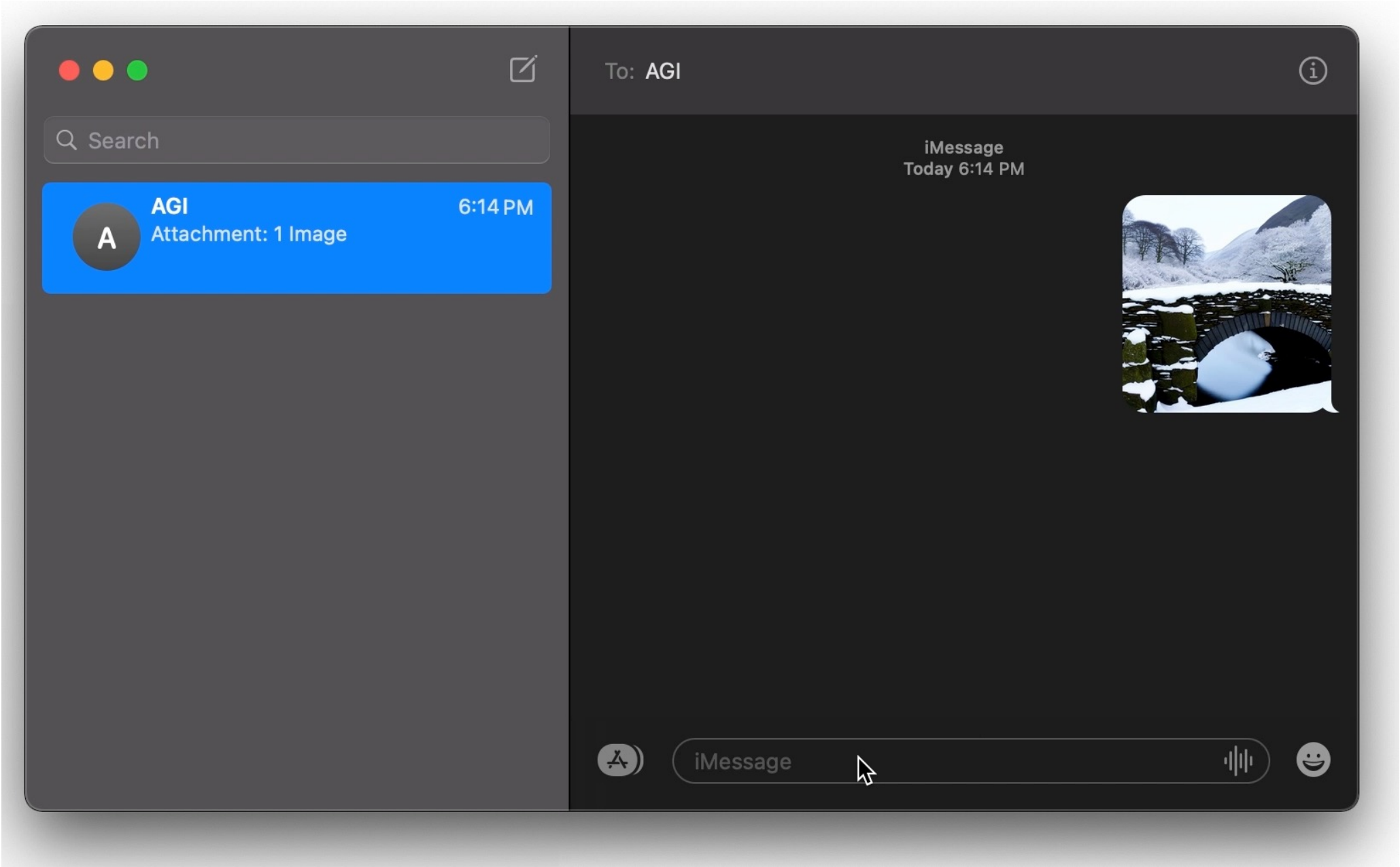
InstructPix2Pix: Learning to Follow Image Editing Instructions

Tim Brooks, Aleksander Holynski, Alexei A. Efros

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

We want an AI image editor that does what we ask



"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"

Editing images goals

- Tell the model exactly what edit to make as a written instruction.
- Require no extra input (full captions, additional images, drawn masks).
- Perform edit in forward pass without need for inversion/finetuning.

Related work: Prompt-to-Prompt

- Only works reliably when both images are generated.
- Requires full written description of both input and output images.

"Photo of a cat riding on a bicycle."

"Photo of a cat riding on a car."

Related work: DreamBooth, Imagic

- Expensive finetuning for every new input image.
- Require full descriptions of the desired output image.

DreamBooth:

Imagic:

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?
- <u>Combine knowledge of large pretrained models to generate training data.</u>

"have her ride a dragon"

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 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).

Generating caption edits with GPT-3

	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.

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."

Generating paired 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 →

Generated training examples:

"have her ride a dragon"

"Color the cars pink"

"Make it lit by fireworks"

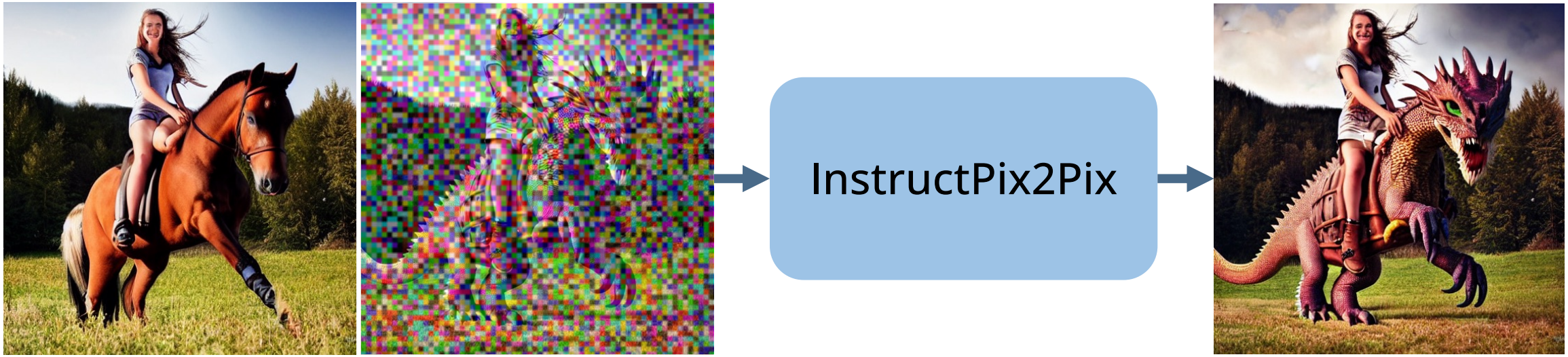
"convert to brick"

...

Training an image editing diffusion model

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

"have her ride a dragon"



Generalization to real images and instructions

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

"Make it a grocery store"

InstructPix2Pix

Edmund Leighton's *Lady in a Garden*

Input

"Add boats on the water"

"Replace the mountains with a city skyline"

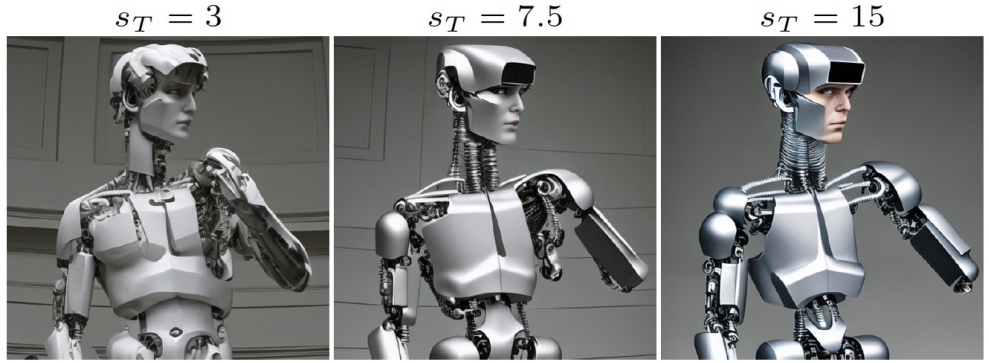
Input

"It is now midnight"

"Add a beautiful sunset"

Classifier-free guidance (CFG) for two conditionings

"Turn him into a cyborg!"

