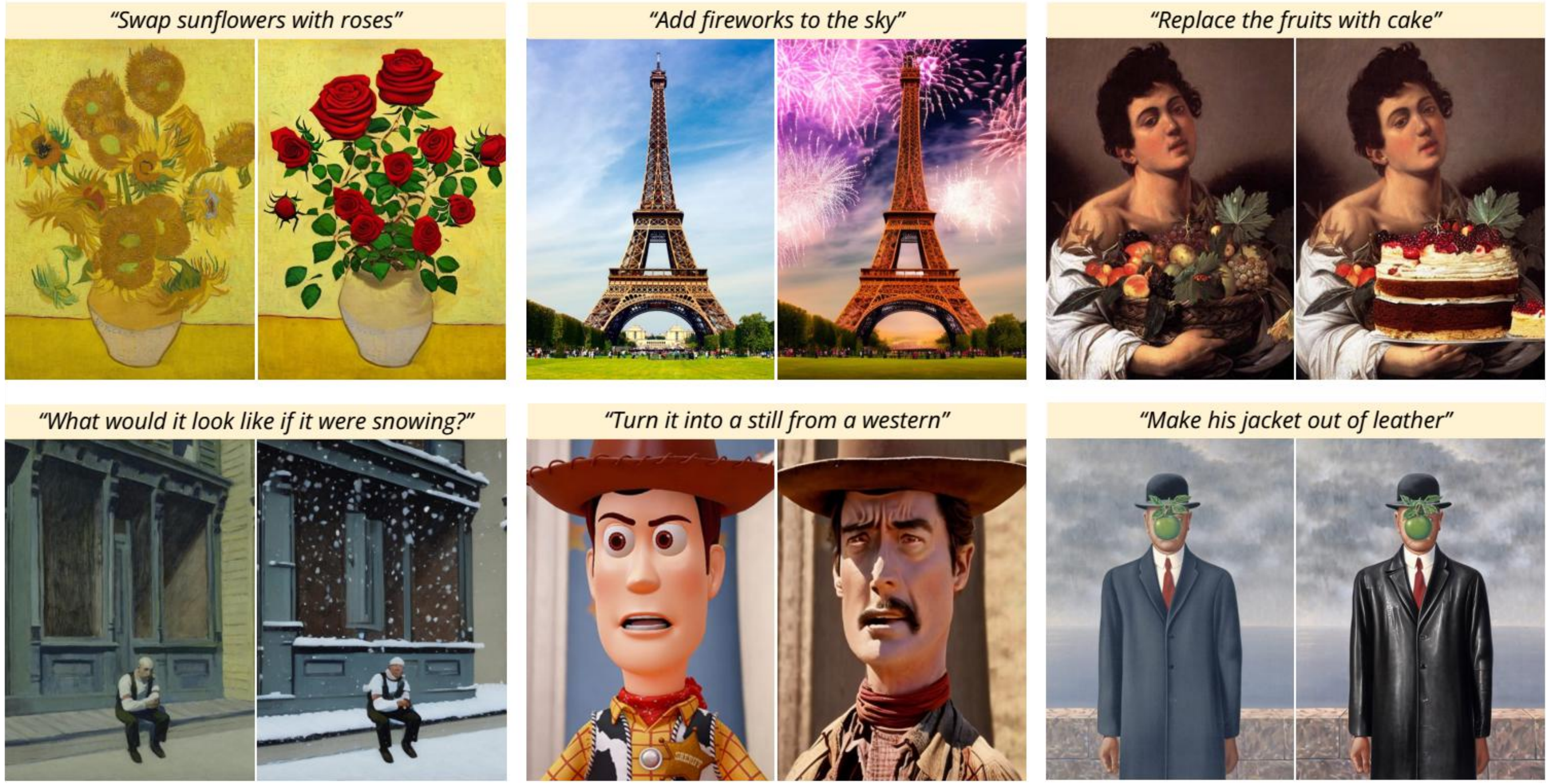
InstructPix2Pix: Learning to Follow Image Editing Instructions

