

InstructPix2Pix

Learning to Follow Image Editing Instructions

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





Search



AGI

Attachment: 1 Image

6:14 PM

To: AGI



iMessage
Today 6:14 PM



iMessage



"Swap sunflowers with roses"



"Add fireworks to the sky"



"Replace the fruits with cake"



"What would it look like if it were snowing?"



"Turn it into a still from a western"



"Make his jacket out of leather"



Our approach

- Leverage large pretrained models to generate supervised training data.
- Train a diffusion model to directly edit images.

"have her ride a dragon"



InstructPix2Pix



Generating training data

(1) Generate text edits:

Input Caption: *"photograph of a girl riding a horse"* →

GPT-3
(finetuned)

Instruction: *"have her ride a dragon"*

Edited Caption: *"photograph of a girl riding a dragon"*

Generating training data

(1) Generate text edits:

Input Caption: *"photograph of a girl riding a horse"* →

GPT-3
(finetuned)

Instruction: *"have her ride a dragon"*

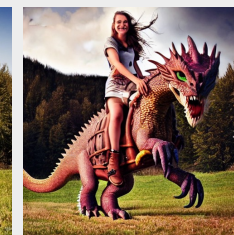
Edited Caption: *"photograph of a girl riding a dragon"*

(2) Generate paired images:

Input Caption: *"photograph of a girl riding a horse"*

Edited Caption: *"photograph of a girl riding a dragon"* →

Stable Diffusion
+ Prompt2Prompt



Generating training data

(1) Generate text edits:



(2) Generate paired images:



Generated examples:

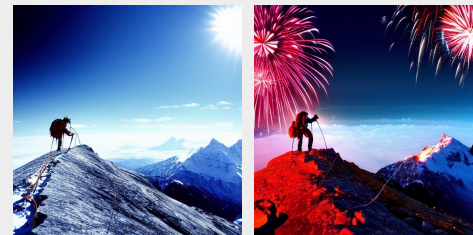
"have her ride a dragon"



"Color the cars pink"



"Make it lit by fireworks"



"convert to brick"



...

Generalization to real images

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

"Make it a grocery store"





Input



"Apply face paint"



"What would she look like as a bearded man?"



"Put on a pair of sunglasses"



"She should look 100 years old"



"What if she were in an anime?"



"Make her terrifying"



"Make her more sad"



"Make her James Bond"



"Turn her into Dwayne The Rock Johnson"



Input



“Add boats on the water”



“Replace the mountains with a city skyline”



Input



“It is now midnight”



“Add a beautiful sunset”

Thank you!



"Make it Paris"



"Make it Hong Kong"



"Make it Manhattan"



"Make it Prague"



"Make it evening"



"Put them on roller skates"



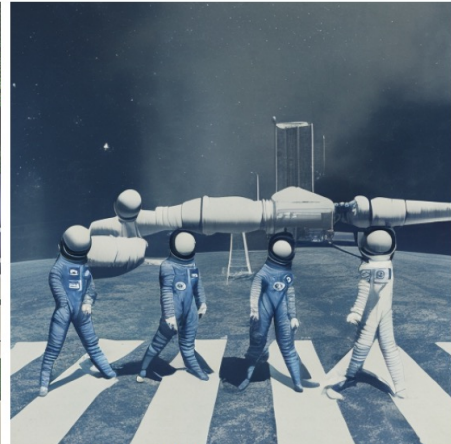
"Turn this into 1900s"



"Make it underwater"



"Make it Minecraft"



"Turn this into the space age"



"Make them into Alexander Calder sculptures"



"Make it a Claymation"





Search



AGI

Attachment: 1 Image

6:14 PM

To: AGI



Today 6:14 PM



Could you make this image look like spring time?

Delivered



iMessage



Related work

SEDIT: GUIDED IMAGE SYNTHESIS AND EDITING WITH STOCHASTIC DIFFERENTIAL EQUATIONS

Chenlin Meng¹ Yutong He¹ Yang Song¹ Jiaming Song¹
Jiajun Wu¹ Jun-Yan Zhu² Stefano Ermon¹
¹Stanford University ²Carnegie Mellon University

PROMPT-TO-PROMPT IMAGE EDITING WITH CROSS-ATTENTION CONTROL

Amir Hertz^{*1,2}, Ron Mokady^{*1,2}, Jay Tenenbaum¹, Kfir Aberman¹, Yael Pritch¹, and Daniel Cohen-Or^{*1,2}

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

Imagic: Text-Based Real Image Editing with Diffusion Models

Bahjat Kawar^{*1,2} Shiran Zada^{*1} Oran Lang¹ Omer Tov¹
Huiwen Chang¹ Tali Dekel^{1,3} Inbar Mosseri¹ Michal Irani^{1,3}
¹Google Research ²Technion ³Weizmann Institute of Science

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

Nataniel Ruiz^{*,1,2} Yuanzhen Li¹ Varun Jampani¹
Yael Pritch¹ Michael Rubinstein¹ Kfir Aberman¹
¹ Google Research ² Boston University

Text2LIVE: Text-Driven Layered Image and Video Editing

Omer Bar-Tal^{1*}, Dolev Ofri-Amar^{1*}, Rafail Fridman^{1*},
Yoni Kasten², and Tali Dekel¹

¹ Weizmann Institute of Science ² NVIDIA Research

Related work: DreamBooth

- Finetunes on many images of a subject or concept.
- Capable of generating new images that include the subject.



