InstructPix2Pix: Learning to Follow Image Editing Instructions

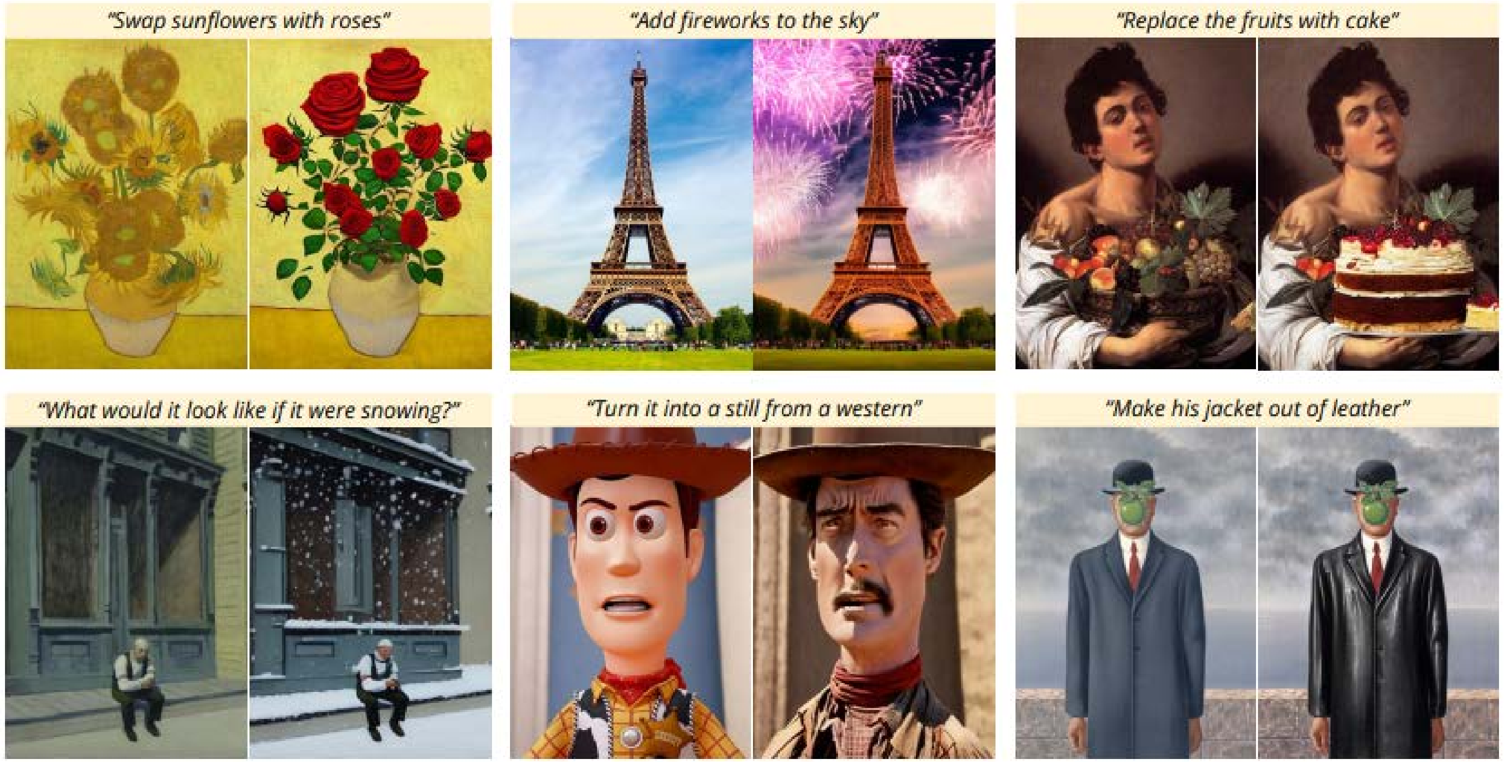
229403g - WAG Weerasundara

Introduction

- Image editing based on human instructions.
- Uses an input image and a written instruction to edit the image following the instruction.
- The training data for the model is generated by combining GPT-3 and Stable Diffusion

How this works

- Provide an initial image
- Give an instruction for how to edit that image
- Instructions are given in natural language


"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"

Examples of the transformation process [2]

[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

Method

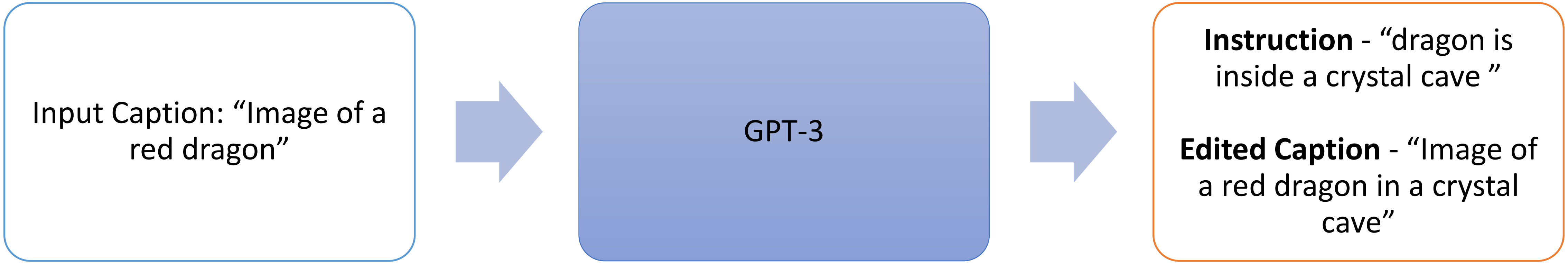
1. Training Data Generation
2. Training the diffusion model

Training Data Generation

- Generate a paired training dataset of text editing instructions and images before/after the edit.
- Then train an image editing diffusion model on this generated dataset.
- **Despite being trained with generated data, the model is able to generalize to real images using arbitrary human-written instructions.**

Training Data Generation Cont…

- Generate text edits



Input Caption: “Image of a red dragon”

