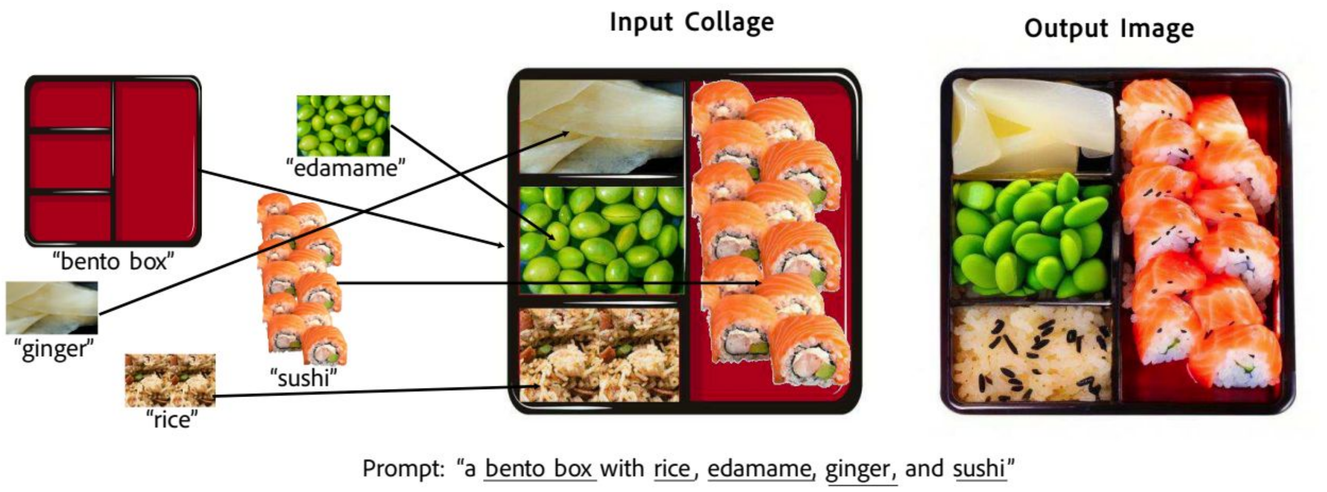
Collage Diffusion

Vishnu Sarukkai, Linden Li∗, Arden Ma∗, Christopher R´e, Kayvon Fatahalian
Stanford University

Introduction

- Diffusion-based text-conditional image generation
- diffusion algorithm that generates novel, high-quality images that
  - have fidelity to the input collage's spatial composition and individual object appearance
  - exhibit global harmonization and visual coherence
- The key challenge in Collage Diffusion is
  - harmonizing an input collage while limiting variation in certain object properties (spatial location, visual characteristics)
  - allowing variation in other object properties (orientation, lighting, perspective, occlusions).