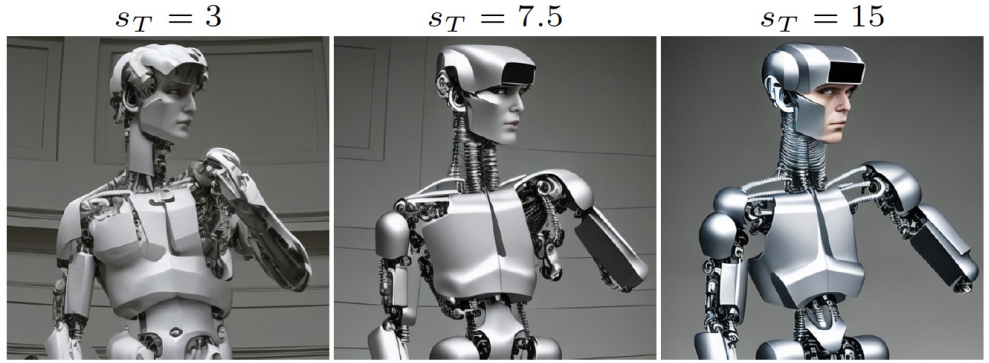


- CFG extrapolates samples toward stronger conditioning:

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

Classifier-free guidance (CFG) for two conditionings

"Turn him into a cyborg!"



- CFG extrapolates samples toward stronger conditioning:

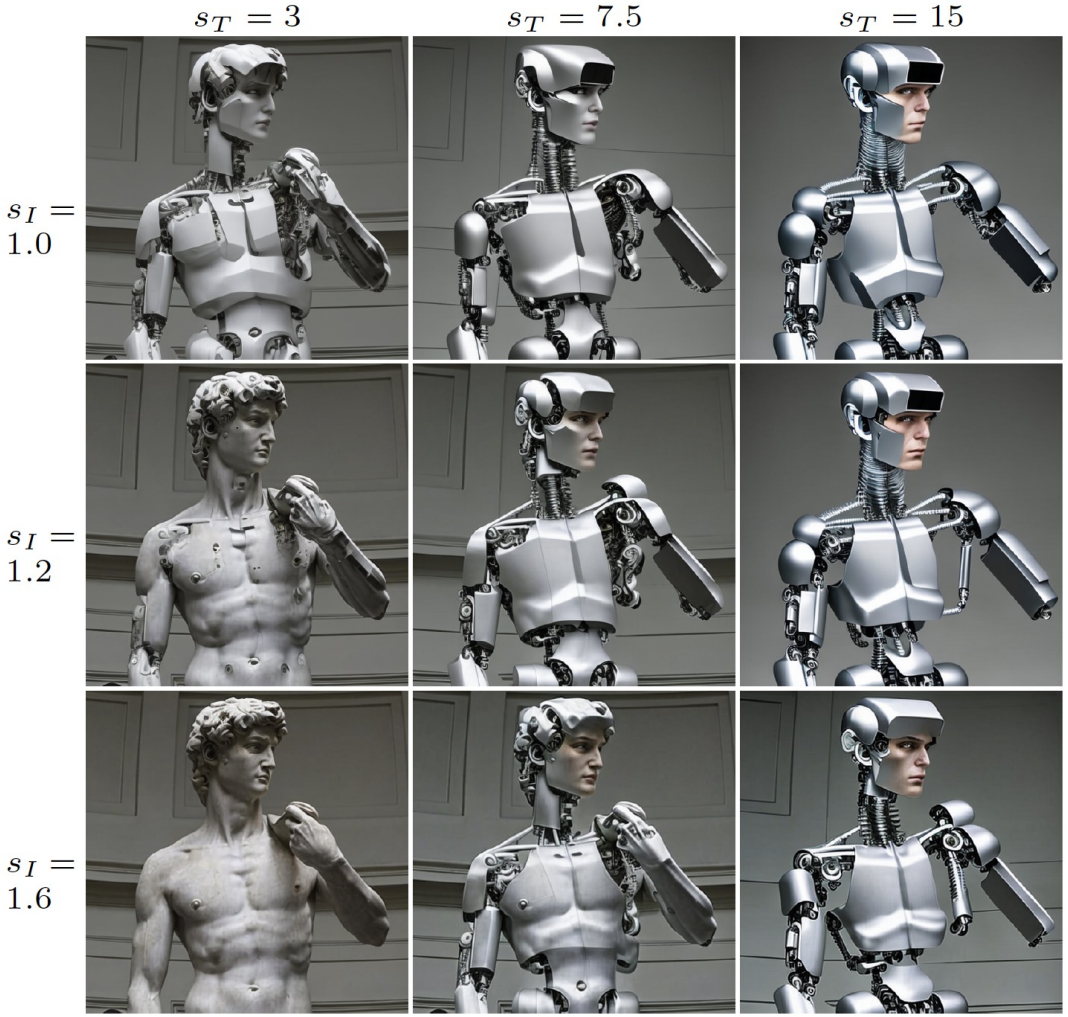
$$\tilde{e_\theta}(z_t, c) = e_\theta(z_t, \varnothing) + s \cdot (e_\theta(z_t, c) - e_\theta(z_t, \varnothing))$$

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

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

Classifier-free guidance (CFG) for two conditionings

"Turn him into a cyborg!"



- CFG extrapolates samples toward stronger conditioning:

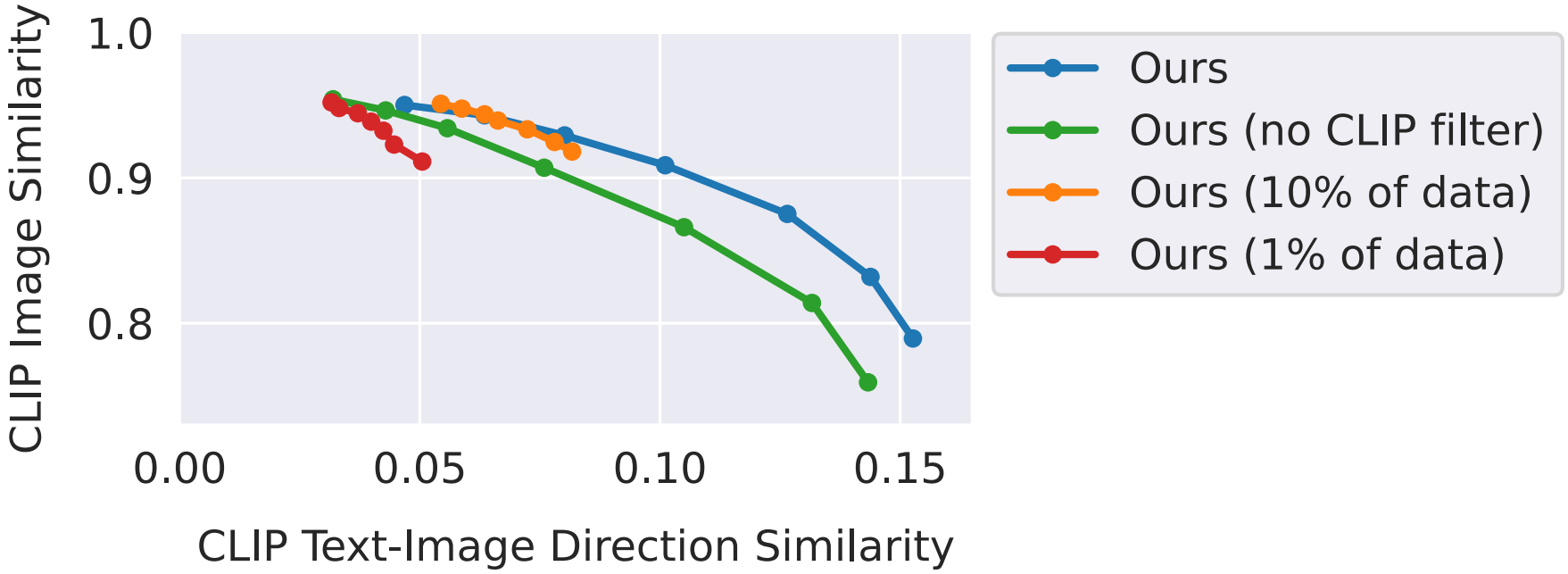
$$\tilde{e}_\theta(z_t, c) = e_\theta(z_t, \varnothing) + s \cdot (e_\theta(z_t, c) - e_\theta(z_t, \varnothing))$$