Related work: Imagic

- Finetunes on a single image.
- Generates a new similar image from a description of the output.



Optimize latents



Finetune model



*“A children’s drawing of
a forest”*



Related work: Prompt-to-Prompt

- Generates two similar images given two prompts as input.
- Copies intermediate attention maps to ensure images look similar.

"Photo of a cat riding on a bicycle."

"Photo of a cat riding on a car."





To: AGI



Search



AGI

6:15 PM

Attachment: 1 Image



And have the sky be blue

make it look more like it's from a fairy tale

Delivered



iMessage



Goals

- Edit real images (not generated).
- Tell the model exactly what edit to make as a written instruction.
- Not require extra input (full descriptions, extra images, drawn masks).
- Perform edit in forward pass without need for inversion/finetuning.

Our approach

- Train a large diffusion model to directly edit images.
- Train on a large supervised dataset of paired images and instructions.

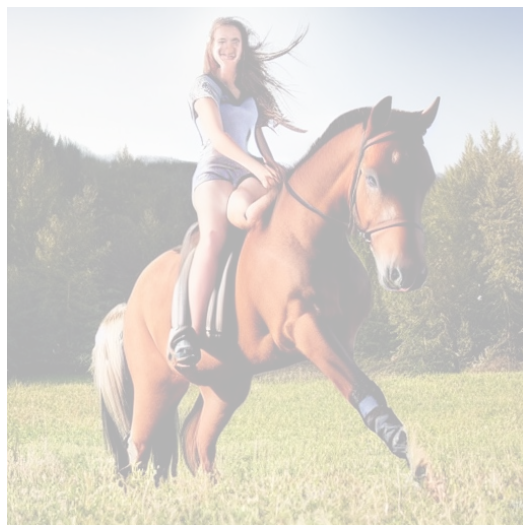
"have her ride a dragon"



Our approach

- Train a large diffusion model to directly edit images.
- Train on a large supervised dataset of paired images and instructions.
- ...but where does this supervised dataset come from?

"have her ride a dragon"



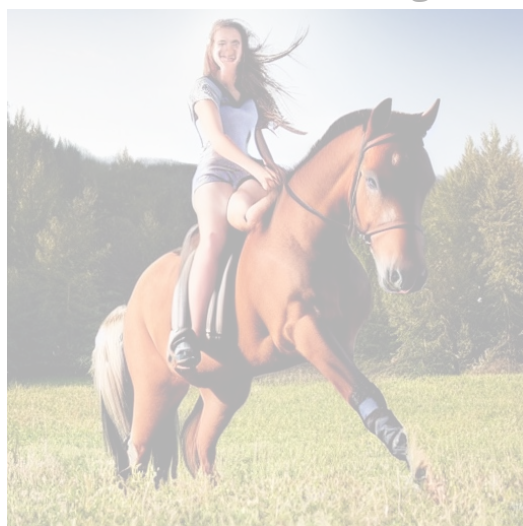
InstructPix2Pix



Our approach

- Train a large diffusion model to directly edit images.
- 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.

"have her ride a dragon"




InstructPix2Pix



Editing captions with GPT-3

← → ↻ beta.openai.com/playground

 Overview Documentation Examples Playground

Playground

Input: Woman with long dark hair sitting in a tree
Edit: Make it a painting by Georges Seurat
Output: A painting of a woman with long dark hair sitting in a tree by Georges Seurat

Input: An image of a person holding a cup of coffee
Edit: Turn the cup of coffee into a bowl of soup
Output: An image of a person holding a bowl of soup

Input: American football player on the field during training
Edit: Have them play soccer
Output: American soccer player on the field during training

Input: Landscape photograph of lake with mirror-like reflection, summer green trees
Edit: Change the season to autumn|

Submit ↻ ↺ ⌛

Generating text edits with GPT-3

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	<i>Yefim Volkov, Misty Morning</i>	<i>make it afternoon</i>	<i>Yefim Volkov, Misty Afternoon</i>
	<i>girl with horse at sunset</i>	<i>change the background to a city</i>	<i>girl with horse at sunset in front of city</i>
	<i>painting-of-forest-and-pond</i>	<i>Without the water.</i>	<i>painting-of-forest</i>

GPT-3 generated (450,000 edits)	<i>Alex Hill, Original oil painting on canvas, Moonlight Bay</i>	<i>in the style of a coloring book</i>	<i>Alex Hill, Original coloring book illustration, Moonlight Bay</i>
	<i>The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it</i>	<i>Add a giant red dragon</i>	<i>The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead</i>
	<i>Kate Hudson arriving at the Golden Globes 2015</i>	<i>make her look like a zombie</i>	<i>Zombie Kate Hudson arriving at the Golden Globes 2015</i>

Highlighted text is generated by GPT-3.

Generating image pairs from captions

- Use a pretrained model (Stable Diffusion) to generate images.
- 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."