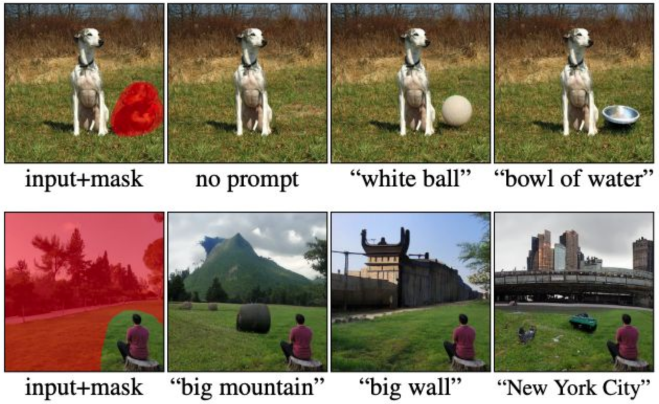


Related Work

Improving Spatial Fidelity

allow users to

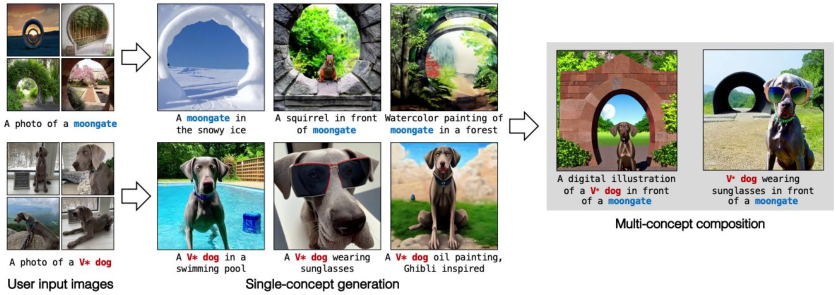
1. define desired spatial layouts of scene objects
2. use diffusion to generate objects according to the desired layout.



Blended Diffusion for Text-driven Editing of Natural Images(CVPR2022)

SpaText: Spatio-Textual Representation for Controllable Image Generation

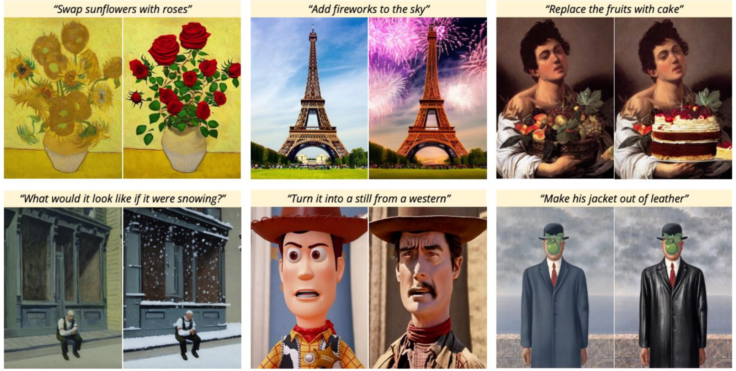
Improving Appearance Fidelity



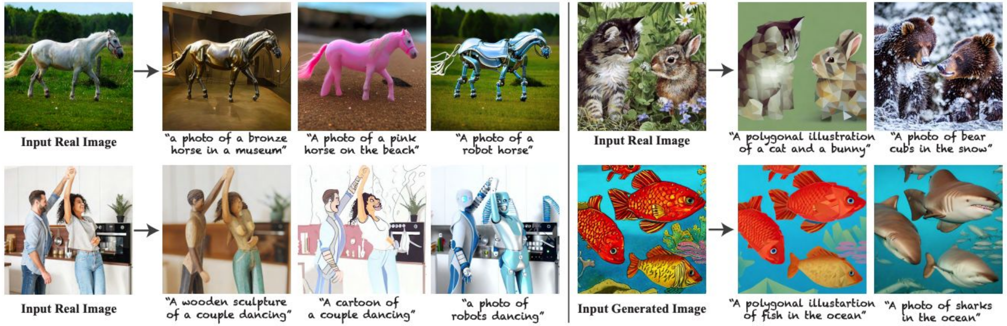
Multi-Concept Customization of Text-to-Image Diffusion

An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Image-to-Image Approaches



InstructPix2Pix: Learning to Follow Image Editing Instructions



Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation

Layered Image and Video Editing

Text2LIVE: Text-Driven Layered Image and Video Editing(ECCV 2022 Oral)

