# A Video of Collage Creation

We have included a video of the collage-creation process in the Supplemental.

# B Quantitative Metrics

Here, we describe the per-layer text-image and image-image similarity metrics in greater detail.

Per-layer text-image similarity aims to measure spatial fidelity, which is defined as having the objects described by layer text $c_i$ matching the correct regions of $x_c^*$. CLIP [22] contrastively learns text and image embeddings such that the similarity of two strings or images can be measured by the cosine similarity between embedding vectors for the two concepts. We multiply $x_c^*$ with layer visibility mask $m_i$, where $m_i$ is 1 where the layer $l_i$ should be visible in the generated image and 0 otherwise (the same visibility computation as in Section 4.2), to generate a new image $x_c^* \odot m_i = x_i^*$. We compute the normalized CLIP text embedding of $c_i$, the normalized CLIP image embedding of $x_i^*$, and compute the cosine similarity of the two vectors as a proxy for spatial fidelity.

Per-layer image-image similarity aims to measure appearance fidelity, which is defined as having the objects shown by layer image $x_i$ sharing visual characteristics of the corresponding region in $x_c^*$. We compute the normalized CLIP image embedding of $x_i$, the normalized CLIP image embedding of $x_i^*$, and compute the cosine similarity of the two vectors as a proxy for appearance fidelity.

# C Qualitative Metrics

We also evaluate collage quality by generating evaluation rubrics to qualitatively measure adherence to the collage diffusion goals specified in section 2, measuring performance along the following three axes:

1. **Image quality**: have we produced a "high-quality", globally-coherent image? (0 for No/1 for Yes)
2. **Spatial fidelity**: for each desired object, have we correctly generated it in the desired position in the image? (0 for No/1 for Yes)
3. **Appearance fidelity**: for each desired object, how closely do its visual features match the visual features of the original image? Note that appearance fidelity requires spatial fidelity as a prerequisite (Construct a list of visual attributes, score between 0.0 and 1.0 per attribute)

Scores are computed for each method on a given scene by using this rubric to "grade" images generated from a large range of random seeds. We obtained 5 human evaluations for each of 10 image seeds for six scenes ("Toys", "Bento Box", "Cake", "Veggie Face", "Striped Sweater", "Ceramic Bowl") across each of four methods (**GH**, **GH+CA**, **GH+CA+TI**, **GH+CA+TI+LN**). We exclude **SA** from the qualitative evaluation because DDIM inversion is deterministic, so we cannot compute averaged scores across many seeds.

| | **GH** | **GH+CA** | **GH+CA+TI** | **GH+CA+TI+LN** |
|---|---|---|---|---|
| Global Harmonization | 0.93 | 0.90 | 0.87 | 0.88 |
| Spatial Fidelity | 0.77 | 0.83 | 0.82 | 0.93 |
| Appearance Fidelity | 0.24 | 0.41 | 0.51 | 0.77 |

Table 2: Averaged qualitative rubric evaluations highlight how **CA** improves spatial fidelity, **TI** improves appearance fidelity, and **LN** improves both spatial and appearance fidelity, all with minimal loss in harmonization. This table presents the averaged rubric results from 5 human evaluators on 10 image seeds for each of 6 seeds, as described in Section C.

The averaged results per method are presented in Table 2. **CA** improves spatial fidelity, with an increase in average score from 0.77 to 0.83. **TI** then improves appearance fidelity, increasing the score from 0.41 to 0.51. Finally, tuning the harmonization-fidelity tradeoff on a per-layer basis with **LN** boosts both spatial fidelity (0.82 to 0.93) and appearance fidelity (0.51 to 0.77). Compared to **GH**, the full *Collage Diffusion*

methodology of **GH+CA+TI+LN** does slightly decrease the harmonization score from 0.93 to 0.88; however, nearly all generated images are still well-harmonized, and in exchange we significantly boost spatial fidelity by 0.16 and appearance fidelity by 0.53.

# D Collage-Conditional Diffusion as Image-To-Image Translation


Figure 8: Image-to-image methods that aim to preserve structure are ineffective at collage-conditional diffusion.