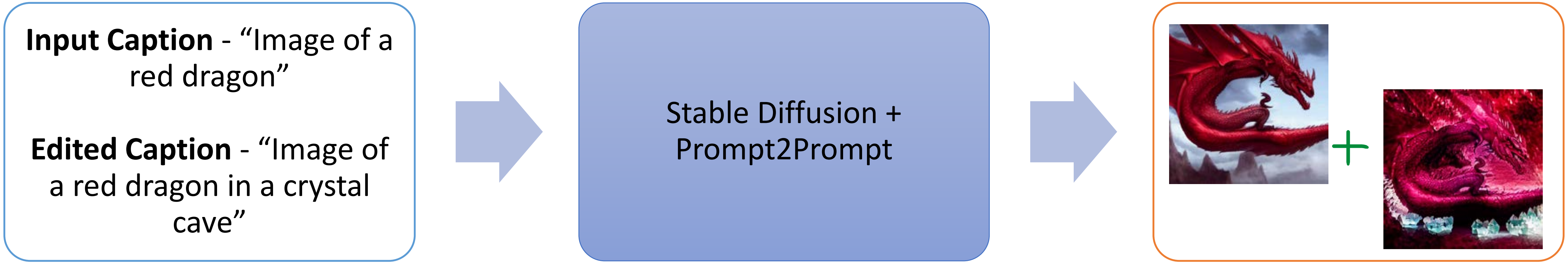
GPT-3

Instruction - “dragon is inside a crystal cave ”

Edited Caption - “Image of a red dragon in a crystal cave”

Training Data Generation Cont…

- Generate paired images

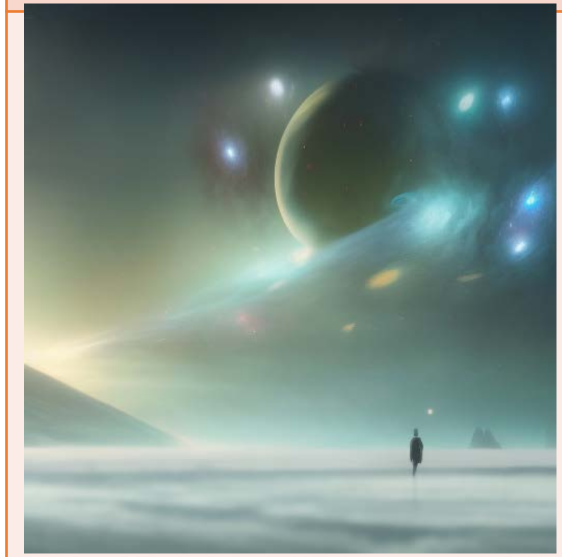
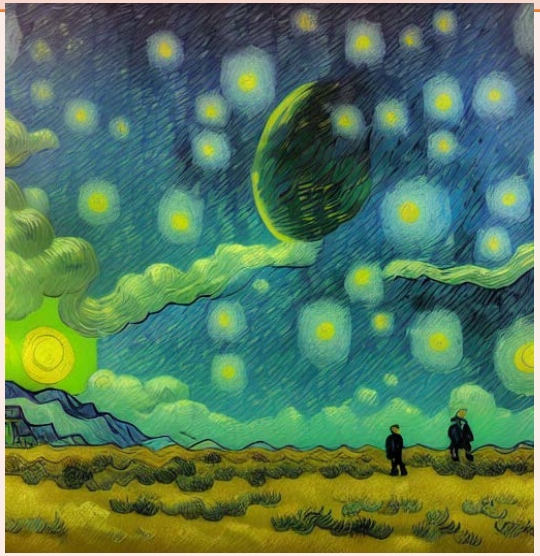
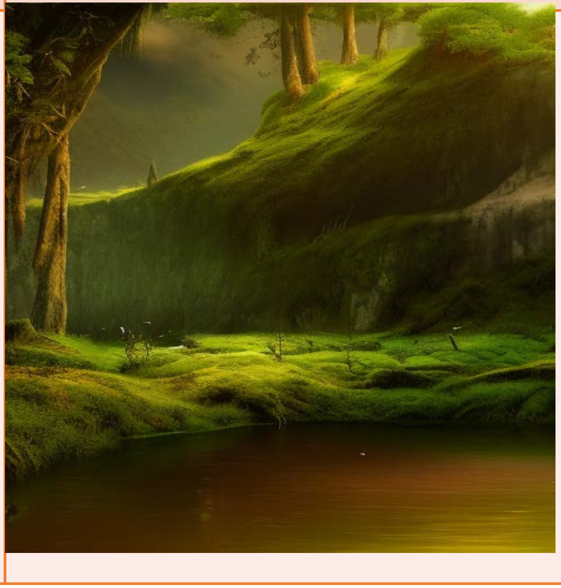
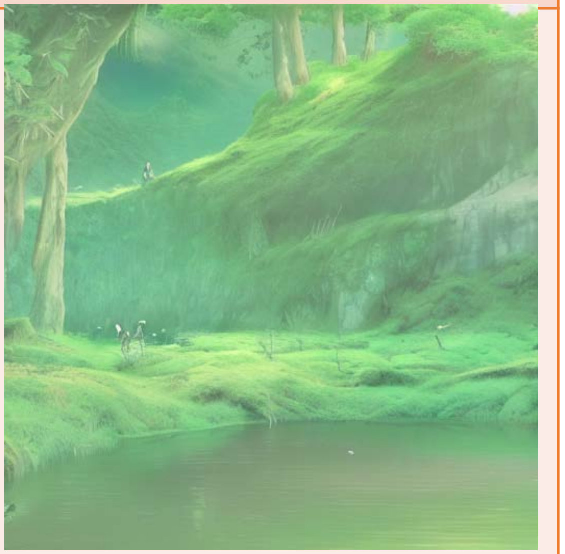


Input Caption - “Image of a red dragon”