Tim Brooks* Aleksander Holynski* Alexei A. Efros
University of California, Berkeley



Wang Dongsheng
2023.3.3

Outline

- ❑ Language instructed image editing
- ❑ Latent diffusion model
- ❑ InstructPix2Pix
- ❑ Experiments

❖ Language instructed image editing

Input

Add boats on the water

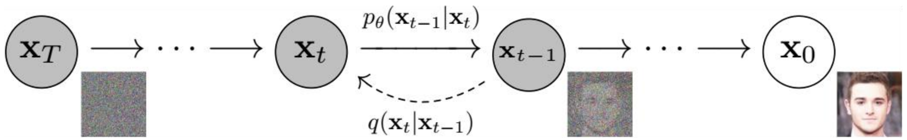
Instruction

Output

- ❑ Understand the human instruction
- ❑ Image consistency

❖ Latent diffusion model (stable diffusion)

❑ Diffusion models



$$\mathbb{E}_q\left[\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\parallel p(\mathbf{x}_T))}_{L_T}+\sum_{t>1}\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)\parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}}\underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}\right]$$

❑ Limitations of diffusion models

- Works on pixel space (low inference speed, very high training costs)
- Low-resolution generation (64*64 images)

❖ Latent diffusion model (stable diffusion)

❑ Framework of the latent diffusion model (two stage piplines)

1. Perceptual image compression (any pretrained autoencoder works)

$$z = \mathcal{E}(x) \qquad z \in \mathbb{R}^{h \times w \times c} \qquad x \in \mathbb{R}^{H \times W \times 3}$$

$$\tilde{x} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x))$$

2. Diffusion model on latent space z

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

$$\Downarrow$$

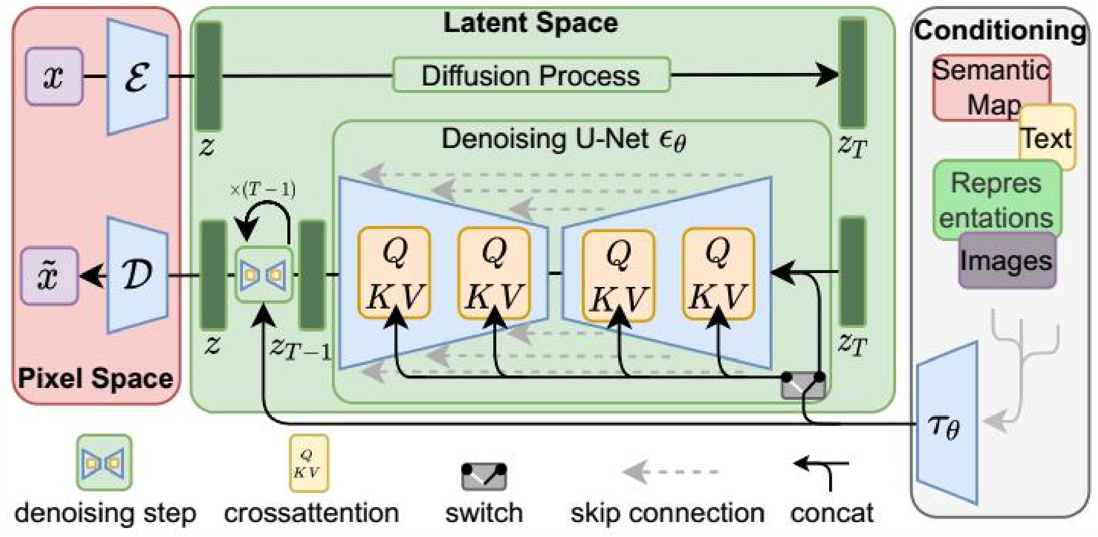
$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon \sim \mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\Big]$$

❖ Latent diffusion model (stable diffusion)

❑ Conditional latent diffusion model via cross-attention layer

How to model $p(x_{t-1}|x_t, t, y)$ given the condition y ? e.g, language prompt.