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

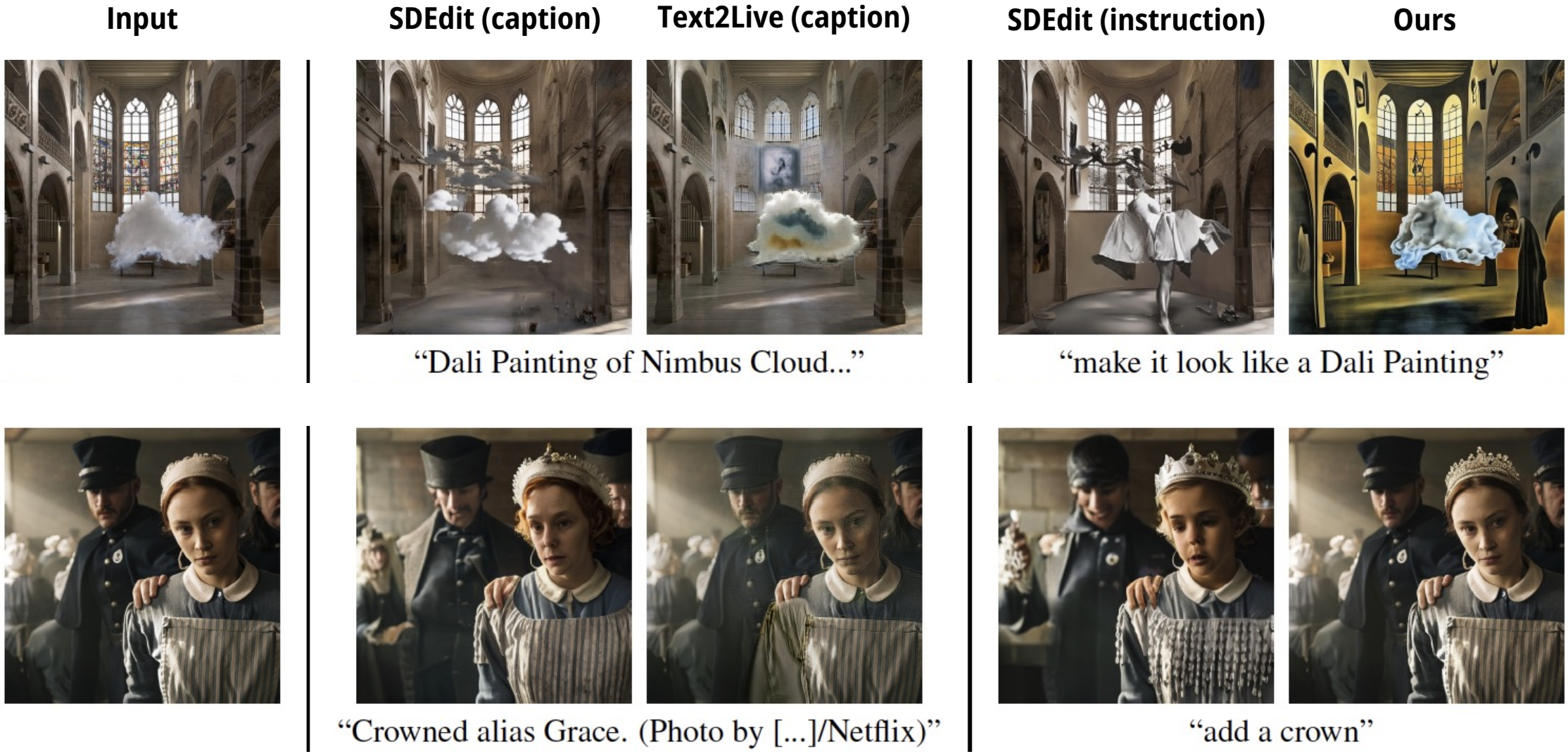
$$\begin{aligned}\tilde{e}_\theta(z_t, c_I, c_T) = e_\theta(z_t, \varnothing, \varnothing) \\ + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))\end{aligned}$$

Data scale and quality is crucial

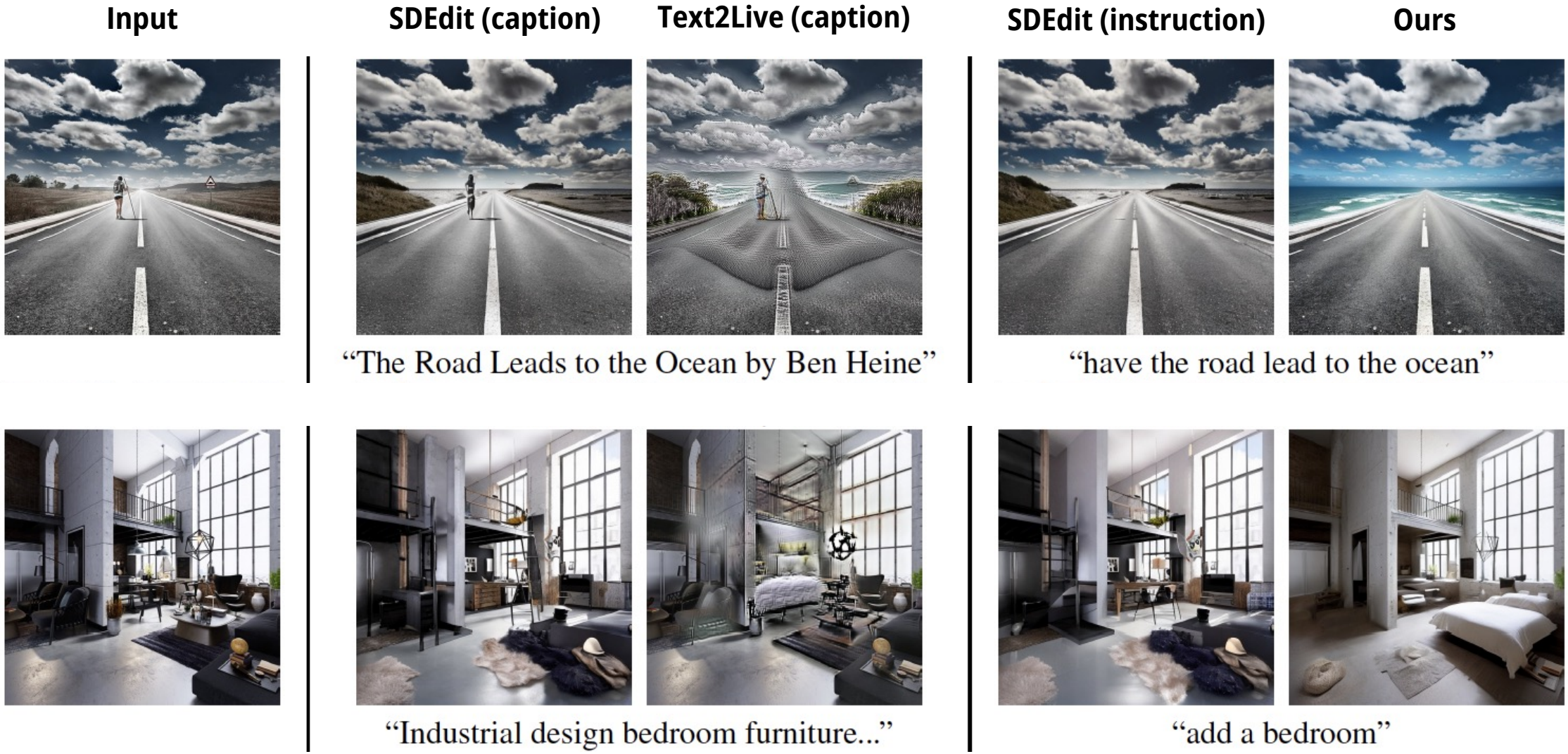
- How well does output image match input image?
- How well does change in images match change in captions?
- Evaluate for a range of guidance scales. Text: 7.5, Image: 1.0-2.2