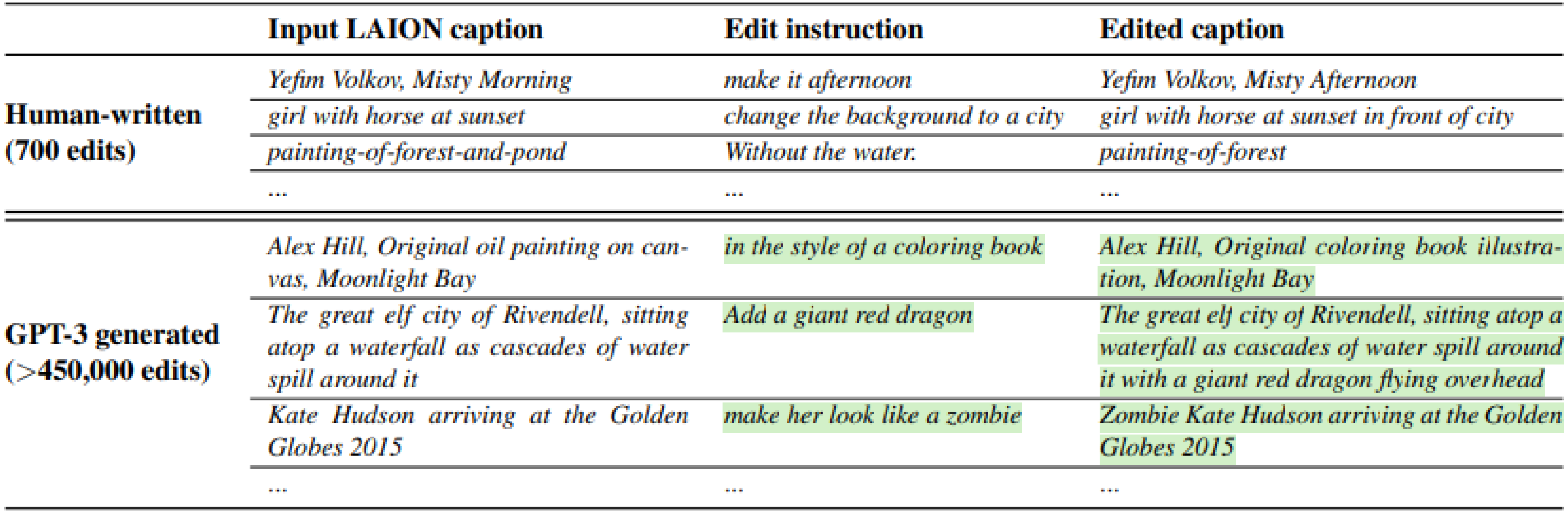
Edited Caption - “Image of a red dragon in a crystal cave”

Stable Diffusion + Prompt2Prompt

Training Data Generation Cont…

A dream of a distant galaxy, concept art, matte painting, HQ, 4k		**photo of a riverbank, concept art, matte painting, HQ, 4k**	
Original	change the style to Vincent van Gogh's style	Original	change the style to a studio Ghibli art
			

Training the diffusion model



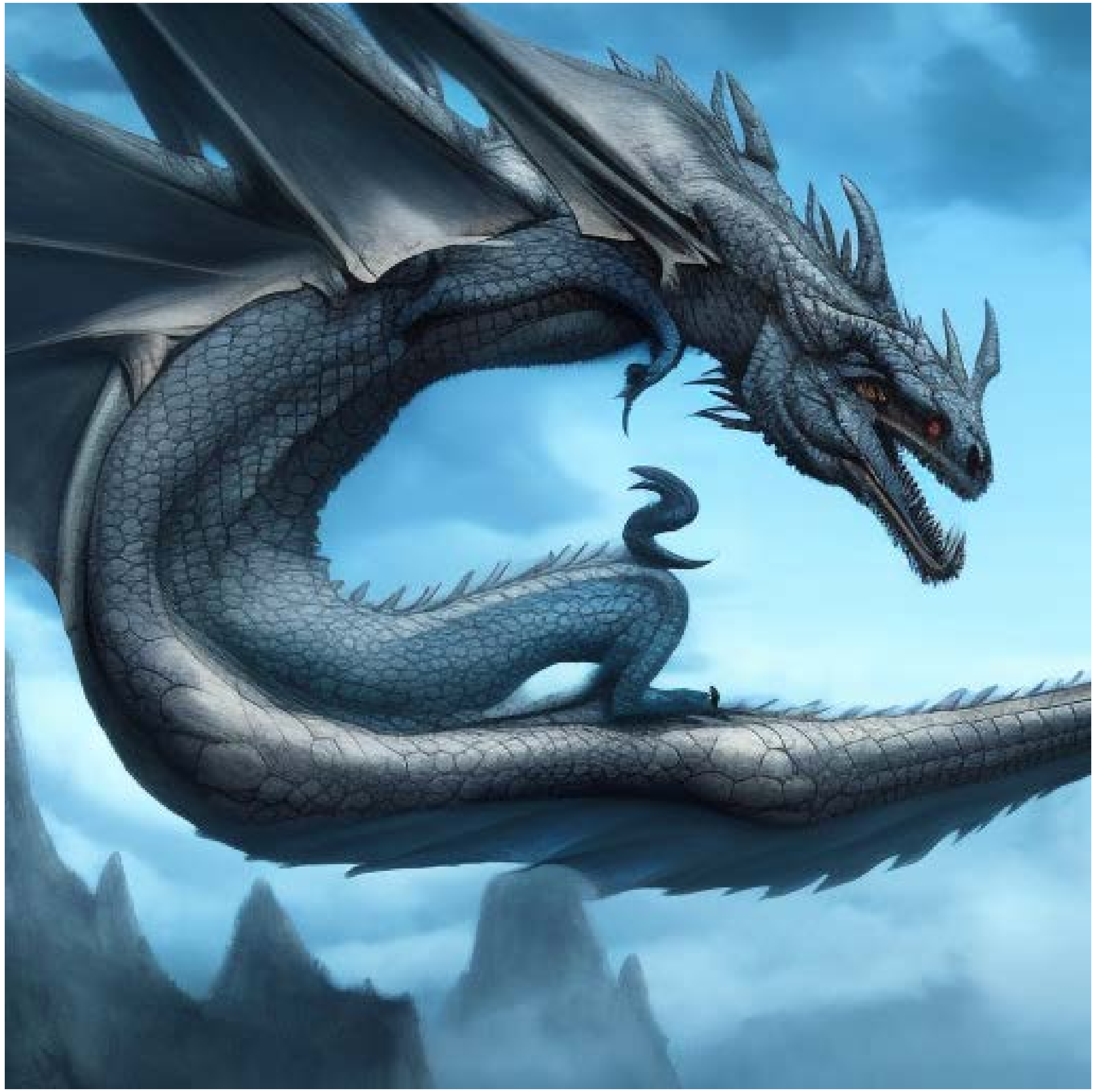
	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	*Yefim Volkov, Misty Morning*	*make it afternoon*	*Yefim Volkov, Misty Afternoon*
	girl with horse at sunset	*change the background to a city*	*girl with horse at sunset in front of city*
	painting-of-forest-and-pond	*Without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

- Used a small dataset of human written data to finetune a GPT-3 model.
- That were used to generate instructions & captions.[2]
- Highlighted text is generated by GPT-3.

