

Lecture 8:

Generative AI for Image Creation (Initial discussion)

**Visual Computing Systems
Stanford CS348K, Spring 2024**

Demo

Many exciting opportunities, but also many issues with emerging class of generative AI technologies

- **Quality/diversity of output images**
- **Performance (cost of training and cost of image generation)**
- **User control and creative workflow**
- **Ethics / social aspects**

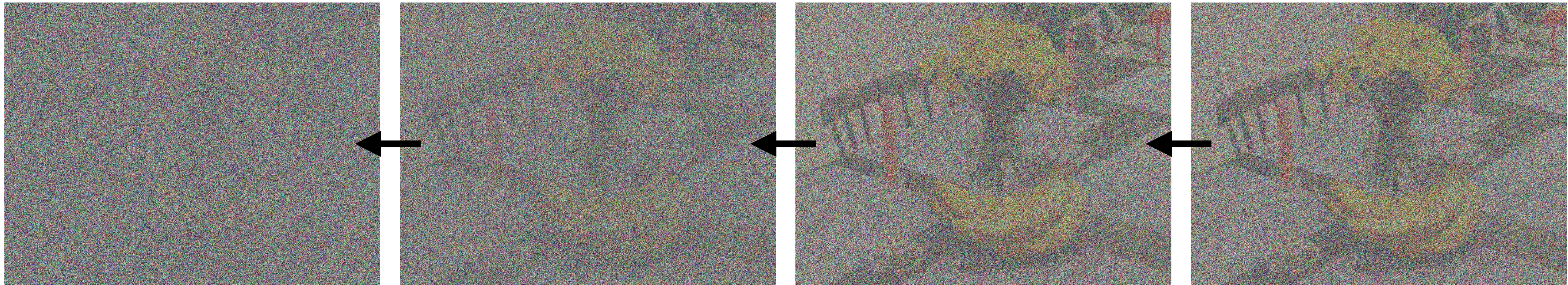
Suppose you are given a data of images x_i

- You'll like to draw a sample according to the underlying data distribution $p(x)$

Diffusion-based image synthesis

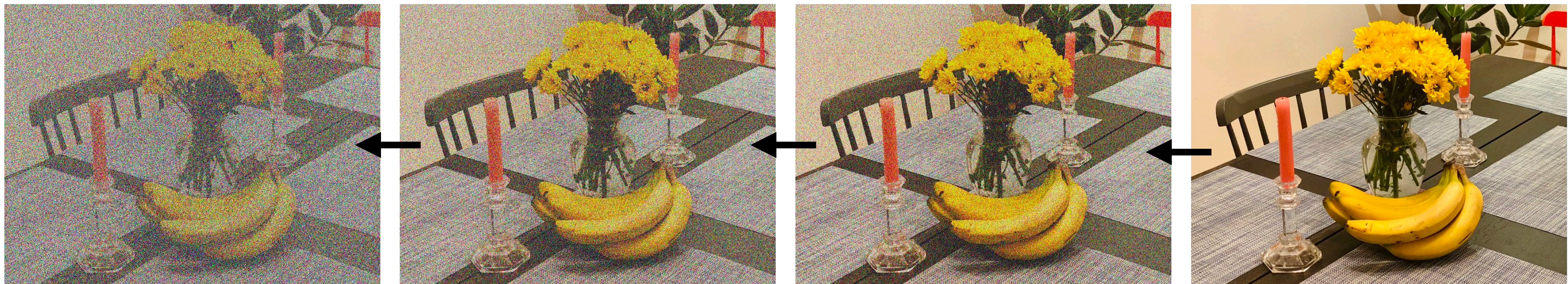
Iterative Markov-chain Monte-carlo (MCMC) process to generate a sample \mathbf{x} (an image) from distribution $p(\mathbf{x})$ of observed images

Forward diffusion: iteratively add noise $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$



\mathbf{x}_T

\mathbf{x}_{T-1}



\mathbf{x}_1

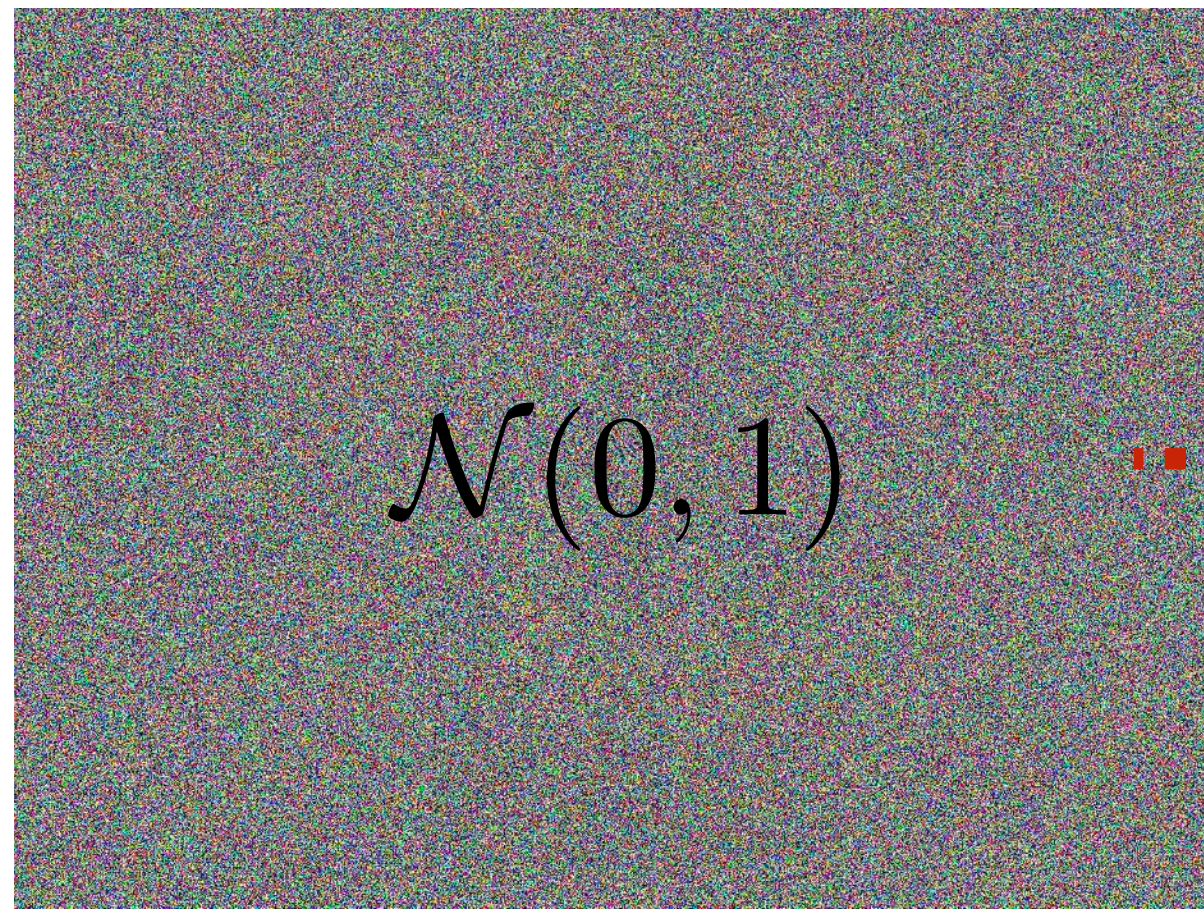
\mathbf{x}_0

Diffusion-based image synthesis

Reverse: iteratively remove noise from random sample to obtain image from $p(\mathbf{x})$

$$\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon} \mathbf{z}_i, \quad i = 0, 1, \dots, T$$

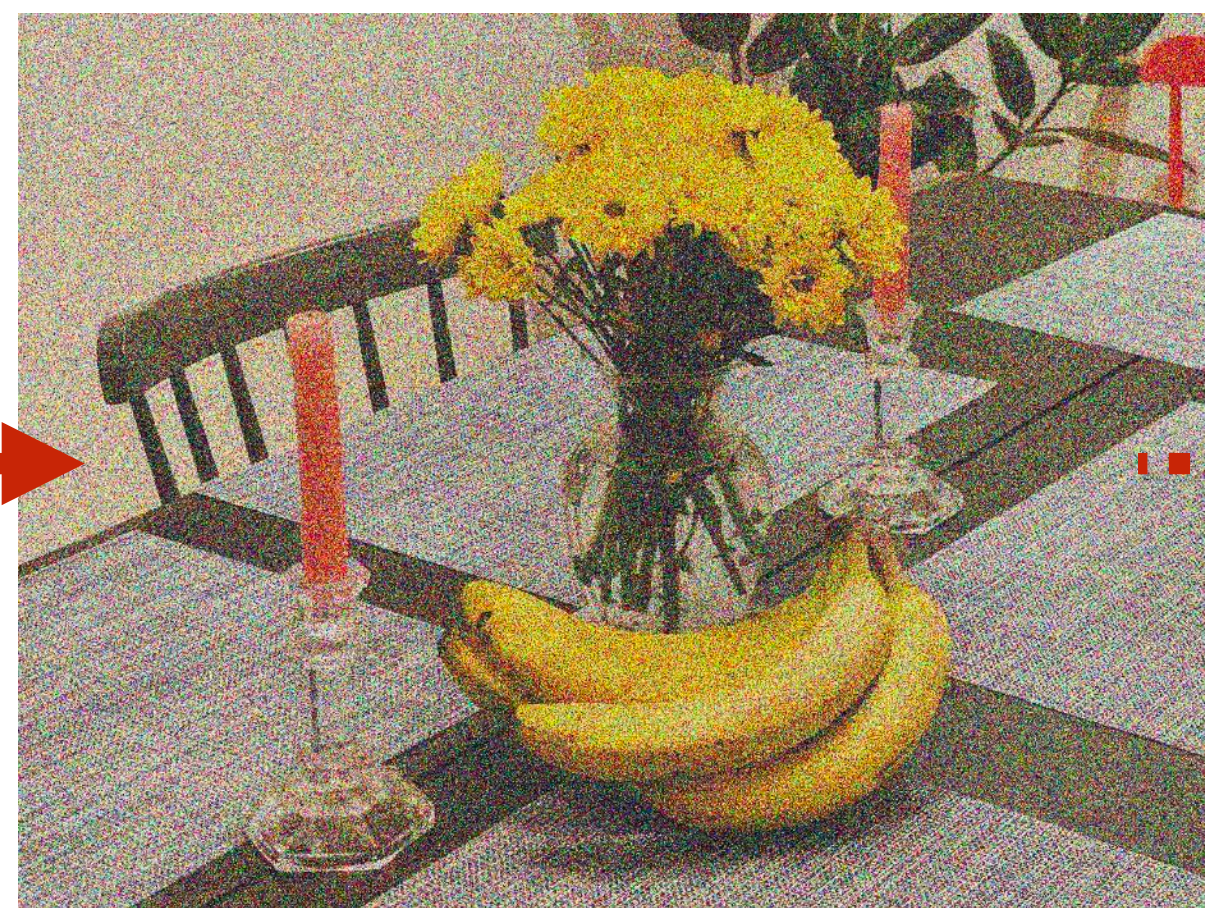
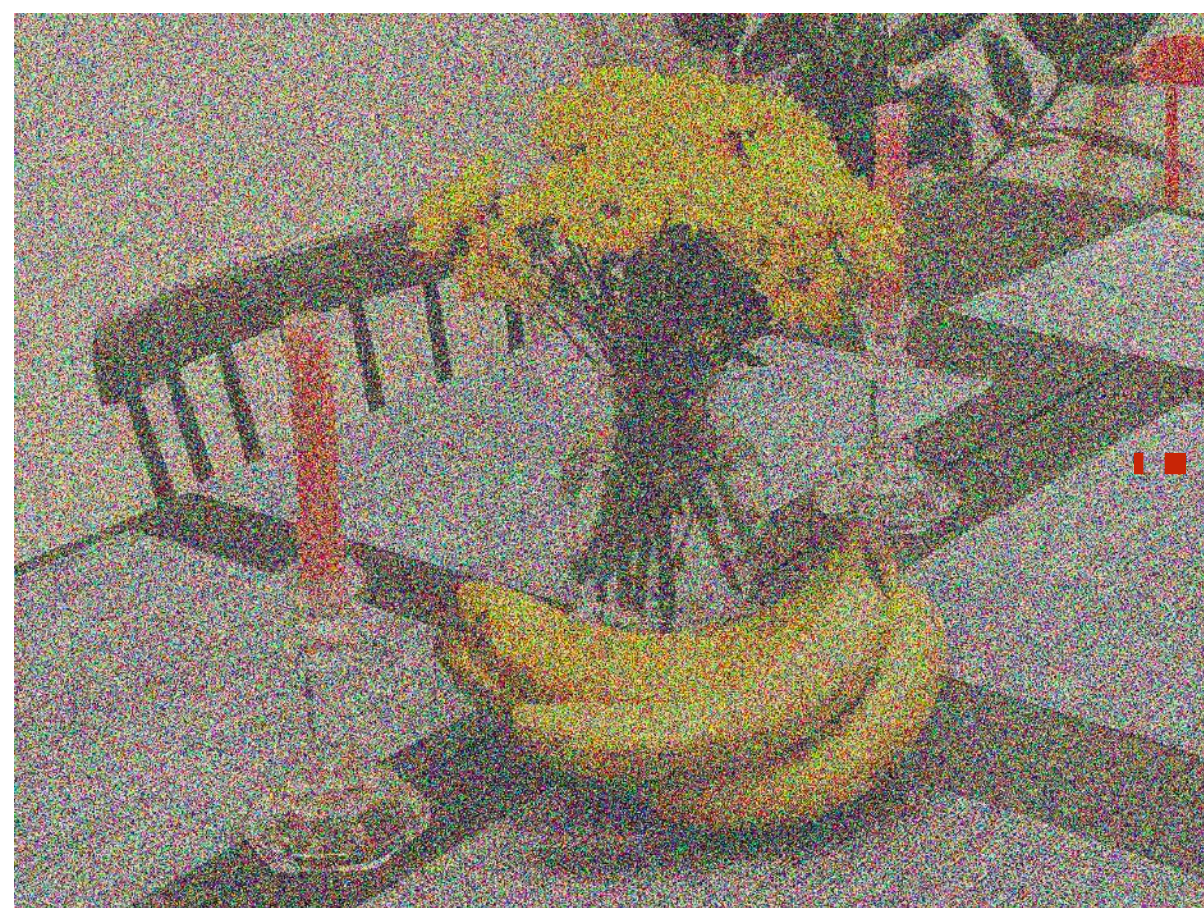
("score function")



\mathbf{x}_0



\mathbf{x}_1



\mathbf{x}_{T-1}



\mathbf{x}_T

Guided diffusion

- Assume we know $p(\mathbf{y} \mid \mathbf{x})$ for random variables \mathbf{x} and \mathbf{y} .
 - Example: \mathbf{x} is an image, \mathbf{y} is a string describing the image
 - Given an image (\mathbf{x}), infer a caption (\mathbf{y})

$$p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{x})p(\mathbf{y} \mid \mathbf{x}) / \int p(\mathbf{x})p(\mathbf{y} \mid \mathbf{x})d\mathbf{x} \quad \text{(Bayes Rule)}$$

