Lecture 16:

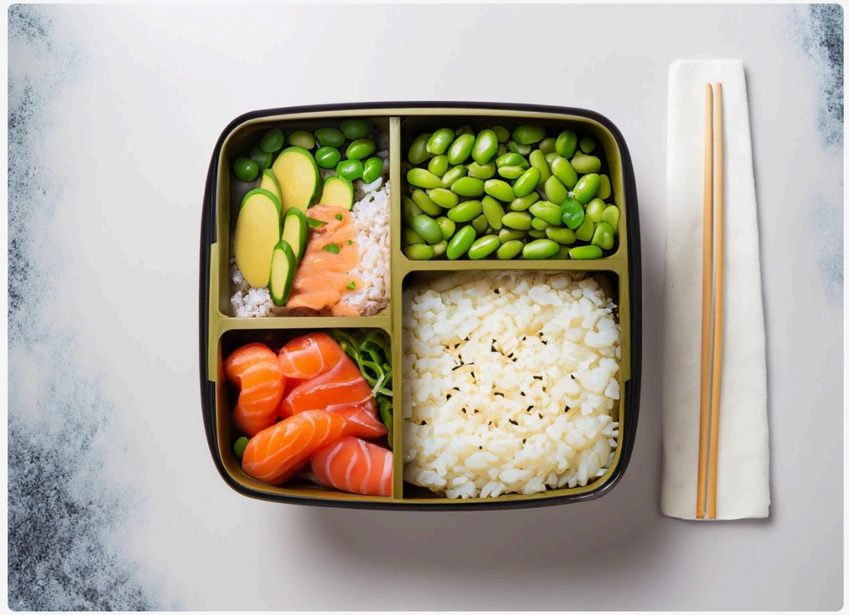
Generative AI for Image Creation (Initial discussion)

Visual Computing Systems
Stanford CS348K, Spring 2023

The Road to Pt. Reyes (Pixar 1983)

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

“A bento box with rice,
edamame, ginger,
and sushi.
Top down view,
white background.
Sushi in right bin of bento box.
Edamame in top left.”



Many issues with emerging class of generative AI technologies

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

Preliminaries

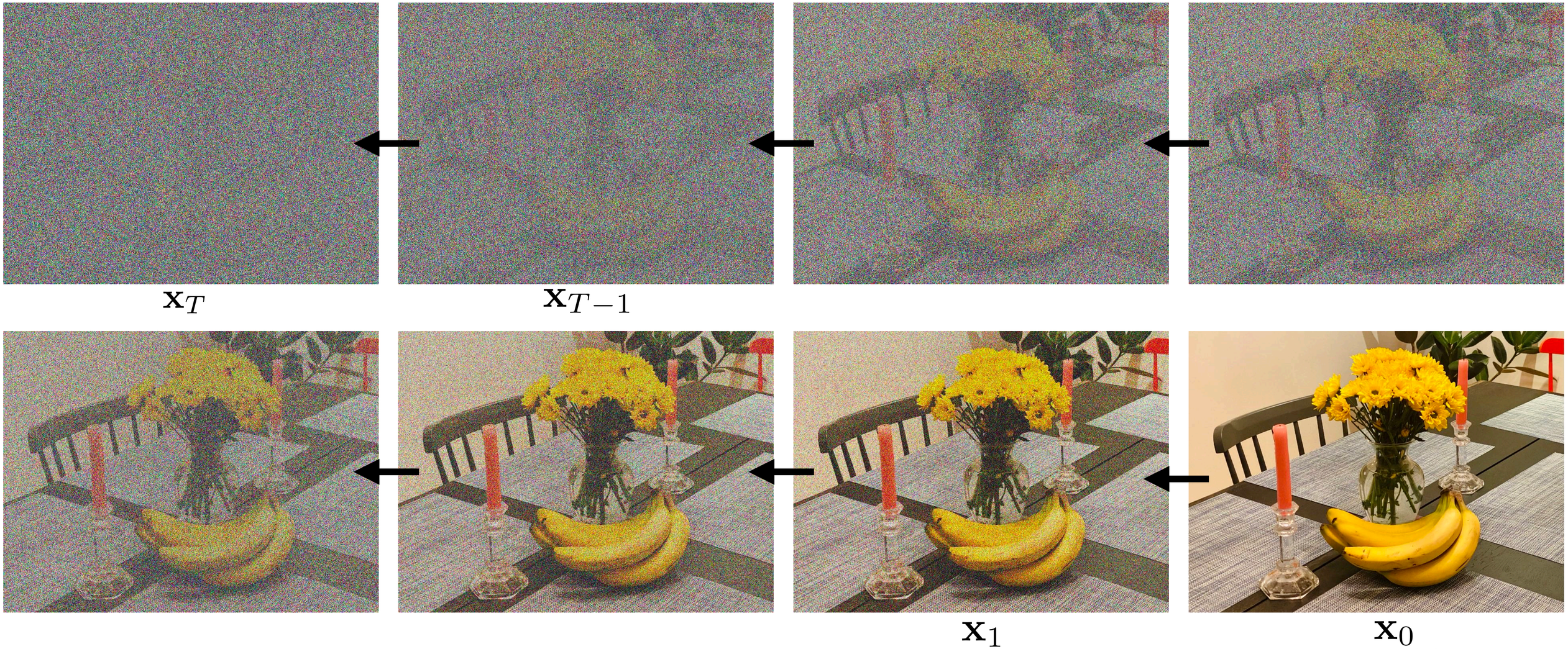
Suppose you are given a data of images x_i

