Collage Diffusion

Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, Kayvon Fatahalian

Stanford University

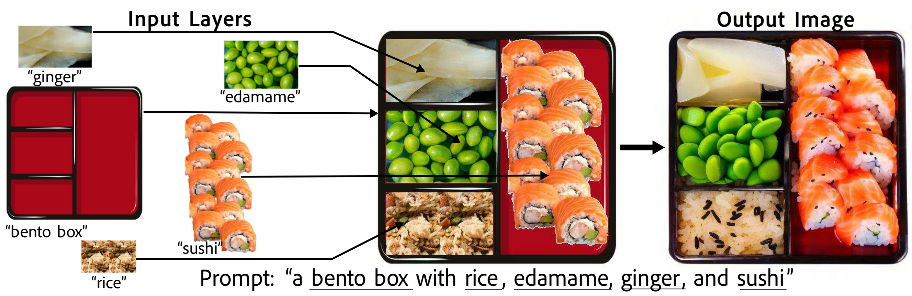


Figure 1: A layer is defined as an image-text pair. Given a sequence of layers and a full-image string, *Collage Diffusion* generates an image that is globally harmonized, yet preserves the locations and key visual characteristics of each input layer.

Abstract

We seek to give users precise control over diffusion-based image generation by modeling complex scenes as sequences of layers, which define the desired spatial arrangement and visual attributes of objects in the scene. Collage Diffusion harmonizes the input layers to make objects fit together—the key challenge involves minimizing changes in the positions and key visual attributes of the input layers while allowing other attributes to change in the harmonization process. We ensure that objects are generated in the correct locations by modifying text-image cross-attention with the layers' alpha masks. We preserve key visual attributes of input layers by learning specialized text representations per layer and by extending prior diffusion-based control mechanisms to operate on layers. Layer input allows users to control the extent of image harmonization on a per-object basis, and users can even iteratively edit individual objects in generated images while keeping other objects fixed. By leveraging the rich information present in layer input, Collage Diffusion generates globally harmonized images that maintain desired object characteristics better than prior approaches.

1. Introduction

Diffusion-based image generation [9, 12, 23, 24, 27, 28] has captured widespread interest with its seemingly magical ability to generate plausible images from a text prompt. Unfortunately, text is a highly ambiguous specification of an image, forcing users to spend significant time tweaking prompt strings to obtain a desired output. A body of recent work has therefore focused on providing more precise controls for scene composition via additional inputs: controlling composition via sketching [3], filling in user-provided segmentation masks [2, 29], providing an image seed for generation [19], etc. Similarly, the desire to precisely dictate object appearance, *"the sushi in THIS reference photo"* rather than *"the sushi"*, has led to approaches that condition generation based on example images [10, 15, 25].

We seek to give users precise control over image output when creating scenes featuring a collection of objects with a specific spatial arrangement. For example, in Figure 1, "A bento box with rice, edamame, ginger, and sushi" neither describes what items go in which Bento bin, nor suggests how each of the items should look. Rather than relying on ambiguous text prompts or forcing the user to sketch scene forms, we return to a traditional and easy-to-create means of expressing artistic intent: defining the composition of a scene and the appearance of individual objects by *making a sequence of layers*. To specify a scene, a user need only acquire reference images of desired scene objects (e.g., via image search or via output from an existing generative model), arrange them on a canvas using a traditional layer-based image editing UI, and pair each object with a text prompt.

Given these layers, we introduce *Collage Diffusion*, a diffusion-based image harmonization algorithm that generates images that 1) have *fidelity* to the input layers' spatial composition and object appearance, but 2) exhibit global *harmonization* and visual coherence that is representative of

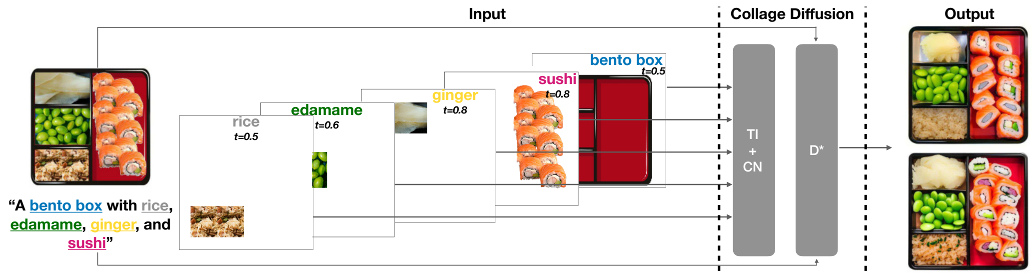


Figure 2. *Collage Diffusion* takes as input a sequence of layers of RGBA images paired with text (the image of sushi and the text "sushi"), along with a full-image text string ("A bento box with rice, edamame, ginger, and sushi"). Layer information enables 1) manipulating cross-attention to map individual layers to the corresponding image regions, creating improved diffusion model D^*, 2) learning layer-specific representations using textual inversion (TI), 3) having the option to preserve per-layer image structures with ControlNet (CN), and 4) harmonizing layers according to per-layer noise levels t_i. *Collage Diffusion* outputs globally-harmonized images that contain objects in the specified locations, and share visual characteristics with the input layer images. ***In the rest of the paper, for brevity we only display the layer composite image and prompt, and we use underlined substrings to indicate contents of individual layers.***

"plausible" real-world images. There is an inherent trade-off between harmonization and fidelity: harmonization involves changing properties of the input layers so that objects "fit together" in a consistent image, while fidelity involves preserving properties of the layers. The key challenge is harmonizing a sequence of layers while limiting variation in certain layer properties (color, texture, edge maps, etc.), but allowing variation in other properties. We tackle this challenge by leveraging the rich information present in layer input—building upon prior diffusion-based techniques for image harmonization, spatial control, and appearance control, we extend them with a focus on mechanisms for per-layer control.