[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

Example – step 1

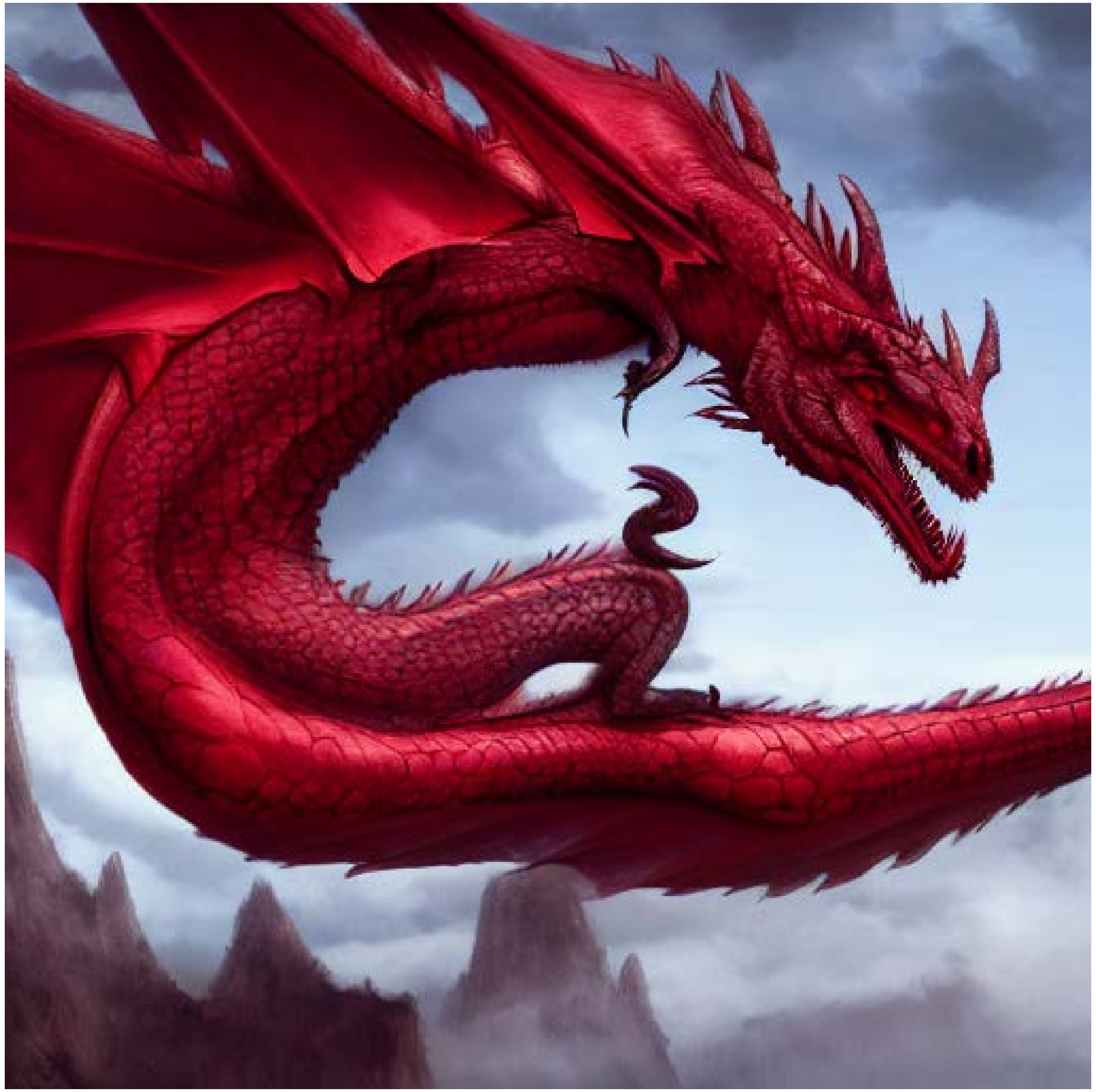
- Image of a dragon, concept art, matte painting, HQ, 4k (Stable diffusion [1])



[1]R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models.” arXiv, 2021. doi: 10.48550/ARXIV.2112.10752.

Example – step 2

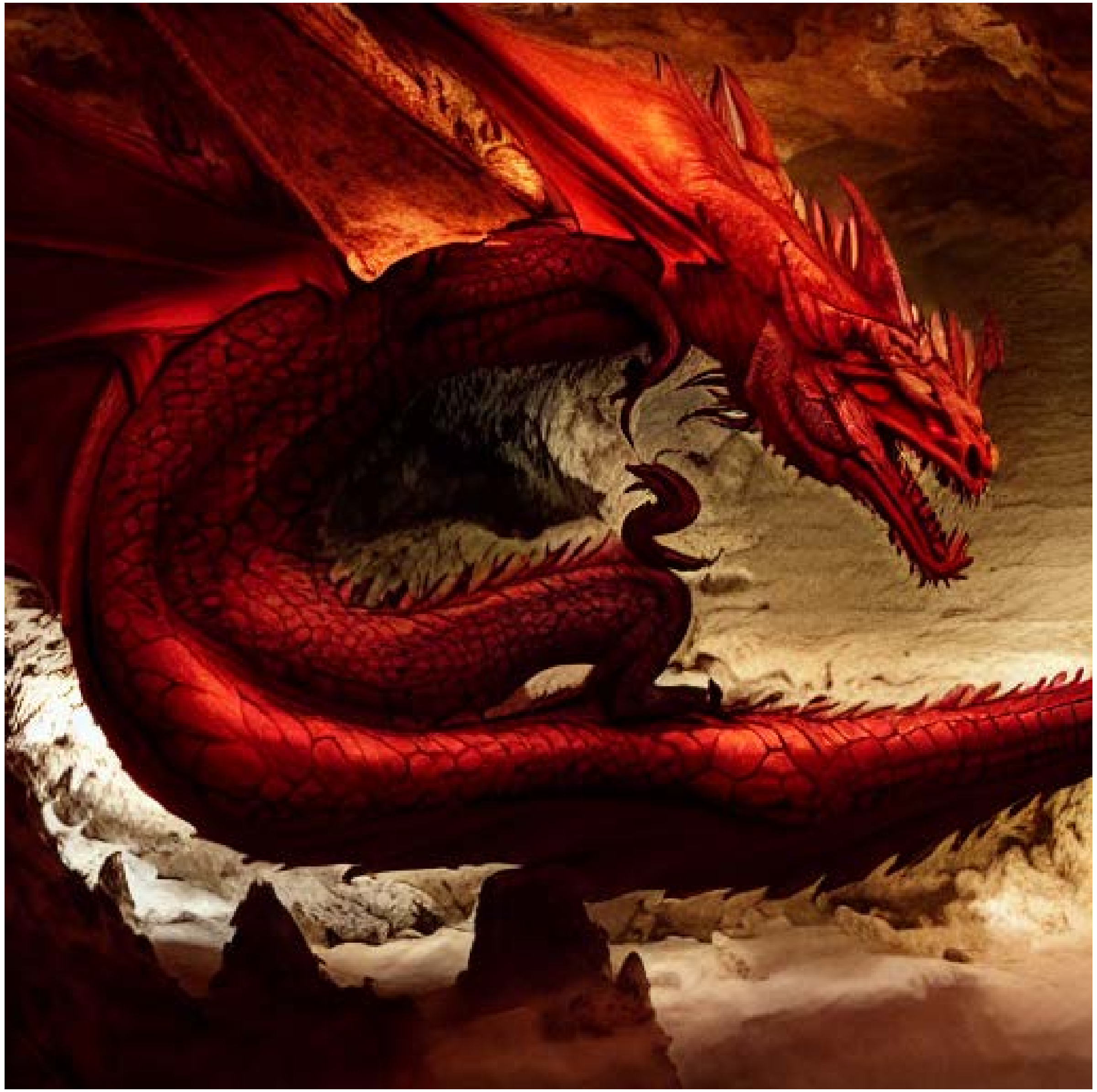
- make only the dragon red (InstructPix2Pix [2])