Collage Diffusion

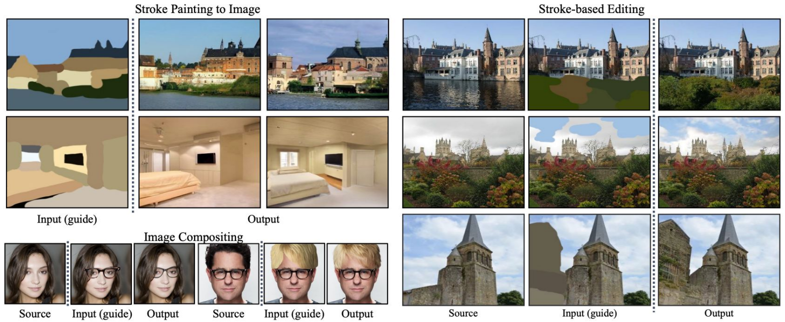
Global image harmonization

- the SDEdit algorithm improves image quality by adding Gaussian noise

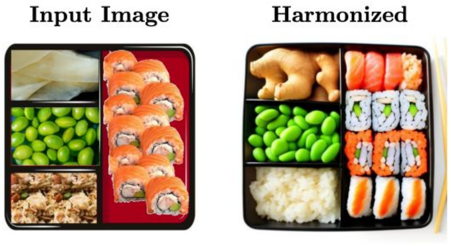
$$x_t = x_c + \mathcal{N}(0, \sigma(t)^2)$$

- using text-conditional diffusion U-Net model

$$D_\theta(x, \sigma(t), c)$$



(a) Prompt: "a poppy plant and a rose plant"



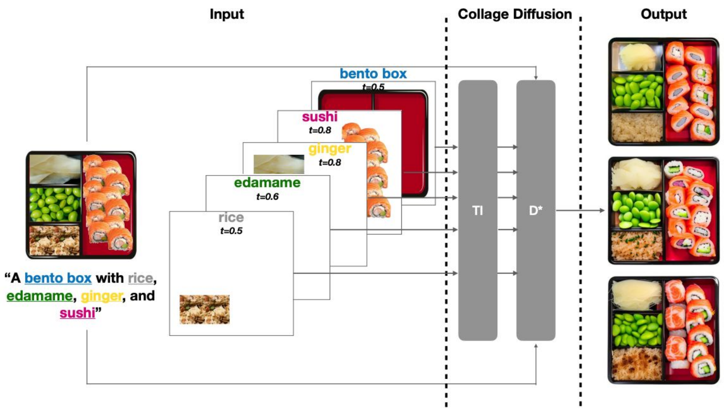
(b) Prompt: "a bento box with rice, edamame, ginger, and sushi"

Spatial fidelity through cross-attention manipulation

- In order to generate an image with the desired objects in the desired locations, Collage Diffusion modifies the text-image cross-attention in the text-conditional U-Net model Dθ

$$j = \max_{k \in 1 \ldots n} (\{k | (x_k^{\alpha})_{ab} > 0\})$$

- Cross-attention in Dθ is computed as softmax

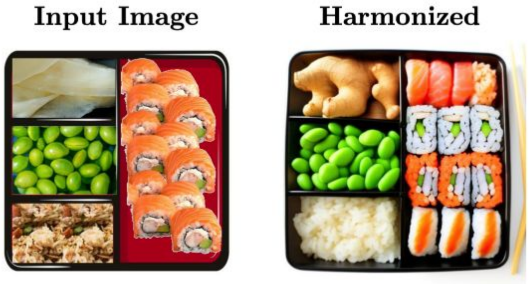


A bento box with rice, edamame, ginger, and sushi

Appearance fidelity through textual inversion

- It is often the case that the layer text c_i for a layer fails to adequately capture the appearance of layer image x_i
  - For instance, for the bento box scene, layer text “ginger” does not capture the fact that the ginger in the bento box is pickled and sliced.

$$a_i^* = \arg\min_{a_i} E_{\epsilon \sim N(0,\sigma)}[x_i^\alpha \cdot (x_{target_i} - D_\theta(x_{target_i} + \epsilon, \sigma, (a_i, c_i)))]$$