Baseline comparisons

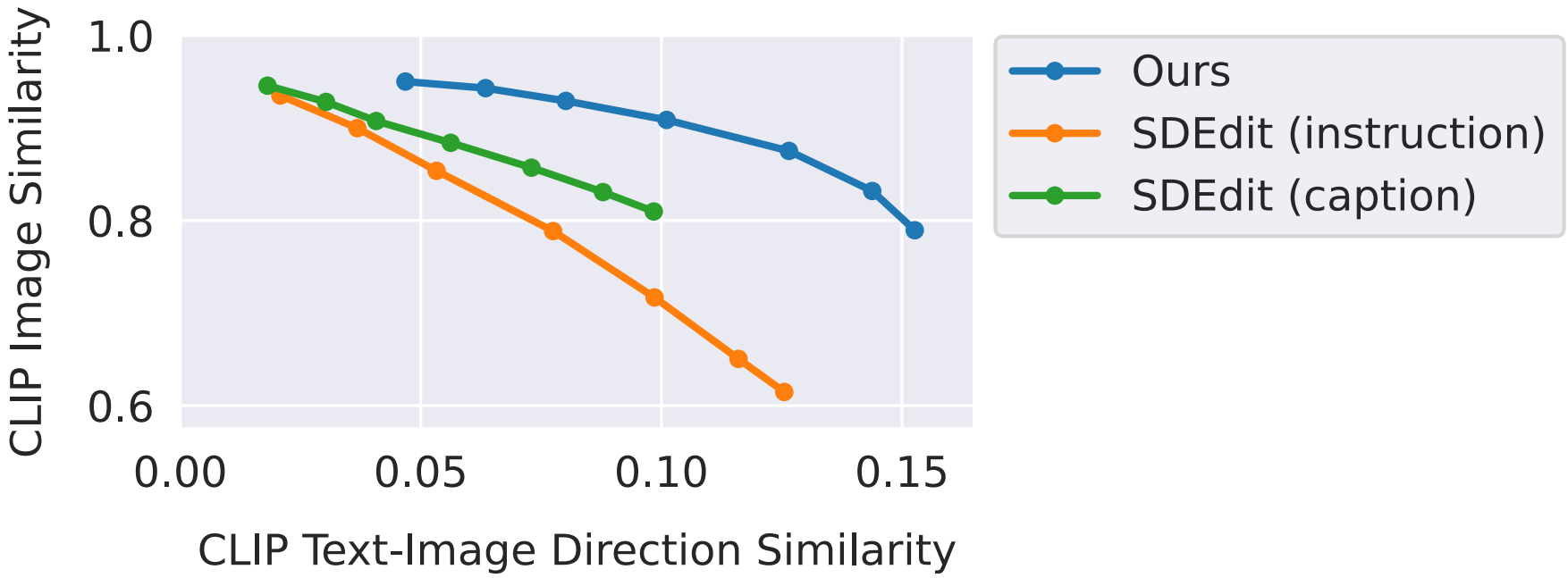


Baseline comparisons



Baseline comparisons

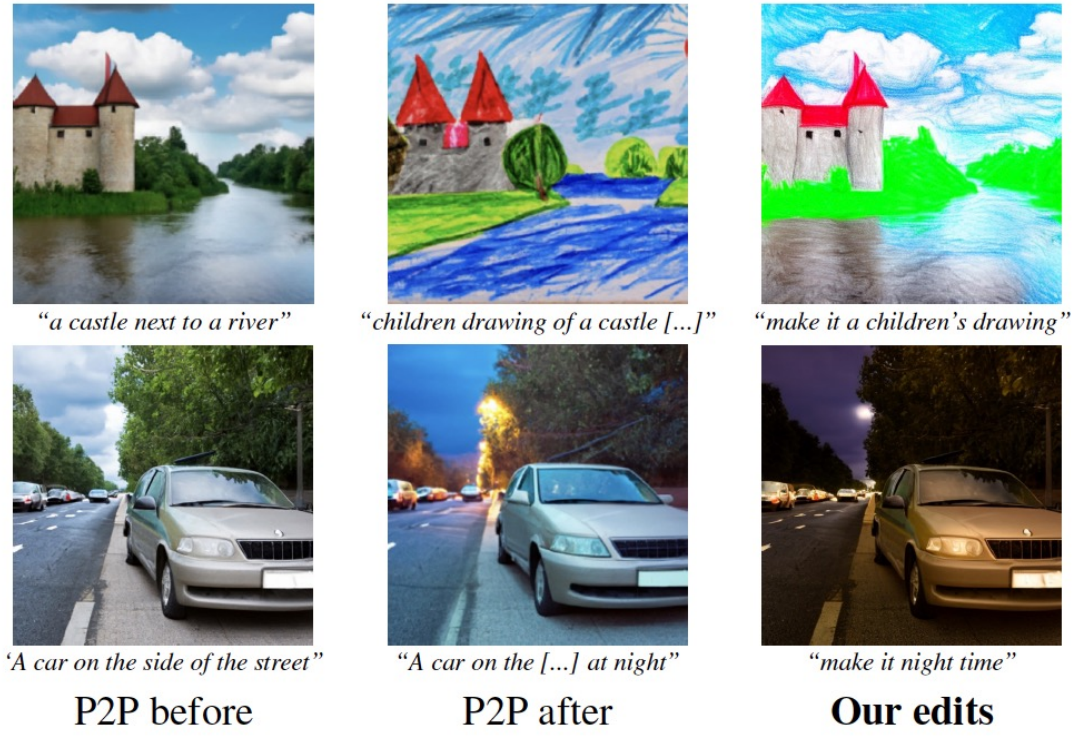
- How well does output image match input image?
- How well does change in images match change in captions?
- Our model achieves a superior tradeoff.