Bayes for score function

$$\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$$

↑
(Unguided score function)

Modify image \mathbf{x} so that image is more likely
[to come from the training set]

← (Prompt guidance)

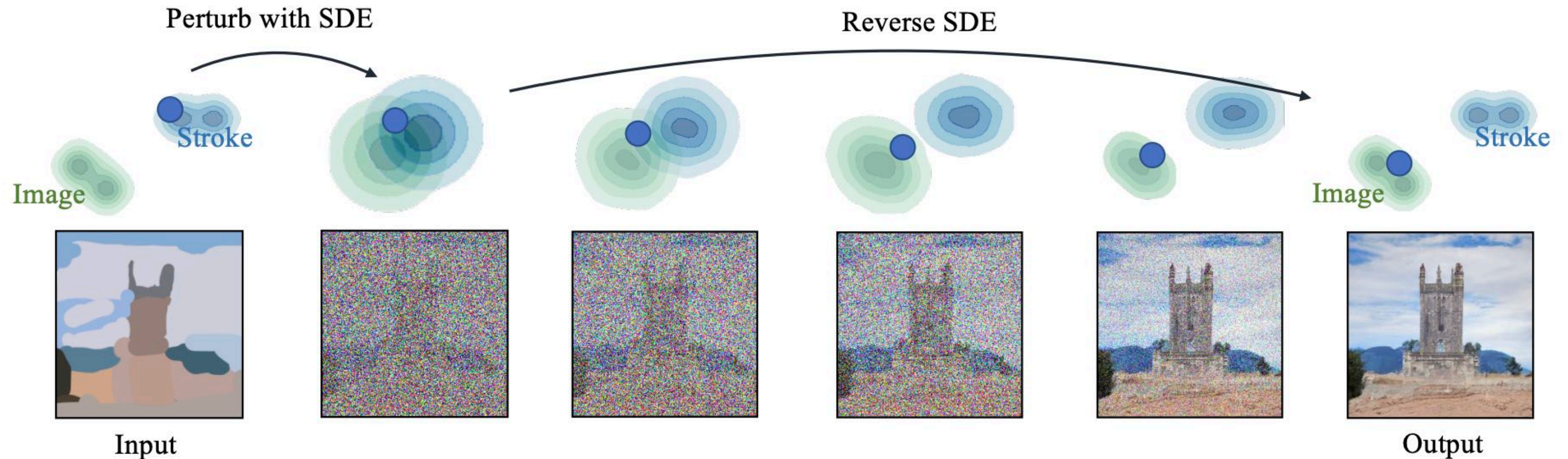
Modify image \mathbf{x} to make the prompt a
more likely description of the image

Controlling the output of diffusion models

img2img (enabling image-based guidance)

1. Start with a guide image (a target)
2. Add “small” amount of noise
3. Iteratively denoise to produce sample from $p(x)$

“Guide toward a visual target”



Inpainting (apply [new] prompt to a region)

User specifies mask for region of interest and text prompt for that region.

Image outside of region remains almost the same.



"bowl of water"



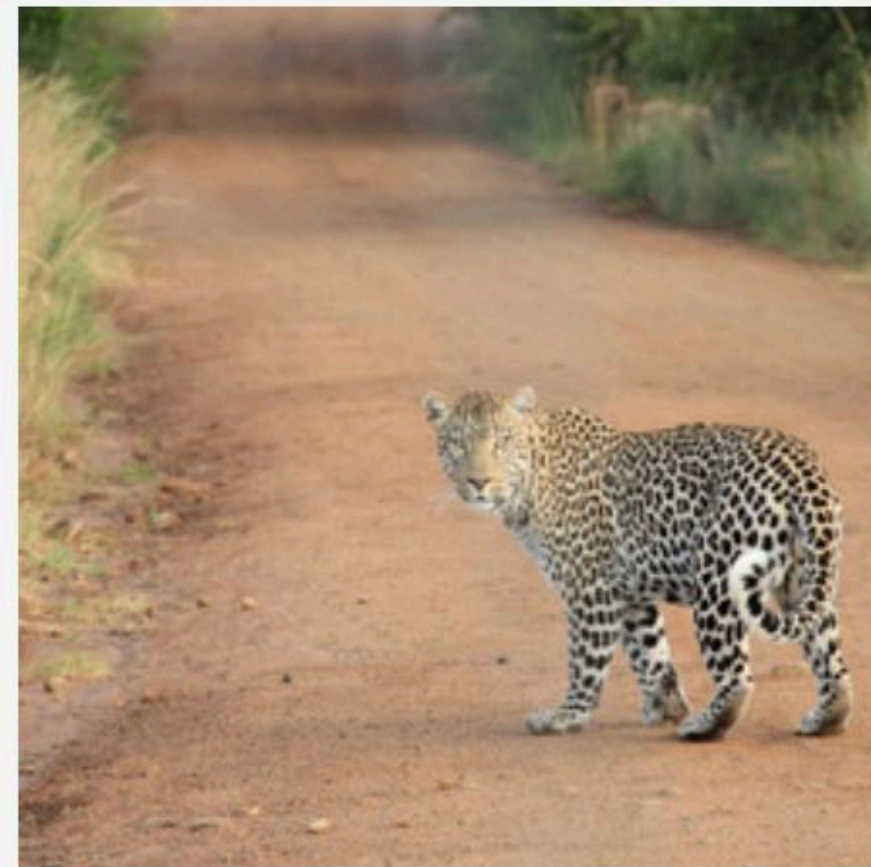
"stool"



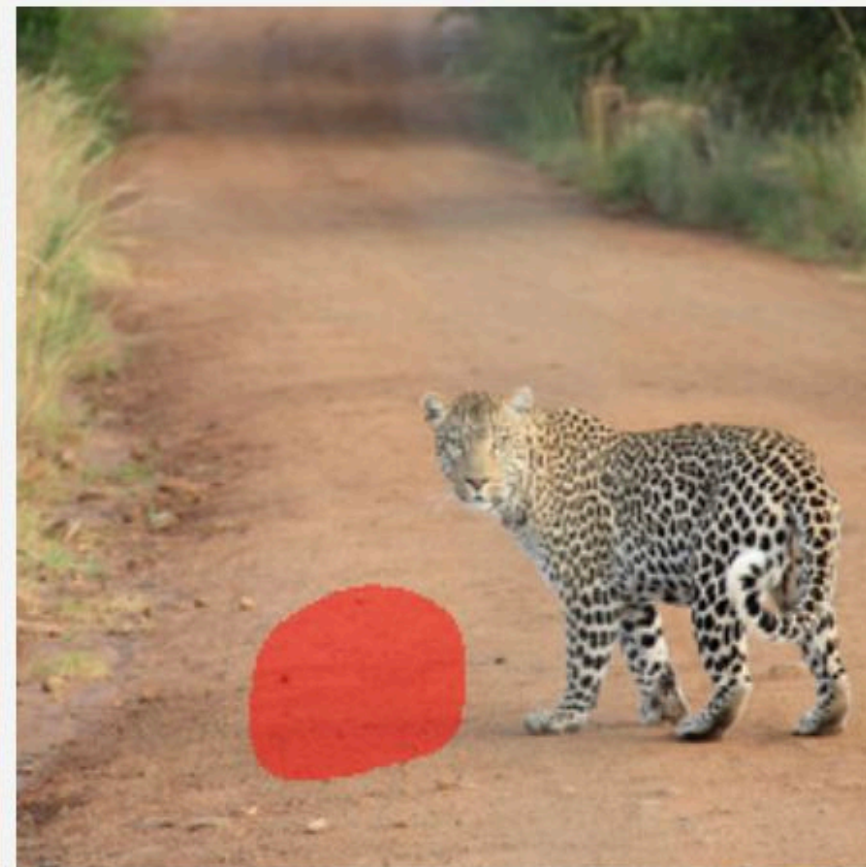
"hole"



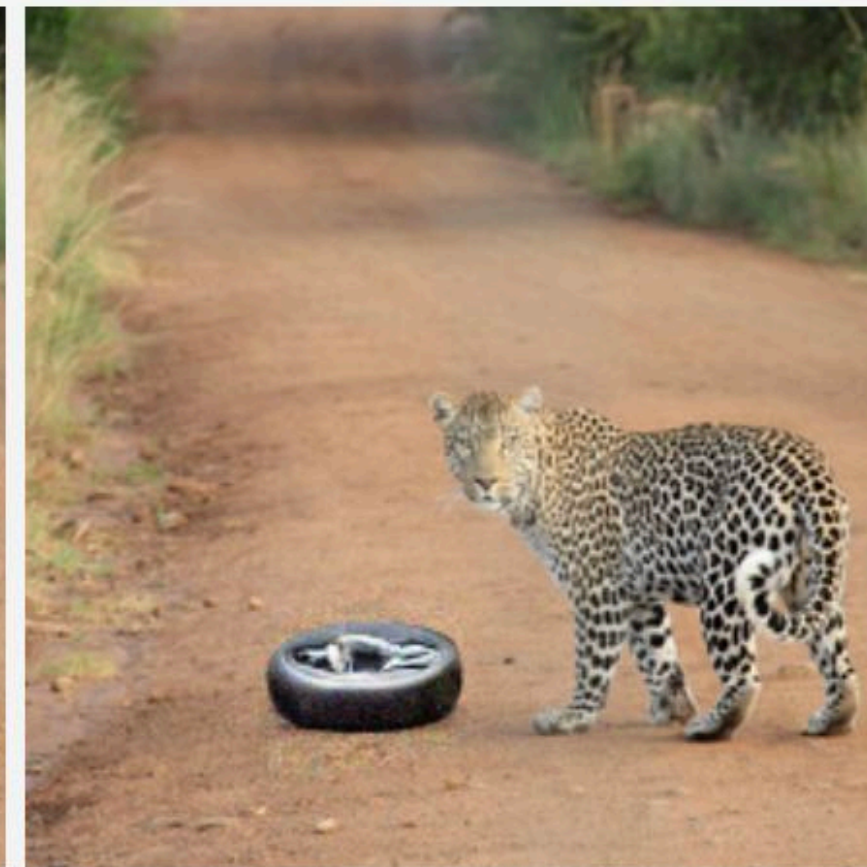
"red brick"



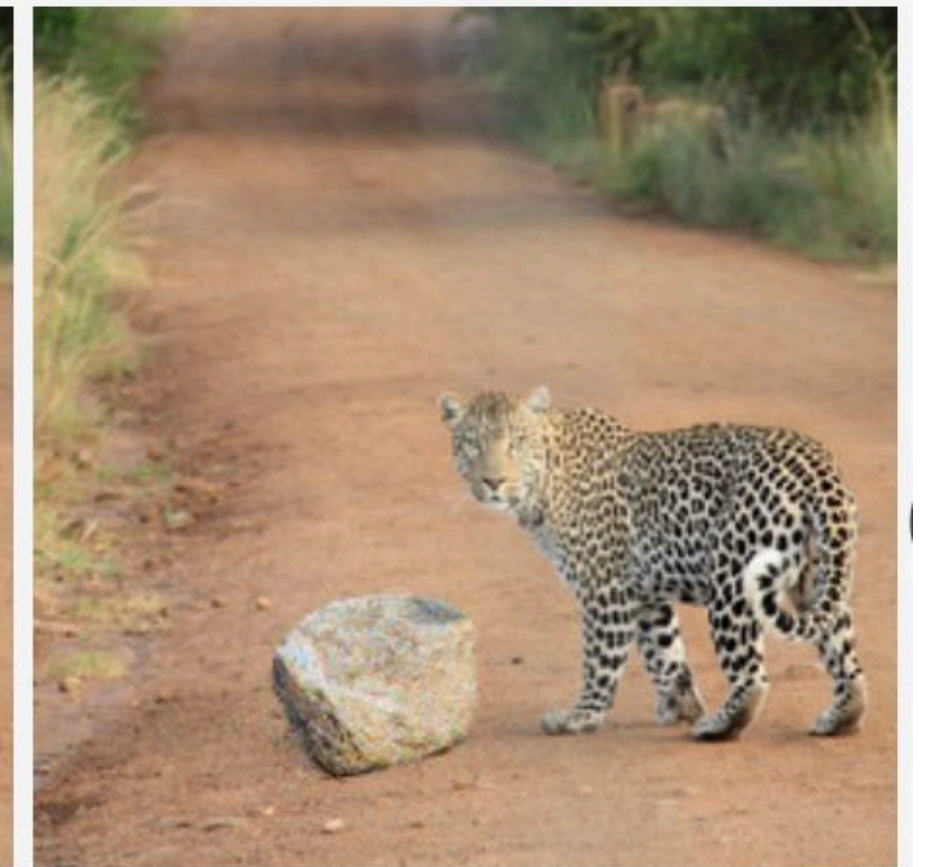
Input image



Input mask



"car tire"



"big stone"

Use change in text prompt to trigger change in image

“A basket full of apples.”