[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

Example – step 3

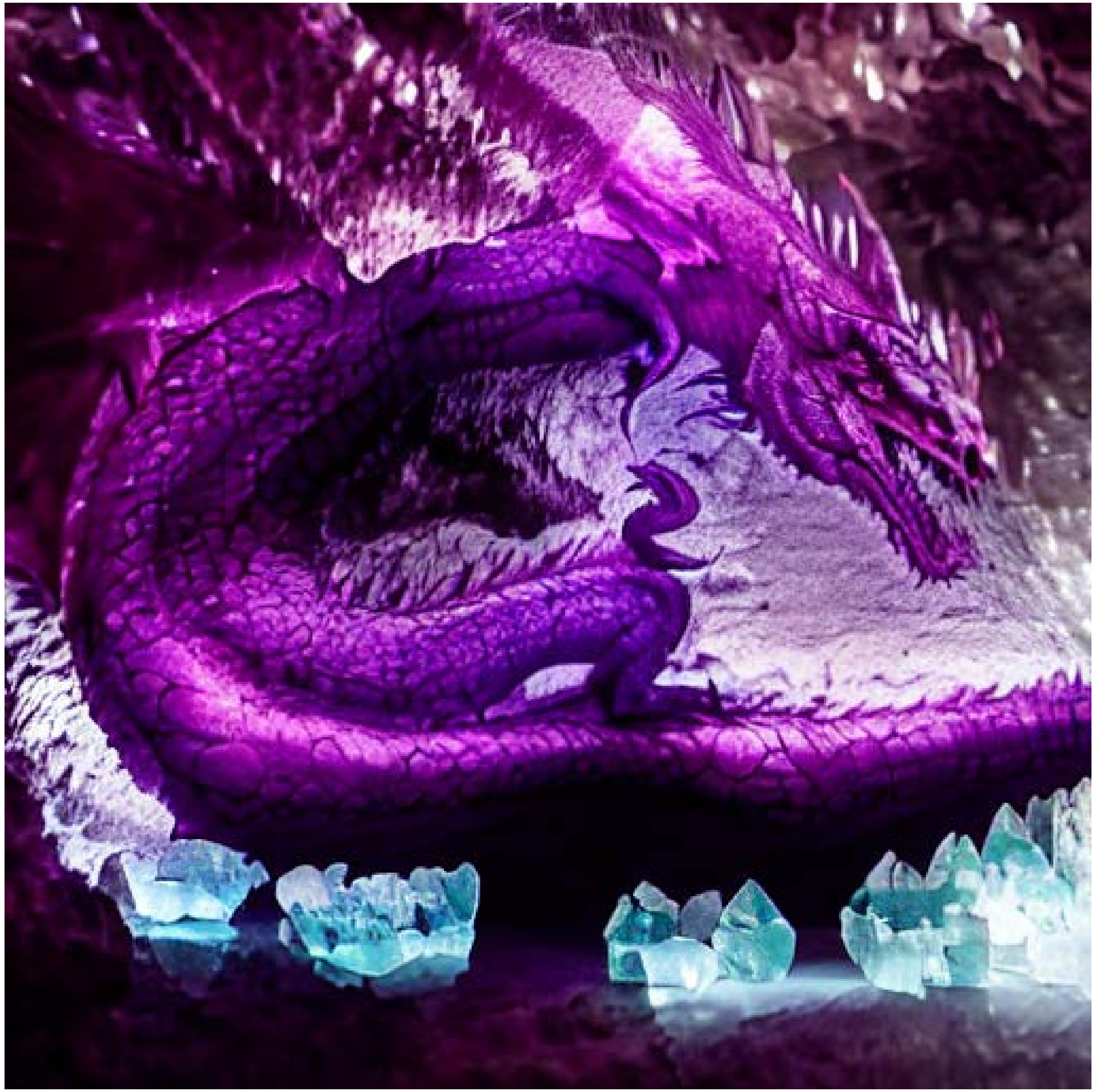
- dragon is inside a cave (InstructPix2Pix [2])