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ K = W_K^{(i)} \cdot \tau_\theta(y),\ V = W_V^{(i)} \cdot \tau_\theta(y). \quad \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$$

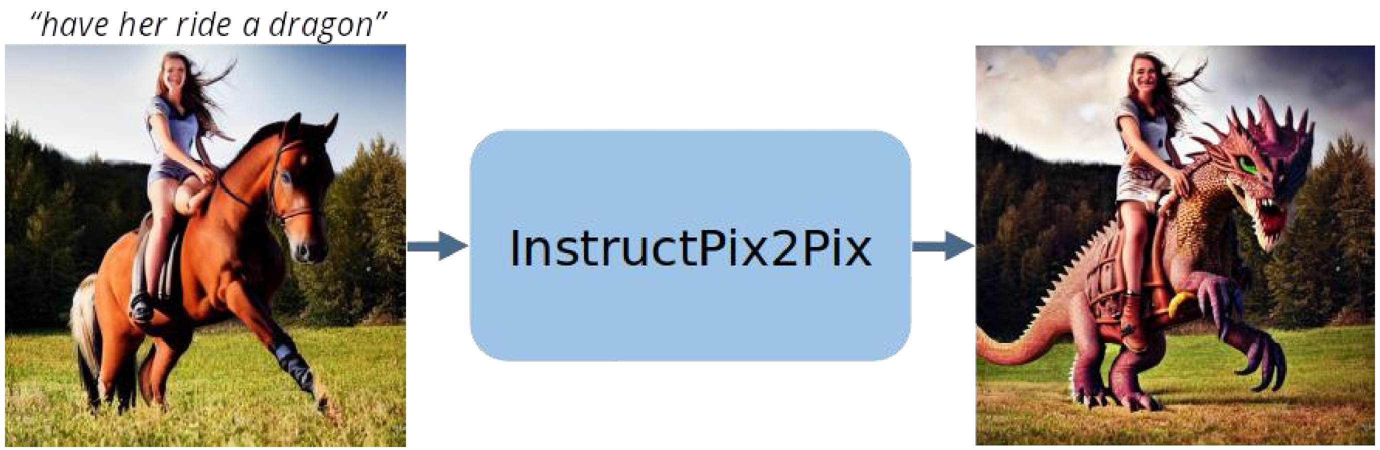


❖ Latent diffusion model (stable diffusion)

❑ Main contributions

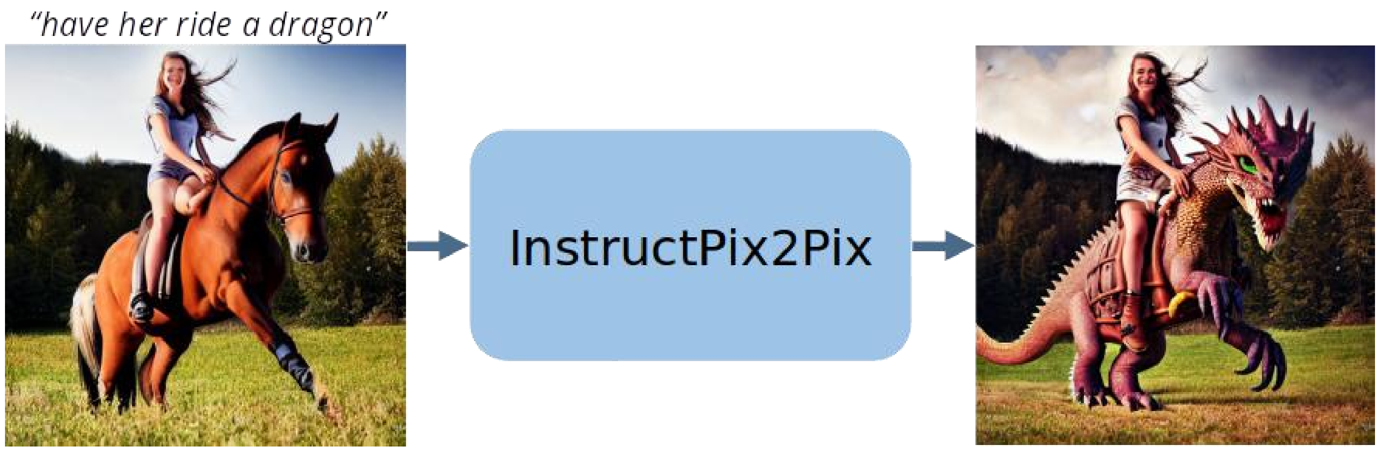
- ✓ Less GPU resource (>11G)
- ✓ pretrained weights and code at github
- ✓ Two stage pipline for training diffustion model at lalent space.