Generating training data

(1) Generate text edits:

Input Caption: *"photograph of a girl riding a horse"* →

GPT-3
(finetuned)

Instruction: *"have her ride a dragon"*

Edited Caption: *"photograph of a girl riding a dragon"*

Generating training data

(1) Generate text edits:

Input Caption: *"photograph of a girl riding a horse"* →

GPT-3
(finetuned)

Instruction: *"have her ride a dragon"*

Edited Caption: *"photograph of a girl riding a dragon"*

(2) Generate paired images:

Input Caption: *"photograph of a girl riding a horse"*

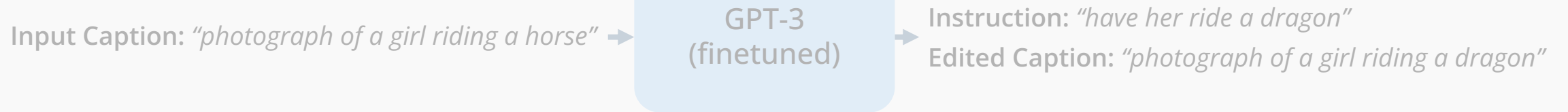
Edited Caption: *"photograph of a girl riding a dragon"* →

Stable Diffusion
+ Prompt2Prompt



Generating training data

(1) Generate text edits:



(2) Generate paired images:



Generated examples:

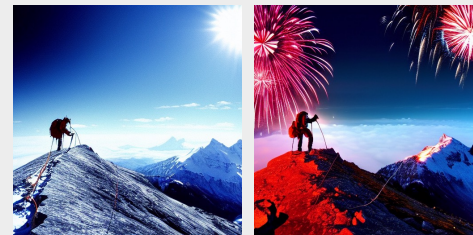
"have her ride a dragon"



"Color the cars pink"



"Make it lit by fireworks"



"convert to brick"

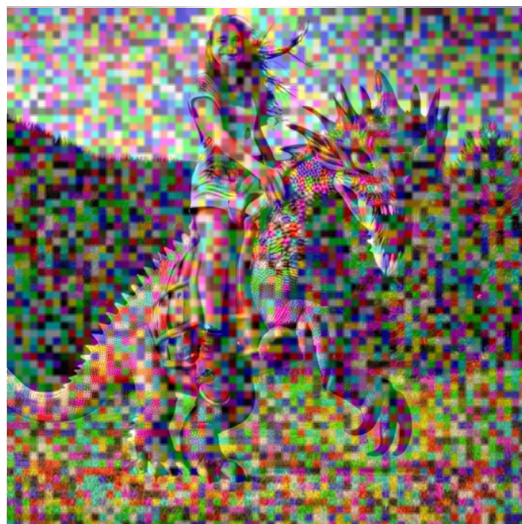


...

Training our model

- Now it is a supervised learning problem!
- Finetune Stable Diffusion on our generated training data.
- Add zero-initialized image conditioning channels.

"have her ride a dragon"



InstructPix2Pix



Generalization to real images

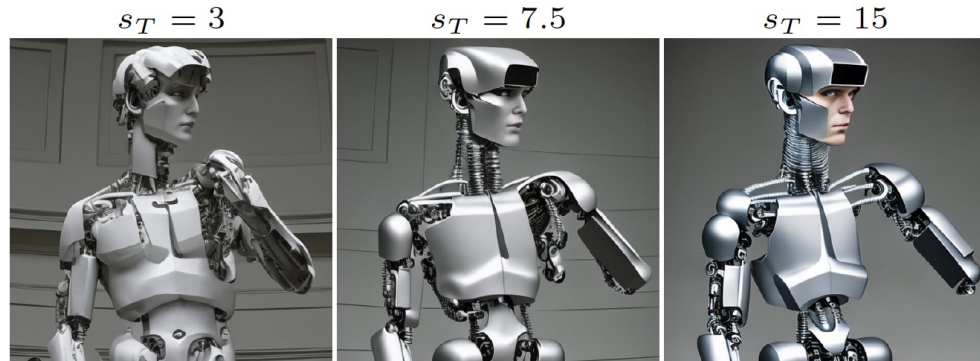
- Trained only on generated images and instructions.
- At inference, generalizes to real images and human-written instructions!

"Make it a grocery store"



Classifier-free guidance for two conditionings

"Turn him into a cyborg!"



- CFG extrapolates samples toward stronger conditioning:

$$\tilde{e}_\theta(z_t, c) = e_\theta(z_t, \emptyset) + s \cdot (e_\theta(z_t, c) - e_\theta(z_t, \emptyset))$$