[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

Example – step 4

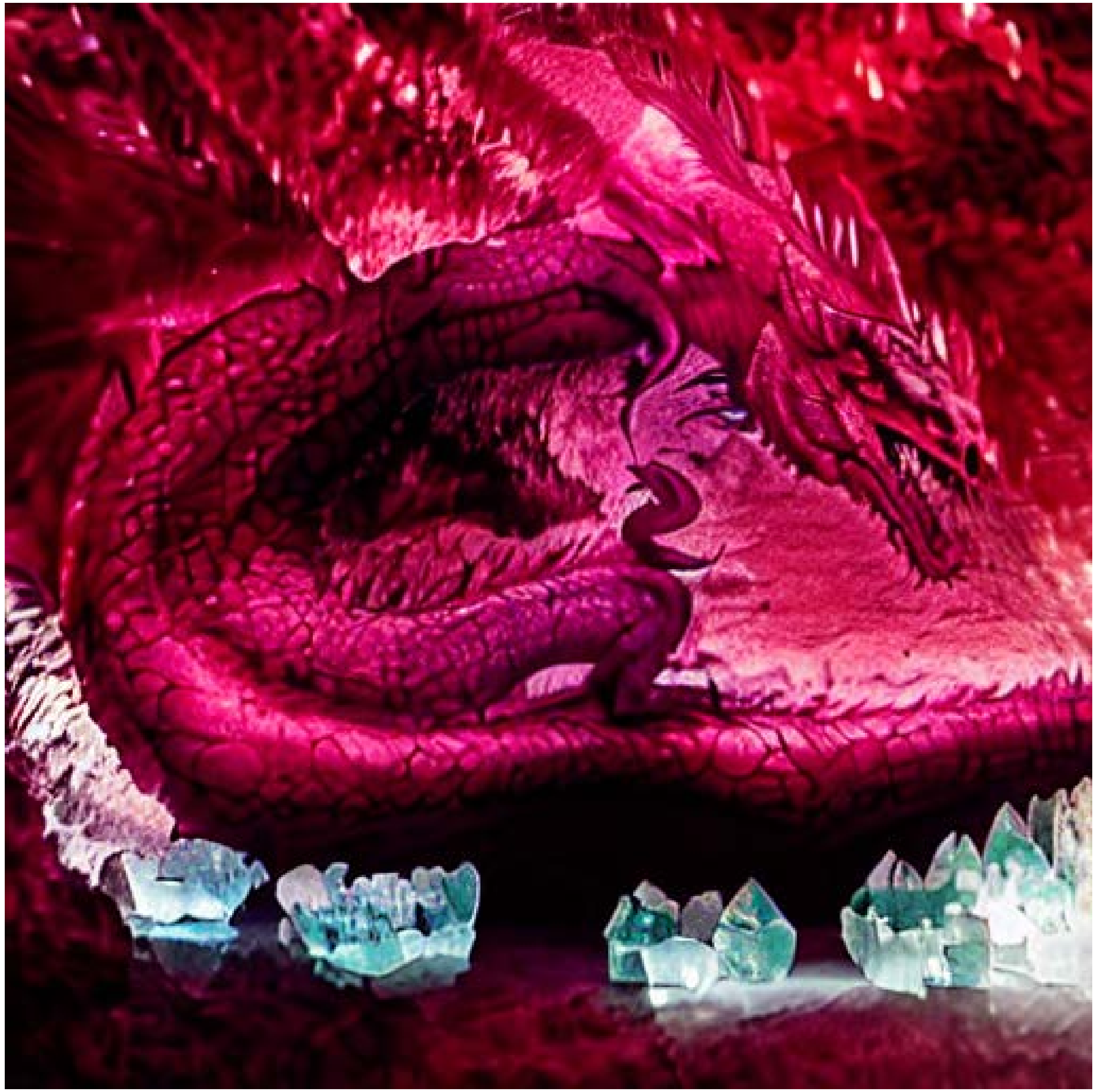
- cave is a crystal cave (InstructPix2Pix [2])



[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

Example – step 5

- make only the dragon red (InstructPix2Pix [2])