Specificially we make the following contributions:

1. We introduce layer-conditioned diffusion, where generation is conditioned on alpha-composited RGBA layers as well as text prompts describing the content of each layer. Sequences of layers can be authored by users in minutes, and *Collage Diffusion* generates high-quality images that respect both the desired scene composition and object appearance, even for complex scenes with many layers.
2. We extend prior diffusion-based control mechanisms [3, 10, 37] to operate on sequences of layers, ensuring that output images adhere to the composition depicted by the layers (cross-attention [3]) and retain salient visual features of objects in each layer (textual inversion [10], ControlNet [37]).
3. We illustrate how layer input allows users to control the harmonization-fidelity tradeoff on a per-layer basis and also enables users to iteratively refine generated images.

2. Problem Definition and Goals

Our goal is to generate globally harmonized images that respect a user's desired scene composition, both in terms of *spatial fidelity*, i.e., preserving the positions and sizes of the desired objects, as well as *appearance fidelity*, i.e., preserving the visual characteristics of the objects. We propose that the user describe their intent by means of a sequence of layers alongside a global text prompt. For brevity, we call this combination a *collage*. We first define a collage, then introduce our goals for collage-conditional generation.

As illustrated in Fig. 2, we define collage C as:

1. A full-image text string c, describing the entire image to be generated ("A bento box with rice, edamame, ginger, and sushi")
2. A sequence of n layers $l_1, l_2, ...l_n$, ordered from back to front, with each l_i having:
   (a) An RGBA image x_i (the alpha-masked input image of sushi), with alpha layer x_i^α
   (b) A text string c_i describing the layer, which is a substring of c ("sushi")

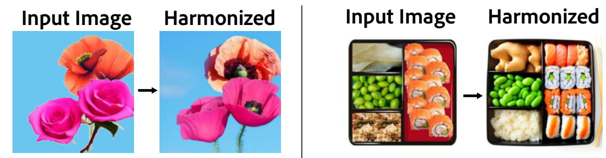
Given input collage C, we seek to generate output image x_c^* with the following properties:

1. *Global harmonization*: x_c^* is a well-harmonized, high-fidelity image. In Figure 1, the output features consistent perspective, lighting, and occlusions among scene objects.
2. *Spatial fidelity*: generated objects are in the correct locations. Specifically, for all layers l_i, the objects described by layer text c_i are generated in the correct regions of x_c^*. In Figure 1, "edamame," "ginger," etc. are all in the same regions of the output image as in the input collage.
3. *Appearance fidelity*: generated objects maintain desired visual characteristics. Specifically, for all layers l_i, in addition to matching layer text c_i, regions of x_c^* that depict the contents of the layer share key visual characteristics with x_i. In Figure 1, the "ginger" in the output image remains sliced sushi ginger (not whole ginger), etc.

In order to achieve the consistency of a real image, we aim to constrain both the spatial layout of generated images and certain aspects of the appearance of individual objects, allowing other aspects to vary in the harmonization process.

3. Related Work

One natural starting point is to "flatten" the input layers into image x_c by alpha-compositing the sequence of



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

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

Figure 3. Without layer information, image harmonization can lead to a loss of spatial and appearance fidelity. Added noise can disrupt object-location mappings—on the left, "poppies" take the place of the "roses." Added noise also can obscure specifics of an object's appearance—on the right, the generated "ginger" is whole instead of sliced.

layer images $x_1, x_2, ...$ into a single image [21], then use diffusion-based image harmonization to improve the visual quality of the image [2, 19, 26]. Diffusion-based approaches can harmonize geometry [19, 29], rather than restricting focus to color and lighting [7, 8, 13, 34]. The problem with this flatten-then-harmonize algorithm is that generated results may diverge from the content of the initial image, undermining user intent. For example, in Fig. 3, noise-based harmonization [19] turns the pink roses into poppies despite the prompt and turns sliced sushi ginger into whole ginger. We seek to better maintain the spatial and appearance fidelity of the initial layers.

Improving Spatial Fidelity Prior work has suggested approaches to (1) define spatial layouts of scene objects, and then (2) generate objects according to the desired layout. Existing techniques define spatial layouts using segmentation maps and bounding boxes, whether defining a region for inpainting [2, 6, 18, 19, 29, 35, 36], providing a full-image segmentation map [1, 3], or using bounding boxes [38]. Inpainting approaches struggle to maintain global coherence with many layers (see Supplemental). Instead of hand-drawing a segmentation map, we see layers as an intuitive, alternative way to specify spatial composition.