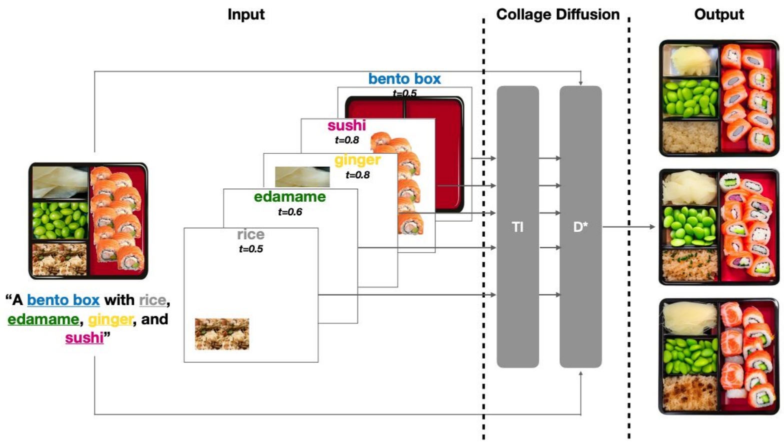
(b) Prompt: “a bento box with rice, edamame, ginger, and sushi”

Controlling the Harmonization-Fidelity Tradeoff with Per-Layer Noise

The content in the input collage layers need to be changed by the Collage Diffusion process to globally harmonize the image, and users may be willing to accept more variation for some objects in the image than for other objects.

$$x'(t-1) = x(t-1) \cdot m(t) + (x_c + \mathcal{N}(0, \sigma(t-1)^2)) \cdot (1 - m(t))$$

$$m_{ab}(t) = \begin{cases} 1 & \text{if } h_{ab} < t \\ 0 & \text{if } h_{ab} \geq t \end{cases}$$