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

Diffusion-based image synthesis

Idea: iterative MCMC process to generate a sample x (an image) from distribution p(x) of observed images

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

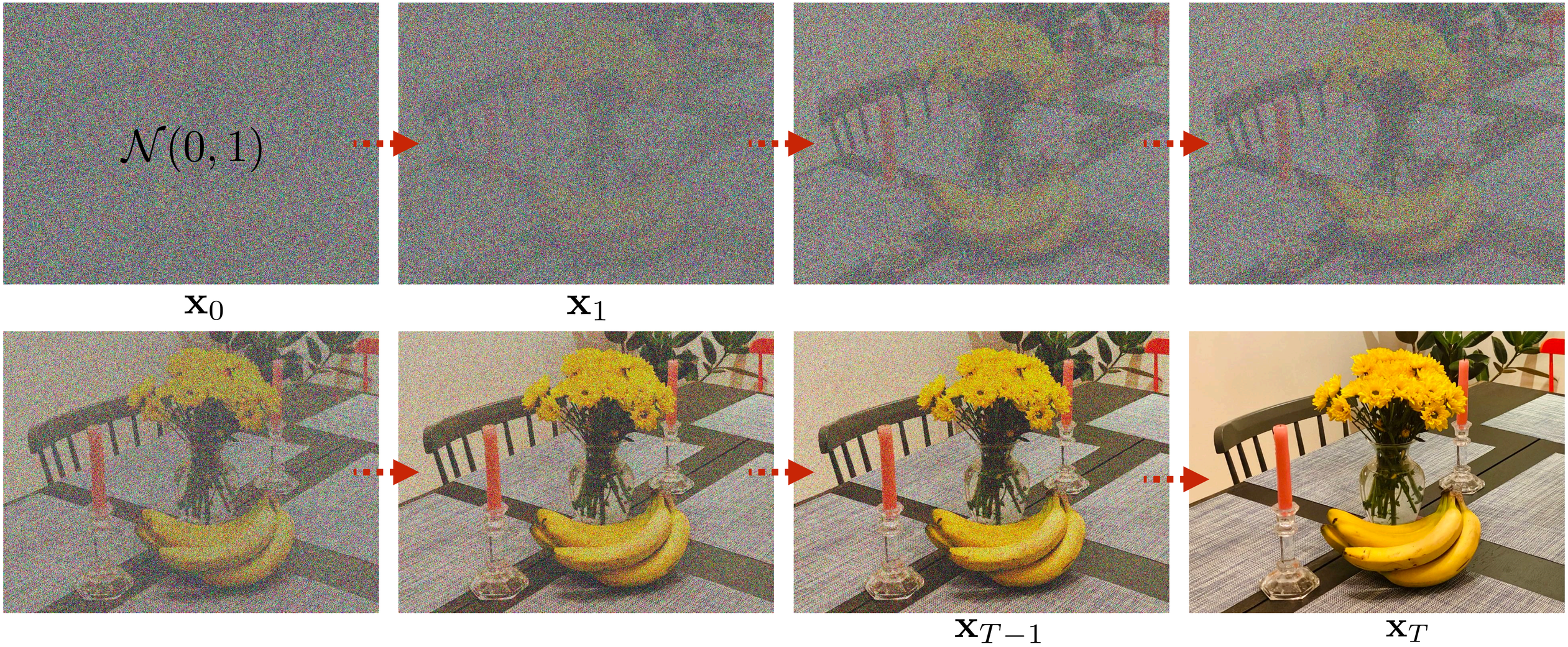


Diffusion-based image synthesis

Reverse: iteratively remove noise noise from random sample to obtain image from p(x)

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

("score function")