Improving Appearance Fidelity In addition to generating objects in the desired locations, we aim to preserve visual characteristics of input layers. Several recent works specialize diffusion models to particular visual concepts (objects, styles, etc.) [10, 15, 25], requiring several input images and either fine-tuning the model [15, 25], learning a specific textual representation for the object [10], pre-training on reference images [33], or reverse-engineering a prompt for a given image [32]. These methods struggle to generate high-quality images of scenes with compositions of many objects [10, 15, 25, 33]. In addition, appraches that fine-tune model weights require either joint multi-concept training or post-hoc combination of model weights, both of which struggle in regimes with several objects [15, 25]. Alternatively, ControlNet [37] enables us to preserve derived features of input layers (edge maps, pose, etc.) without learning a visual concept personalized to the specific object.

We address the goal of appearance fidelity by extending both textual inversion [10] and ControlNet [37] for performance with individual layers. We find that the learned representations are effective for maintaining key visual characteristics of input layers when paired with techniques for spatial control. When preserving an image structure from an input layer such as an edge map, our extension of ControlNet is effective.

Image-to-Image Approaches Constrained image harmonization can also be framed as image stylization: from low-quality layer composite to high-quality harmonized output. Stylization can be approached using existing methods for controlled image-to-image diffusion [5, 11, 20, 30, 37]. Derived features (canny edges, pose, etc.) can provide control [37], but fails to constrain scene composition—the locations of objects are not preserved. Other methods directly [11, 20, 30] or indirectly [5] manipulate U-Net attention layers (cross-attention [5, 11, 20] and self-attention [30]) to maintain image structure while making either local edits (adding/removing/modifying objects) or global edits (style, lighting). Unfortunately, this approach is insufficient for layer-conditional diffusion. *Input layers often need to be changed significantly* to fit together in a harmonized image, as objects may need to be rotated, partially occluded, etc. (see the orientation of the sushi in Fig. 2). This is difficult when preserving the "structure" of the input image. We evaluate against one constrained image-to-image approach [30], and discuss additional baselines in the Supplemental. Less constrained harmonization techniques [19] serve as a more useful starting point for *Collage Diffusion* since they allow the desired flexibility in image structure.

Layered Image and Video Editing Layer-based image and video editing is well-established in computer graphics [21, 31] and is being increasingly adopted in machine learning-driven methods [4, 14, 16, 17]. Layered representations allow modification of individual components in images [4, 16] and in video [4, 14, 17]. This process often requires generating a layered representation from a single input video or image. In contrast, we assume that layered information is provided as input, using machine learning to synthesize image output from the layers.

4. Collage Diffusion

To frame discussion of layer-based image harmonization, we first recap how text-conditioned diffusion models can perform image harmonization by leveraging added noise. Then, we describe how *Collage Diffusion* leverages additional information from individual layers to increase both spatial and appearance fidelity for harmonized output.

4.1. Global image harmonization

Leveraging only layer composite image x_c and full-image string c, the SDEdit algorithm [19] improves image quality by adding Gaussian noise with standard deviation $\sigma(t)$ to x_c, then denoising the noised image $x_t = x_c + \mathcal{N}(0, \sigma(t)^2)$ to generate output image x_c^*, using a text-conditional diffusion U-Net $D_\theta(x, \sigma(t), c)$ as an image prior [24] (x is a noisy input image, $\sigma(t)$ is the noise level at time t, and c is the text conditioning). Unfortunately, added noise can make it difficult to map objects to the correct image regions and can obscure key visual details, reducing spatial and appearance fidelity to the original layers (Fig. 3). Layer input, with text c_i and image x_i corresponding to each region of the image, provides additional information facilitating more precise control over individual components of the generated image.

4.2. Spatial fidelity: cross-attention manipulation

To generate an image with the desired objects in the desired locations, *Collage Diffusion* modifies the text-image cross-attention in text-conditional U-Net model D_θ. Not all tokens in full-image input text c correspond to layer strings c_i—the start token, end token, several words in the input string, and padding tokens lack specific regional influence. We refer to these tokens as "global" tokens, while layer-specific tokens are "layer" tokens. For instance, in Fig. 2, "with" is a global token and "rice" is a layer token. *Collage Diffusion* constrains image generation by restricting the influence of layer tokens to the regions of the image where the corresponding layer is visible.The visible layer at pixel coordinate (a, b) is defined as $j = \max_{k \in 1 \ldots n} (\{k | (x_k^\alpha)_{ab} > 0\})$, where j is the layer index of the highest of the n layers with non-zero alpha at pixel coordinate (a, b).

Cross-attention in D_θ is computed as $\mathrm{softmax}(\frac{QK^T}{\sqrt{d}})V$, where Q is a matrix of query embeddings from image tokens, K is a matrix of key embeddings from text tokens, V is a matrix of value embeddings from text tokens, and d is the embedding dimensionality. To increase or decrease the influence of a particular token on a part of the image, *Collage Diffusion* alters QK^T, an approach similar to the mechanism proposed by eDiffI [3]. Like eDiffI, *Collage Diffusion* uses positive attention map A^{pos} to increase the influence of layer tokens on a region relative to global tokens, but unlike eDiffI, *Collage Diffusion* also constructs negative map A^{neg} to prevent layer tokens from influencing regions outside the desired location.