Prompt2Prompt comparisons

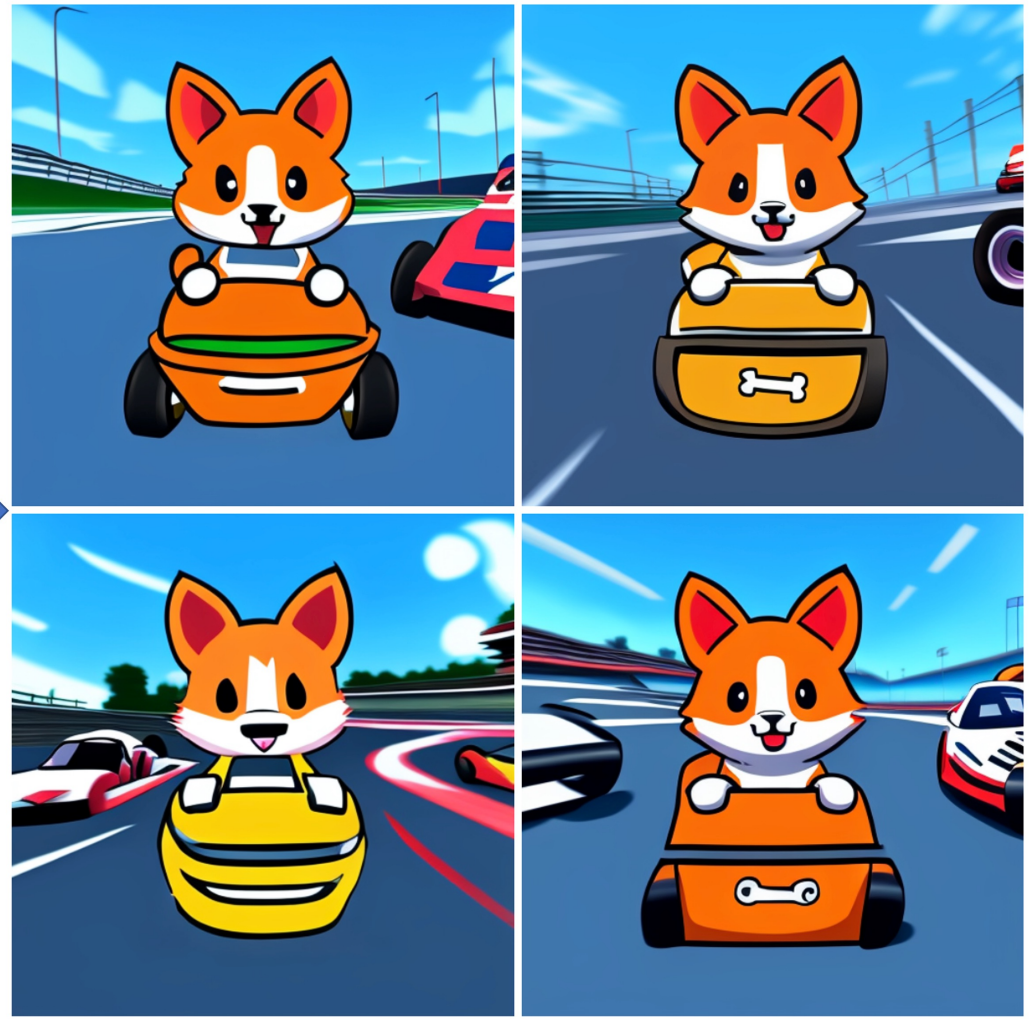
Editing real images

Editing generated images

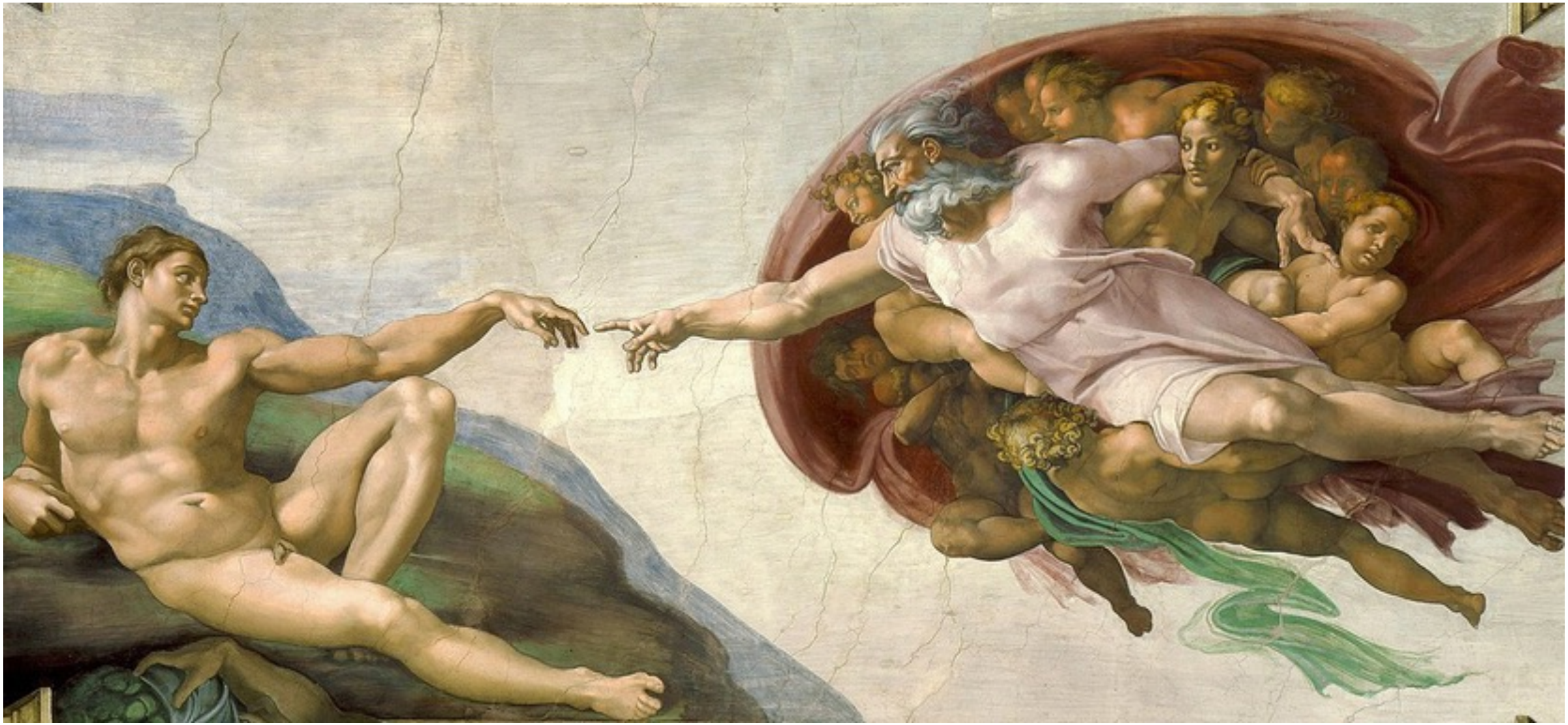


Varying latent noise produces diverse samples

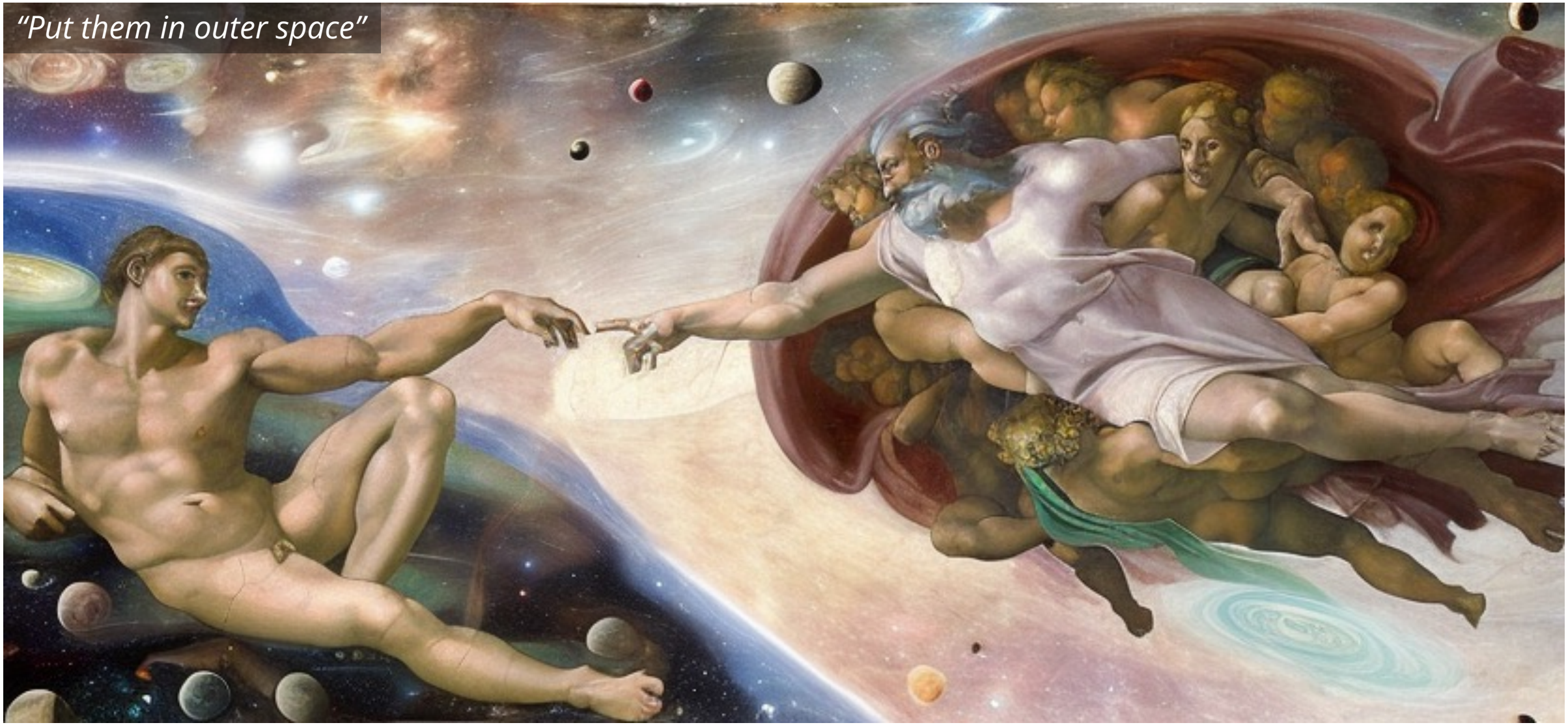
"in a race car video game"