Classifier free guidance

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

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

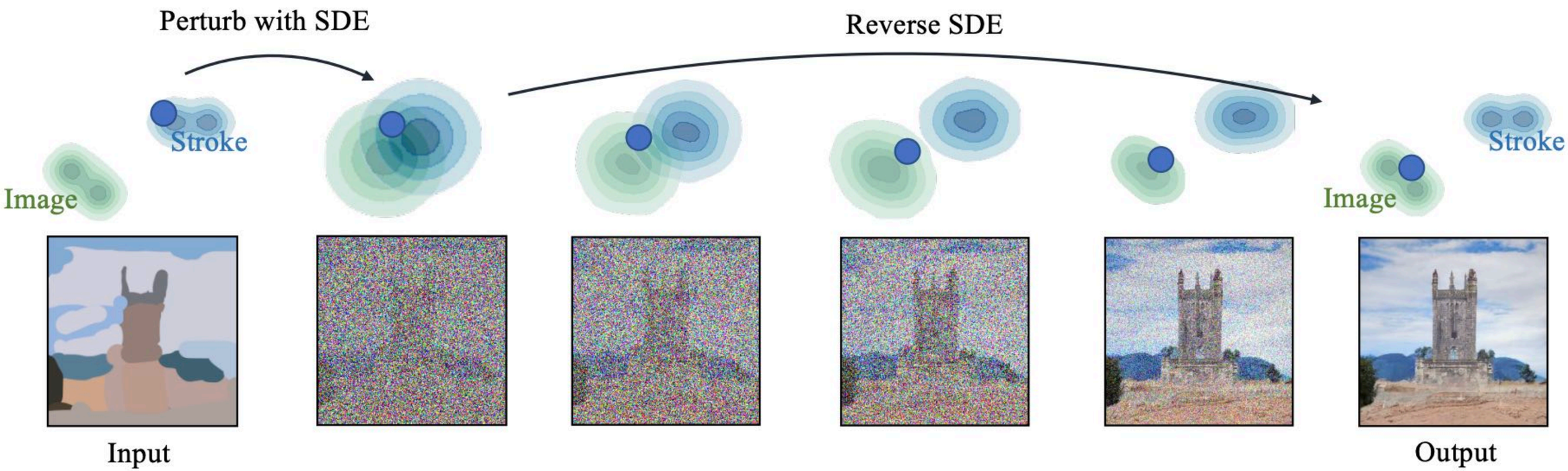
