z_t, \emptyset, \emptyset) \\
 & + s_{image} \cdot \epsilon_{\theta}(z_t, C_{image}, \emptyset) - \epsilon_{\theta}(z_t, \emptyset, \emptyset) \\
 & + s_{text} \cdot \epsilon_{\theta}(z_t, C_{image}, C_{text}) - \epsilon_{\theta}(z_t, C_{image}, \emptyset)
 \end{aligned}$$

More on controllable image generation

Awesome Controllable T2I Diffusion Models






 **Awesome-Controllable-T2I-Diffusion-Models** Public Watch 45 Fork 26 Star 889

main 1 Branch Tags

 **caopulan** update 2024-10-07 56e2727 · 3 weeks ago 140 Commits

assets	update 2024-10-07	3 weeks ago
CITATION.cff	add CITATION.cff	7 months ago
LICENSE	update license	last year
README.md	update 2024-10-05	3 weeks ago

README MIT license


    

About
A collection of resources on controllable generation with text-to-image diffusion models.


[awesome](#) [personalization](#)
[awesome-list](#) [text-to-image](#)
[diffusion-models](#) [controllable-generation](#)
[spatial-controls](#) [multi-concept](#)

- Readme
- MIT license
- Cite this repository
- Activity
- Custom properties

Awesome Text-to-Image Studies

 **awesome-text-to-image-studies** Public Watch 12 Fork 22 Star 395

main 1 Branch Tags

 **AlonzoLeeeooo** Update EmoGen abf20c4 · last week 111 Commits

github-materials	Update topic: Diffusion Models Meet Federated Learning	6 months ago
topics	Update LinFusion	last month
LICENSE	Initial commit	8 months ago
README.md	Update EmoGen	last week
reference.bib	Update EmoGen	last week

About
A collection of awesome text-to-image generation studies.

[artificial-intelligence](#) [text-to-image](#)
[diffusion-models](#) [text-to-image-diffusion](#)
[text-to-image-ai](#)

- Readme
- MIT license
- Activity
- 395 stars