Source image



apples → cookies



basket → bowl



basket → box



basket → nest



apples → oranges

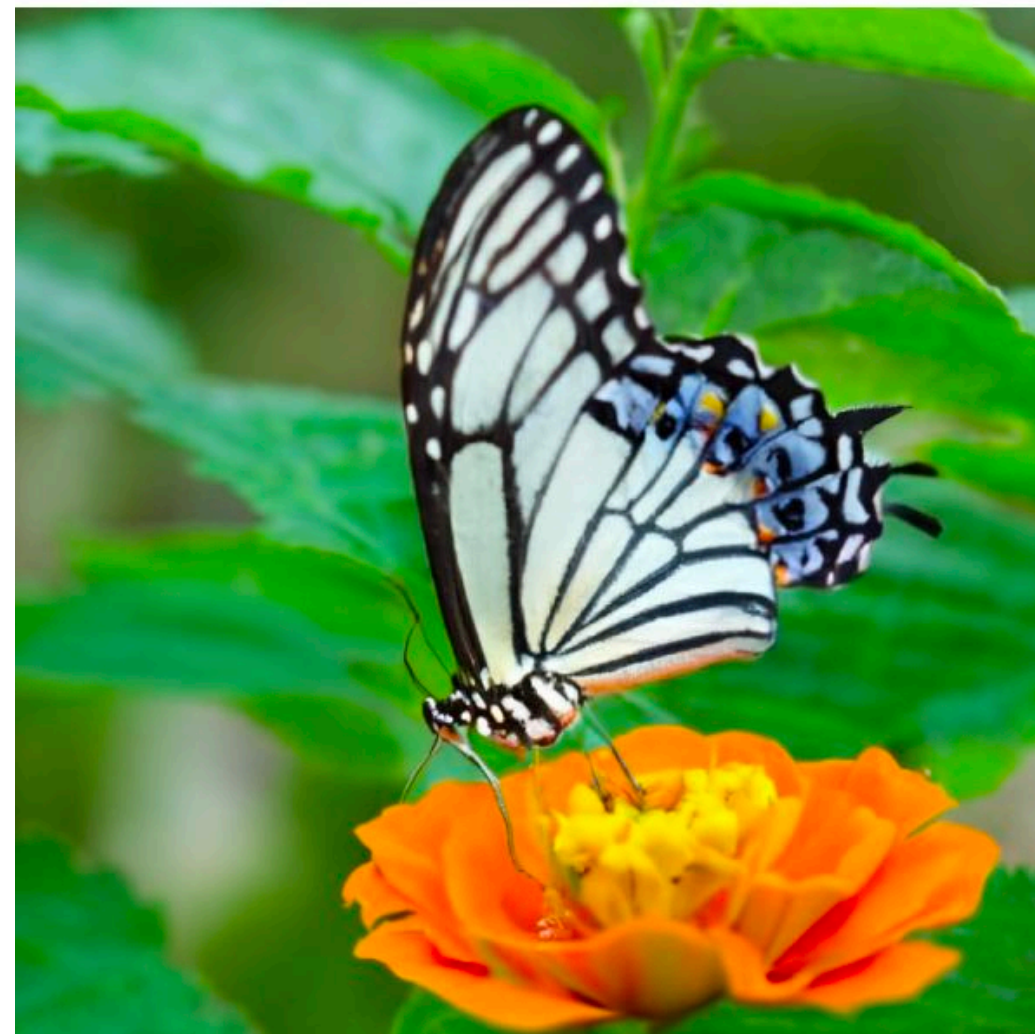


apples → chocolates



apples → kittens

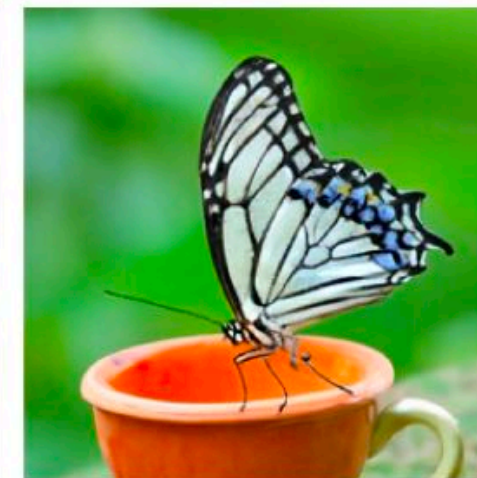
“A photo of a butterfly on a flower.”



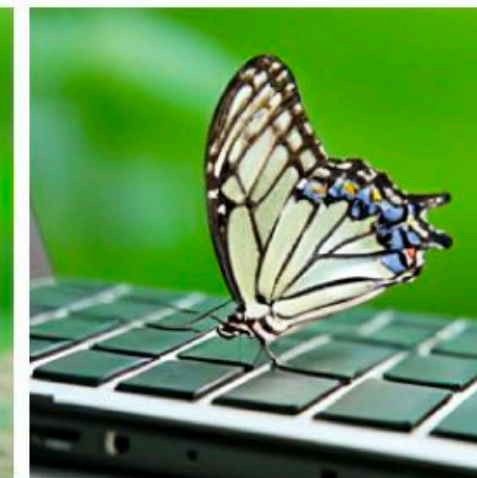
Source image



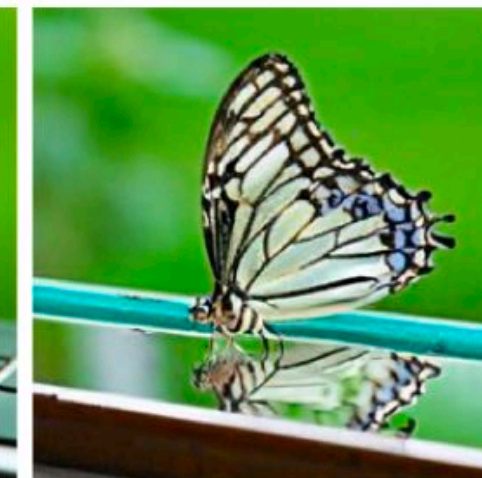
flower → bread



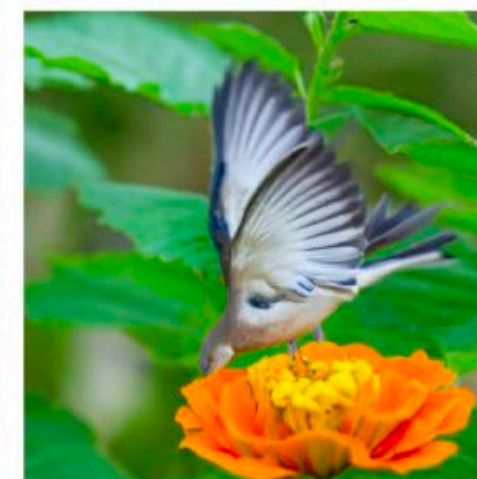
flower → mug



flower → computer



flower → mirror



butterfly → bird



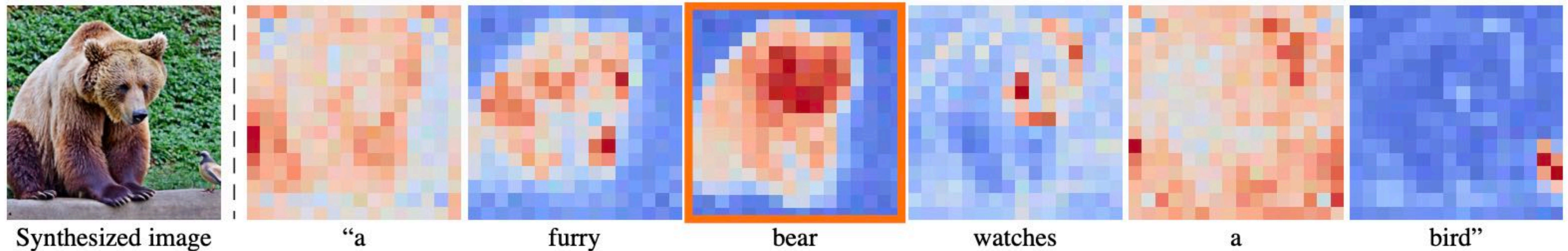
butterfly → snail



butterfly → drone

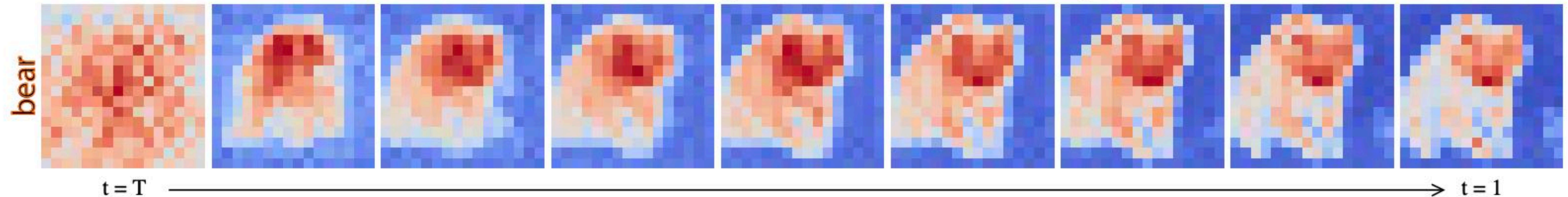
“Masks” come from learned attention

- Use attention masks from original generation process to constrain what pixels can change after prompt is edited



Average cross-attention maps across all timestamps

Cross-attention maps for individual timestamps



Using text to describe how to change the image

"Swap sunflowers with roses"



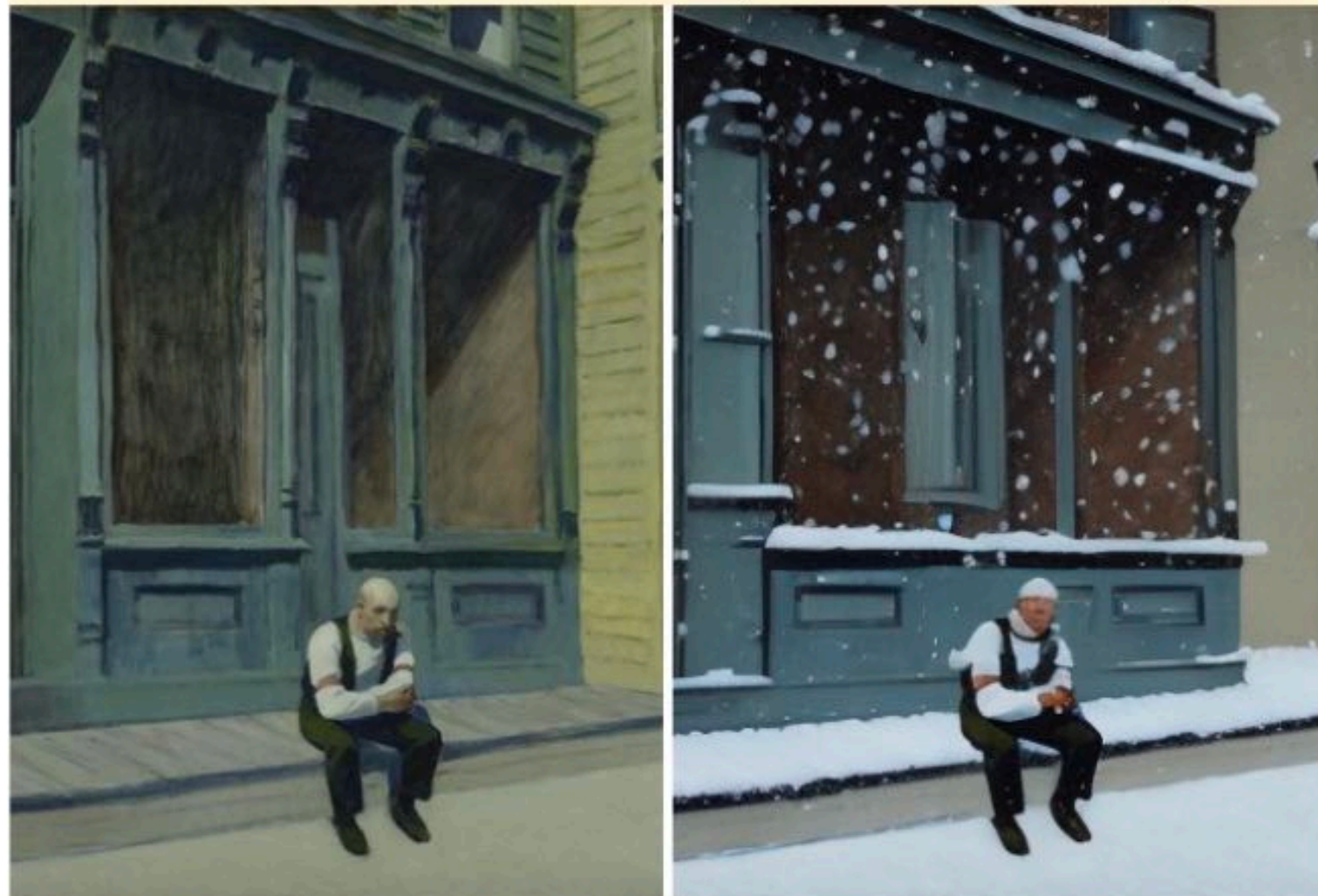
"Add fireworks to the sky"



"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"