Bayes for score function

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

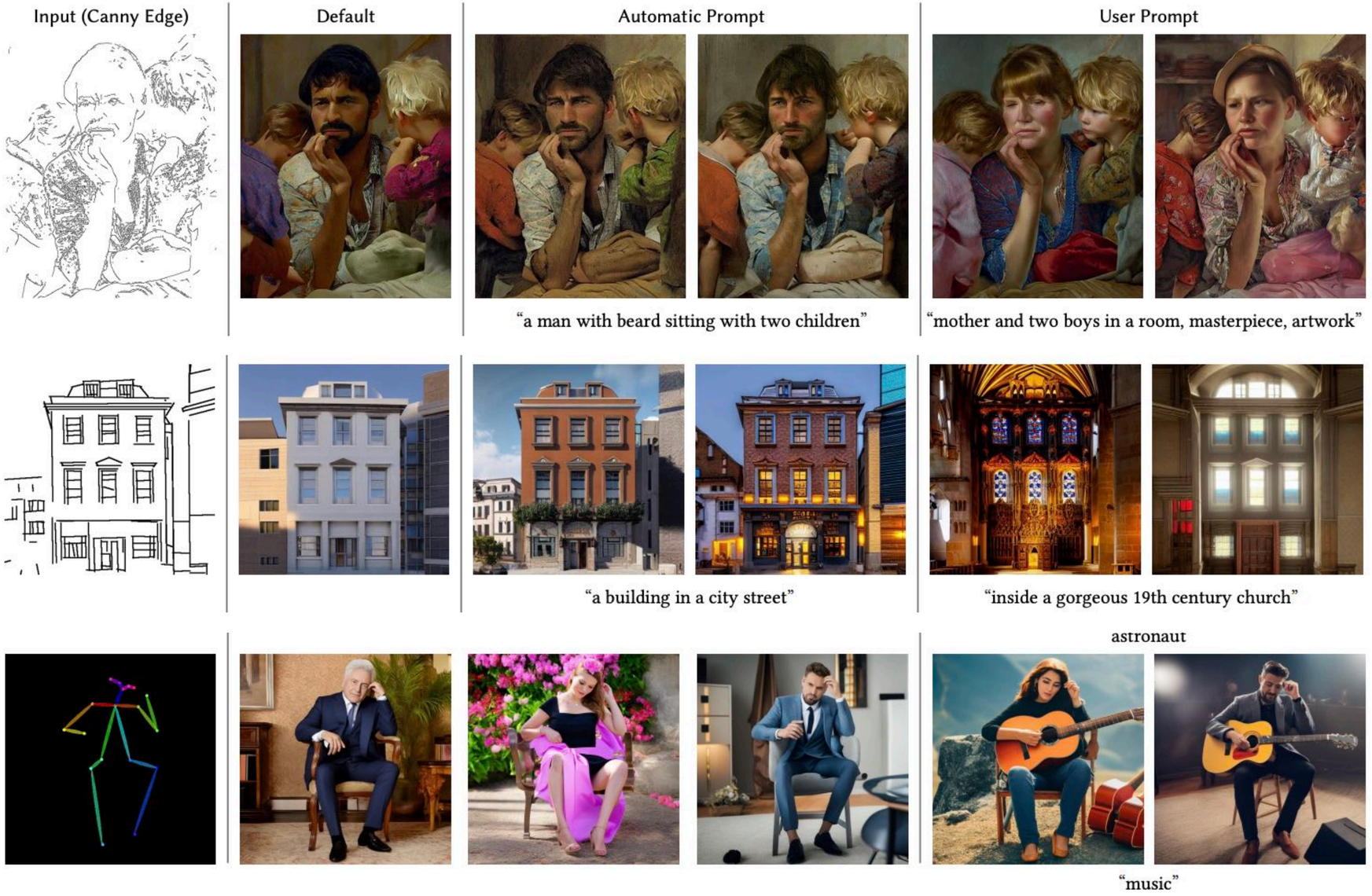
Controlling diffusion

img2img (enabling image-based guidance)