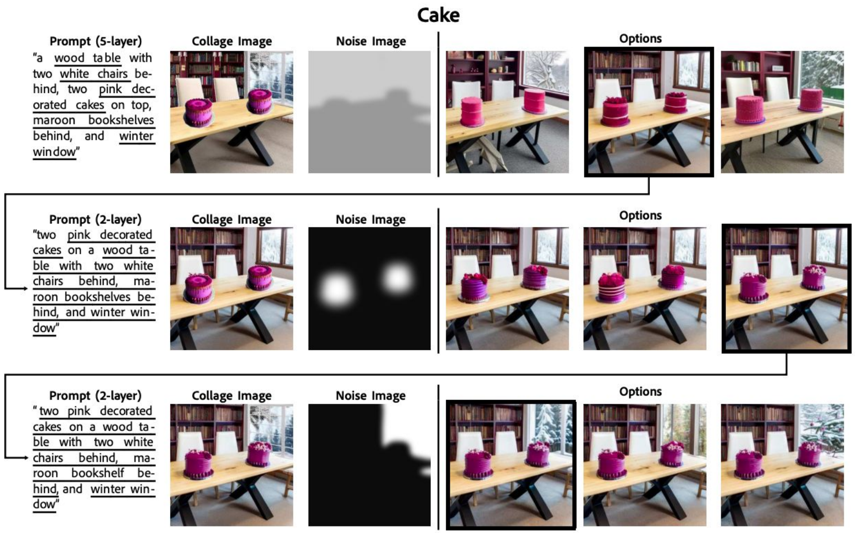


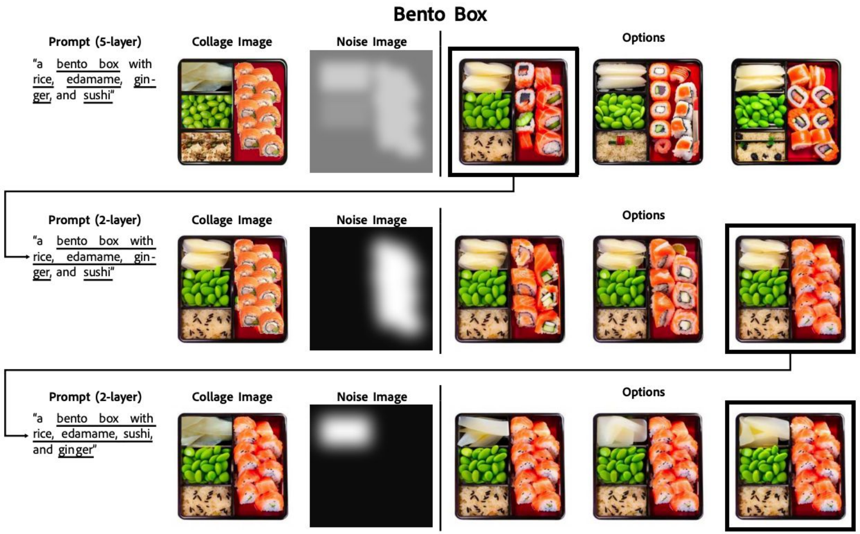
Experimental setup

Experimental setup

- Interactive Editing
  - generating 10 images using different random seeds
  - allowing the user to select the image they like the most
  - selecting an object in the selected image that they would like to re-generate
- Non-Interactive Generation
  - CA: composite image, with negative prompt "A collage"
  - GH: applying the SDEdit algorithm to composite image.
  - GH+CA: using the collage information to improve spatial fidelity, but lacks any specific mechanism to improve appearance fidelity.
  - GH+CA+TI: learned per-layer representations via Textual Inversion. This leverages collage information to improve both spatial and appearance fidelity.
  - GH+CA+TI+LN: This leverages collage information to improve both spatial and appearance fidelity, and allows user control over the harmonization-fidelity tradeoff on a per-layer basis.

Interactive Editing





Non-Interactive Generation

Bento Box

“a bento box with rice, edamame, ginger, and sushi”

Collage (5-Layer)	SA	GH	GH+CA	GH+CA+TI	GH+CA+TI+LN

Collage (5-Layer)	SA	GH	GH+CA	GH+CA+TI	GH+CA+TI+LN
	Sushi orientation and shading not harmonized, edamame in place of ginger on the top left	Harmonized image, sushi in place of ginger in the top left, wasabi in place of rice in bottom left, no sushi in bottom right	Harmonized image, ginger paste instead of sliced sushi ginger in the top left	Harmonized image, sliced sushi ginger in the top left, darker rice in the bottom left, sushi on right more similar to collage	Harmonized image, sliced sushi ginger in the top left, dark rice in bottom left, sushi on right very similar to collage

Toys

“a teddy bear, a wood train, and an american football, in front of a tan background”

Collage (4-Layer)	SA	GH	GH+CA	GH+CA+TI	GH+CA+TI+LN

Collage (4-Layer)	SA	GH	GH+CA	GH+CA+TI	GH+CA+TI+LN
	Issues with harmonization on the football and merged teddy bears, no wood train in the bottom left	Harmonized image, no wood train in the bottom left	Harmonized image, wood train in the bottom left	Harmonized image, wood train with styling of wood closer to the starting image, white face and tie of teddy bear preserved	Harmonized image, wood train very similar to the original train, red tie of teddy bear preserved

Red Skirt

"a person wearing a <u>patterned red skirt</u>, <u>buttoned blue blouse</u>, and <u>pink summer coat</u>, in front of a <u>gray background</u>"

Collage (4-Layer)	SA	GH	GH+CA	GH+CA+TI	GH+CA+TI+LN
	Image artifact on the sleeve, all objects correctly mapped to the desired locations, collage image structure preserved	Harmonized image, all objects correctly mapped to the desired locations	Harmonized image, no additional benefit from CA	Harmonized image, TI introduces folds in the skirt	No further changes with **LN**

Conclusion

- Collage Diffusion introduces a new form of control in the form of a collage, a combination of images that expresses both a user’s desired spatial layout as well as details of the visual characteristics of the individual objects in the generated image.
- One key insight in using collage input is that users can easily express compositional intent, a key element of content generation across a variety of domains—video is composed of various moving and stationary objects.