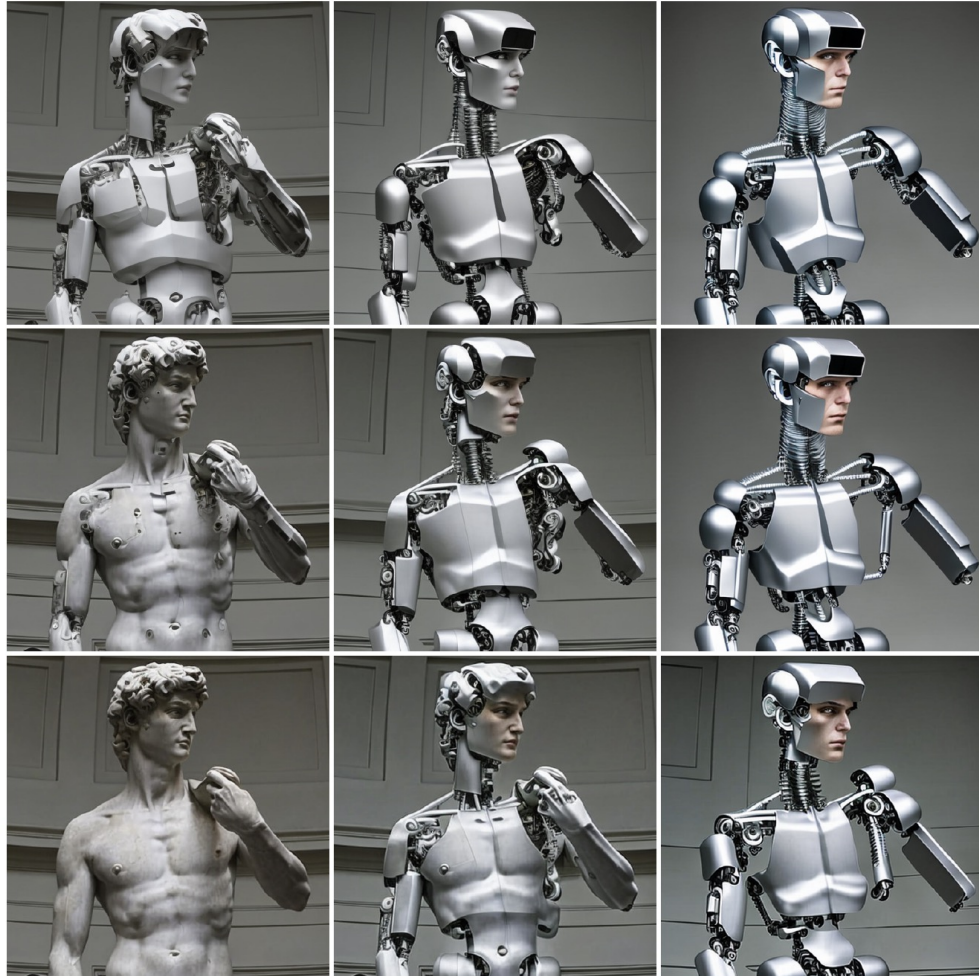
Classifier-free guidance for two conditionings

"Turn him into a cyborg!"

$s_T = 3$

$s_T = 7.5$

$s_T = 15$



- CFG extrapolates samples toward stronger conditioning:

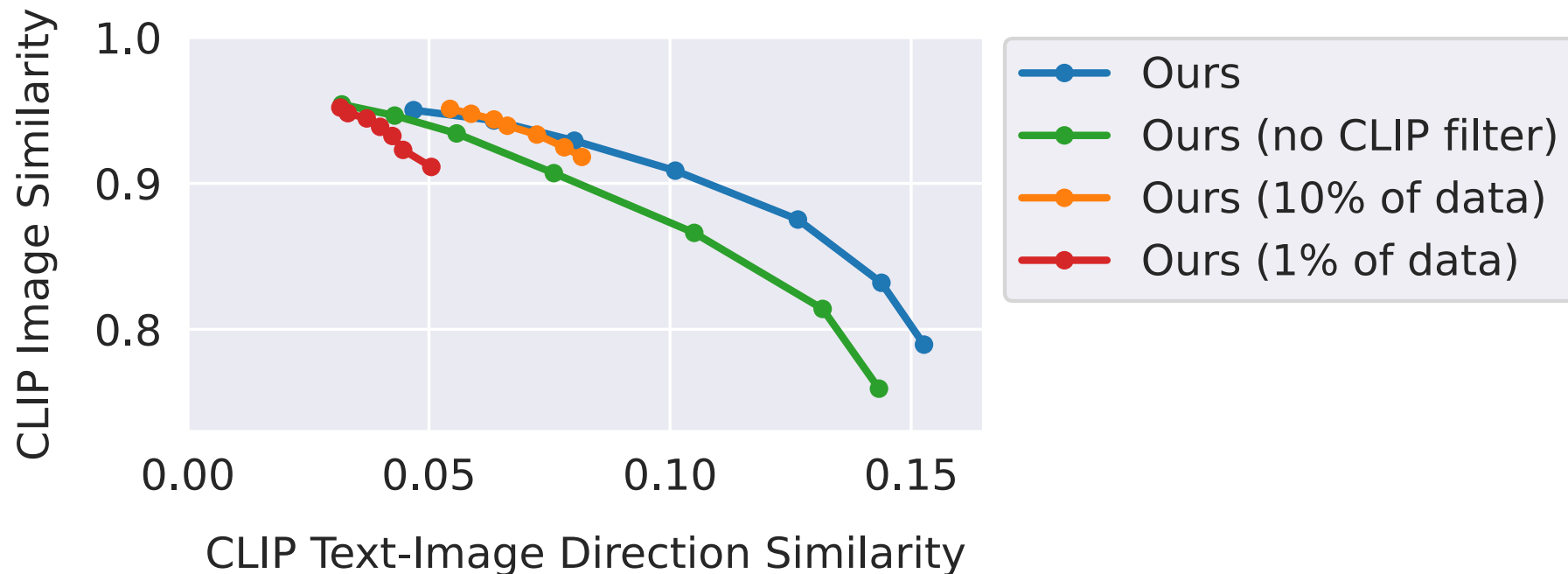
$$\tilde{e}_\theta(z_t, c) = e_\theta(z_t, \emptyset) + s \cdot (e_\theta(z_t, c) - e_\theta(z_t, \emptyset))$$

- We apply CFG with separate scales for image and text conditionings:

$$\begin{aligned} \tilde{e}_\theta(z_t, c_I, c_T) &= e_\theta(z_t, \emptyset, \emptyset) \\ &+ s_I \cdot (e_\theta(z_t, c_I, \emptyset) - e_\theta(z_t, \emptyset, \emptyset)) \\ &+ s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \emptyset)) \end{aligned}$$