1. Start with a guide image (a target)
2. Add "small amount of noise"
3. Denoise to produce sample from p(x)



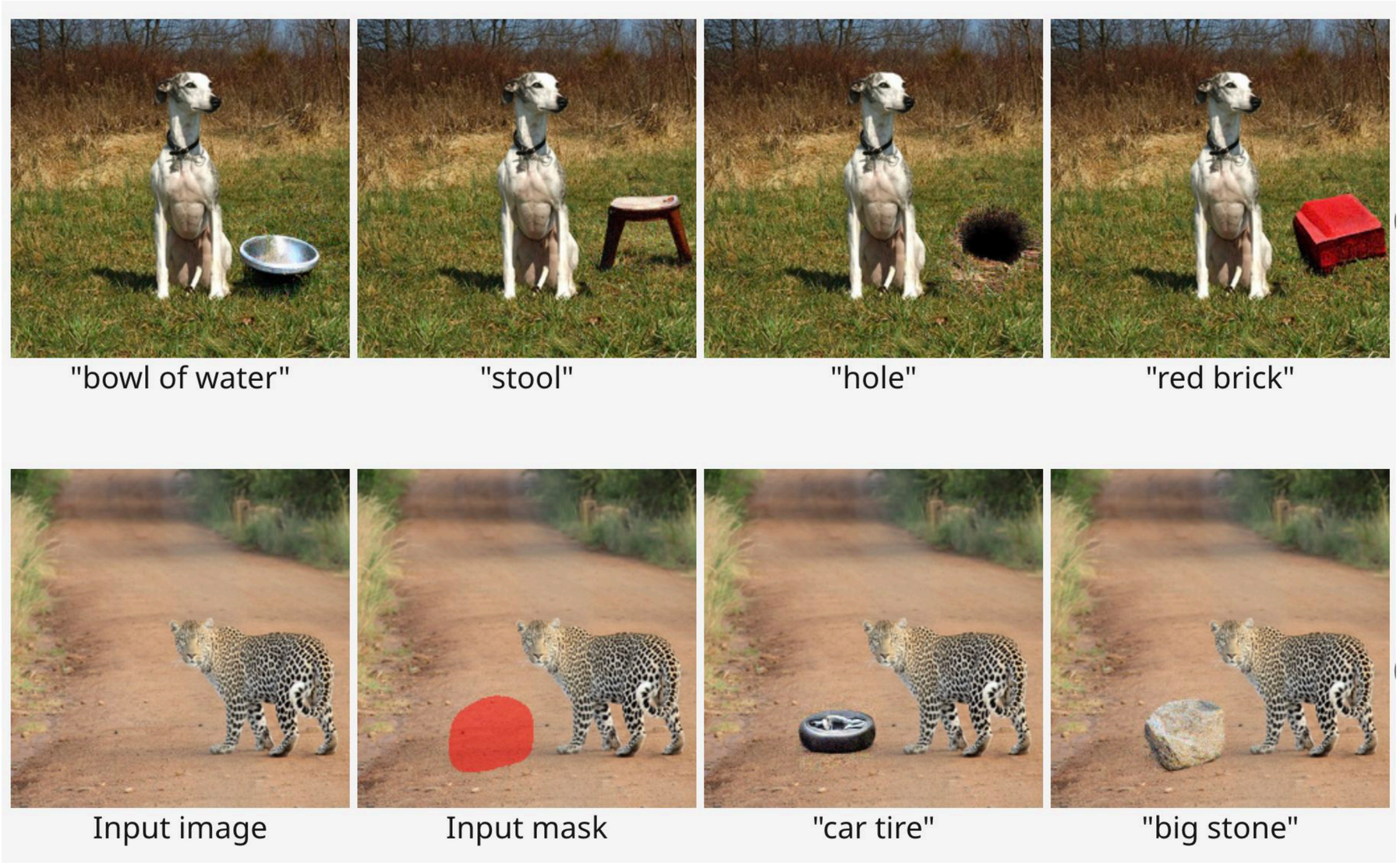
Other forms of guidance (not just text)



Inpainting (apply [new] prompt to a region)