[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

Comparison

Image of a red dragon in a crystal cave, concept art, matte painting, HQ, 4k

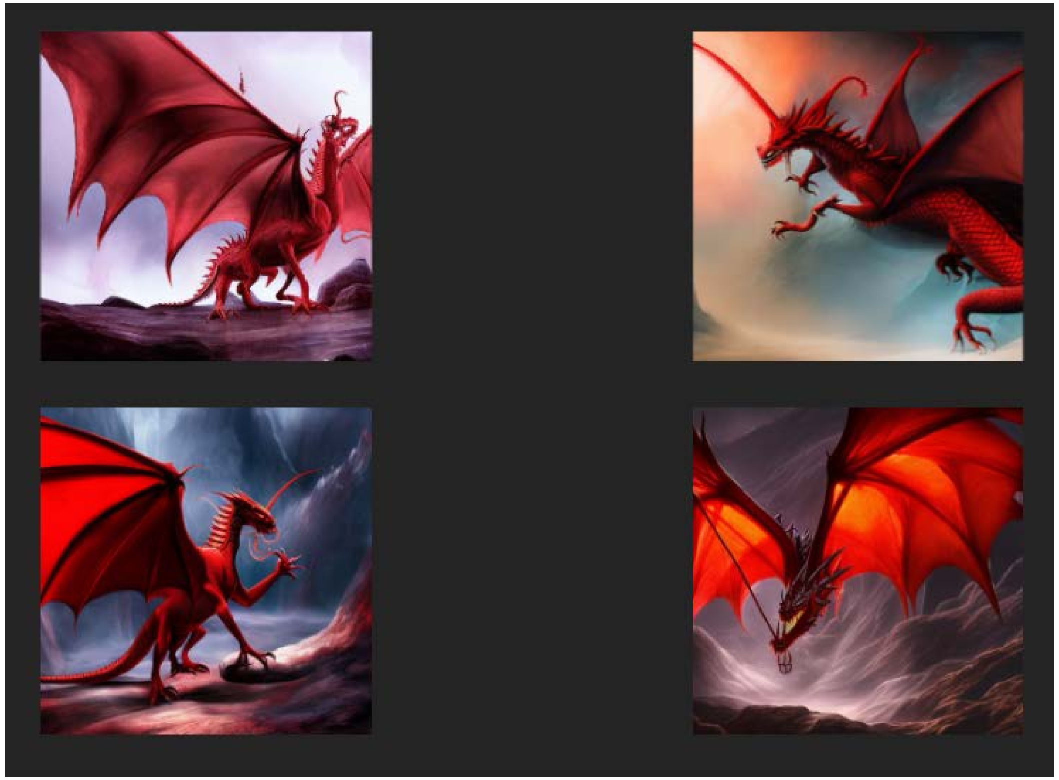
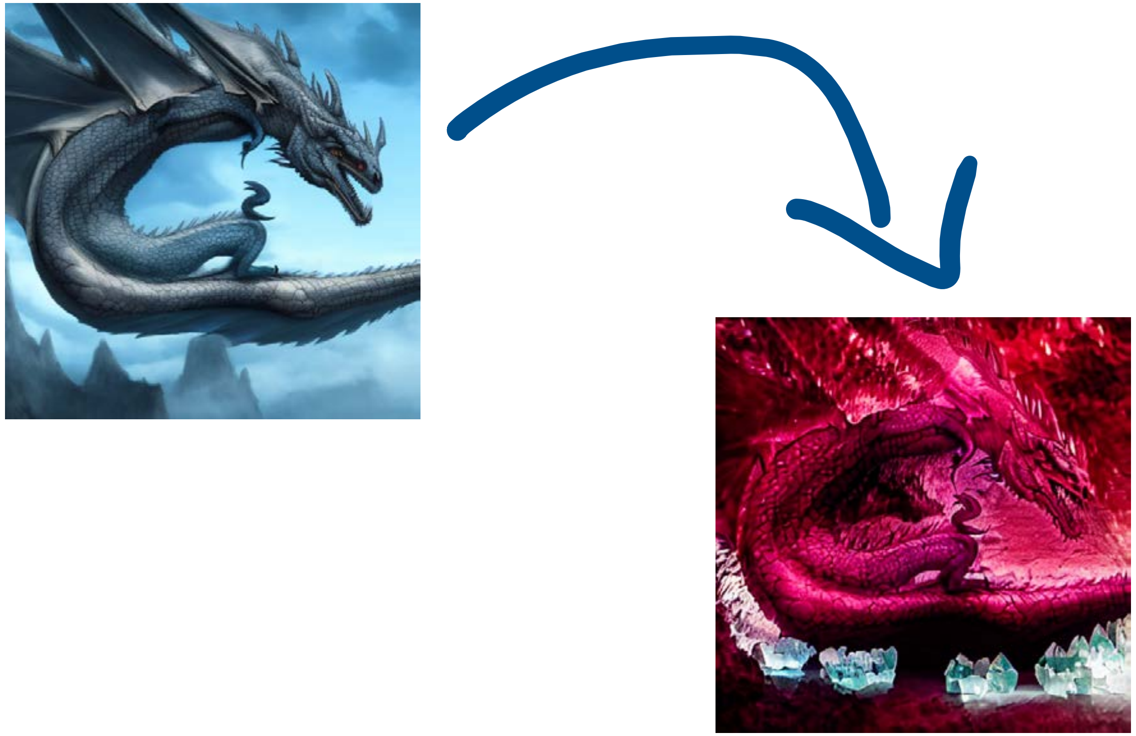
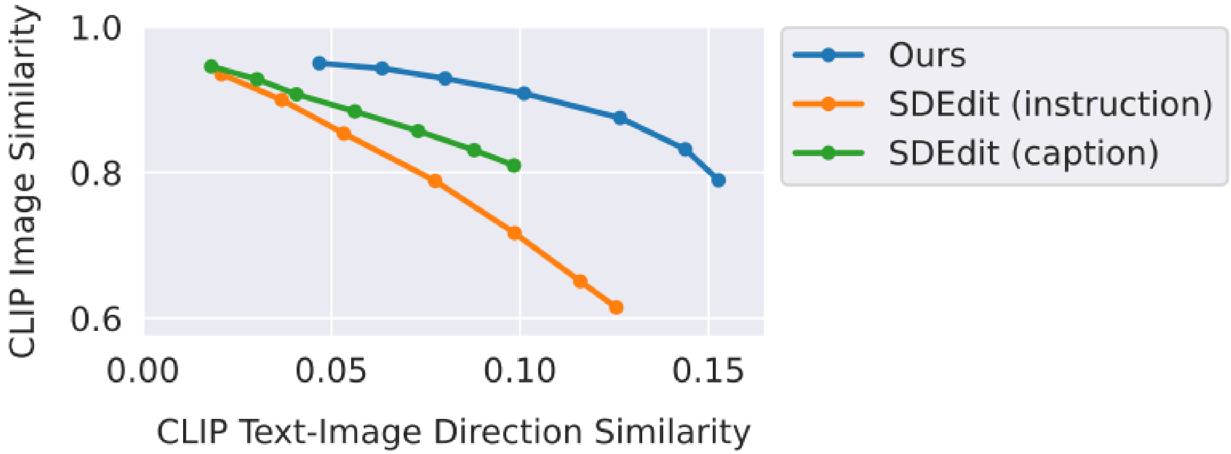


Image of a dragon, concept art, matte painting, HQ, 4k



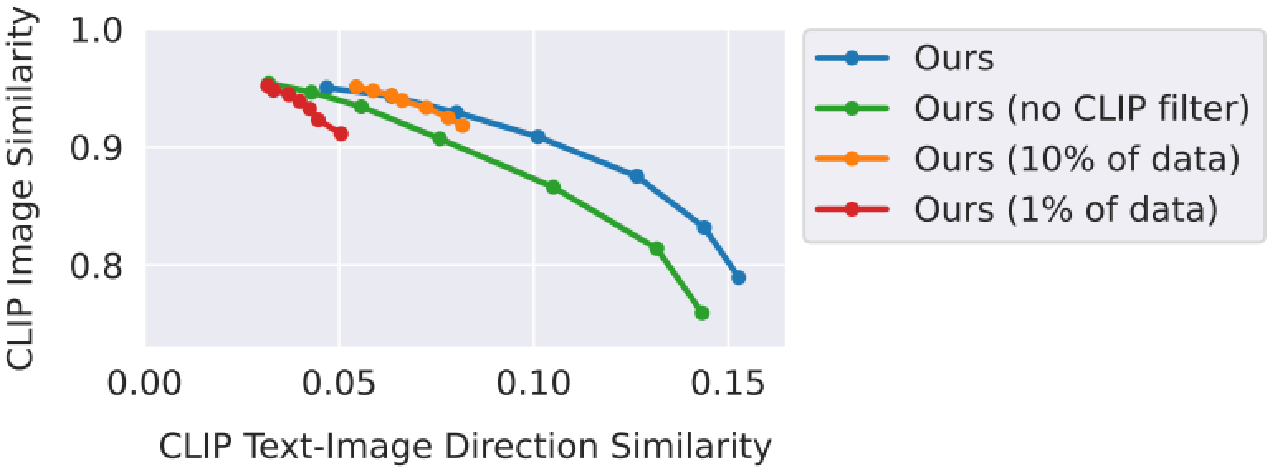
Results



- The trade-off between consistency with the input image (Y-axis) and consistency with the edit (X-axis). [2]
- For both metrics, higher is better.