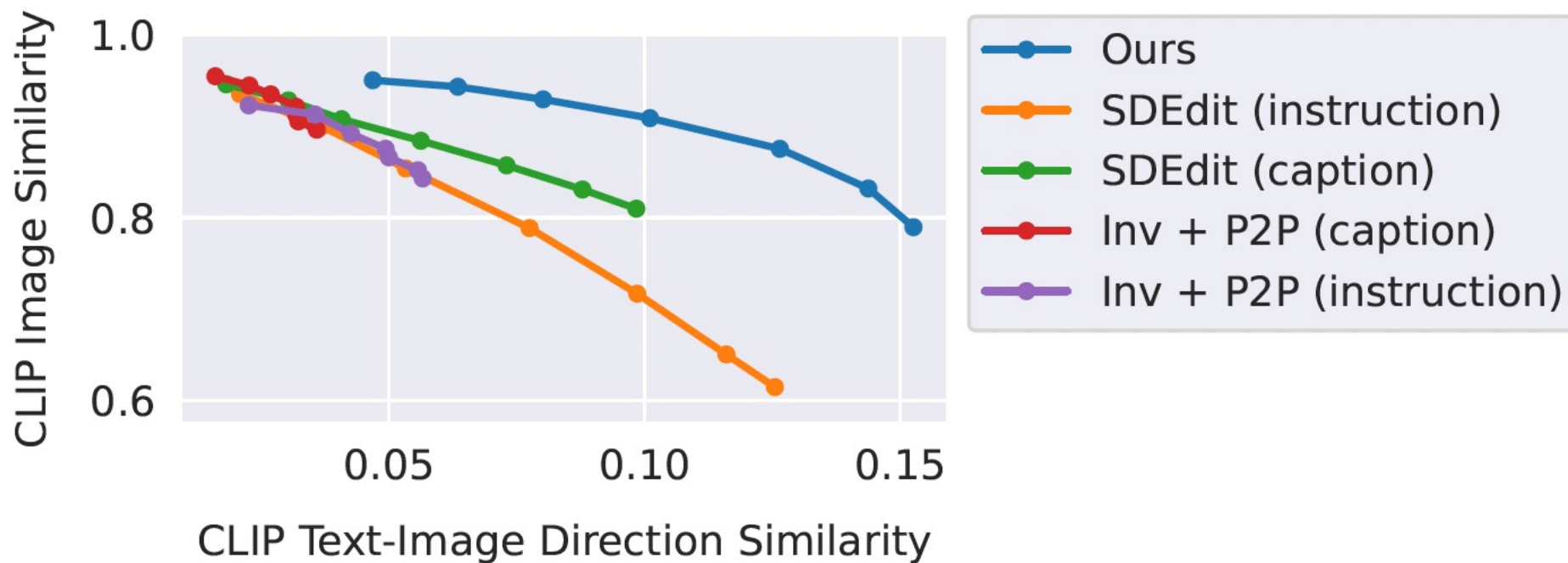
Data scale and quality is crucial

- How well does output image match input image?
- How well does change in images match change in captions?
- Evaluate for a range of guidance scales. Text: 7.5, Image: 1.0-2.2



Baseline comparisons

- Our model outperforms SDEdit and Prompt-to-Prompt.



Baseline comparisons

Input



SDEdit



Text2Live



Prompt2Prompt

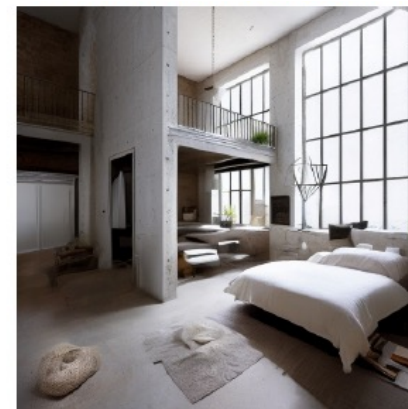
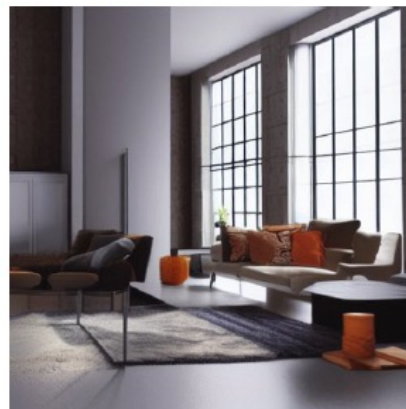


Ours



"Dali Painting of Nimbus Cloud..."

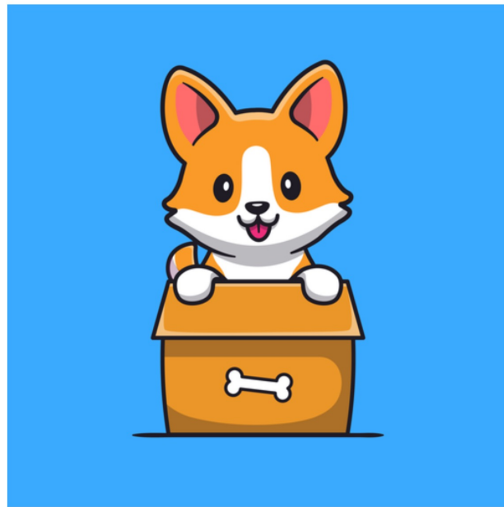
"make it look like a Dali painting"



"Industrial design bedroom furniture..."