Leveraging layer information to control composition

Prompt + a rgba per layer

Ginger



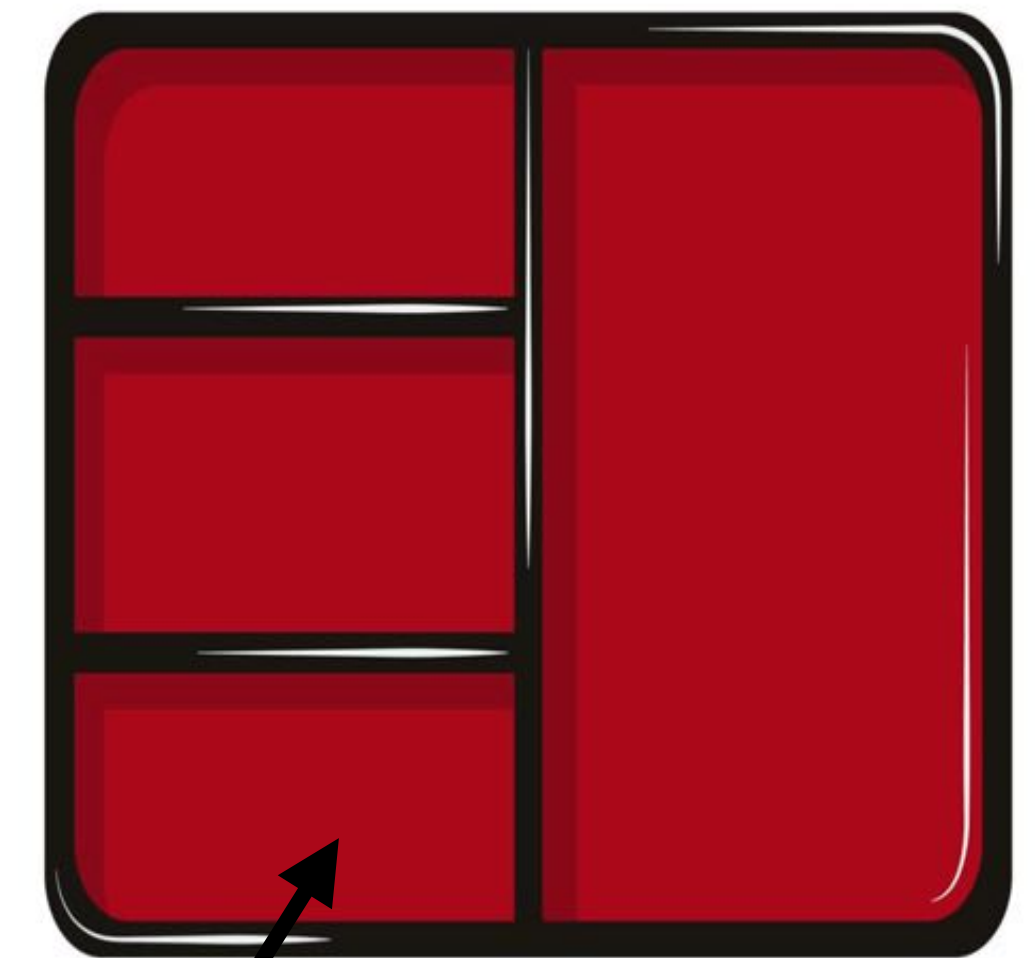
Edamame



Rice



Sushi

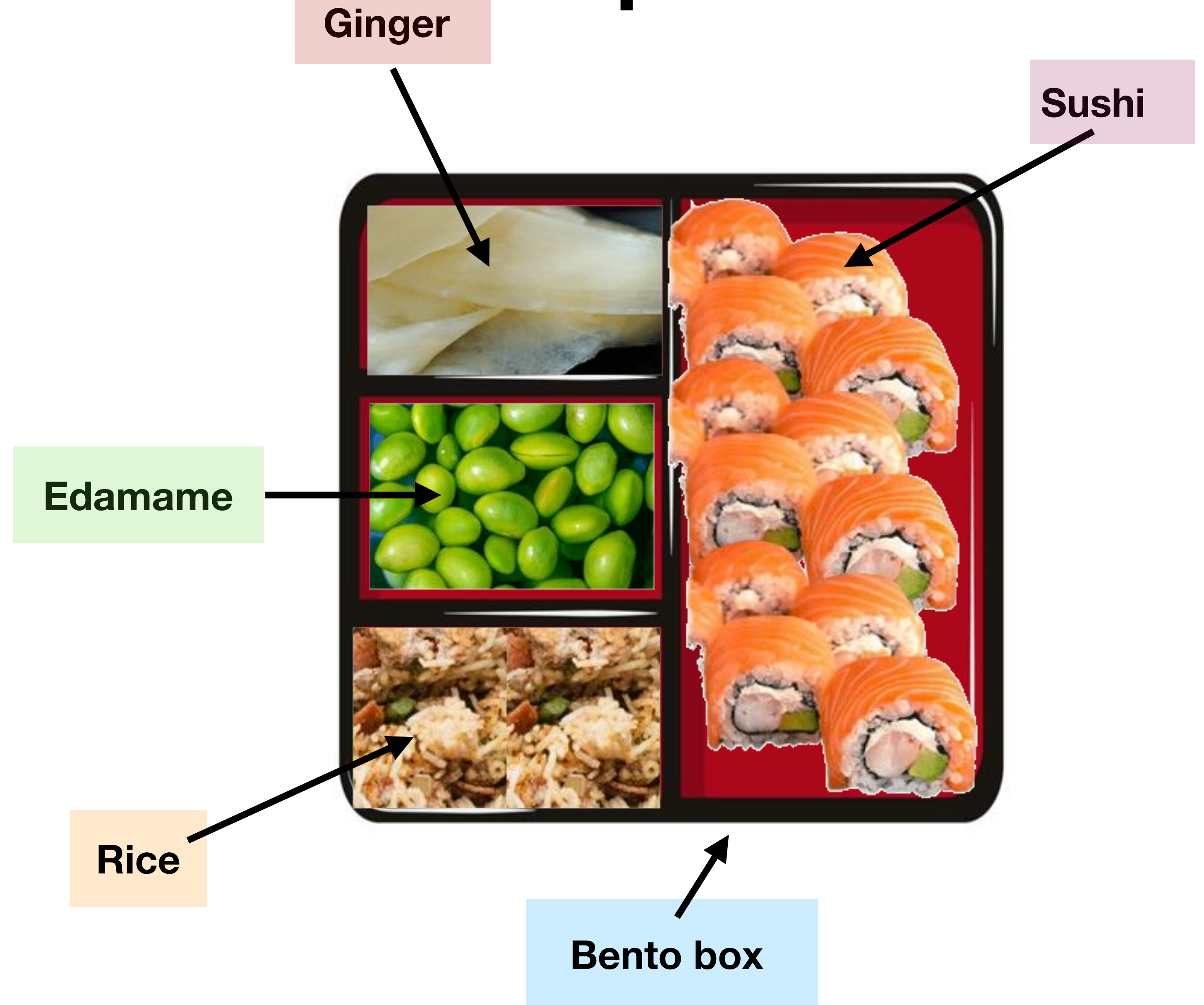


Bento box

- **Key idea: layer-based editing is a tried and true way to manipulate an image's composition.**
 - User manipulates layers
 - Model receives per-layer information, and leverages this information to generate a globally harmonious image

Leveraging layer information to control composition

“A bento box with
rice,
edamame,
ginger, and
sushi.”



Leveraging layer information to control composition

“A bento box with

rice,


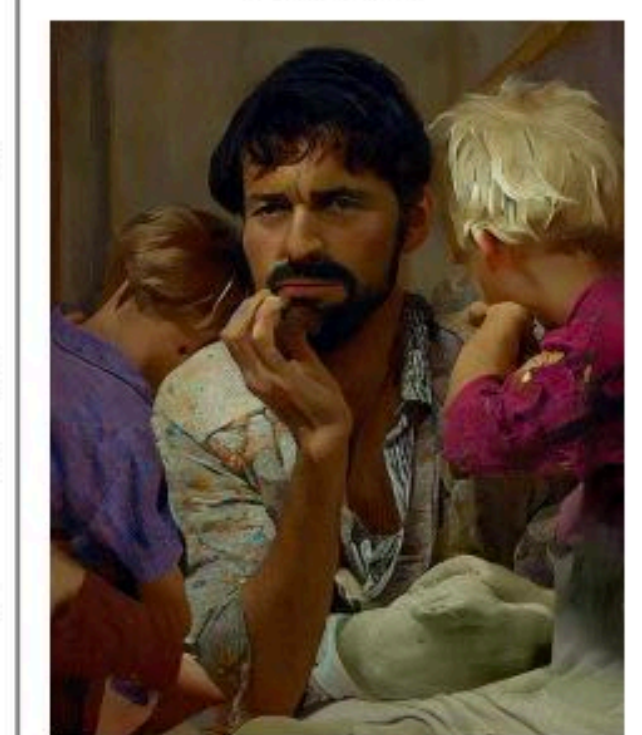
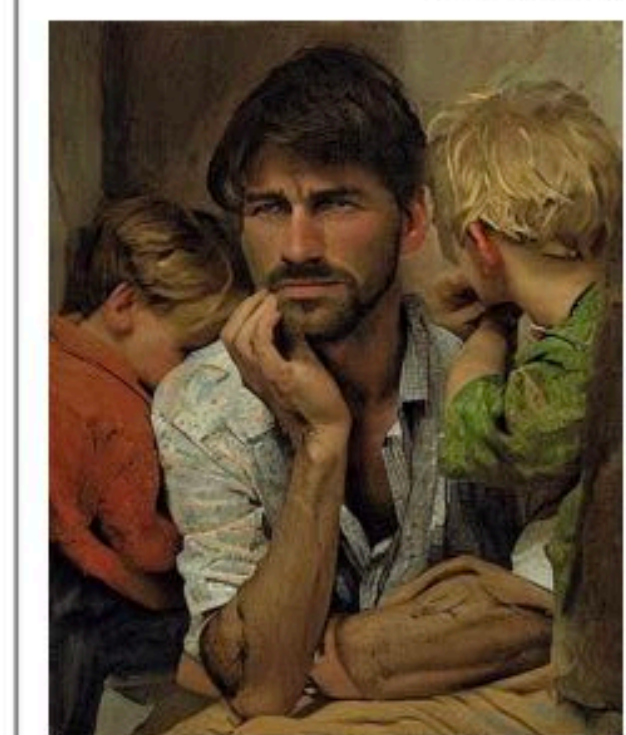

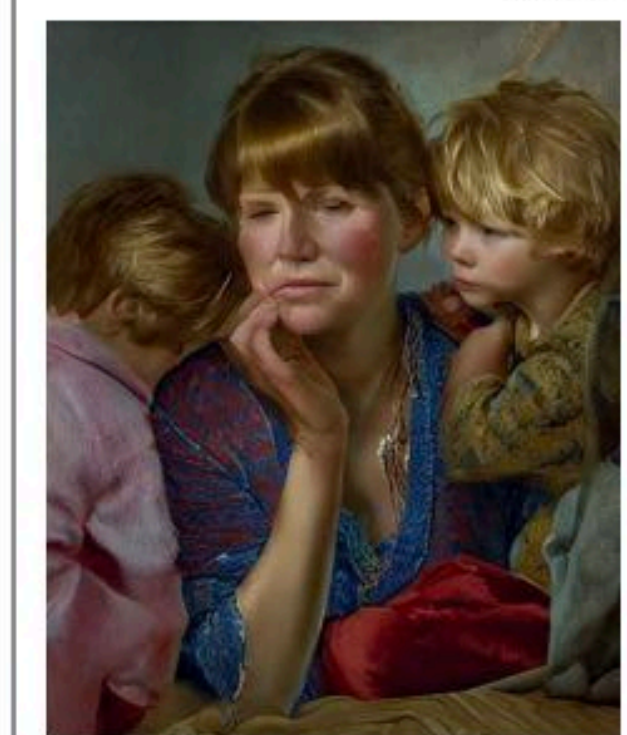


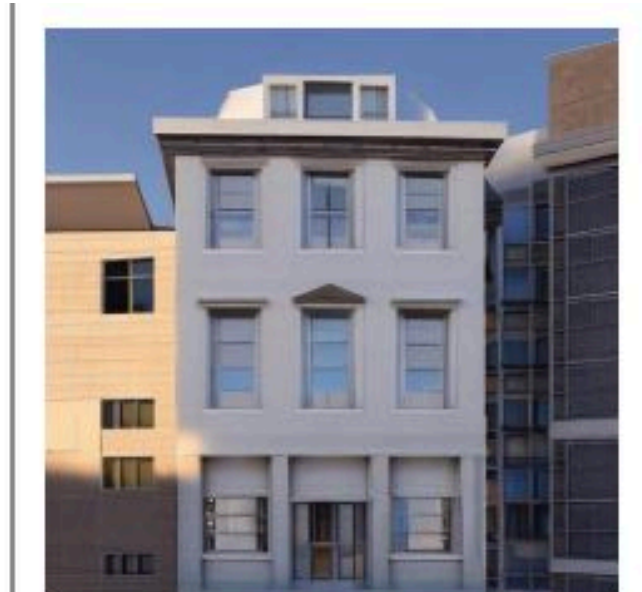
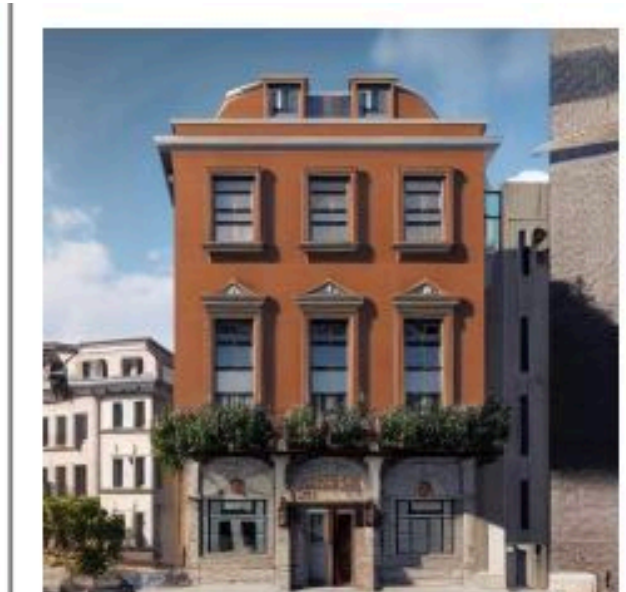
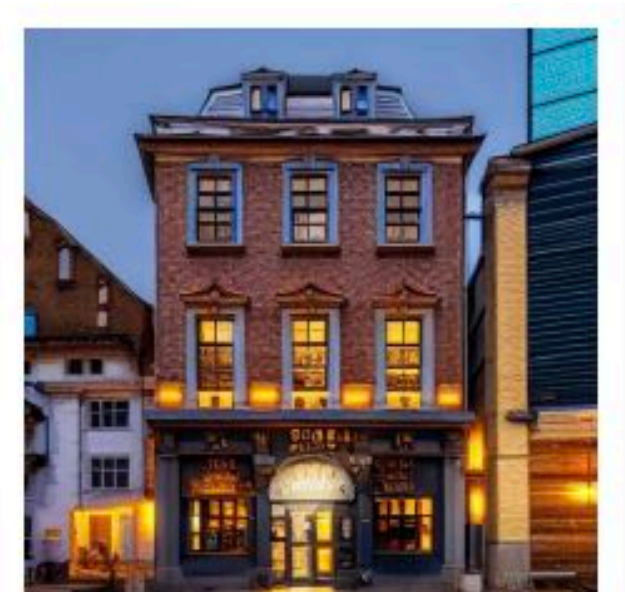
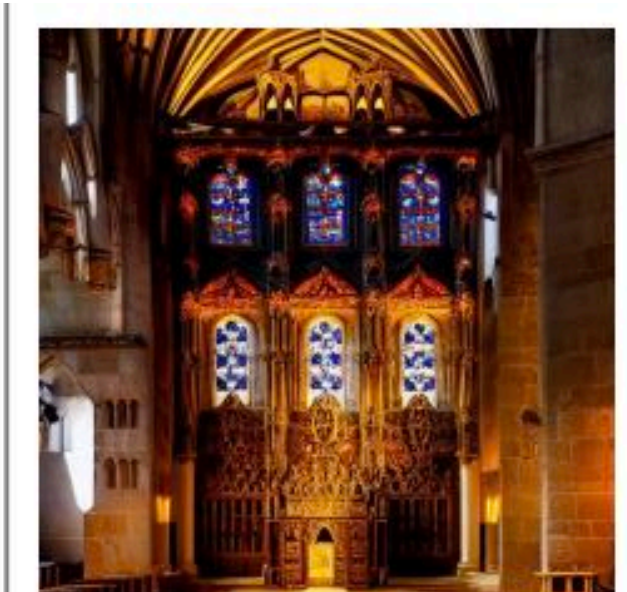
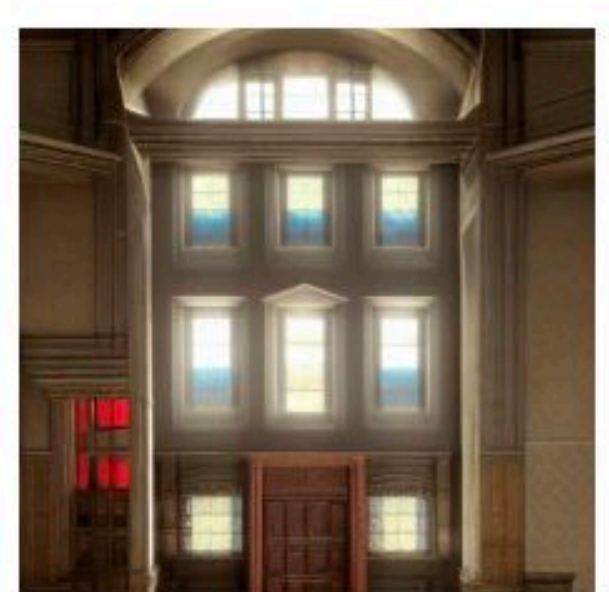
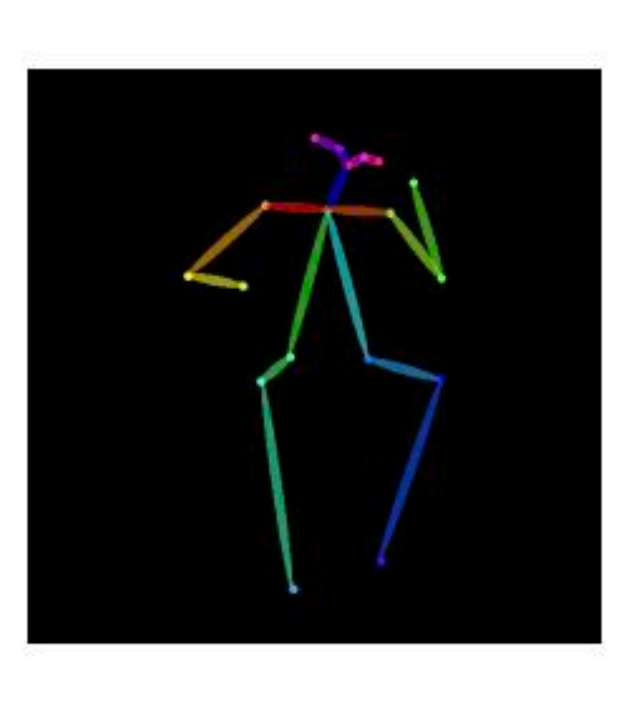
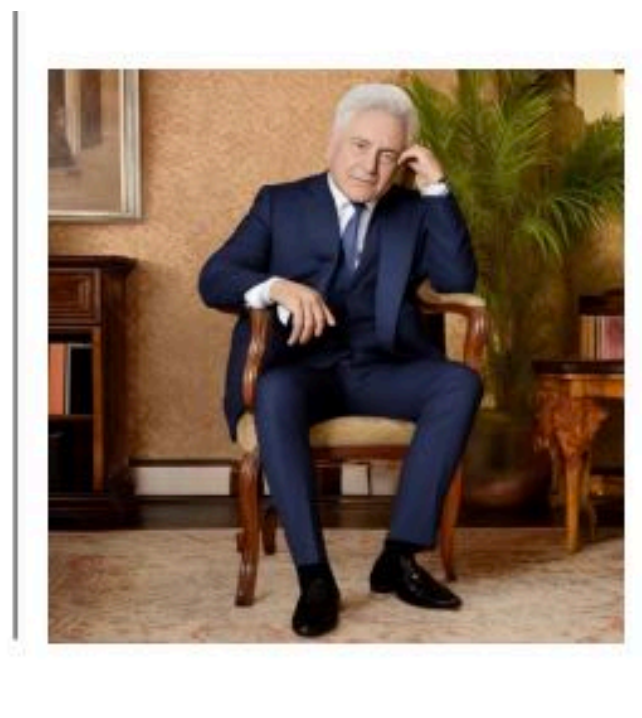