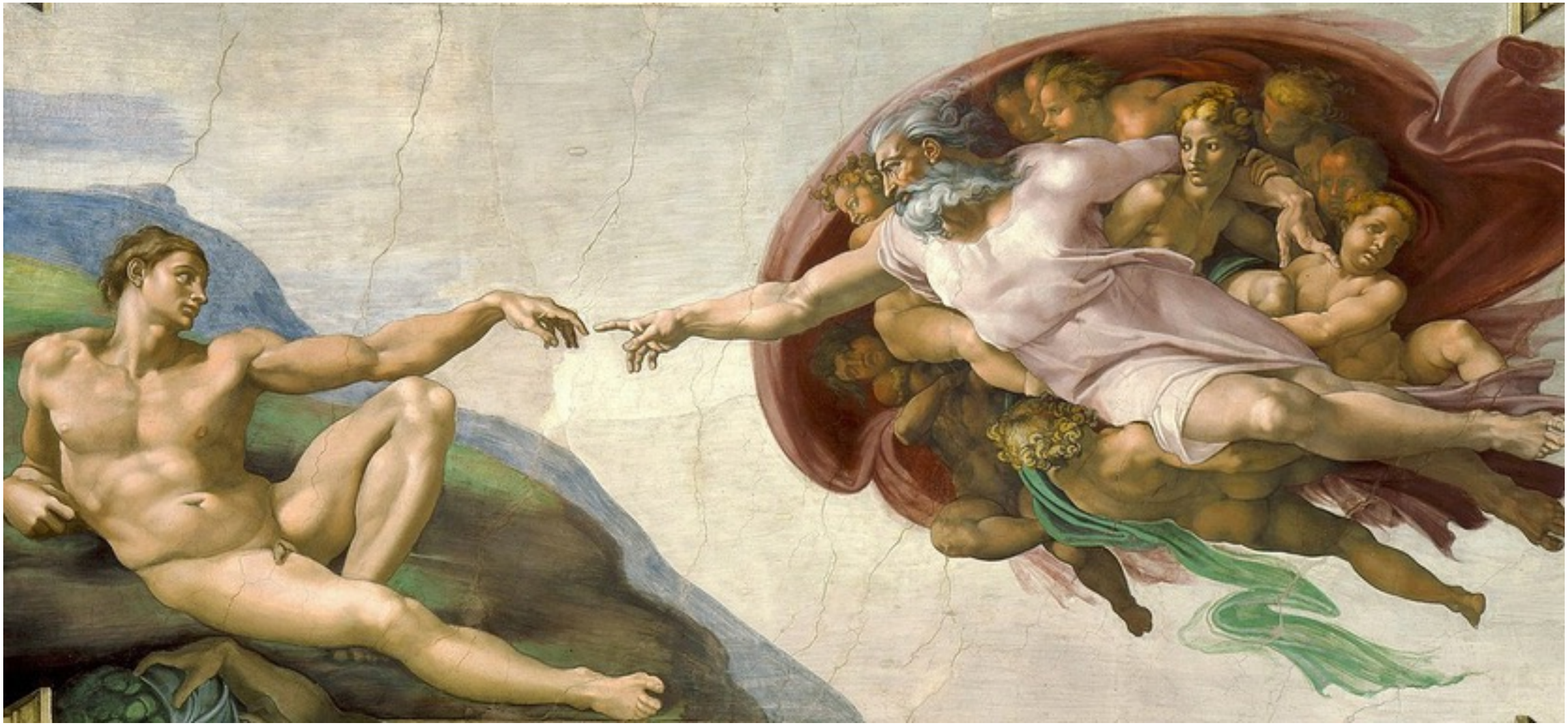
Our model generalizes to high resolutions



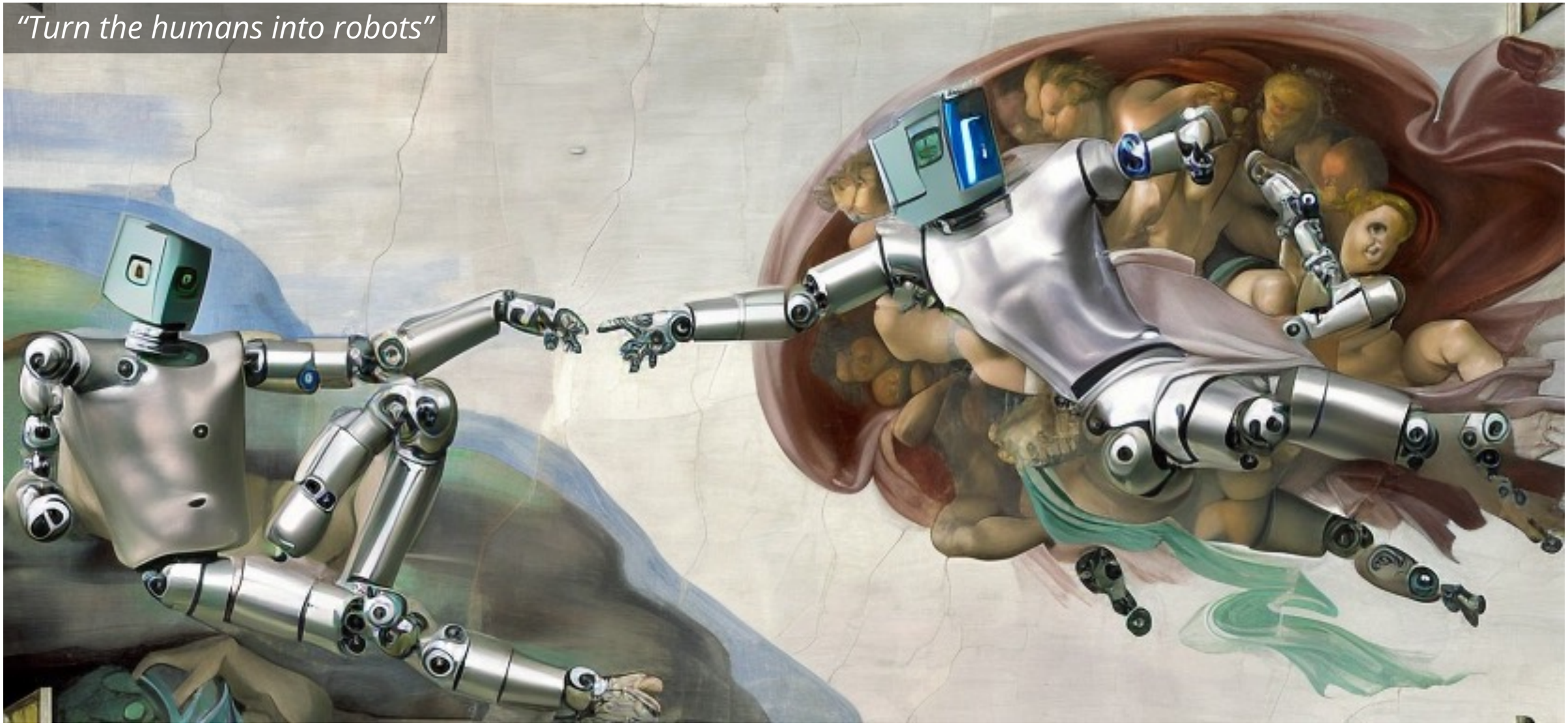
Our model generalizes to high resolutions



Our model generalizes to high resolutions



Our model generalizes to high resolutions



Fast models enable 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.