- Customize big model in our cases (InstructPix2Pix)
  - Train on a large supervised dataset of paired images and instructions
  - ... but where does this supervised dataset come from?



❖ Customize big model in our cases (InstructPix2Pix)

- ❑ Train on a large supervised dataset of paired images and instructions
- ❑ ... but where does this supervised dataset come from
- ❑ Combine knowledge of large pretrained models to generate training data



❖ Customize big model in our cases (InstructPix2Pix)

❑ Generating caption edits with GPT-3

- Finetune GPT-3 to generate instructions and before/after captions.
- Train on 700 human-written image editing instructions.
- Then generate >450,000 examples (providing LAION captions as input).

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	*Yefim Volkov, Misty Morning*	*make it afternoon*	*Yefim Volkov, Misty Afternoon*
	girl with horse at sunset	*change the background to a city*	*girl with horse at sunset in front of city*
	painting-of-forest-and-pond	*Without the water.*	*painting-of-forest*
	...	...	...
GPT-3 generated (450,000 edits)	*Alex Hill, Original oil painting on canvas, Moonlight Bay*	*in the style of a coloring book*	*Alex Hill, Original coloring book illustration, Moonlight Bay*
	The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it	*Add a giant red dragon*	*The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead*
	Kate Hudson arriving at the Golden Globes 2015	*make her look like a zombie*	*Zombie Kate Hudson arriving at the Golden Globes 2015*
	...	...	...

Highlighted text is generated by GPT-3.

❖ Customize big model in our cases (InstructPix2Pix)

- Generating pairs of images from captions
  - Use a pretrained text-to-image model to generate examples.
  - Leverage Prompt-to-Prompt method to make images look similar.