As mentioned in Section 3, it is possible to frame collage-conditional diffusion as a controlled image-to-image task—manipulating individual objects or the overall style of an image while keeping the image structure as fixed as possible. Given access to these methods, is it even necessary to leverage individual layer information for collage-conditional diffusion? Testing both InstructPix2Pix [5] and Plug-and-Play Diffusion [30] (the **SA** method in Section 5.3), Figure 8 highlights how both methods have a minimal impact in terms of harmonizing the input bento box image—the sushi still aren't oriented in a way that fits the bento box, etc.—and Plug-and-Play Diffusion accidentally removes both the ginger and parts of the sushi from the image.

We also test the capacity of InstructPix2Pix and Plug-and-Play Diffusion to map complex compositional prompts to the appropriate regions of the image by attempting to replace the edamame in the bento box with black beans. Figure 9 highlights the failure of both techniques for the task—InstructPix2Pix replaces the the rice and parts of the sushi with a rice-bean hybrid, while Plug-and-Play Diffusion turns the edamame into a green chopped vegetable while turning the ginger into rice.

# E Note on Iterative Inpainting

Iterative inpainting-based algorithms such as "Paint by Example"[35] have spatial fidelity due to the provided inpainting masks, and can have appearance fidelity to the input layers, but struggle with harmonizing many input layers; In Fig. 10, the orientation and lighting of the cakes, chairs, table, etc. do not fit together. On the other hand, in *Collage Diffusion*, SDEdit-style denoising helps enforce global harmonization not provided by iterative inpainting approaches.

# F Additional Experimental Details

We leverage the following negative prompts by scene:

**Input Image**

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

**InstructPix2Pix**

Prompt: "replace edamame with black beans"

**PnP Diffusion**

Prompt: "a bento box with rice, black beans, ginger, and sushi"
Negative Prompt: "a collage"

Figure 9: Image-to-image methods that aim to preserve structure struggle to handle prompts with many objects.

**Input Layers (5)**

**Paint By Example**

Figure 10: Iterative inpainting-based algorithms struggle with harmonizing many input layers, as the orientations and lighting of the generated objects don't quite fit together.

- “Toys”: ‘collage’
- “Bento Box”: ‘collage’
- “Cake”: ‘collage, warped’
- “Veggie Face”: ‘collage, plastic, bowl’
- “Striped Sweater”: ‘collage, backlit, ugly, disfigured, deformed’
- “Ceramic Bowl”: ‘collage, ugly, disfigured, warped’
- “Red Skirt”: ‘’

In Figures 11 through 14, we illustrate individual layers for some of the collages tested:

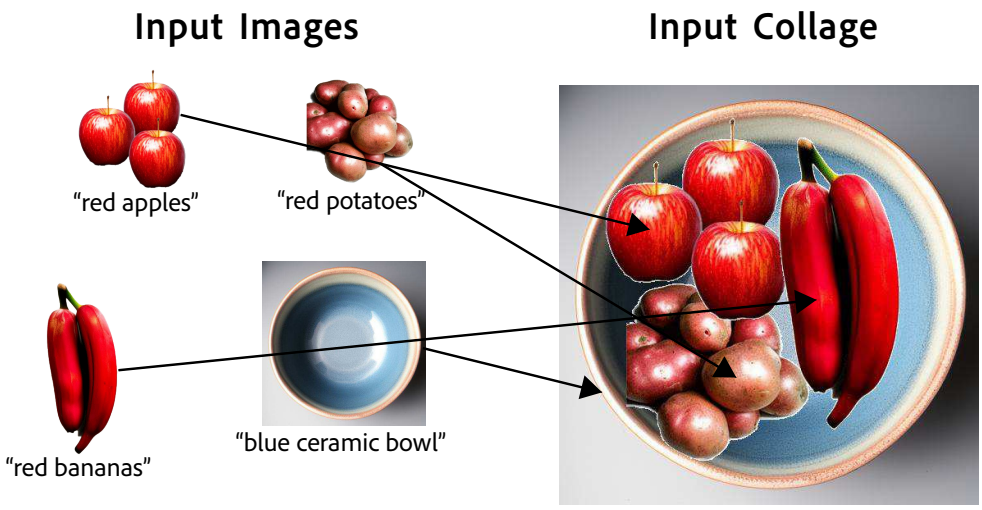


Figure 11: **Fruit** collage definition

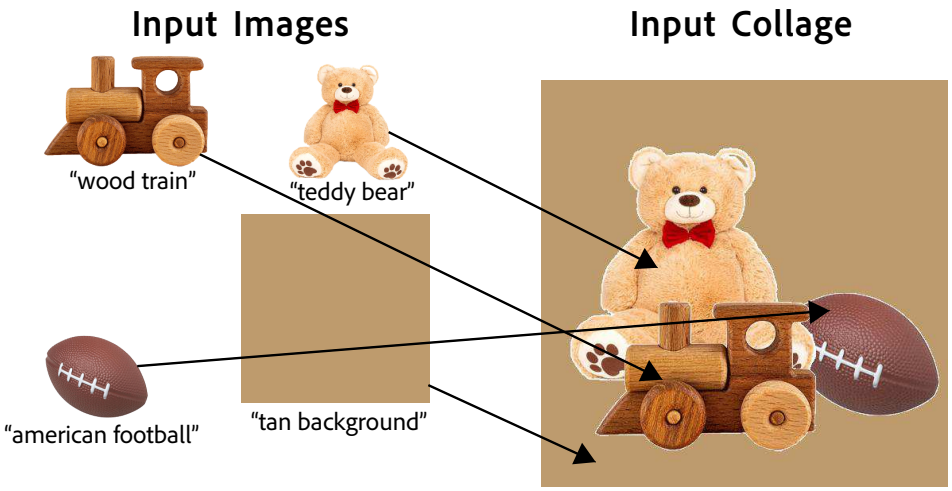


Figure 12: **Toys** collage definition

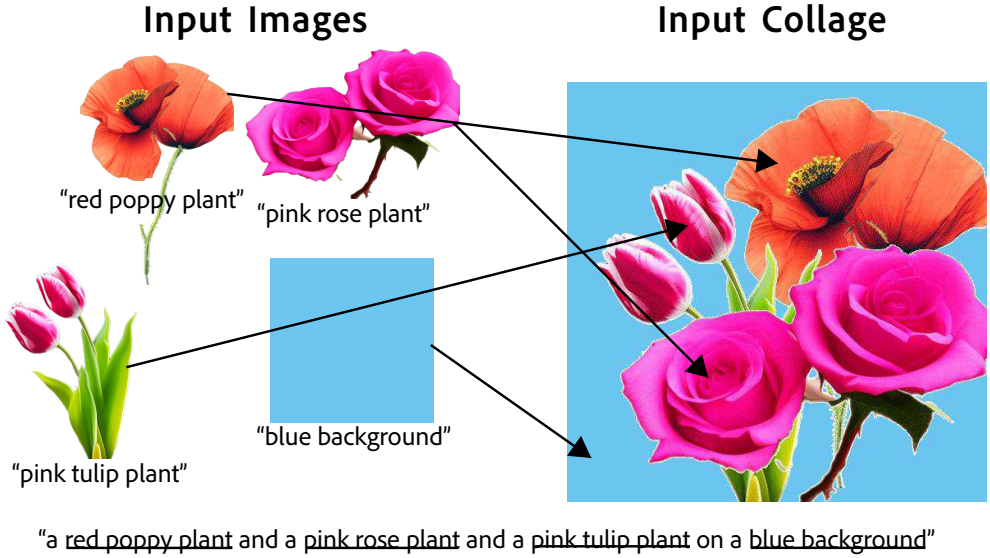


Figure 13: **Flowers** collage definition

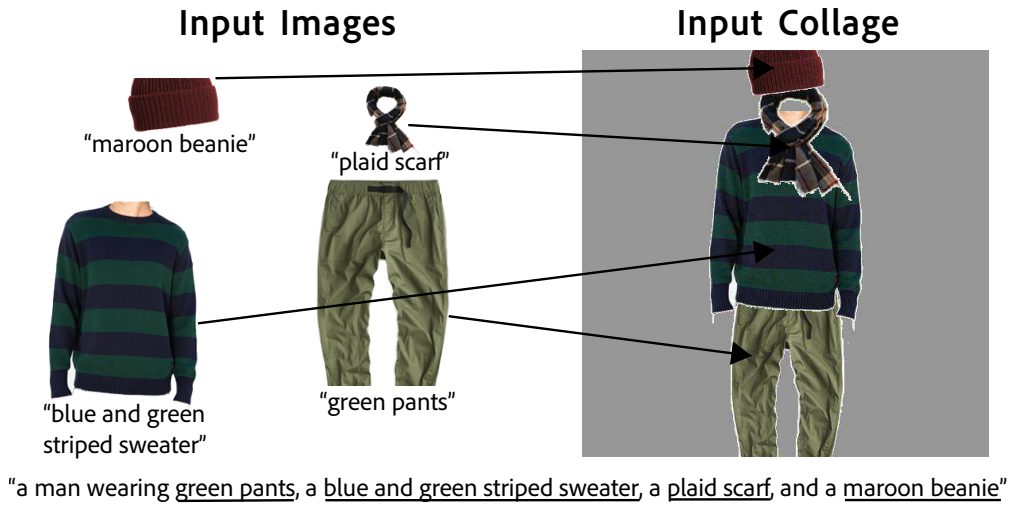


Figure 14: **Striped Sweater** collage definition

# G Note on generation speed

Efficient inference is invaluable for interactive applications. Note that without the use of ControlNet and with pre-computed Textual Inversion, *Collage Diffusion* adds negligible inference cost to the default Stable Diffusion 2.1 generation pipeline; computational cost does not increase with increasing layers! On the other hand, ControlNet does add additional computational costs per control signal used due to the additional auxiliary network; see [37] for more details.

Our experiments with *Collage Diffusion* achieve roughly 25 it/sec on a single NVIDIA V100 GPU for a $512 \times 512$ image. For our layer-based image editing UI, we leverage 4 V100 GPUs simultaneously and generate one image from each GPU using 50 diffusion solver steps, enabling users to view 4 possible harmonizations of the input layers within 2 seconds.

# H Automatic parameter tuning

We devise an automatic heuristic-based parameter adjustment algorithm to aid the user in navigating the design space of parameters governing spatial fidelity, appearance fidelity, and harmonization. We utilize the following heuristics in the algorithm:

1. The noise strength is set to a lower value for foreground objects and a higher value for background objects. Conversely, canny edge strength via ControlNet is set to a higher value for foreground objects and a lower value for background objects. These heuristics are chosen because users tend to prefer preserving the visual appearance of subjects in the foreground while trading off visual fidelity in the background for harmonization. Foreground objects and background objects are determined via layer order.

2. Cross attention strengths are set to a higher values for foreground and smaller objects; they are set to lower values for background and larger objects. Objects in the foreground and smaller objects tend to be omitted with low cross attention strengths. Users generally care more about these layers since they tend to be important components of the image composition. We determine whether an object is "small" or "large" by evaluating the scale of the size relative to the entire canvas.

In Fig. 15, the automatic parameter adjustment algorithm is able to generate compelling images with high spatial fidelity, appearance fidelity, and harmonization for scenes of varying complexity. Even for the