Human 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”


"Make it Paris"


"Make it Hong Kong"


"Make it Manhattan"


"Make it Prague"

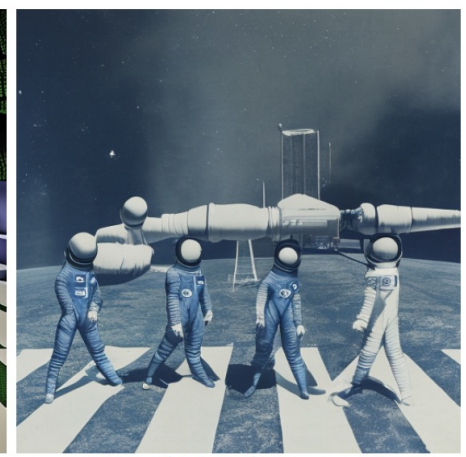

"Make it evening"


"Put them on roller skates"


"Turn this into 1900s"


"Make it underwater"


"Make it Minecraft"

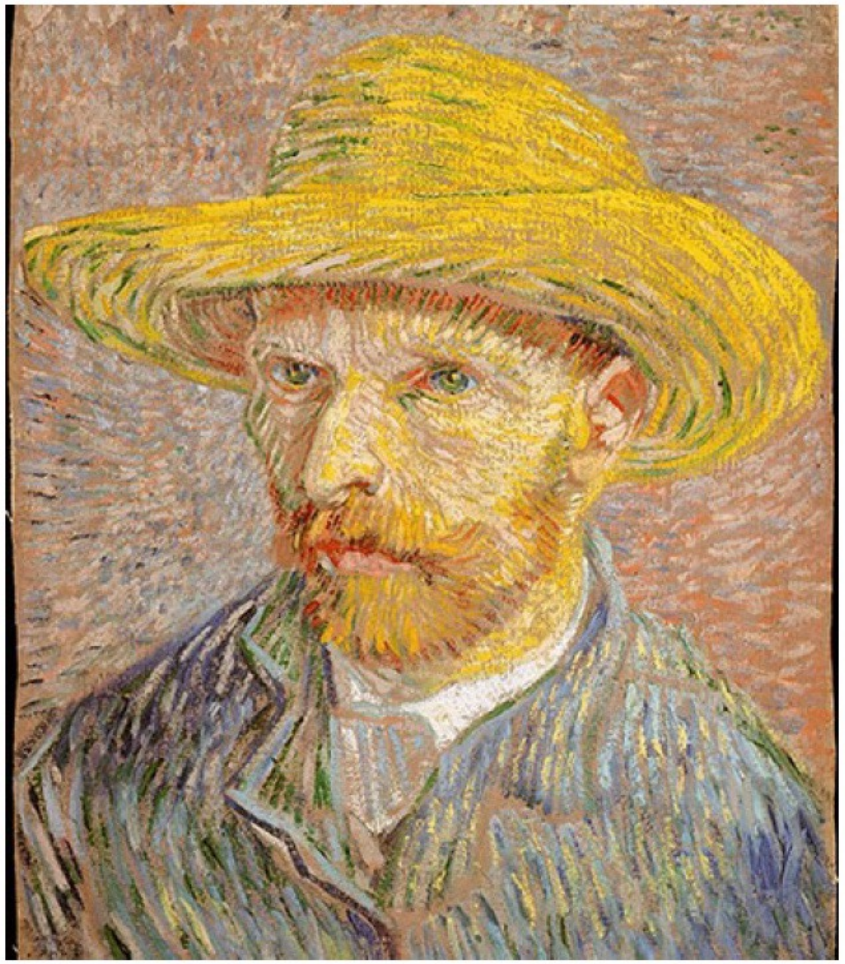

"Turn this into the space age"


"Make them into Alexander Calder sculptures"


"Make it a Claymation"

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

"Convert to a realistic photo"

"Turn into a 3D model"

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"

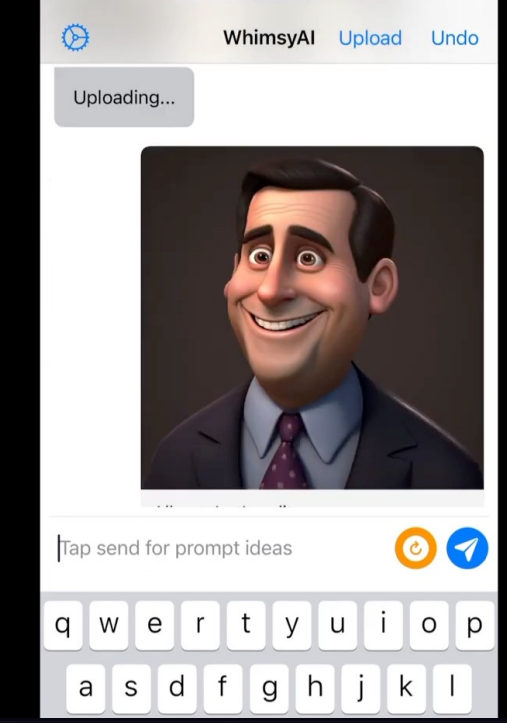
"Move it to Mars"

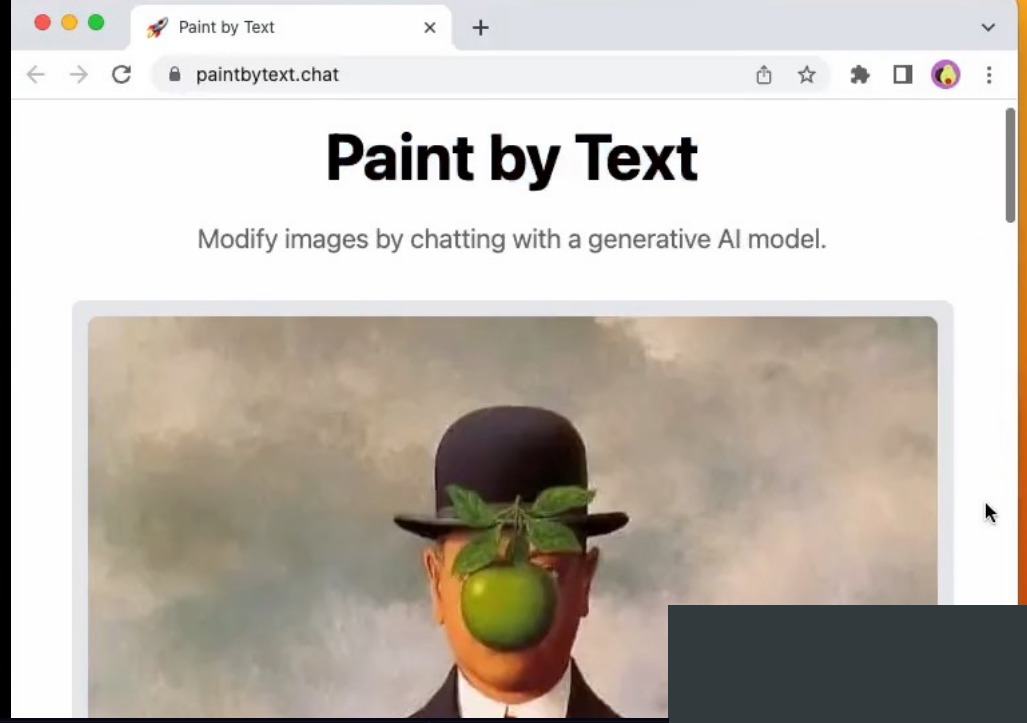
"Color the tie blue"