To alter QK^T, *Collage Diffusion* constructs attention maps $A^{pos}, A^{neg} \in \mathbb{R}^{N_v \times N_t}$, where N_v is the number of image tokens and N_t is the number of text tokens, and each column A_j^{pos}, A_j^{neg} is a flattened alpha mask dependent on visibility of text token j. $A_{ij} = 0$ for all global tokens j. $A_{ij}^{pos} = 1$ if image token i corresponds to a region of the image that layer token j should influence, and $A_{ij}^{neg} = 1$ if image token i corresponds to a region of the image that layer token j should not influence. Along with scalar weights w^{pos} and w^{neg}, attention maps A^{pos} and A^{neg} are incorporated into the softmax operation: $\mathrm{softmax}(\frac{QK^T + w^{pos}A^{pos} - w^{neg}A^{neg}}{\sqrt{d}})V$. With larger weights w^{pos} and w^{neg}, the influence of attention maps A^{pos} and A^{neg} on image layout is greater. Weights w^{pos} and w^{neg} vary dependent on noise level $\sigma(t)$ throughout the diffusion process: $w^{pos} = v^{pos} \cdot y(t)$ and $w^{neg} = v^{neg} \cdot y(t)$, where $y(t) = \log(1 + \log(1 + \sigma(t))) \cdot \max(QK^T)$, and v^{pos} and v^{neg} are scalars specified by the user. Denote this modified diffusion model as D_θ^*.

4.3. Appearance fidelity: inversion and ControlNet

Layer text c_i for a given layer often fails to adequately capture the intended appearance of layer image x_i. For instance, in Fig. 2, layer text "ginger" does not capture that the ginger is pickled and sliced. Starting image x_c provides some guidance on the desired look of each layer, but the influence of x_c is reduced when noise is added to the image. Therefore, we would like additional control over the appearance of generated content corresponding to individual layers. We offer per-layer control over two aspects of the input layer: the unique attributes of the real-world object, such as colors, textures, and shape, as well as the image structure, including edges and poses.

To preserve attributes of the real-world object in the layer, *Collage Diffusion* builds upon Textual Inversion [10]: layer text c_i is specialized to image x_i by learning a *modifier* token a_i per layer, prepended to the layer text: (a_i, c_i). a_i serves as an adjective describing the object in layer l_i, subject to the constraints of the existing layer description c_i. For instance, string "ginger" is modified into new string "$\langle a_i \rangle$ ginger". The embedding for a_i is learned by optimizing the following loss:

$$
\begin{aligned}
a_i^* = \arg\min_{a_i} & E_{\epsilon \sim N(0,\sigma)} \\
& [x_i^\alpha \cdot (x_{target_i} - D_\theta(x_{target_i} + \epsilon, \sigma, (a_i, c_i)))]
\end{aligned} \tag{1}
$$

target image x_{target_i} is constructed by alpha-compositing the first i layers $l_1 \ldots l_i$, and layer alpha mask x_i^α restricts the loss to the relevant region of x_{target_i}. Textual Inversion [10] learns token a_i as a standalone prompt, and performs optimization using several images of the same object that communicate invariances in pose, lighting, etc. *Collage Diffusion* operates in a single-image setting, where x_{target_i} is the only reference for learning a_i. Therefore, it leverages the layer textual description c_i to help regularize optimization.

To preserve image structure, we extend ControlNet [37] to enable per-layer controls. The ControlNet auxiliary network outputs 2-d feature maps $m_k \in R^{h,w,c}$ from its zero convolutions, where h is height, w is width, and c is number of channels. In standard ControlNet, we multiply feature maps m_k by scalar ControlNet weight $w_{all} \in [0, 1]$ that controls the "strength" with which ControlNet influences the generated image. We replace w_{all} with weight map

w_{layer}: the user sets ControlNet weights w_i for each layer l_i, and the w_i are converted into single-channel weight map w_{layer}: $w_{layer_{ab}} = t_j$, where $j = \max_{k \in 1...n}(\{k|(x_k^\alpha)_{ab} > 0\})$ is the layer index of the highest of the n layers with nonzero alpha for pixel coordinate (a, b), and $w_{layer_{ab}}$ is the value of w_{layer} at pixel (a, b). We resize w_{layer} to $[0, 1]^{h,w}$ using bilinear interpolation, then elementwise-multiply $w_{layer} * m_k$ to produce re-weighted ControlNet outputs. Now, the user can control the influence of ControlNet on regions corresponding to each layer with per-layer weights w_i.