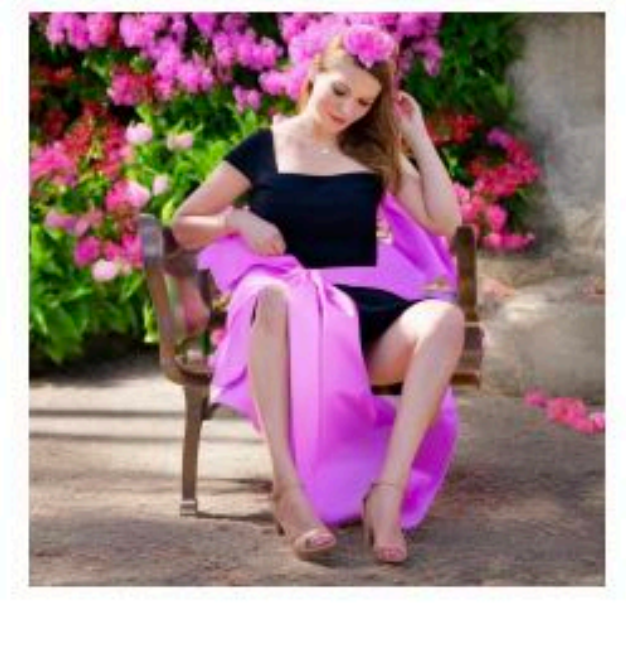
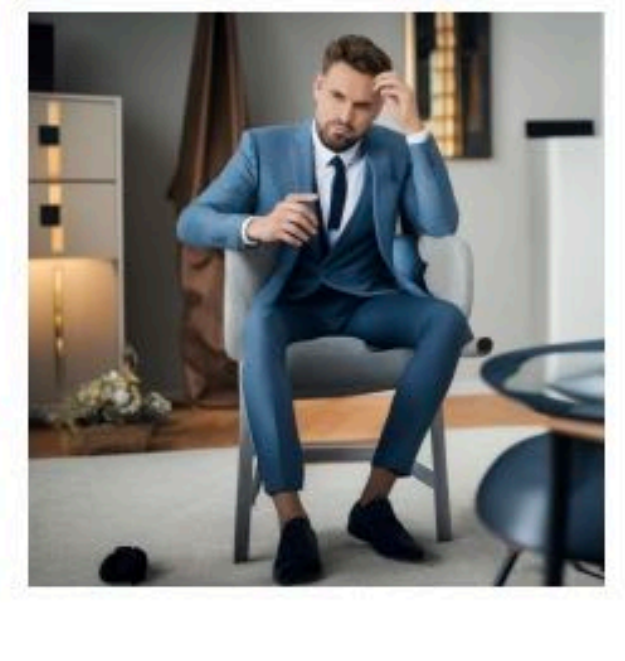

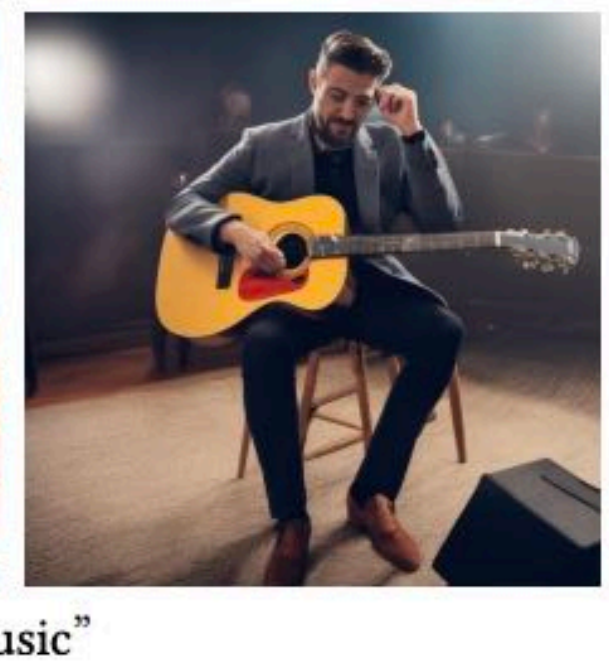
edamame,

ginger, and

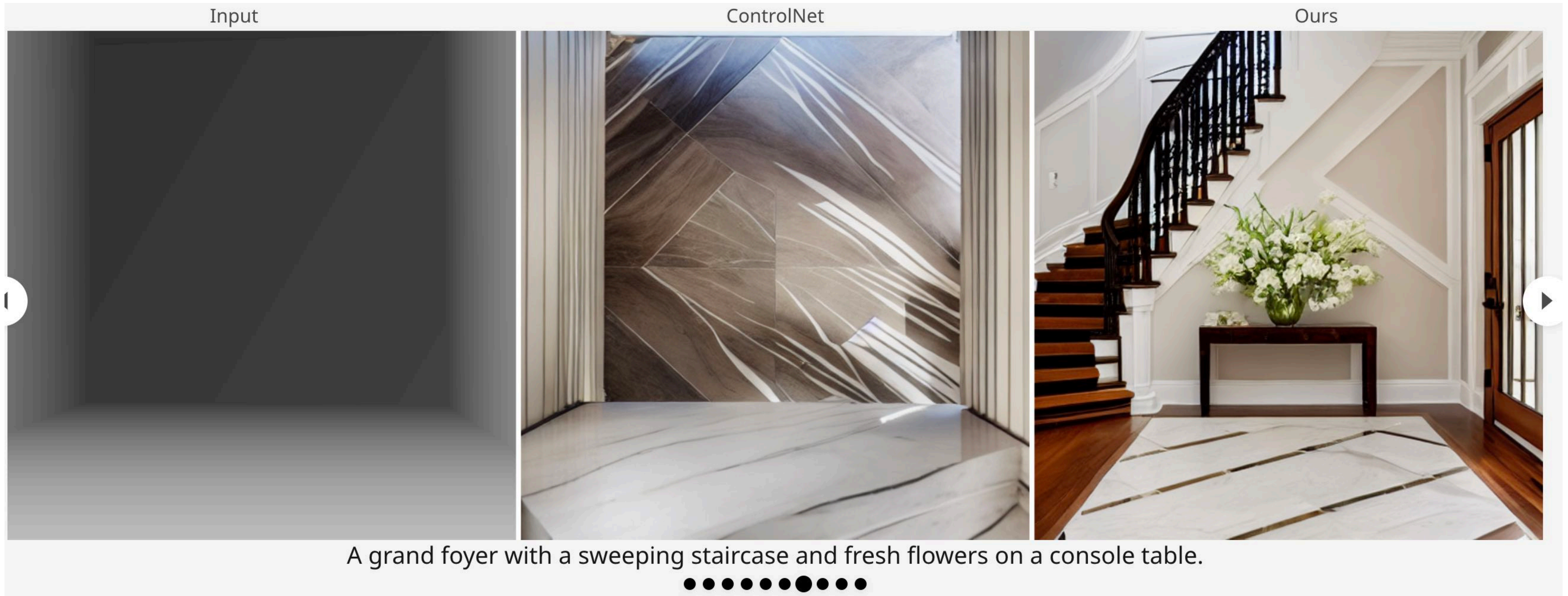
sushi.”



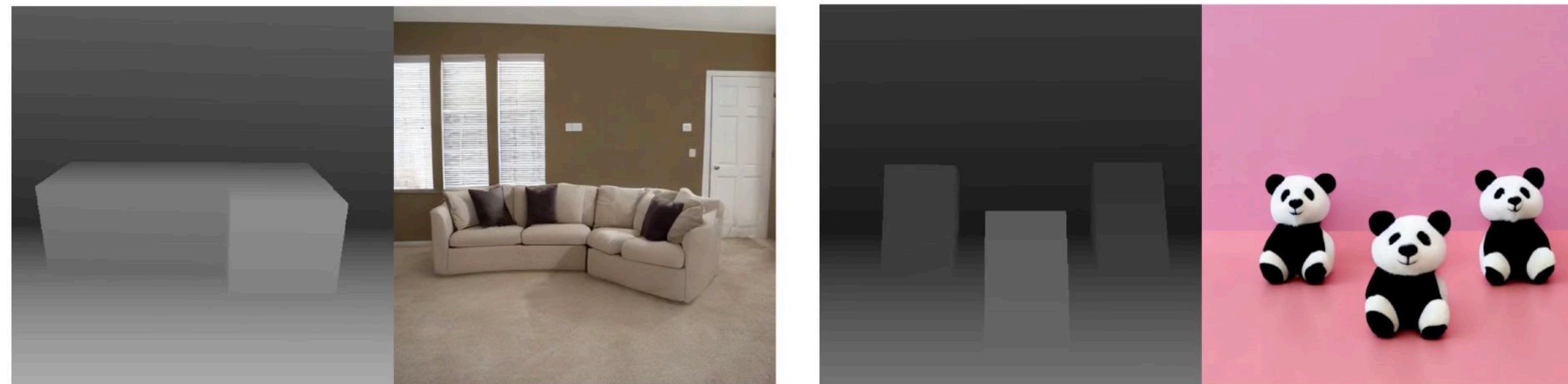
ControlNet: compositional control via images

Input (Canny Edge)	Default	Automatic Prompt		User Prompt	
					
		"a man with beard sitting with two children"		"mother and two boys in a room, masterpiece, artwork"	
					
		"a building in a city street"		"inside a gorgeous 19th century church"	
				astronaut	
					
				"music"	

