Collage Diffusion

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

Gravity: “a bento box with rice, edamame, ginger, and sushi”
Output Image
“bento box”
“ginger”
“rice”
“sushi”
“edamame”

Input Layers
Prompt: “a bento box with rice, edamame, ginger, and sushi”
Output Image

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
“plausible” real-world images. There is an inherent tradeoff 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.

Specifically 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 $l_1, l_2, \ldots, l_n$, 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 $x_\alpha^i$
   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 $x^n_*$ with the following properties:

1. **Global harmonization**: $x^n_*$ 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 $x^n_*$. 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 $x^n_*$ 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
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 im-
age 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 $x^*_t$, using a text-conditioned diffusion U-Net $D_\theta(x, \sigma(t), c)$ as an image prior [24] ($x$ is a noised 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 $D_\theta$. 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 $j = \max \{ \{ k | (x_{ab}^k) > 0 \} \}$, 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 $D_\theta$ is computed as $\text{softmax}(\frac{QK^T}{\sqrt{d}}) V$, 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 $A^{pos}, A^{neg} \in \mathbb{R}^{N_c \times N_t}$, where $N_c$ is the number of image tokens and $N_t$ is the number of text tokens, and each column $A^{pos}_j, A^{neg}_j$ is a flattened alpha mask dependent on visibility of text token $j$. $A^{pos}_{ij} = 0$ for all global tokens $j$, $A^{pos}_{ij} = 1$ if image token $i$ corresponds to a region of the image that layer token $j$ should influence, and $A^{neg}_{ij} = 1$ 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: $\text{softmax}(\frac{QK^T+w^{pos}A^{pos}+w^{neg}A^{neg}}{\sqrt{d}}) V$. 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: $w^{pos} = q^{pos} \cdot y(t)$ and $w^{neg} = q^{neg} \cdot y(t)$, where $y(t) = \log(1+\log(1+\sigma(t))) \cdot \max(QK^T)$, and $q^{pos}$ and $q^{neg}$ are scalars specified by the user. Denote this modified diffusion model as $D_\theta^*$.

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 “(a_i) ginger”. The embedding for $a_i$ is learned by optimizing the following loss:

$$a^*_i = \arg \min_{a_i} E_{c \sim N(0, \sigma)} \left[ x^*_i \cdot (x_{\text{target}_i} - D_\theta(x_{\text{target}_i} + \epsilon, c_i(a_i, c_i))) \right]$$

(1)

target image $x_{\text{target}_i}$ is constructed by alpha-compositing the first $i$ layers $l_1 \ldots l_i$, and layer alpha mask $x^*_i$ restricts the loss to the relevant region of $x_{\text{target}_i}$. 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 $x_{\text{target}_i}$ 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 $m_k \in \mathbb{R}^{h,w,c}$ 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

$$w_{all} = \left\{ \begin{array}{ll}
1 & \text{if } t = t_{\text{image}} \\
\left( \frac{t - t_{\text{image}}}{t_{\text{image}} - t_{\text{image}}} \right) & \text{otherwise}
\end{array} \right.$$

where $t_{\text{image}}$ is the time step at which the image is generated.
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 trade-off 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 \( j = \max_k \{ k | (x_k^a)_{ab} > 0 \} \) is the layer index of the highest of the \( n \) layers with nonzero alpha for pixel coordinate \((a, b)\), and \( w_{layer,b} \) is the value of \( w_{layer} \) at pixel \((a, b)\). We resize \( w_{layer} \) to \([0, 1]^b\) using bilinear interpolation, then element-wise-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 \( w_i \).

4.6. 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 difficulties 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 images are present. We also evaluate the performance of the following methods for a range of scenes:

1. SA: Image generation with Self-Attention 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: Global Harmonization by applying SDEdit [19] (Sec. 4.1) to composite image \( x_c \). This is another base-
line that does not leverage layer information.
3. **GH+CA**: GH with modified Cross-Attention (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 Textual Inversion [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-Layer Noise 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 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.

---

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 base model as $D_b$ 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].
**Veggie Face**

“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”

**Striped Sweater**

“a man wearing green pants, a blue and green striped sweater, a plaid scarf, and a maroon beanie”

**Ceramic Bowl**

“a blue ceramic bowl with red potatoes, red apples, and red bananas”

**Red Skirt**

“a person wearing a patterned red skirt, buttoned blue blouse, and pink summer coat, in front of a gray background”

---

**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.

<table>
<thead>
<tr>
<th></th>
<th>GH</th>
<th>GH+CA</th>
<th>GH+CA</th>
<th>GH+CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>+TI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+TI+LN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

↑Txt-Img. Sim. | 0.215 | 0.236 | 0.233 | 0.238 |
↑Img-Img. Sim. | 0.846 | 0.867 | 0.877 | 0.893 |

---

**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 $x_v^i$ 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 $x_v^i$. Appearance fidelity aims for layer image $x_i$ to match the visual content in $x_v^i$ where layer $i$ is visible—we measure this by computing CLIP image-image similarity between $x_i$ and the corresponding region of $x_v^i$. We include additional details on metrics in the Supplemental.

---

**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 $x_v$ 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, 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 “Vegetable 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 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 edge maps for the ships, rocks, and lighthouse. Second row: high ControlNet weights preserve edge maps for the house and the backyard.

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


[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