4.4. Tuning the Harmonization-Fidelity Tradeoff

The content in the input layers must be modified to globally harmonize the image, and users may be willing to accept more variation for some objects than others. Layer input allows users to control the harmonization-fidelity tradeoff on a per-object basis by having users specify the desired level of harmonization per layer. The user sets noise levels t_i for each layer l_i, and the t_i are converted into single-channel noise image h: $h_{ab} = t_j$, where $j = \max_{k \in 1...n}(\{k|(x_k^\alpha)_{ab} > 0\})$ is the layer index of the highest of the n layers with nonzero alpha for pixel coordinate (a, b), and h_{ab} is the value of h at pixel (a, b). A Gaussian blur is applied to h to smooth boundaries where the noise level changes sharply. Building upon Blended Diffusion [2], *Collage Diffusion* modifies the diffusion process so that different levels of noise are added to different regions of the image according to h, controlling the harmonization-fidelity tradeoff per layer:

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

$$m_{ab}(t) = \begin{cases} 1 & \text{if } h_{ab} < t \\ 0 & \text{if } h_{ab} \geq t \end{cases} \tag{3}$$

where $x(t)$ is the original solver output at time t, $x'(t)$ is the modified solver output at time t, and $m(t)$ is a binary mask computed at time t based on the noise image h. For instance, in Fig. 2, $t_i = 0.5$ for both the “bento box” and “rice” layers, $t_i = 0.6$ for the “edamame” layer, and $t_i = 0.8$ the “sushi” and “ginger” layers, indicating that the user would like a greater level of harmonization for the ginger and sushi than for the bento box and the rice.

4.5. Editing Individual Layers in Generated Images

Per-layer noise controls also enable layer-by-layer image editing. Especially for scenes with many objects, it can be difficult to look through large output galleries to find an example where all objects in the scene look *exactly* as desired. Rather, the user can simply select a generated image where nearly all objects look as desired, then refine the image by generating alternate possibilities for the remaining objects.

Per-layer noise controls enable users to keep a part of an input collage “fixed” by setting the noise level to $t = 0$ for the layers that should remain constant. Having generated an image using *Collage Diffusion*, an individual object may be edited by creating a new two-layer collage, where the generated image is the background layer, and the object to be re-generated is the foreground layer. Setting per-layer noise $t = 0$ to the background layer, a variety of possibilities are generated for the foreground layer, harmonized and combined with the fixed background layer. Especially for complex scenes, a small part of a generated image might not quite look right. Here, iterative, layer-driven editing can be the difference between obtaining a final image that is *nearly* satisfactory and one that precisely satisfies the user’s image generation goals. *Collage Diffusion*’s generation speeds support interactive editing workflows; see the Supplemental for additional discussion.

4.6. Auto-adjust parameters

The additional parameters provided for tuning spatial and appearance fidelity substantially improve user control over the image harmonization process, but can pose difficulty for novice users to tune. Therefore, we introduce a heuristic-based algorithm that automatically generates parameters that qualitatively produce aesthetically pleasing images. We discuss our parameter-setting algorithm in detail in the Supplemental.

5. Evaluation

We evaluate the value of layer information in terms of supporting iterative editing workflows as well as how that information can meet our fidelity and image harmonization goals. We choose to focus on qualitative evaluation because our goals are primarily visual and because generative metrics for distributional comparison (FID, etc.) are not applicable in the layer-conditional setting where no ground-truth test dataset for “the perfect output” exists. Nevertheless, we also present a short quantitative study that mirrors our qualitative observations.

5.1. Experimental Setup

We evaluate the capacity of *Collage Diffusion* to generate images without a user in the loop against two prior work baselines that do not use layer information. We also ablate *Collage Diffusion* to create (1) a baseline that omits textual inversion but does modify cross-attention using layer information, and (2) a baseline that modifies cross-attention, leverages textual inversion, but does not enable per-layer control over harmonization. We evaluate the performance of the following methods for a range of scenes:

1. **SA**: Image generation with **S**elf-**A**ttention control via Plug-and-Play Diffusion [30] applied to composite image x_c, with negative prompt “A collage”. This is a baseline that does not leverage layer information, but maintains the image structure of x_c via self-attention control.
2. **GH**: **G**lobal **H**armonization by applying SDEdit [19] (Sec. 4.1) to composite image x_c. This is another base-

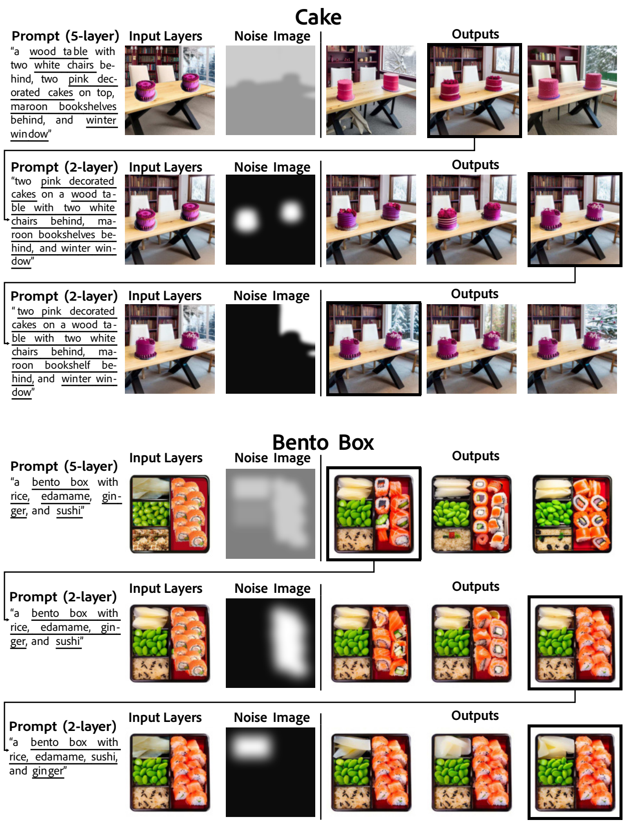


Figure 4. An iterative editing workflow where the user modifies individual layers of generated images for the **Cake** and **Bento Box** scenes. In each example, the user generates an initial image using *Collage Diffusion*, then improves the images using two refinement iterations, re-generating one of the original input layers in each refinement iteration.