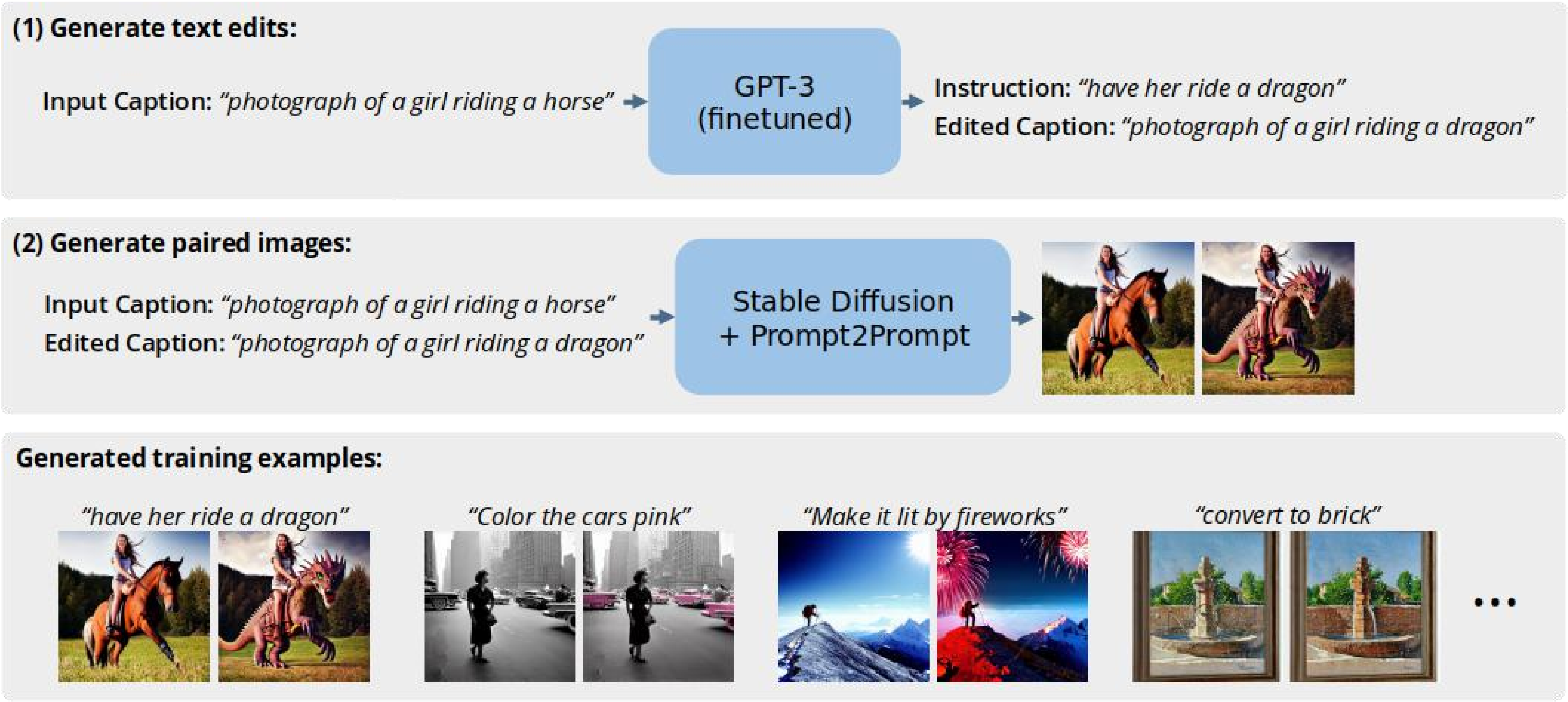
"Photo of a cat riding on a bicycle."

"Photo of a cat riding on a car."

→

❖ Customize big model in our cases (InstructPix2Pix)

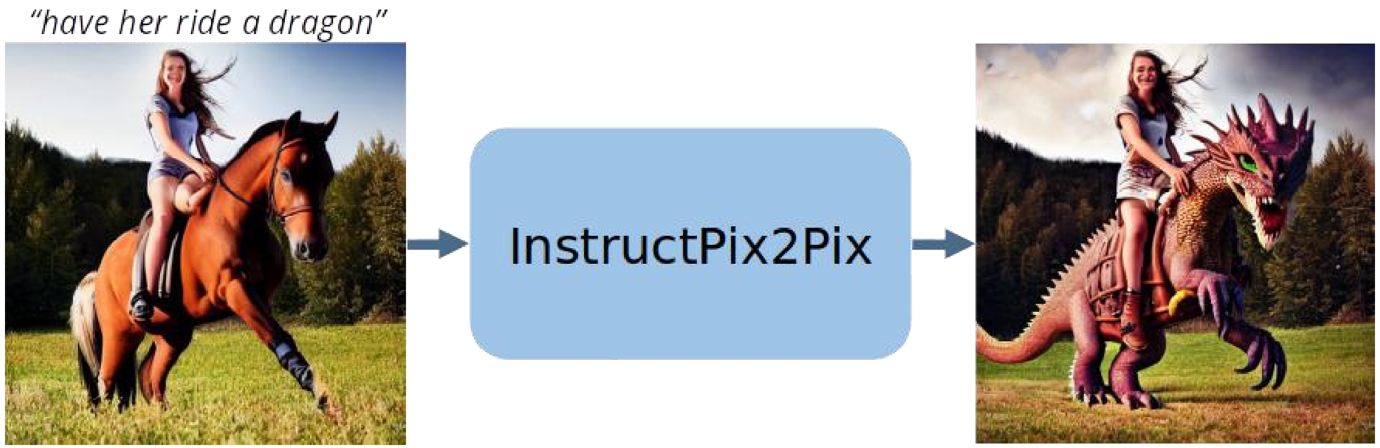
- ❑ Generating paired training data (pre-trained GPT-3, pre-trained text-image diffusion model, 700 human-written edits)