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

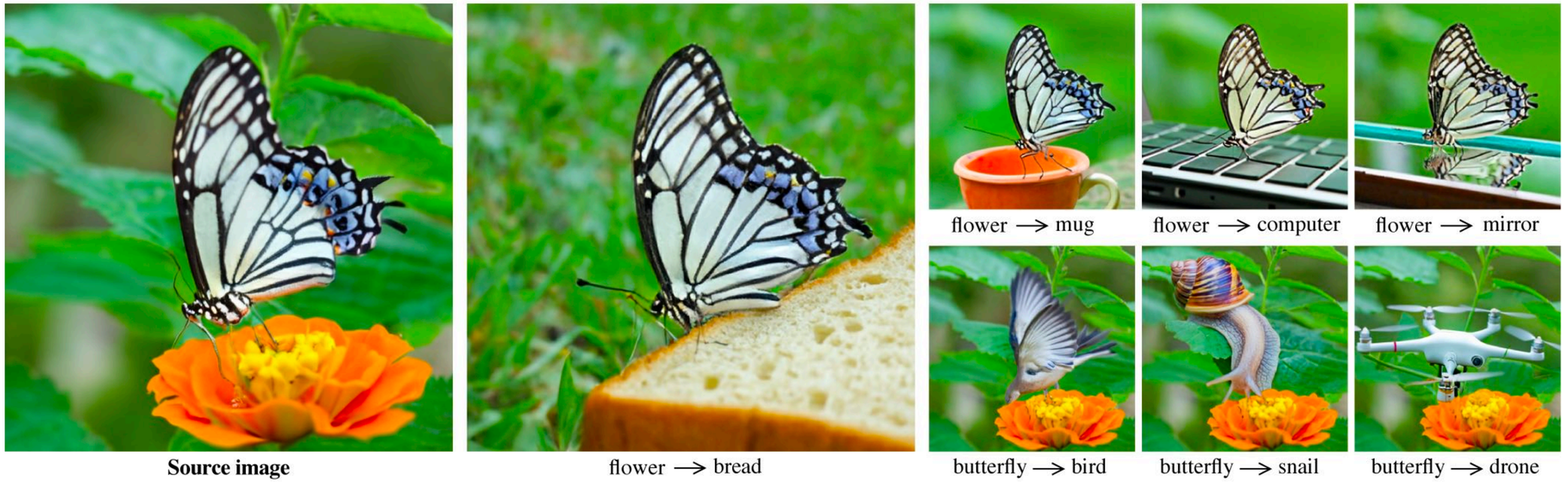
Image outside of region remains almost the same.



Use change in text prompt to trigger change in image

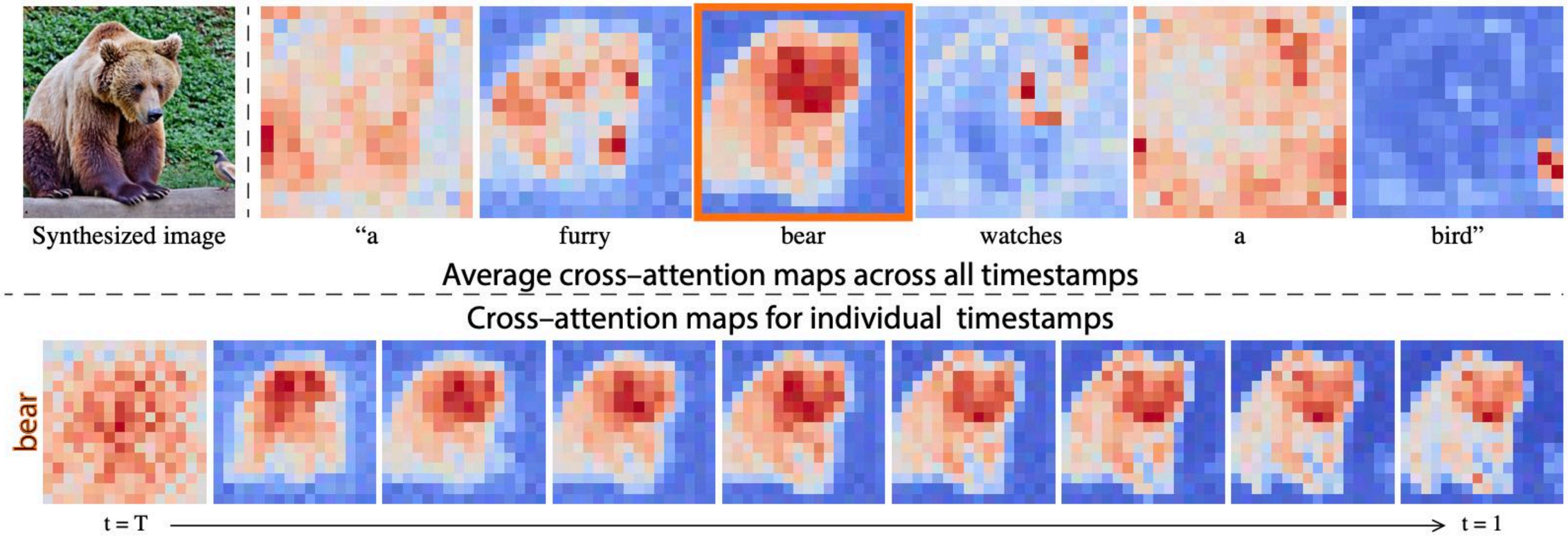
"A basket full of apples."