line that does not leverage layer information.

3. **GH+CA**: **GH** with modified **C**ross-**A**ttention (Sec. 4.2). This builds upon **GH** by using layer information to improve spatial fidelity, but lacks specific mechanisms to improve appearance fidelity.
4. **GH+CA+TI**: **GH** applied to composite image x_c with both **CA** learned per-layer representations via **T**extual **I**nversion [10] (Sec. 4.3). This leverages layer information to improve both spatial and appearance fidelity.
5. **GH+CA+TI+LN** (*Collage Diffusion*): **GH** applied to composite image x_c with both **CA** and **TI**, with per-**L**ayer **N**oise control (Sec. 4.4). This leverages layer information to improve both spatial and appearance fidelity, and allows user control over the harmonization-fidelity tradeoff on a per-layer basis.

Controlled image-to-image techniques [5, 11, 20, 30] adhere too closely to starting image structure, as discussed in Sec. 3, resulting in performance worse than the **GH** baseline. To illustrate this, we evaluate against one of these methods in **SA** [30]; see the Supplemental for additional discussion.

Figure 5. (Part 1) By leveraging layer information, *Collage Diffusion* generates images with greater spatial and appearance fidelity than the baseline **GH** approach. For each scene above, there are several aspects in which **CA**, **TI**, and **LN** improve fidelity; we comment on some of these aspects in each row. Compared to **GH**, **SA** fails to effectively harmonize input layers; we comment on issues with harmonization in each row.

Scene construction. We evaluate *Collage Diffusion* on seven diverse scenes created using an interactive layer editor UI that provides controls similar to those in popular layer-based image editing software. Creating a scene using the UI is simple and straightforward–see the Supplemental for a video example.

Model and optimization. We use the Stable Diffusion [24] 2.1 base model as D_θ for **GH**, **GH+CA**, **GH+CA+TI**, and **GH+CA+TI+LN**, and generate images using the Euler ancestral solver with 50 steps. For each scene, we tune the noise added to the image to qualitatively optimize the harmonization-fidelity tradeoff; values are between $t = 0.7$ and $t = 0.8$ for all scenes tested. We use the official PyTorch implementation of **SA** [30].