"add a bedroom"

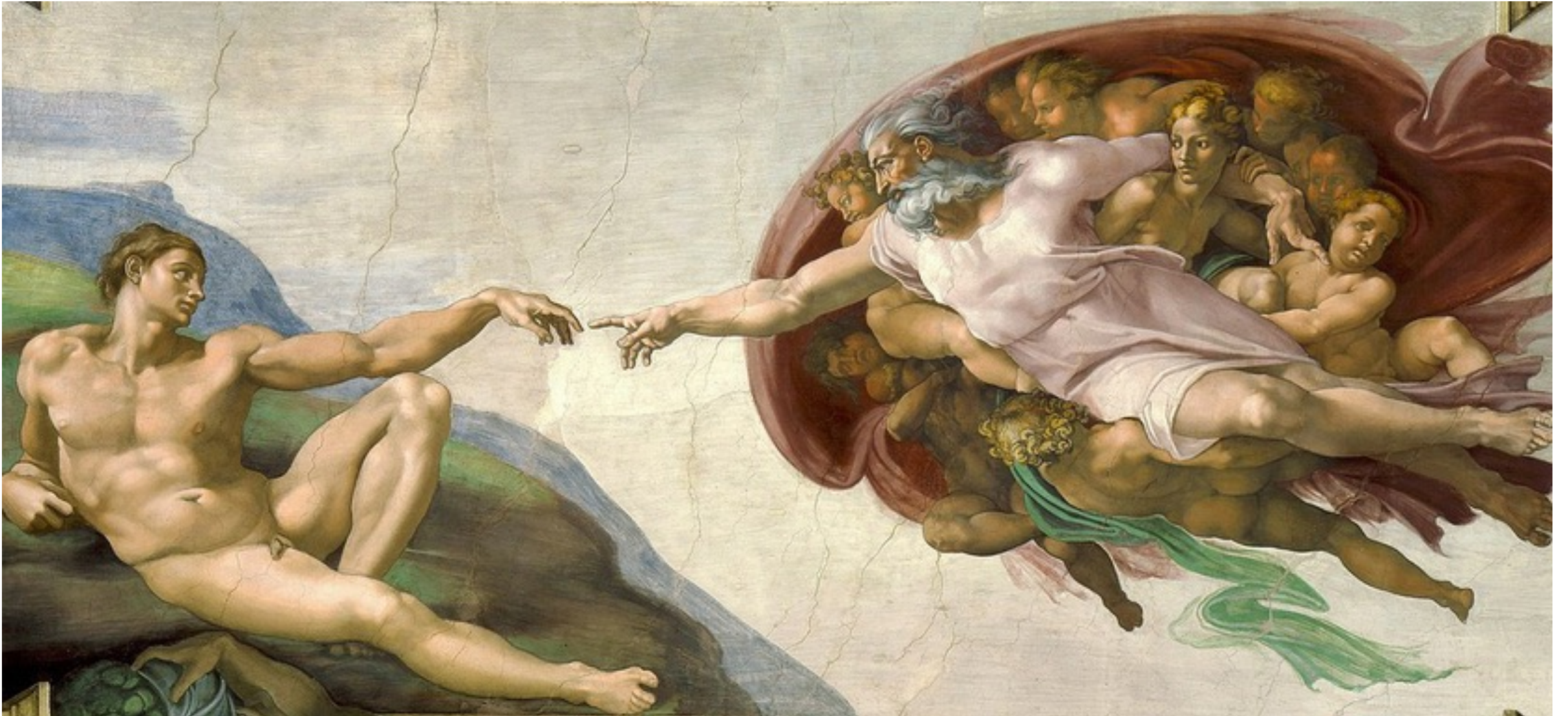
Varying latent noise produces diverse samples



“in a race car video game”