❖ Customize big model in our cases (InstructPix2Pix)

❑ Traing an image editing diffusion model

- Now it is a supervised learning problem!
- Finetune Stable Diffusion on generated training data.



$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(c_I),c_T,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T))\|_2^2\Big]$$

cI: image condition

c_T: instruction condition

❖ Customize big model in our cases (InstructPix2Pix)

❑ And generalization to real images and instructions

- Trained only on generated images and instructions.
- At inference, generalizes to real images and human written instructions

❖ Customize big model in our cases (InstructPix2Pix)

- ❑ And generalization to real images and instructions

Input

"Add boats on the water"

"Replace the mountains with a city skyline"

Input

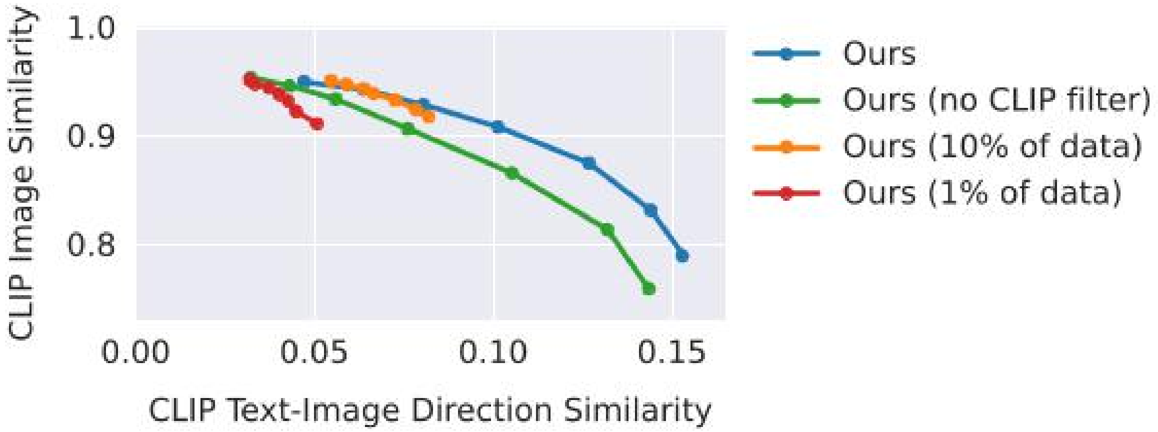
"It is now midnight"

"Add a beautiful sunset"

❖ Customize big model in our cases (InstructPix2Pix)

❑ Data scale and quality

- How well does output image match input image?
- How well does change in images match change in captions?



❖ Customize big model in our cases (InstructPix2Pix)

❑ Bias in generated images

- InstructPix2Pix learns biases such as correlations between profession and gender.

Input | *"Make them look like flight attendants"* | *"Make them look like doctors"*

❖ Customize big model in our cases (InstructPix2Pix)

❑ Failure cases

- Unable to alter viewpoint or spatial layout.
- Too significant of change (needs tuning CFG to prevent).
- Difficulty isolating objects.

"Zoom into the image"

"Move it to Mars"

"Color the tie blue"

"Have the people swap places"