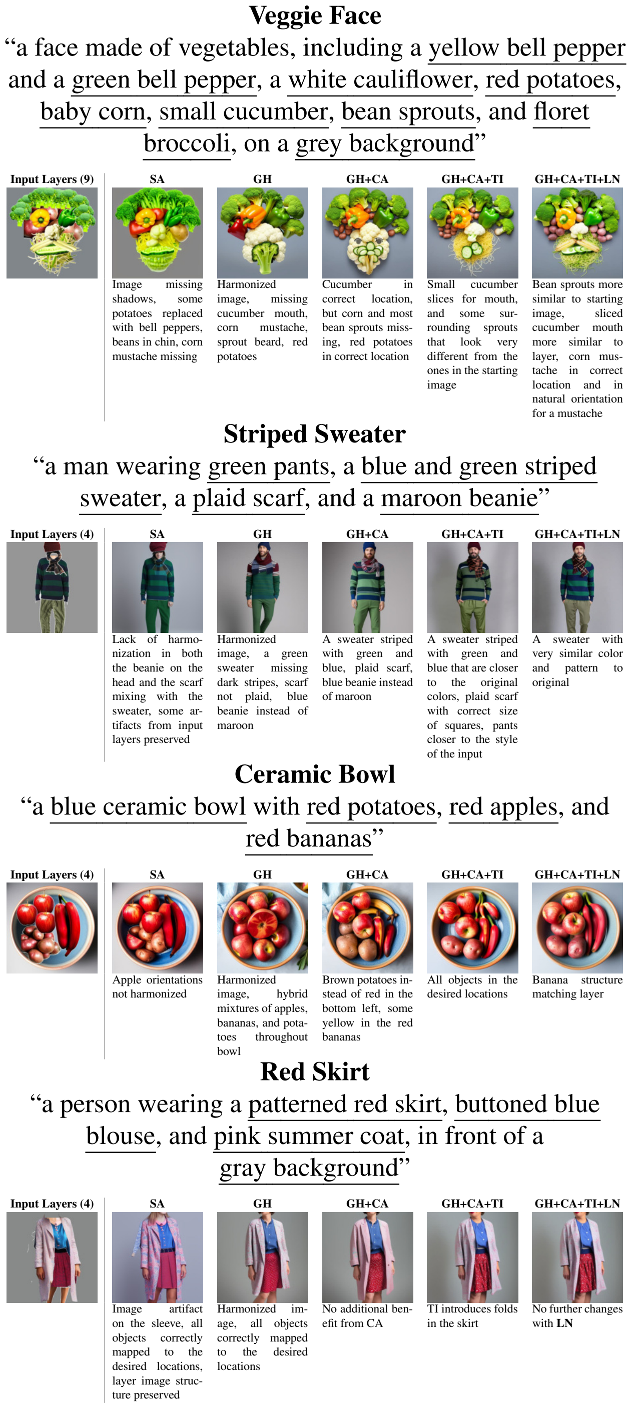


Figure 6. (Part 2) By leveraging layer information, *Collage Diffusion* generates images with greater spatial and appearance fidelity than the baseline **GH** approach. See Fig. 5 caption for more detail.

Metrics We use the following metrics for quantitative evaluation. Our spatial fidelity goals aim for layer text c_i to match the visual content in x_c^* in regions where layer i is visible—we measure this by computing CLIP [22] text-image similarity between c_i and the corresponding region of x_c^*. Appearance fidelity aims for layer image x_i to match the visual content in x_c^* where layer i is visible—we measure this by computing CLIP image-image similarity between x_i and the corresponding region of x_c^*. We include additional details on metrics in the Supplemental.

	GH	**GH+CA**	**GH+CA +TI**	**GH+CA +TI+LN**
↑Txt-Img. Sim.	0.215	0.236	0.233	0.238
↑Img-Img. Sim.	0.846	0.867	0.877	0.893

Table 1. **CA**, **TI**, and **LN** help *Collage Diffusion* improve both spatial fidelity, as measured by per-layer text-image similarity with the input layers, and appearance fidelity, as measured by per-layer image-image similarity with the input layers. Metrics are averaged across 10 image seeds and all layers for seven scenes.

5.2. Interactive Editing

We illustrate interactive editing with *Collage Diffusion* by repeatedly (1) generating 10 images using different random seeds, (2) allowing the user to select the image they like the most, and (3) selecting an object in this image that they would like to re-generate. This process continues until the user is satisfied with all aspects of the generated image.

Fig. 4 illustrates the value of *Collage Diffusion* for interactively authoring complex scenes. For the "Cake" scene, the user generates a final image in three steps: (1) generating an initial collection of images from the input layers, (2) exploring different options for the cake, and (3) exploring different options for the winter window. Similarly, for "Bento Box," the user generates a final image in three steps: (1) generating an initial collection of images from the input layers, (2) exploring different options for the sushi, and (3) exploring different options for the ginger. We successfully preserve all previously-generated objects while providing a diverse set of options for each modified object that match the layer specifications. This interactive refinement procedure is valuable for ensuring that the user is satisfied with all parts of the generated image.

5.3. Non-Interactive Generation

Collage Diffusion is a combination of several components: **GH**, **CA**, **TI**, and **LN**, as outlined in Sec. 5.1. Fig. 5 and 6 illustrate how all of these components contribute to our harmonization and fidelity goals. We did not cherry-pick the individual image seeds for each scene—additional examples from each test scene are included in the Supplemental, and reflect the same overall trends.

GH generates globally-harmonized images, while SA struggles with harmonization. Comparison of the **SA** and **GH** columns in Fig. 5 and 6 illustrates the capacity of **GH** to generate a harmonized image from input x_c while highlighting the downsides of manipulating self-attention to preserve image structure in **SA**. When image harmonization requires altering the orientations of objects in the scene—the sushi in "Bento Box," the cakes in "Cake," the apples in

"Ceramic Bowl," etc.—**SA** fails to harmonize the image due to the constraints placed on the self-attention maps. In contrast, **GH** reliably generates globally-harmonized images: the images have consistent perspective and lighting, with fewer artifacts. Note that **GH** still inherits the limitations of Stable Diffusion 2.1–the harmonization capacity is limited by the quality of the underlying diffusion prior.

CA consistently improves spatial fidelity across scenes. Comparison of the **GH** and **GH+CA** columns in Fig. 5 and 6 illustrates the benefits of layer-based cross-attention control. In "Bento Box," using **CA** results in ginger and rice in the appropriate locations in the generated output. **CA** also helps preserve the table legs in "Cake," maps the correct fruits to the correct parts of "Ceramic Bowl," etc. This trend is also reflected quantitatively: in Tab. 1, **GH+CA** has a higher average per-layer text-image similarity than **GH**, indicating better spatial fidelity.

TI consistently improves appearance fidelity across scenes. Having mapped the desired concepts to the desired locations, comparison of the **GH+CA** and **GH+CA+TI** columns in Fig. 5 and 6 illustrates the benefits of layer-based textual representations. **TI** helps generate a wood train with similar style to the starting image in "Toys," the right type of sushi ginger in "Bento Box," the proper legs for the table in "Cake," the correct color and shape for the potatoes in "Ceramic Bowl," the proper saturation of colors and presence of wrinkles in "Clothing," etc. This trend is also reflected quantitatively: in Tab. 1, **GH+CA+TI** has a higher average per-layer image-image similarity than **GH+CA**, indicating better appearance fidelity.