Loose control



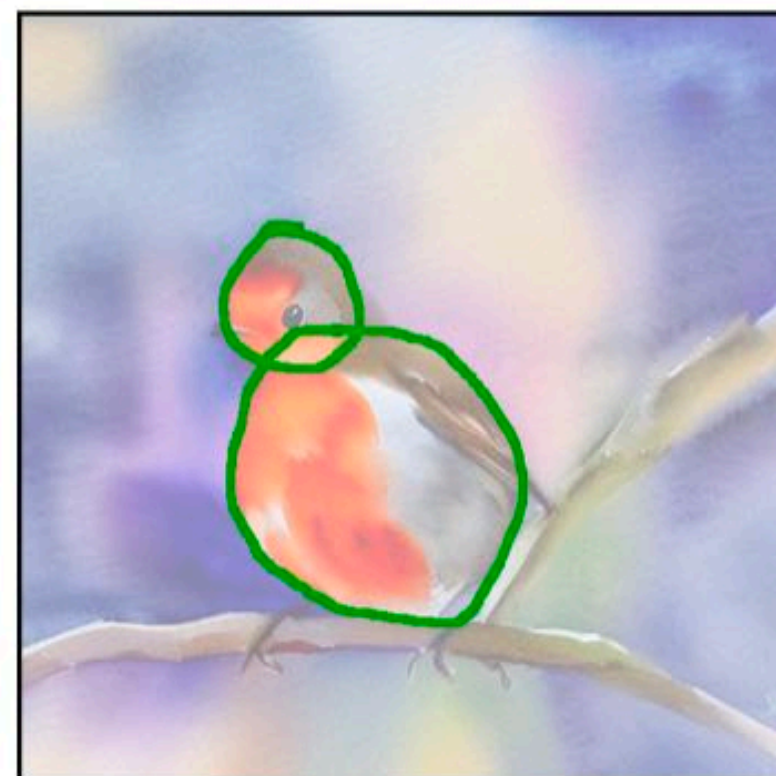
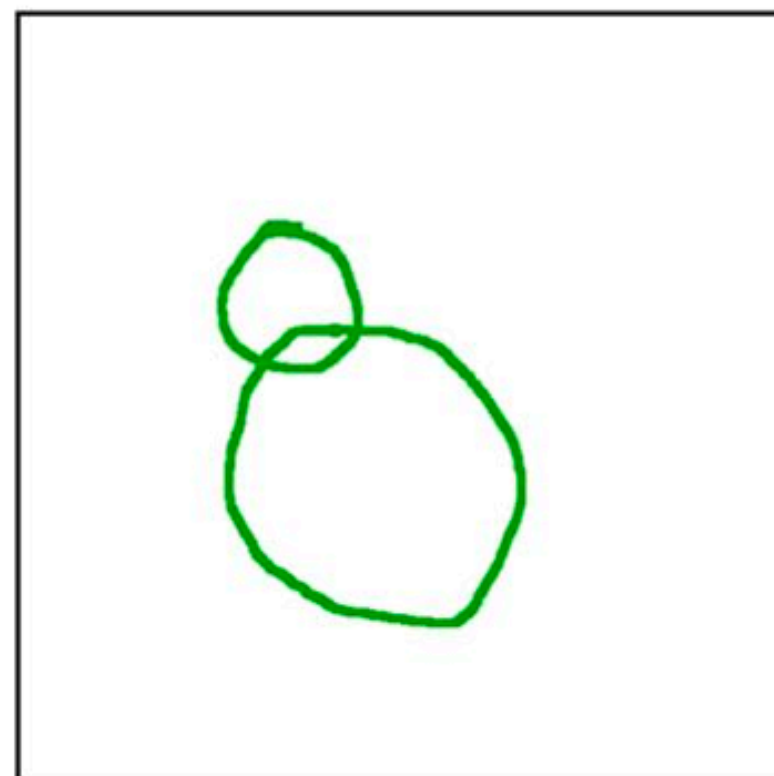
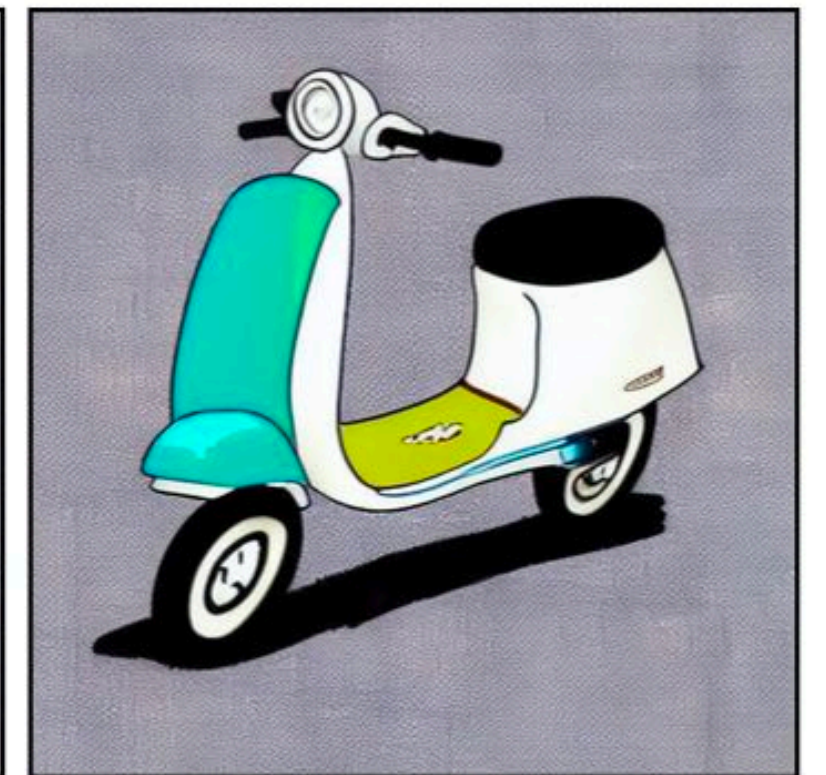
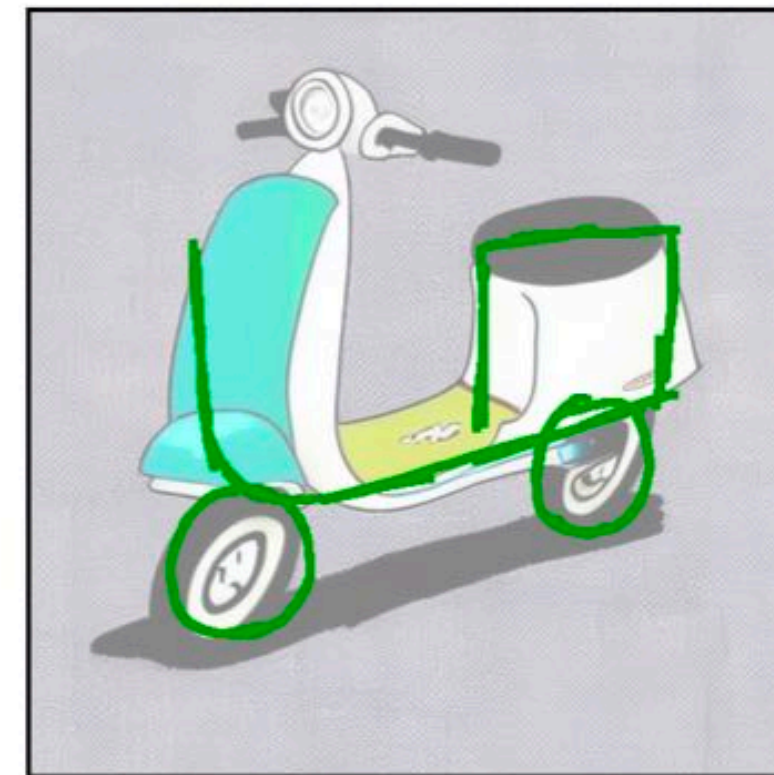
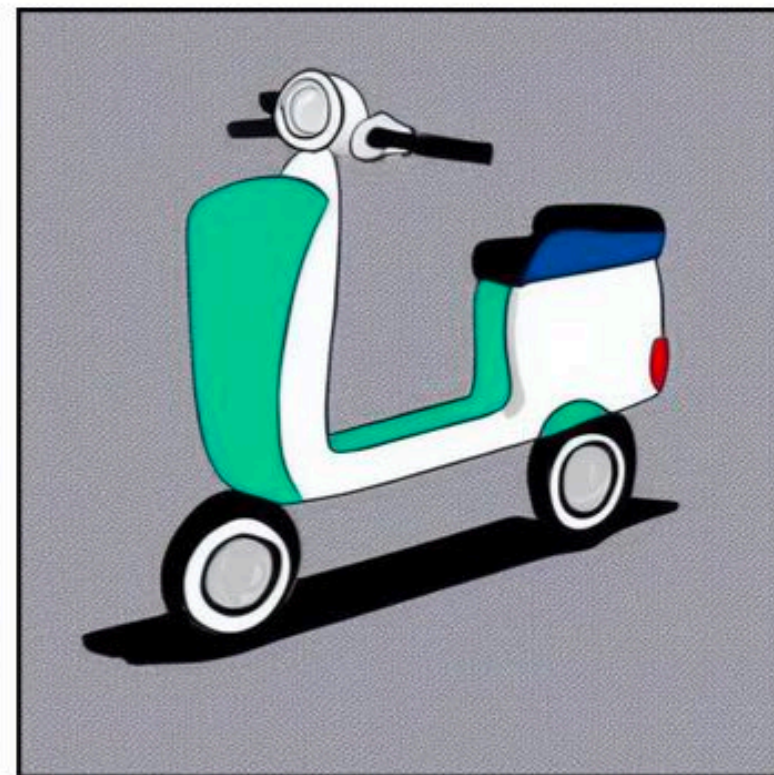
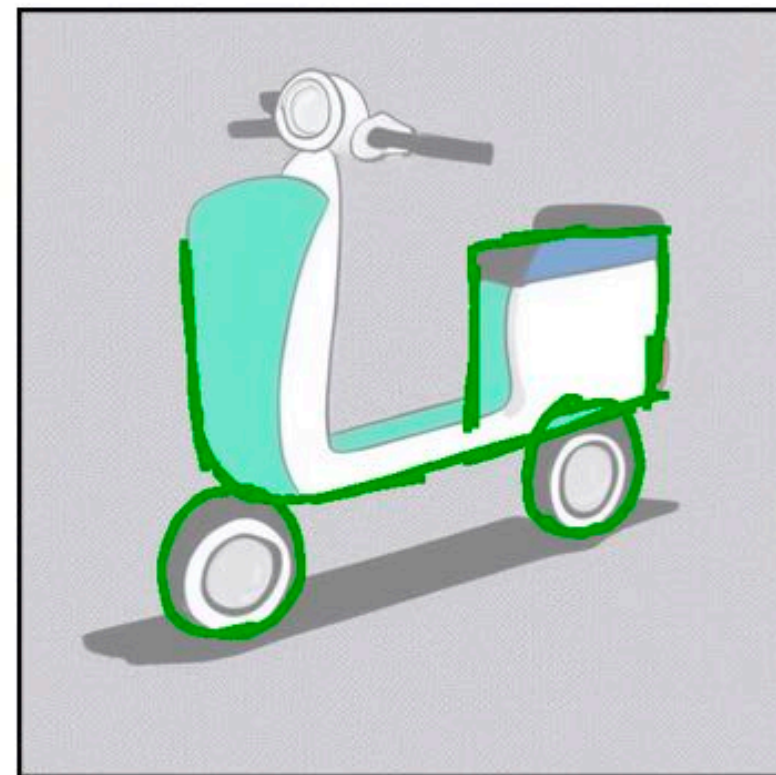
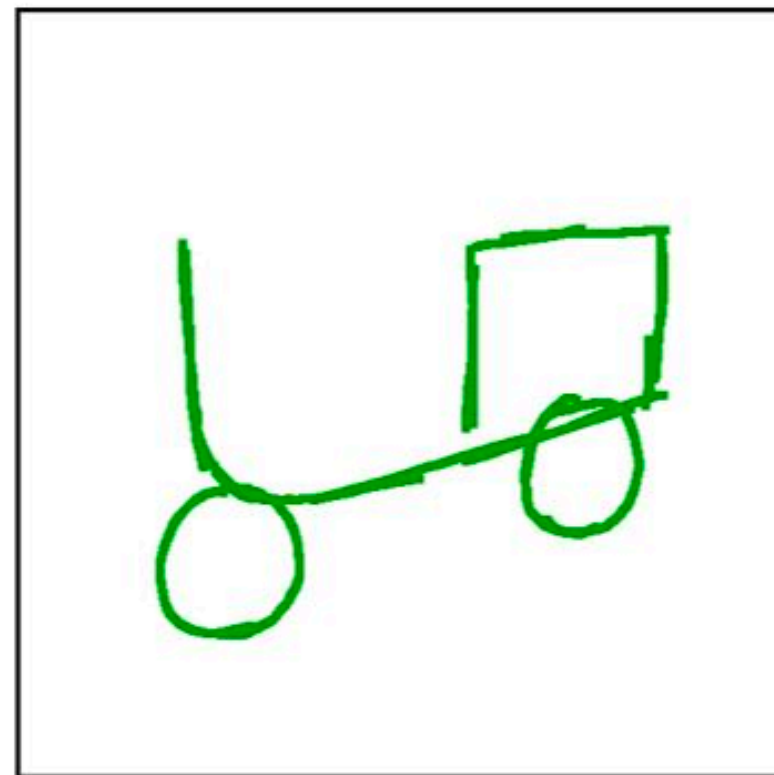
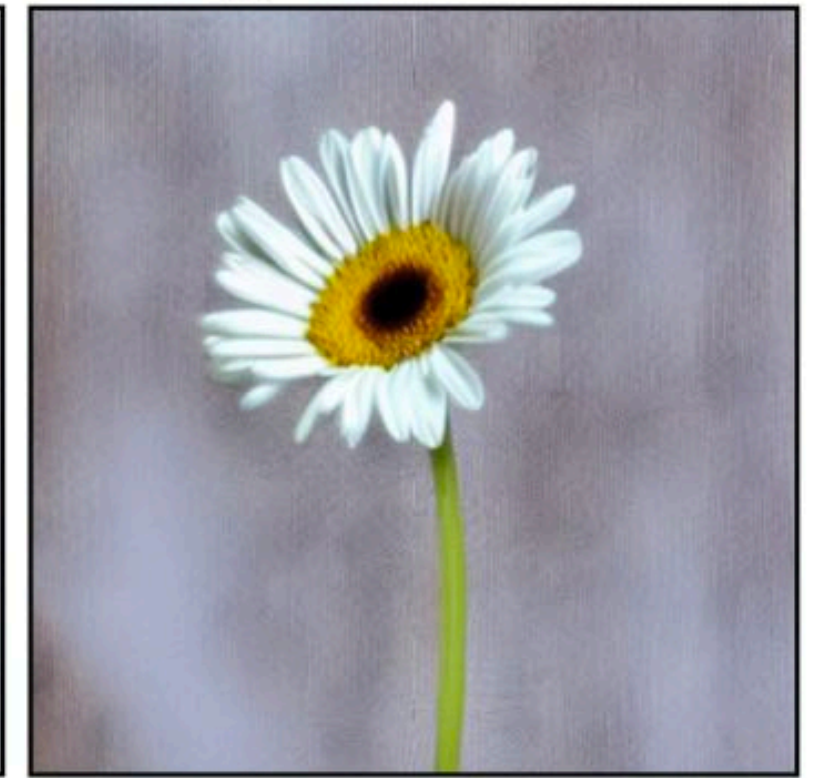
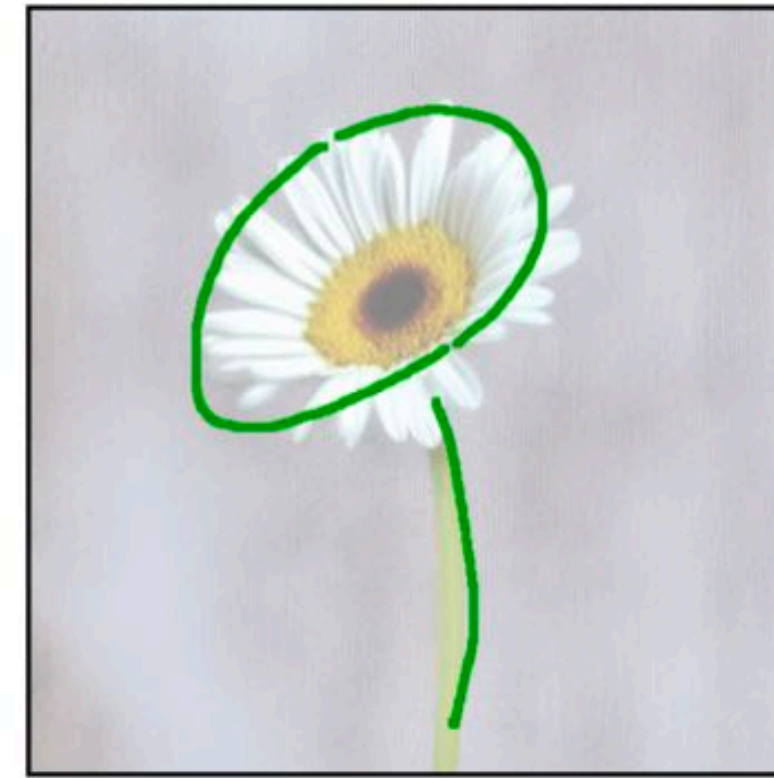
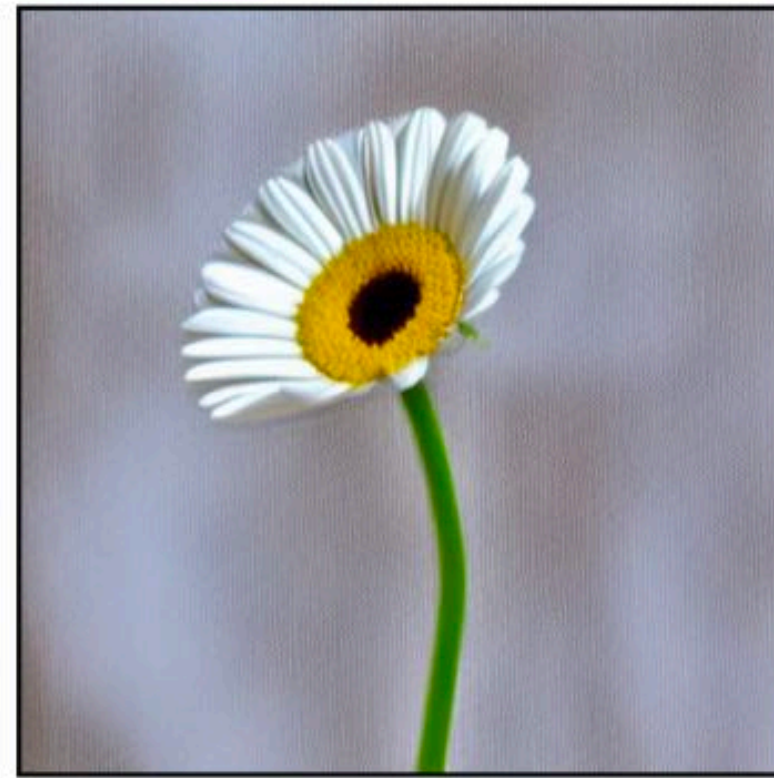
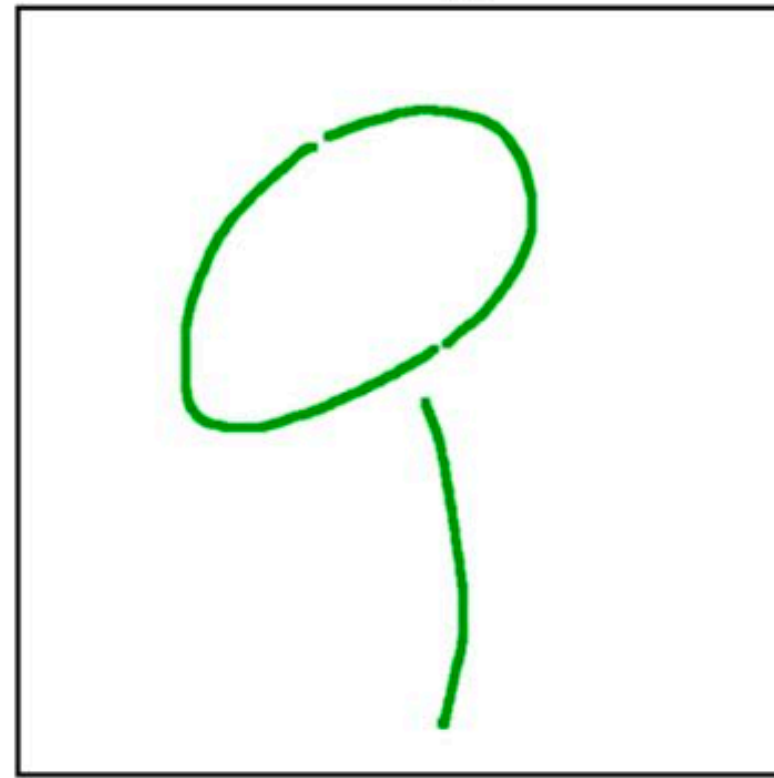
Key idea: user does not want to (or may not have capability to) specific visual controls precisely. Just have the user “block out” the basic shape of the scene.