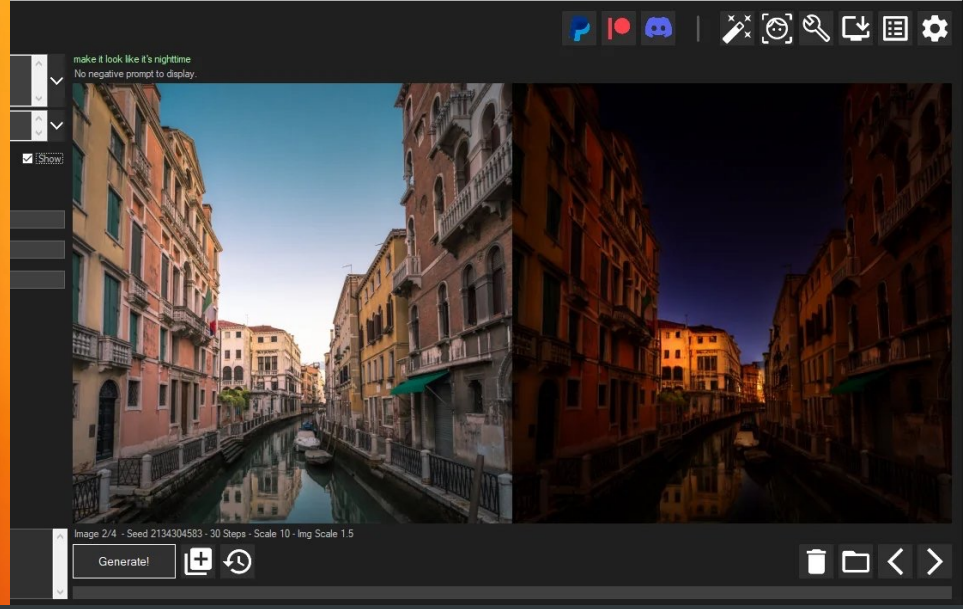
"Have the people swap places"

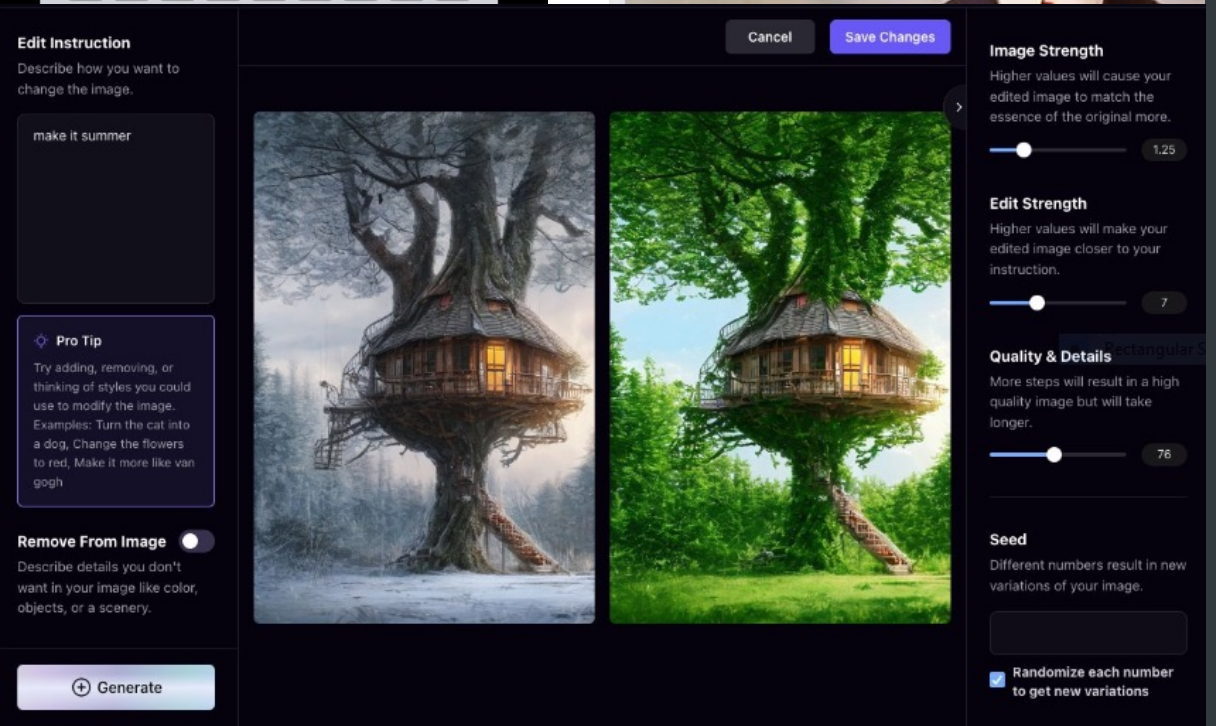
InstructPix2Pix for video editing

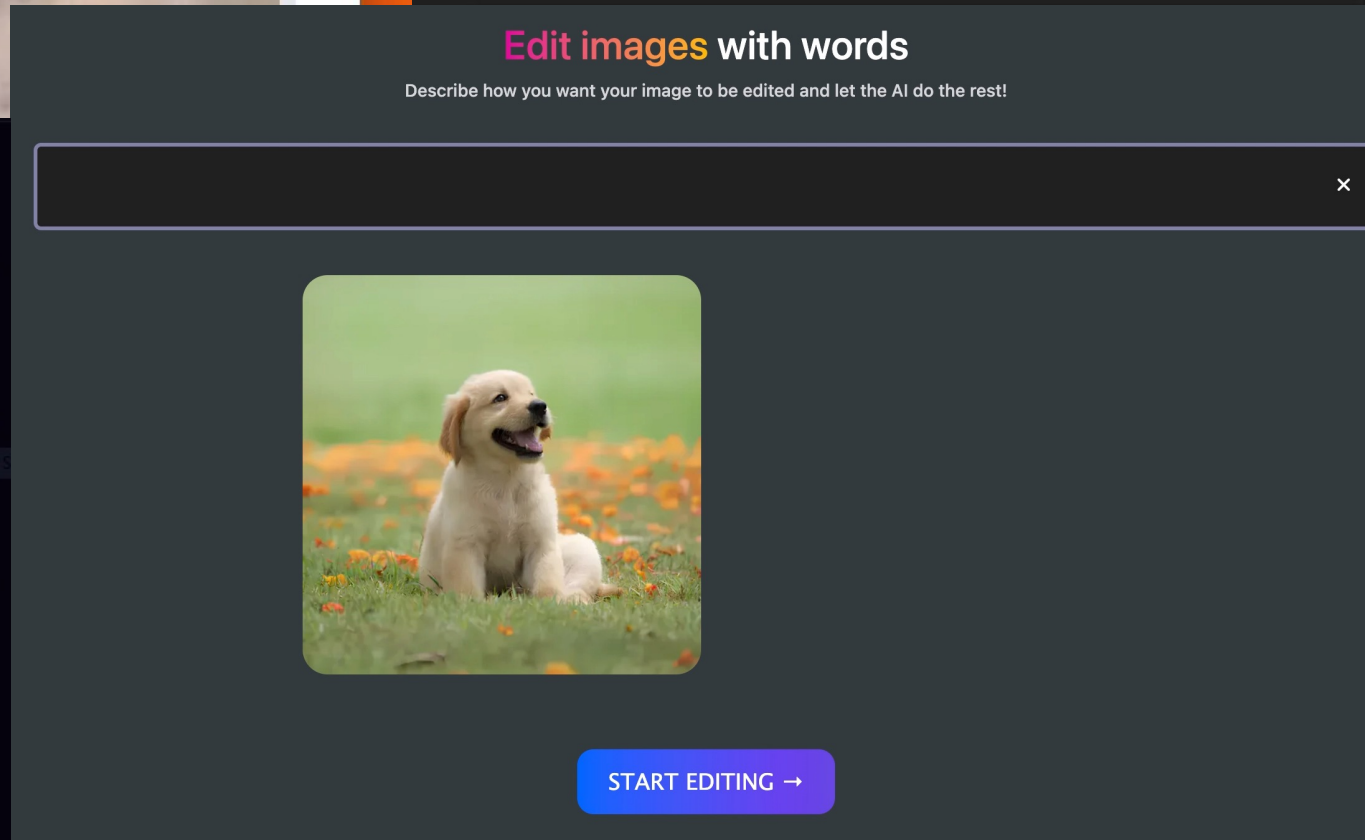
+EBSynth


WhimsyAI
Upload
Undo
Uploading...
Tap send for prompt ideas


Paint by Text
paintbytext.chat
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 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
Cancel
Save Changes
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


Edit images with words
Describe how you want your image to be edited and let the AI do the rest!
START EDITING →

Thank you!