[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

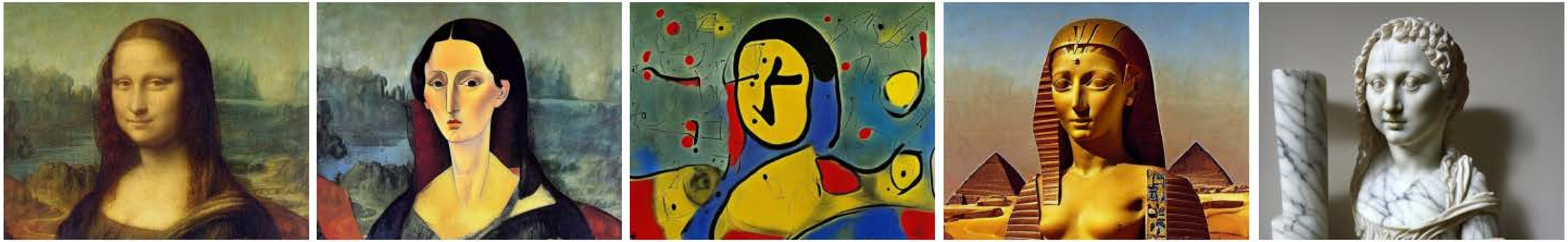
Results Cont...



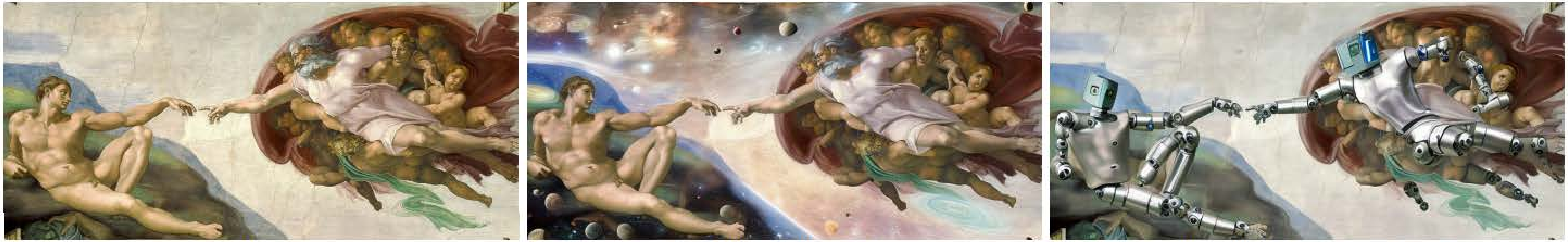
Compare ablated variants of their model (smaller training dataset, no CLIP filtering)[2]

[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

Results Cont…



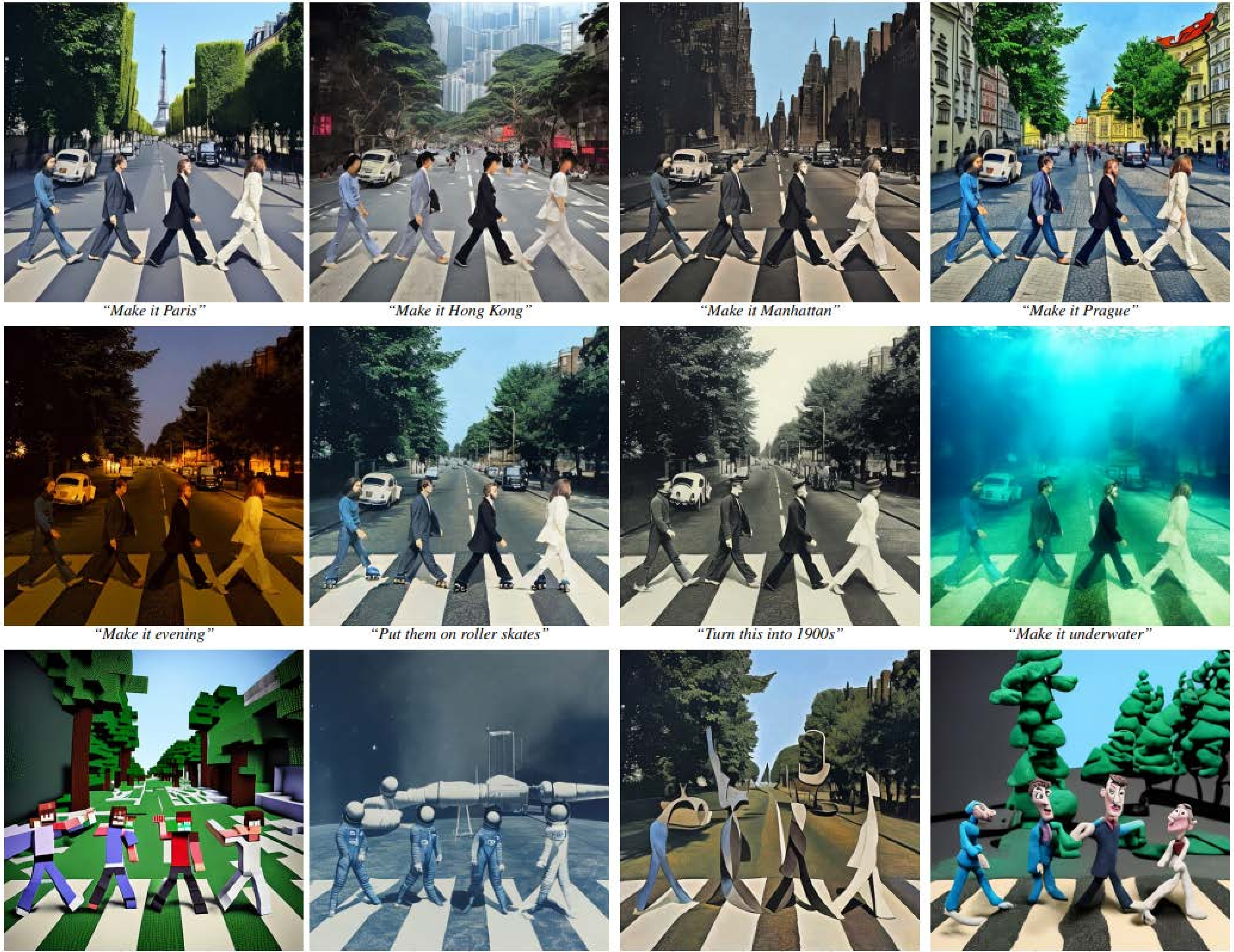
Input *"Make it a Modigliani painting"* *"Make it a Miro painting"* *"Make it an Egyptian sculpture"* *"Make it a marble roman sculpture"*



Input *"Put them in outer space"* *"Turn the humans into robots"*

[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

Results Cont…



"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"

[2]Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.

References

- R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models.” arXiv, 2021. doi: 10.48550/ARXIV.2112.10752.
- Brooks, T., Holynski, A., & Efros, A. A. (2022). Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800.