Sketch-to-image

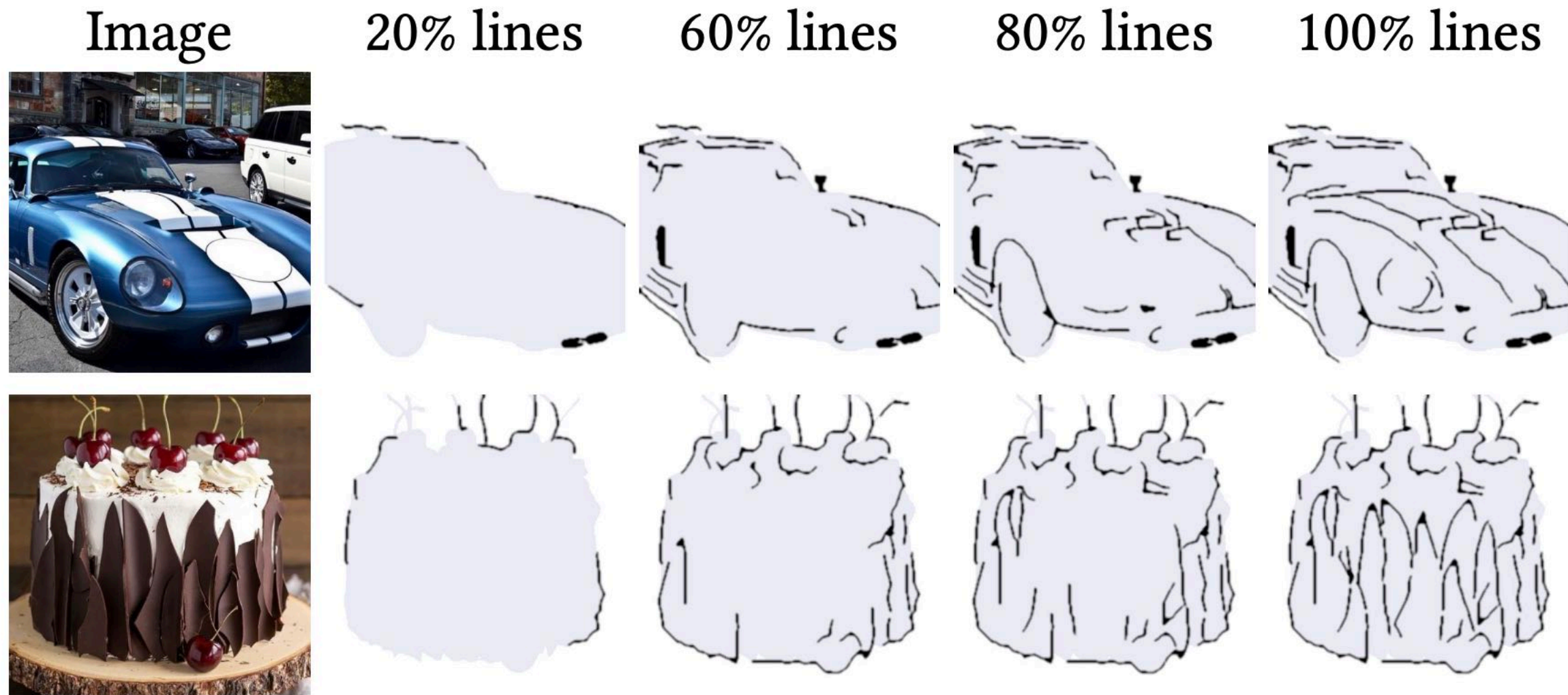
Tight control (ControlNet)

Looser control (Blended Renoising)



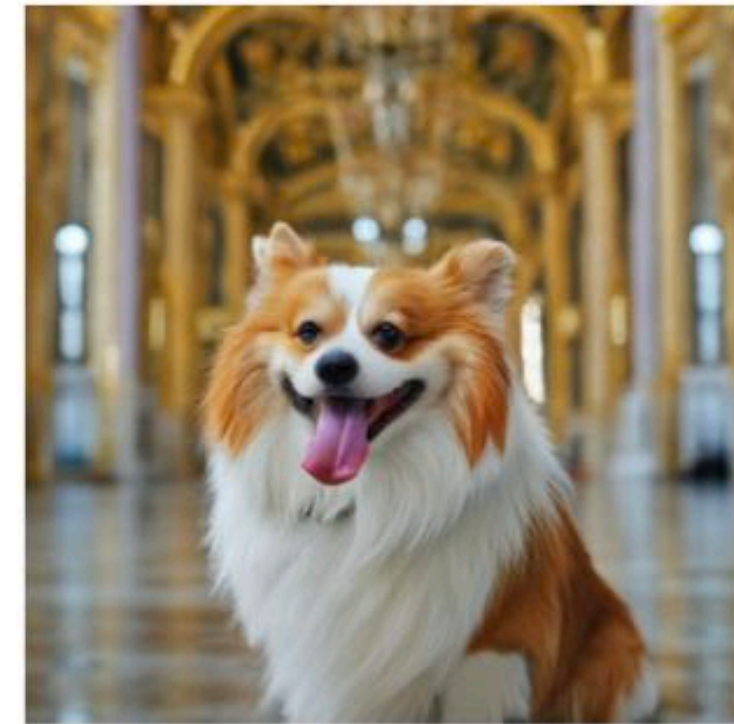
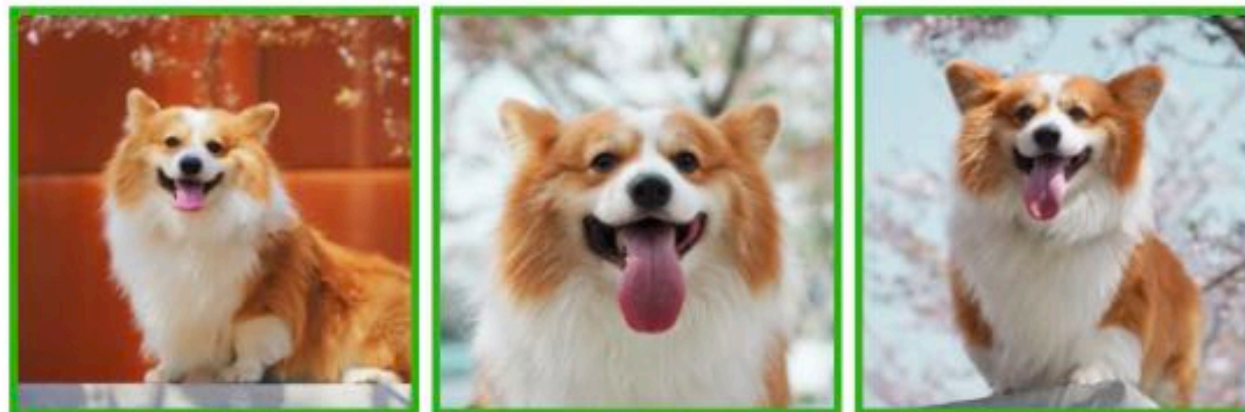
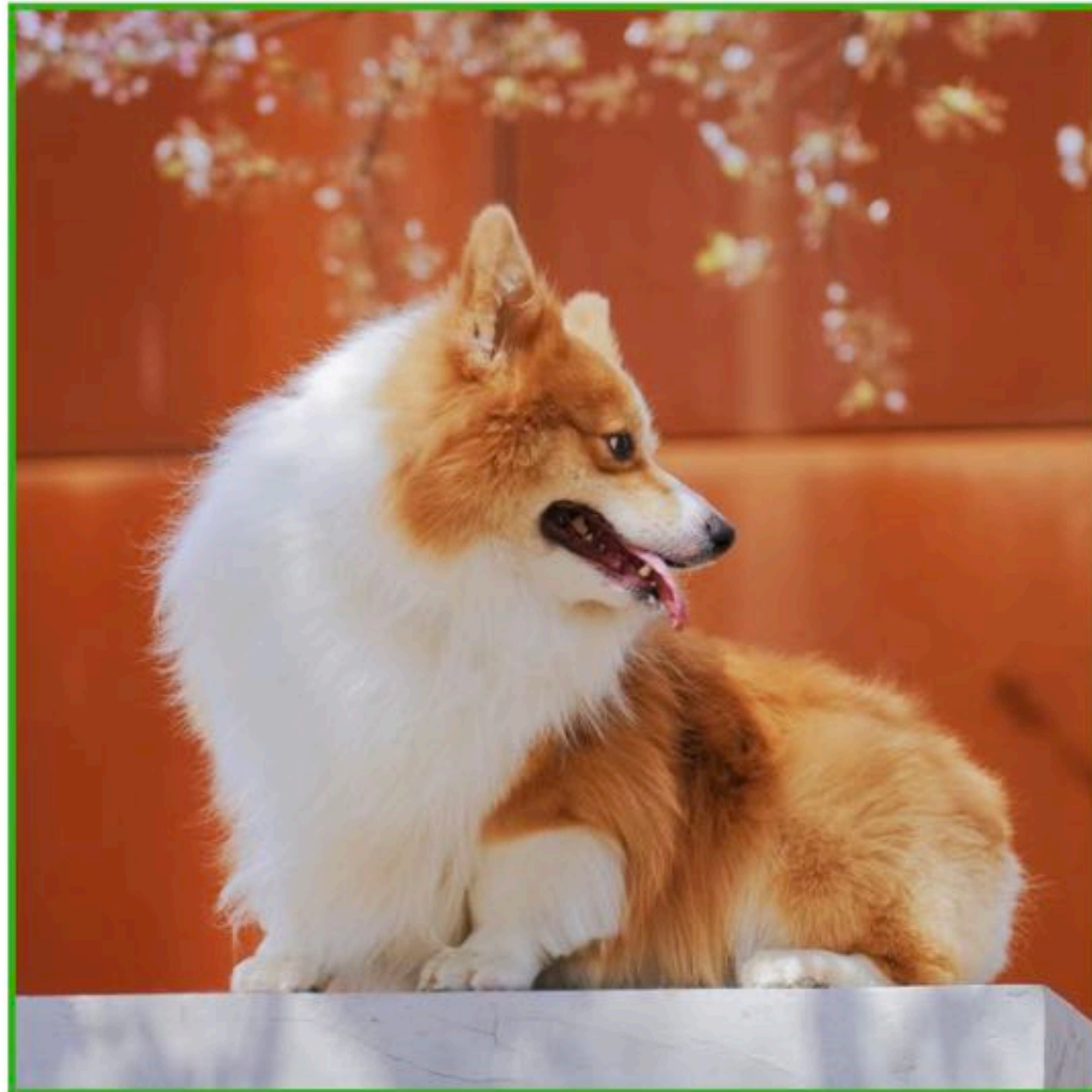
Partial-sketch to image