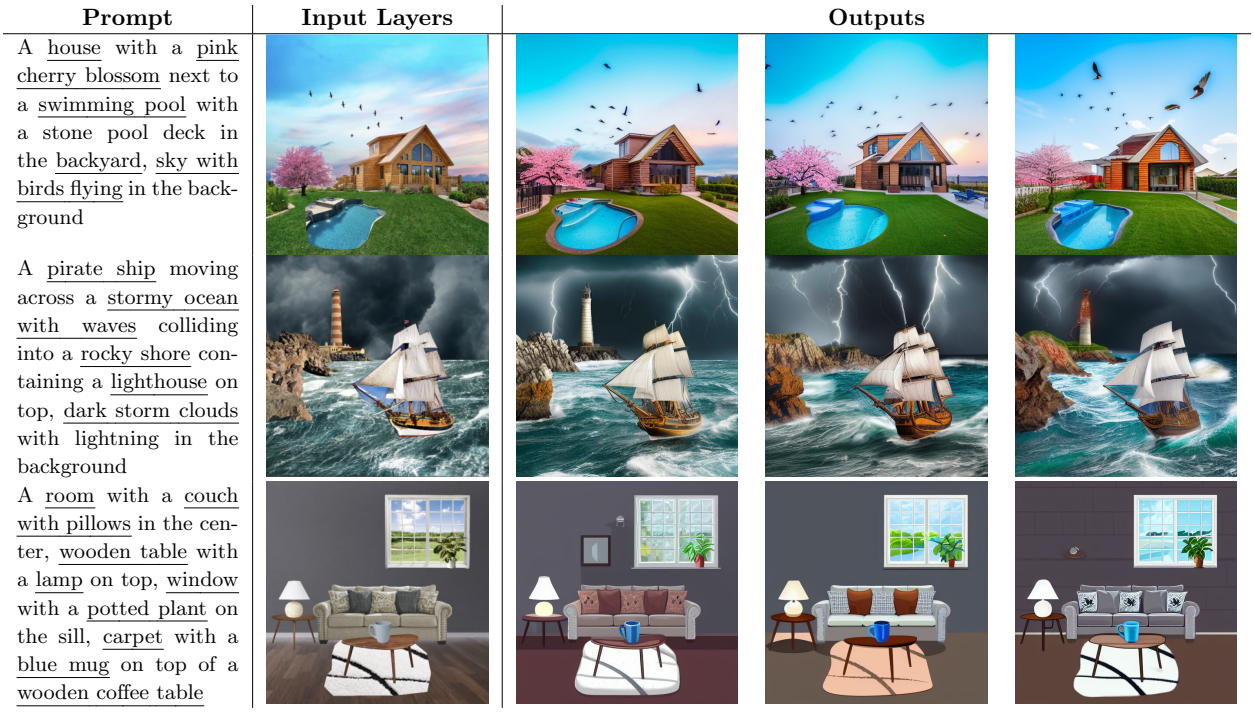


Figure 15: Our automated heuristic-based parameter adjustment algorithm generates compelling images across several seeds for all three scenes. Even for the living room scene, which contains nine layers of varying sizes, the algorithm is able to generate a high-quality image with automated parameters.

living room scene, which contains nine layers of varying sizes, the algorithm is able to generate a high-quality image with automated parameters.

# I Robustness to Random Seed

Figures 16 through 22 contain additional results with different noise seeds for **GH**, **GH+CA**, and **GH+CA+TI**, highlighting that the trends of **CA** improving spatial fidelity and **TI** improving appearance fidelity hold across noise seeds for the collages tested. Additional results are not included for **SA** because the Plug-and-Play algorithm generates noise seeds through DDIM inversion, not at random [30].


Figure 16: "a teddy bear, a wood train, and an american football, in front of a tan background"


Figure 17: "a bento box with rice, edamame, ginger, and sushi"


Figure 18: Prompt: "a face made of vegetables, including a yellow bell pepper and a green bell pepper, a white cauliflower, red potatoes, baby corn, small cucumber, bean sprouts, and floret broccoli, on a grey background"


Figure 19: "a wood table with two white chairs behind, two decorated cakes on top, maroon bookshelves behind, and winter window"


Figure 20: "a blue ceramic bowl with red potatoes, red apples, and red bananas"


Figure 21: "a man wearing green pants, a blue and green striped sweater, a plaid scarf, and a maroon beanie"


Figure 22: "a red poppy plant and a pink rose plant and a pink tulip plant on a blue background"