LN consistently helps optimize the harmonization-fidelity tradeoff across scenes Having mapped the desired concepts to the desired locations, with textual inversion to increase appearance fidelity, comparison of the **GH+C+TI** and **GH+CA+TI+LN** columns in Fig. 5 and 6 illustrates the benefits of control over per-layer noise. **LN** increases the preservation of the structure of the wood train in "Toys", the salmon on the sushi in "Bento Box", the books on the bookshelves in "Cake", the shape of the bananas in "Ceramic Bowl", the stripes of the sweater in "Striped Sweater", the corn and cucumber in "Veggie Face," etc. For all these scenes, the quality of image harmonization is maintained across **GH+C+TI** and **GH+CA+TI+LN**. This trend is also reflected quantitatively: in Tab. 1, **GH+CA+TI+LN** has higher average per-layer text-image and image-image similarity than **GH+CA+TI**, indicating better spatial and appearance fidelity.

Where is layer-driven harmonization most helpful? To understand the situations where layer information is most valuable, we highlight the "Red Skirt" (Fig. 6) and "Cake" (Fig. 5) scenes as examples at either end of the range of difficulty where layers are valuable. When harmonization

Figure 7. ControlNet lets users preserve image structures, rather than unique object identity, on a per-layer basis. First row: high ControlNet weights preserve edge maps for the ships, rocks, and lighthouse. Second row: high ControlNet weights preserve edge maps for the house and the backyard.

requires limited changes to image structure, **SA** can be suitable—while **SA** still produces artifacts on "Red Skirt", the approach is more effective than on other scenes because fewer changes in image structure are required to harmonize the image. When objects are easy to discriminate even after noise is added (large objects with distinct colors), **GH** performs well, and **GH+CA** provides negligible added value. If the visual attributes that the user cares to preserve in the layer are well-described by the layer prompt, **TI** may be unnecessary—in Fig. 6, the only added benefit in "Red Skirt" comes from the preservation of the folds on the skirt and the dark band around the waist.

On the other end of the spectrum, when the user is particular on the *exact* appearance of many complex layers, even *Collage Diffusion* may struggle to satisfy user intent across all objects in the scene. For instance, in "Cake," the user may want a specific color and icing pattern on the cake, a snowy pine outside the window, a full bookshelf, etc. For these situations, our iterative editing workflow is valuable, as highlighted in Sec. 5.2 and Fig. 4.

5.4. Flexible per-layer controls with ControlNet

One of the key benefits of per-layer control is that we can vary the definition of appearance fidelity on a per-layer basis. In Fig. 7, our ControlNet extension enables users to preserve image structures, rather than unique object identity, on a per-layer basis. In the first row, high ControlNet weights preserve the edge maps of the ships, rocks, and lighthouse (note that the colors/textures of the rocks and lighthouse vary). The generated images have more structural variation in the ocean and sky. In the second row, high ControlNet weights strictly preserve the structure of the house, while loosely preserving the layout of the backyard, and allowing variation in the pool shape and pattern of birds in the sky.

6. Conclusion

In this paper, we show the value of maintaining an explicit notion of scene layers for AI-based image generation. Per-layer editing provides users the ability to precisely control image output, and the additional information afforded by layer-based representations can be leveraged by the image generation process to more closely match user intent.

Acknowledgement: This work was supported by a gift from Meta.

References

[1] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. *arXiv preprint arXiv:2211.14305*, 2022. 3

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1, 3, 5

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1, 2, 3, 4

[4] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 707–723. Springer, 2022. 3

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 3, 6

[6] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 3

[7] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 3

[8] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8394–8403, 2020. 3

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2, 3, 4, 6

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3, 6

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[13] Yan Hong, Li Niu, and Jianfu Zhang. Shadow generation for composite image in real-world scenes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 914–922, 2022. 3

[14] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM Transactions on Graphics (TOG)*, 40(6):1–12, 2021. 3

[15] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 1, 3

[16] Ting-Hsuan Liao, Songwei Ge, Yiran Xu, Yao-Chih Lee, Badour AlBahar, and Jia-Bin Huang. Text-driven visual synthesis with latent diffusion prior, 2023. 3

[17] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. *arXiv preprint arXiv:2009.07833*, 2020. 3

[18] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 3

[19] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1, 3, 5

[20] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 3, 6

[21] Thomas Porter and Tom Duff. Compositing digital images. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pages 253–259, 1984. 3

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 4, 6

[25] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 1, 3

[26] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[28] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[29] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Generative object compositing. *arXiv preprint arXiv:2212.00932*, 2022. 1, 3

[30] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 3, 5, 6

[31] John YA Wang and Edward H Adelson. Representing moving images with layers. *IEEE transactions on image processing*, 3(5):625–638, 1994. 3

[32] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023. 3

[33] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 3

[34] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *European Conference on Computer Vision*, pages 300–316. Springer, 2022. 3

[35] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3

[36] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model. *arXiv preprint arXiv:2308.10040*, 2023. 3

[37] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2, 3, 4

[38] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 3