Generalization to high resolution

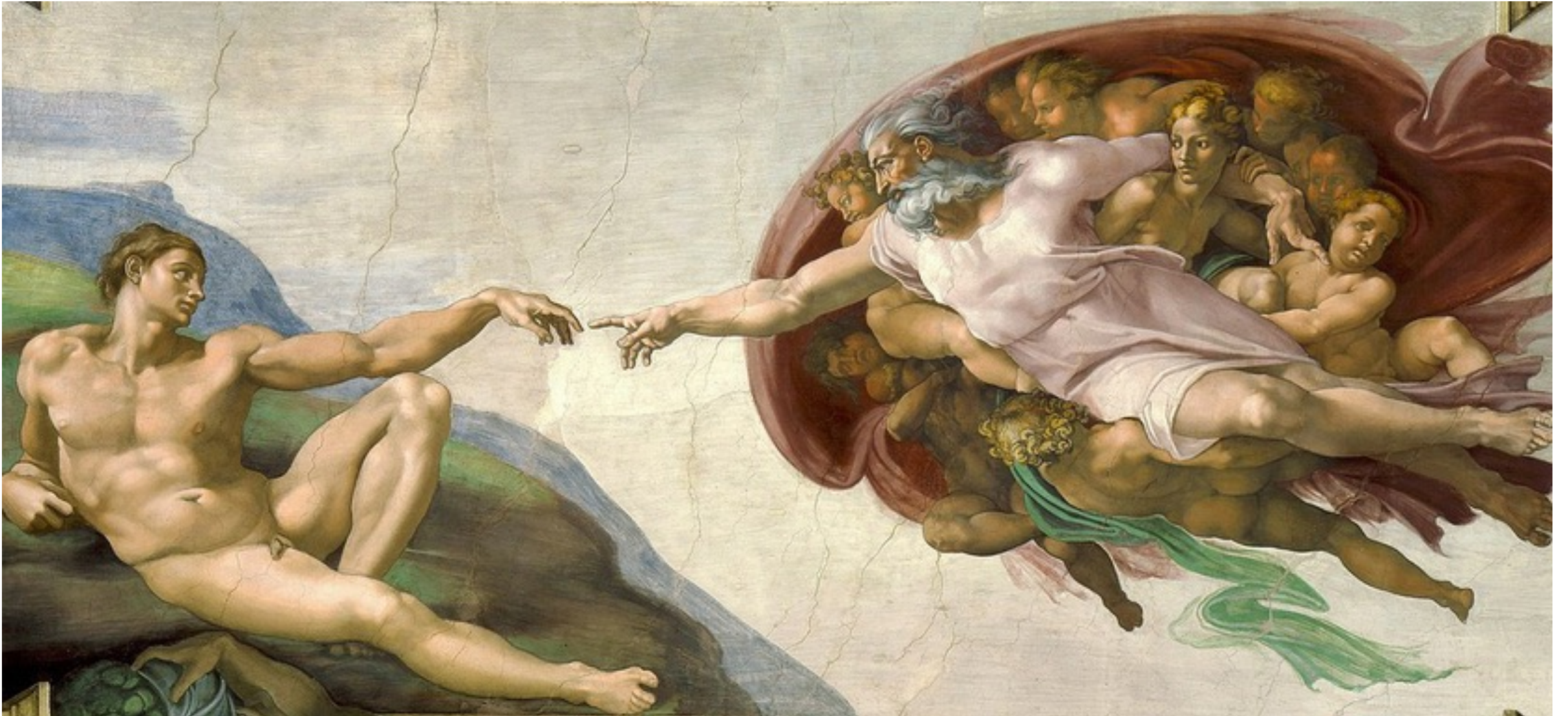


Generalization to high resolution

"Put them in outer space"

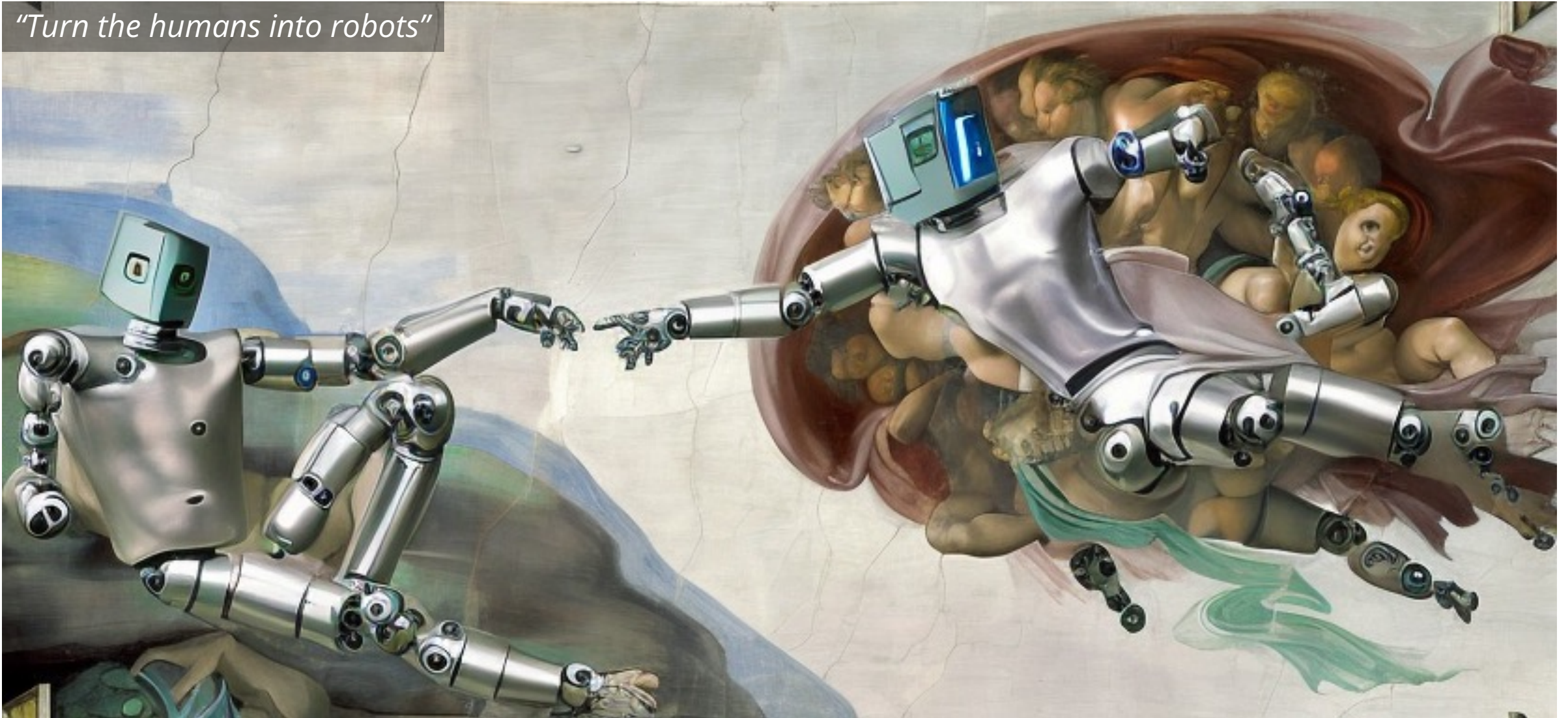


Generalization to high resolution



Generalization to high resolution

"Turn the humans into robots"