Creating "partial-sketch" training data

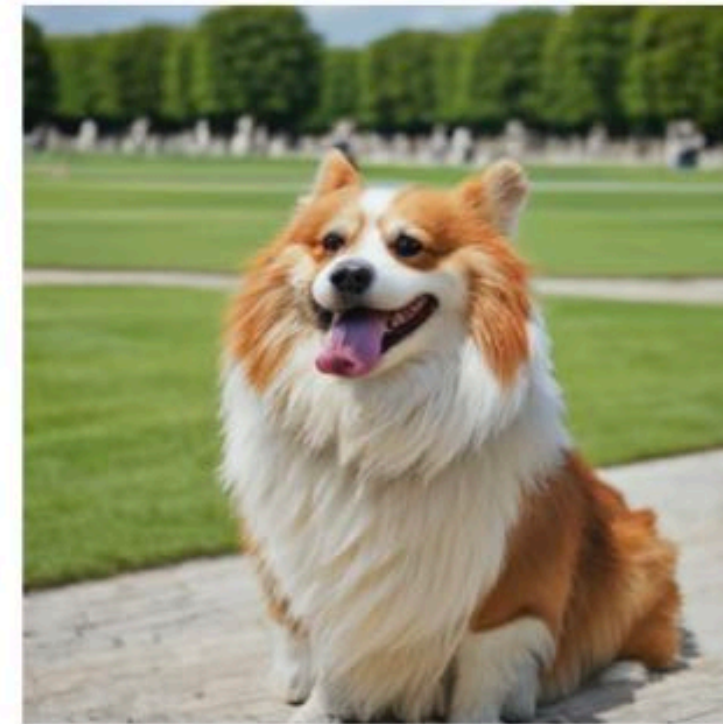


Specialization to a concept

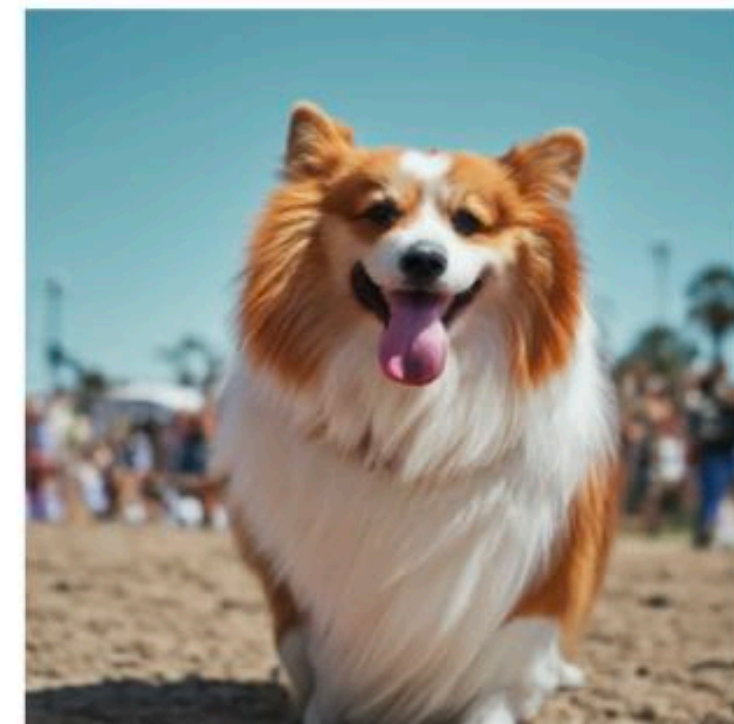
Input images



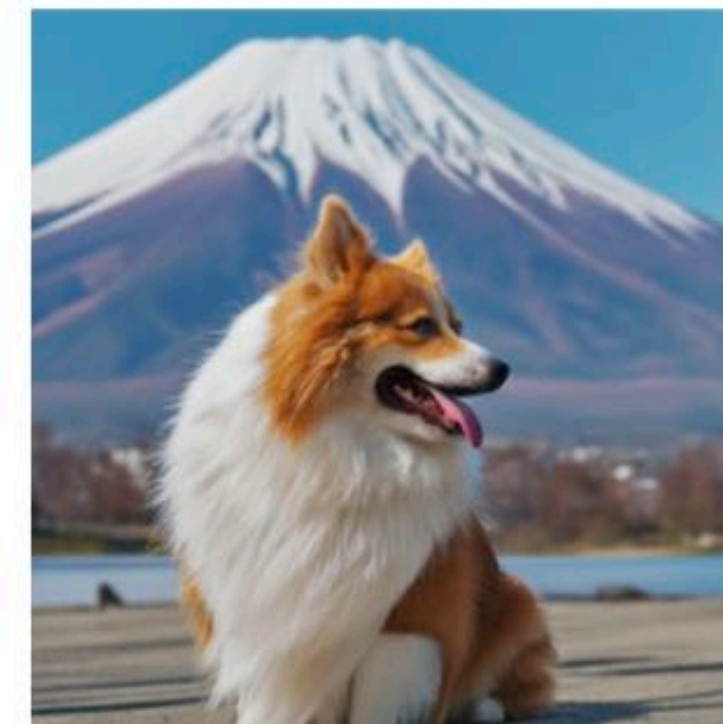
A [V] dog in the Versailles hall of mirrors



A [V] dog in the gardens of Versailles



A [V] dog in Coachella



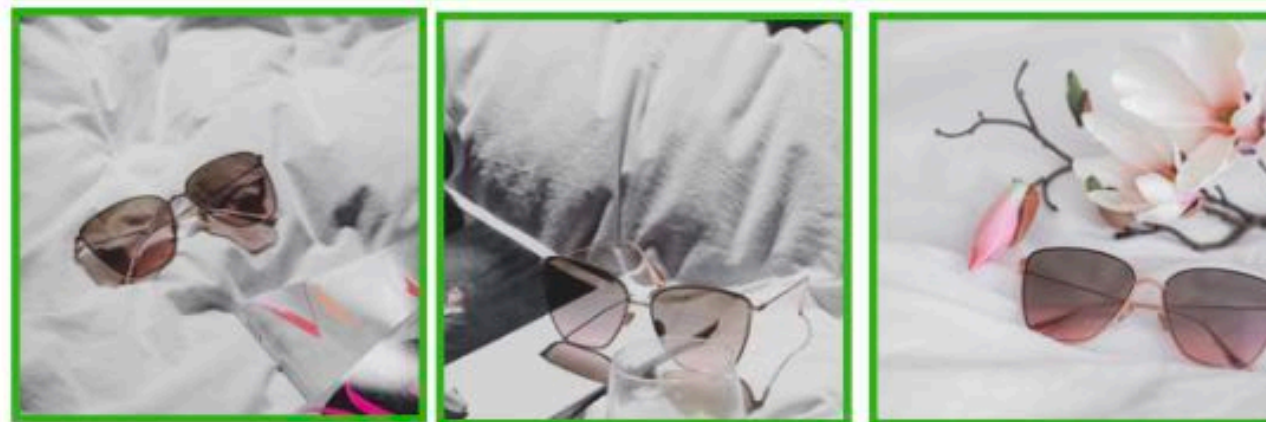
A [V] dog in mountain Fuji



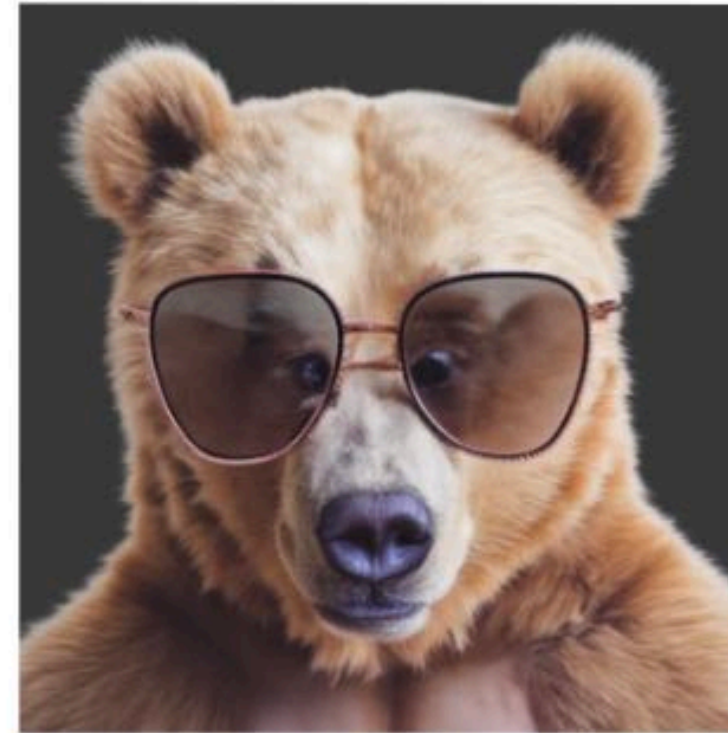
A [V] dog with Eiffel Tower in the background

Specialization to a concept

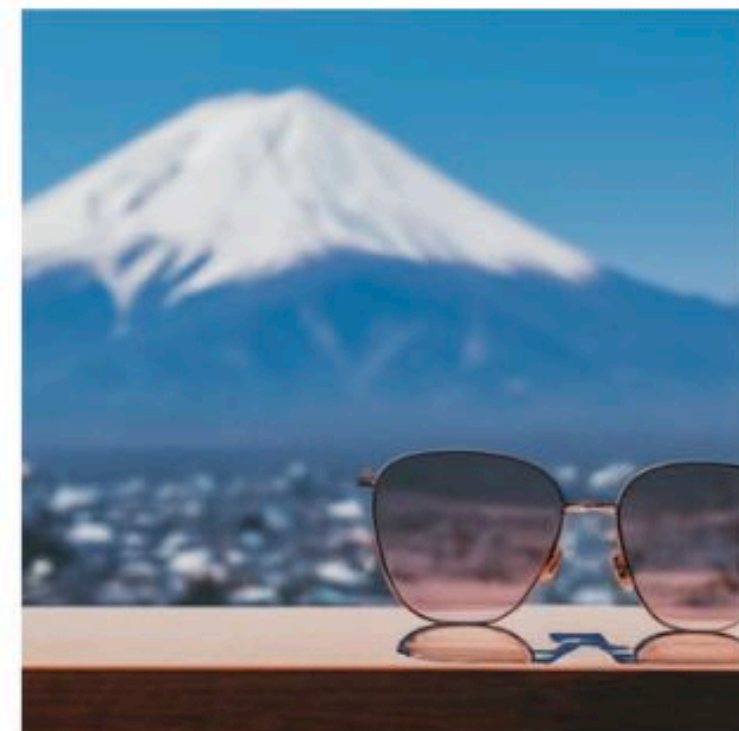
Input images



A [V] sunglasses in the jungle



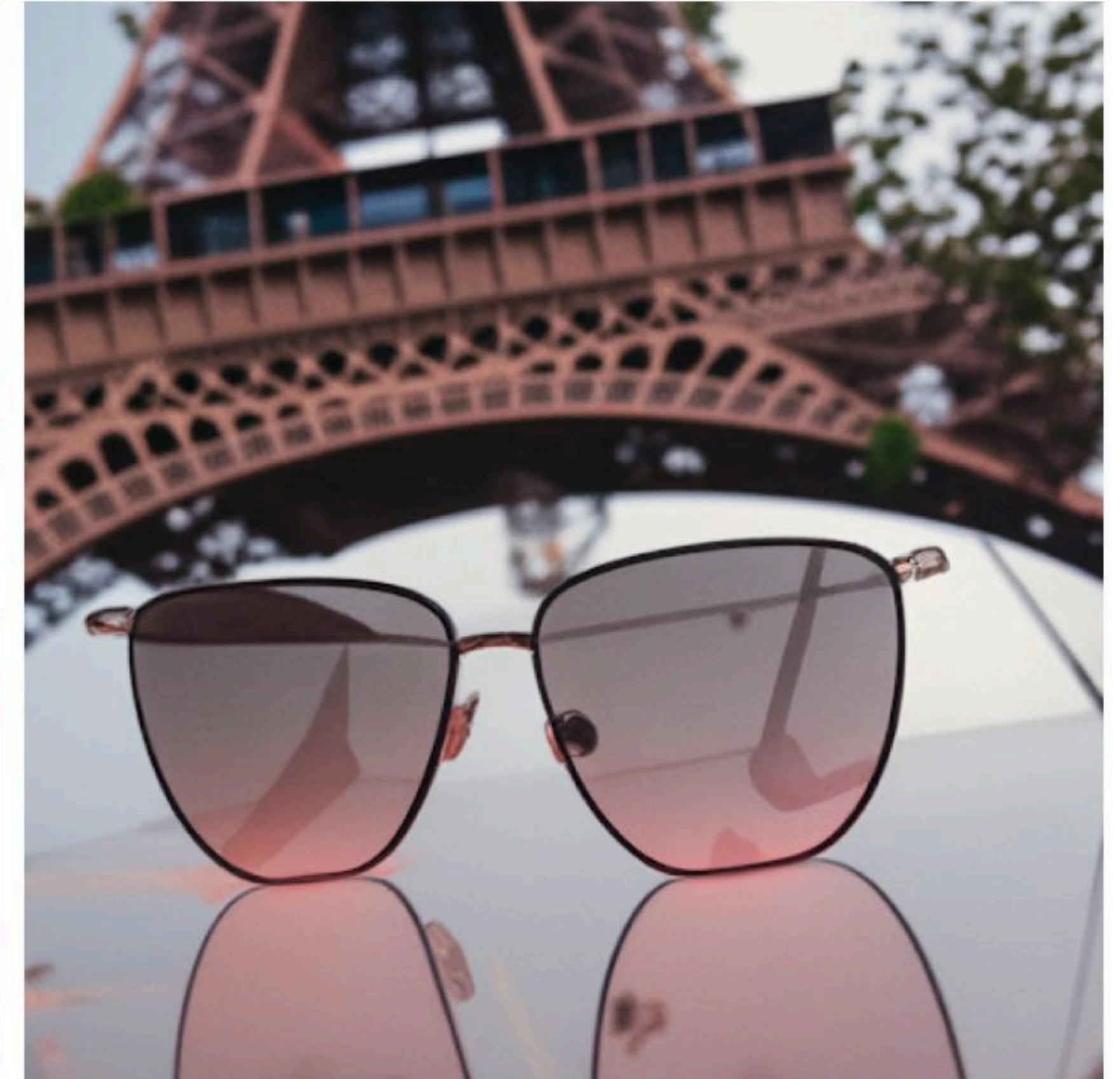
A [V] sunglasses worn by a bear



A [V] sunglasses at Mt. Fuji



A [V] sunglasses on top of snow



A [V] sunglasses with Eiffel Tower in the background

Reminder: key aspect in the design of any system

Choosing the “right” representations for the job

- **Good representations are productive to use:**
 - Embody a “preferred” way of thinking about a problem
- **Good representations enable the system to provide **useful services**:**
 - Validating/providing certain guarantees (correctness, resource bounds, conversion of quantities, type checking)
 - Performance optimizations (parallelization, vectorization, use of specialized hardware)
 - Implementations of common, difficult-to-implement functionality (complex array indexing code, texture mapping in 3D graphics, auto-differentiation, etc.)

Key takeaways

- **What is the type of control that aligns with the users thought process / mental model**
 - **Text is often an ambiguous, imprecise, or flat out inefficient way to describe visual intent**
- **Examples:**
 - **Users want to control spatial composition**
 - **“Dog on the left” vs. dragging a layer to the right location**
 - **Users want to “block out” an idea, and have the diffusion model “fill in the details”, “correct proportions”, “harmonize the image”**
 - **Users want to express intent via an example: “I want it to look LIKE THIS!”**

Key takeaways

- **What is the type of control that aligns with the users thought process / mental model**
 - **Text is often an ambiguous, imprecise, or flat out inefficient way to describe visual intent**
- **Much active research on teaching a model to follow a specific intent**
 - **One key strategy: dataset engineering to create pairs (control input, expected output)**
 - **Image analysis used to create the “control input” part of the pair: depth image, partial sketch, etc.**

Next time

- **Visual design tools are often most useful if they provide immediate / interactive feedback**
- **Efficiency and performance matter!**