"A photo of a butterfly on a flower."



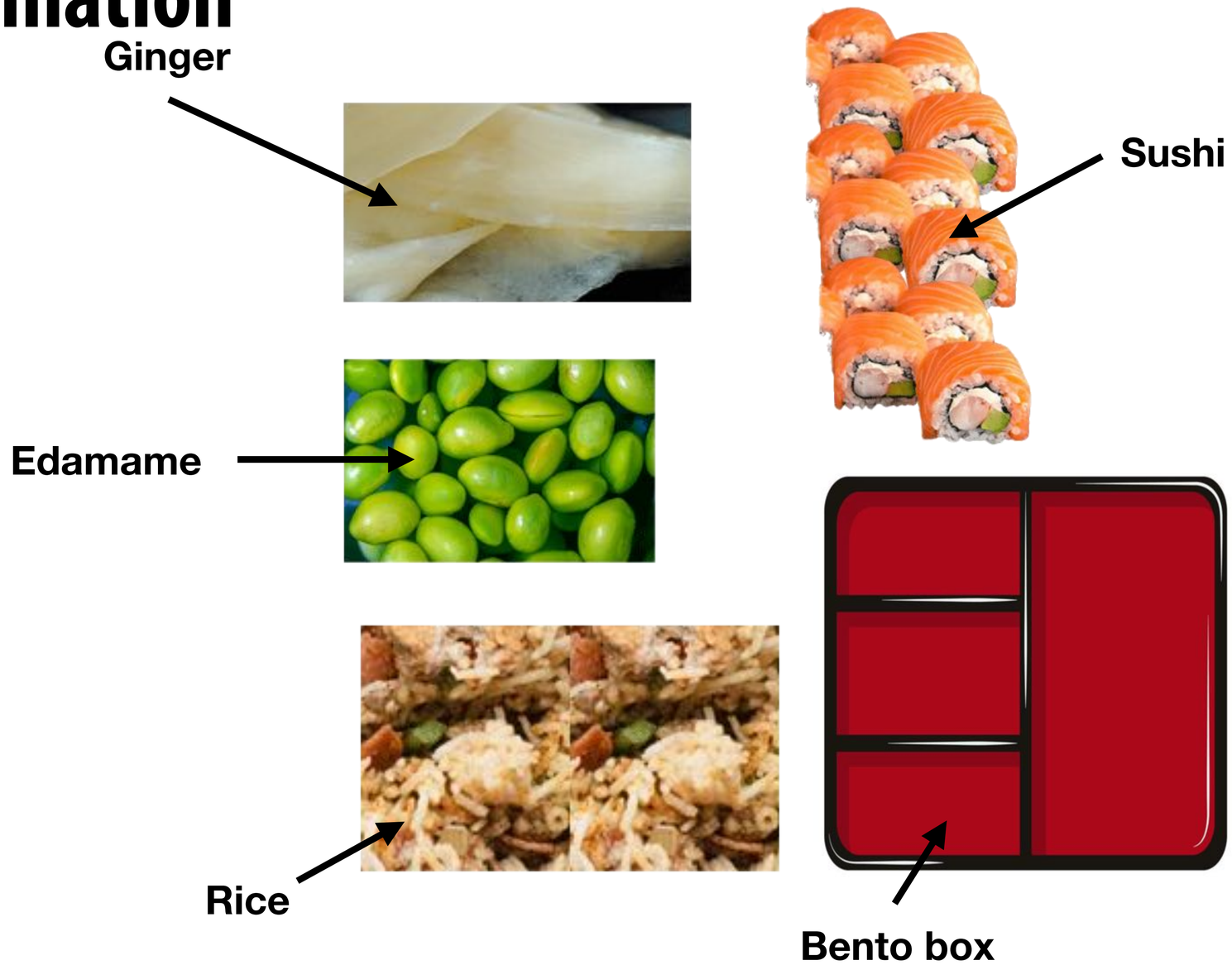
“Masks” come from learned attention

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



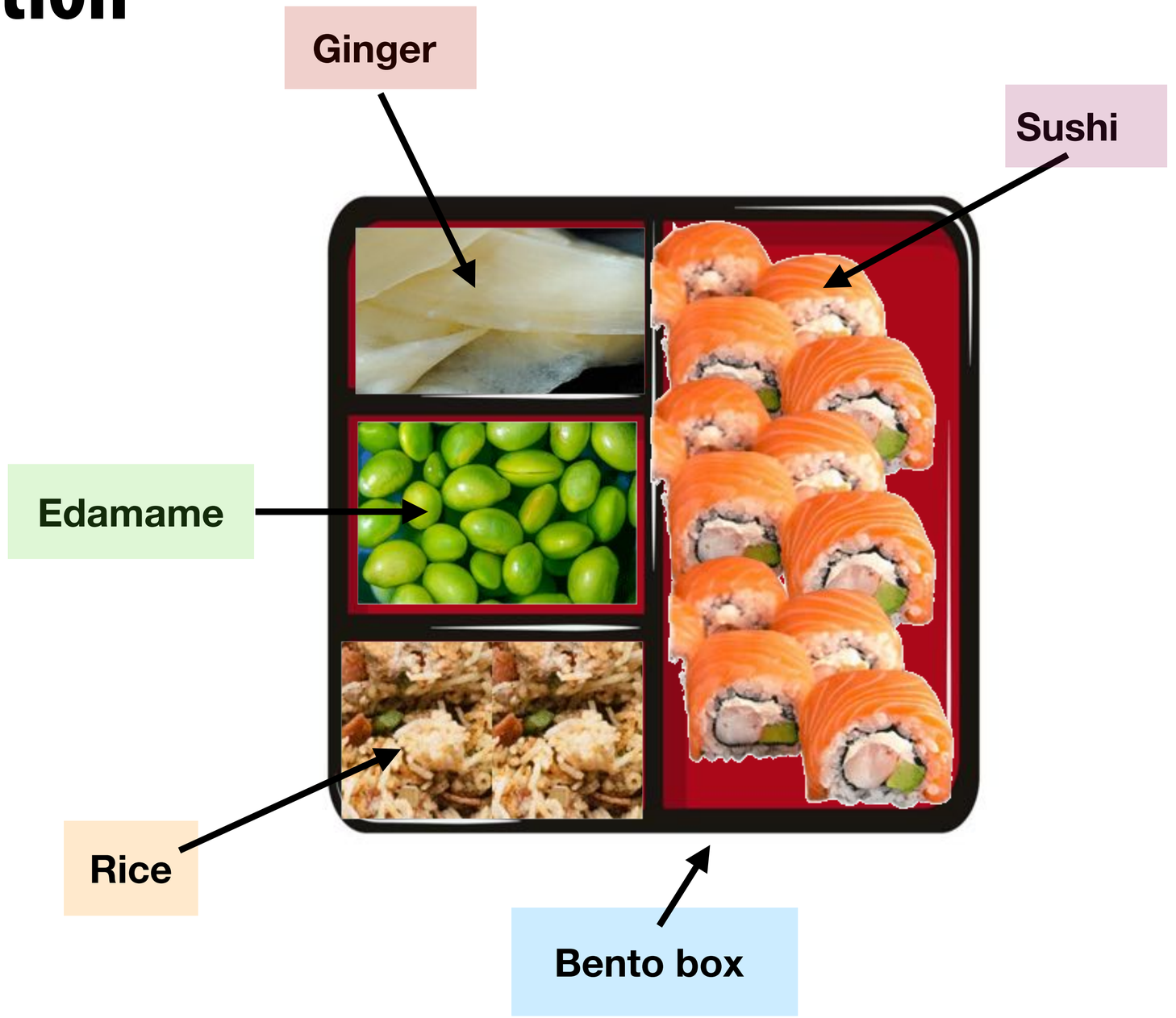
Leveraging layer information

Prompt + a rgba per layer



Leveraging layer information

"A bento box with rice, edamame, ginger, and sushi."



Leveraging layer information

"A bento box with rice, edamame, ginger, and sushi."

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"