Fast model enables iterative editing

- Can easily apply edits in a sequence.
- Benefit of our model being feed-forward (no inversion/finetuning).
- Inference takes < 10s per edit of a 512x512 image.



"Insert a train."



*"Add an eerie
thunderstorm."*



*"Turn into an oil
pastel drawing."*



*"Give it a dark
creepy vibe."*



Identity preservation

- Reasonably capable at preserving identity.
- Requires tuning CFG for specific images/edits.

Input



"Have them wear brown leather jackets"



"Replace the background with a fancy party"



Bias in generated images

- Our model learns biases such as correlations between profession and gender.



Input



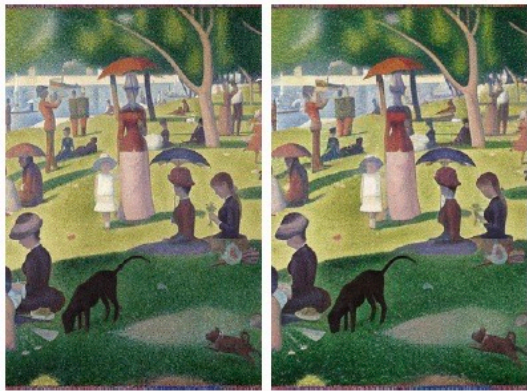
“Make them look like flight attendants”



“Make them look like doctors”

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”

WhimsyAI Upload Undo

Uploading...




Tap send for prompt ideas

q w e r t y u i o p
a s d f g h j k l

Paint by Text

Modify images by chatting with a generative AI model.



make it look like it's nighttime
No negative prompt to display.

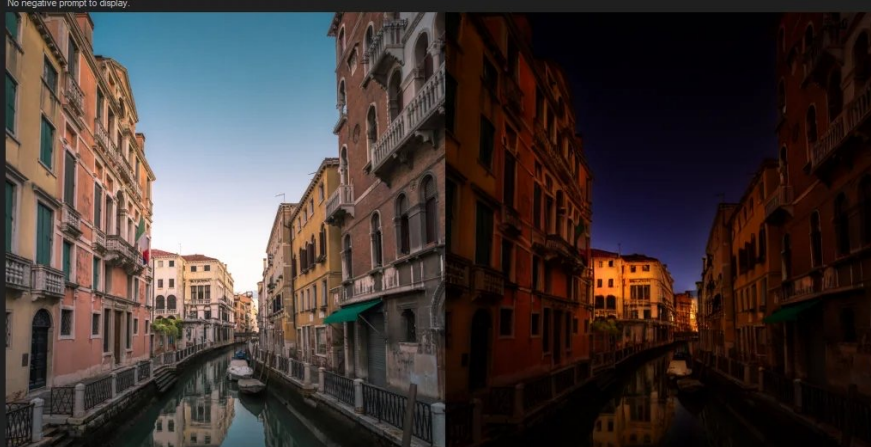
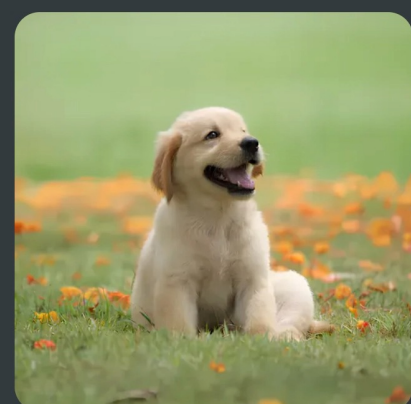


Image 2/4 - Seed 2134304583 - 30 Steps - Scale 10 - Img Scale 1.5

Generate!

Edit images with words



Describe how you want your image to be edited and let the AI do the rest!



START EDITING →

Edit Instruction
Describe how you want to change the image.

make it summer



Pro Tip
Try adding, removing, or thinking of styles you could use to modify the image.
Examples: Turn the cat into a dog, Change the flowers to red, Make it more like van gogh

Remove From Image

Describe details you don't want in your image like color, objects, or a scenery.

Generate

Image Strength
Higher values will cause your edited image to match the essence of the original more.
1.25

Edit Strength
Higher values will make your edited image closer to your instruction.
7

Quality & Details
More steps will result in a high quality image but will take longer.
76

Seed
Different numbers result in new variations of your image.

Randomize each number to get new variations

Takeaways

1. Image generation models can be made more useful by teaching them to follow instructions.
2. We can use large pretrained models to generate training data for new multimodal tasks.

"Swap sunflowers with roses"



"Add fireworks to the sky"



"Replace the fruits with cake"



"What would it look like if it were snowing?"



"Turn it into a still from a western"



"Make his jacket out of leather"

