Composite Diffusion: whole \( \geq \sum \) parts

Vikram Jamwal  
TCS Research, India  
vikram.jamwal@tcs.com

Ramaneswaran S*  
NVIDIA, India  
s.ramaneswaran2000@gmail.com

Abstract

For artists or graphic designers, the spatial arrangement of a scene is a critical design choice. However, existing text-to-image diffusion models provide limited support for incorporating spatial information. This paper introduces Composite Diffusion as a means for artists to generate high-quality images by composing from sub-scenes. The artists can specify the arrangement of the sub-scenes through a free-form segment layout and can describe the content of each sub-scene using natural text and additional control inputs. We provide a comprehensive and modular framework for Composite Diffusion that enables alternative ways of generating, composing, and harmonizing sub-scenes. We further argue that existing image quality metrics lack a holistic evaluation of image composites. To address this, we propose novel quality criteria especially relevant to composite generation. We believe that our approach provides an intuitive method of art creation. Through extensive user surveys and quantitative and qualitative analysis, we show how it achieves greater spatial, semantic, and creative control over image generation. In addition, our methods do not need to retrain or modify the architecture of the base diffusion models and can work in a plug-and-play manner with the fine-tuned models.

1. Introduction

Recent advances in diffusion models [9], such as Dalle-2 [18], Imagen [21], and Stable Diffusion [19] have enabled artists to generate vivid imagery by describing their envisioned scenes with natural language phrases. However, it is cumbersome and occasionally even impossible to specify spatial information or sub-scenes within an image solely by text descriptions. Consequently, artists have limited or no direct control over the layout, placement, orientation, and properties of the individual objects within a scene. These creative controls are indispensable for artists seeking to express their creativity [24] and are crucial in various content creation domains, including illustration generation, graphic design, and advertisement production. Frameworks like Controlnets [27] offer exciting new capabilities by training parallel conditioning networks within diffusion models to support numerous control conditions. Nevertheless, as
we show in this paper, creating a complex scene solely based on control conditions can still be challenging. As a result, achieving the desired imagery may require several hours of labor or maybe only be partially attainable through pure text-driven or control-condition-driven techniques.

To overcome these challenges, we propose Composite-Diffusion as a method for creating composite images by combining spatially distributed segments or sub-scenes. These segments are generated and harmonized through independent diffusion processes to produce a final composite image. The artistic intent in Composite Diffusion is conveyed through the following two means:

(i) Spatial Intent: Artists can flexibly arrange sub-scenes using a free-form spatial layout. A unique color identifies each sub-scene.

(ii) Content intent: Artists can specify the desired content within each sub-scene through text descriptions. They can augment this information by using examples images and other control methods such as scribbles, line drawings, pose indicators, etc.

We believe, and our initial experience has shown, that this approach offers a powerful and intuitive method for visual artists to stipulate their artwork.

This paper seeks to answer two primary research questions: First, how can native diffusion models facilitate composite creation using the diverse input modalities we described above? Second, how do we assess the quality of images produced using Composite Diffusion methods? Our paper contributes in the following novel ways:

1. We present a comprehensive, modular, and flexible method for creating composite images, where the individual segments (or sub-scenes) can be influenced not only by textual descriptions, but also by various control modalities such as line art, scribbles, human pose, canny images, and reference images. The method also enables the simultaneous use of different control conditions for different segments.

2. Recognizing the inadequacy of existing image quality metrics such as FID (Frechet Inception Distance) and Inception Scores [11, 23] for evaluating the quality of composite images, we introduce a new set of quality criteria. While principally relying on human evaluations for quality assessments, we also develop new methods of automated evaluations suitable for these quality criteria.

We rigorously evaluate our methods using various techniques including quantitative user evaluations, automated assessments, artist consultations, and qualitative visual comparisons with alternative approaches.

2. Related work

In this section, we discuss the approaches that are related to our work from multiple perspectives.

The work that comes closest to our approach in diffusion models is inpainting where the model edits a portion of an image specified by a segment mask (and an optional textual description). Almost all the popular diffusion models such as Dalle-2 [18], Imagen [21], and Stable Diffusion [19], and frameworks like Controlnets [27] support some form of inpainting such as repaint [15], blended-diffusion [3], latent-blended diffusion [1], and RunwayML [19]. As we show in this paper, one can conceive of an approach for Composite Diffusion by repeatedly applying inpainting. However, this approach has some drawbacks such as it requires a suitable background image (refer to Supp Sec. 4.2).

Some works look at the composition or editing of images through a different lens. These include prompt-to-prompt editing [8, 10, 16], composing scenes through composable prompts [13, 14], and methods for personalization of subjects in a generative model [20]. We concentrate specifically on composing the spatial segments specified via a spatial layout. So our methods can be complementary to these techniques.

Some related concurrent works such as SpaText [2], eDiff-I [4], and Multi-diffusion [5] provide some methods for composing images from spatially free-form layouts with natural text descriptions. SpaText [2] achieves spatial control by training the model to be space-sensitive by additional CLIP-based spatial-textual representation. eDiff-I [4] proposes a method called paint-with-words which exploits the cross-attention mechanism of U-Net in the diffusion model to specify the spatial positioning of objects. Multi-diffusion [5] proposes a mechanism for controlling the image generation in a region by providing the abstraction of an optimization loss between an ideal output by a single diffusion generator and multiple diffusion processes that generate different parts of an image. We utilize a pre-trained text-conditioned diffusion model or a control-conditioned model without the need to retrain them (unlike SpaText [2]), and without the need for architectural modification (unlike eDiff-I [4]). Refer to Supp Sec. 6 for a more detailed discussion.

In comparison to all the above approaches, we have a fundamental difference in the artistic approach as we lay emphasis on composing through sub-scenes. Unlike all these approaches we achieve additional control over the orientation and placement of objects within a segment through reference images and control conditions specific to the segment. Further, our approach is more generic, has a wider scope, and provides alternative ways of composition in its two-stage composition process.
3. Our Composite Diffusion method

We present our method for Composite Diffusion. We first formally define our goal. We define a subscene as a scene making up part of a larger scene. We will use the term ‘segment’ to particularly denote a sub-scene.

We want to generate an image \( x \) which is composed entirely based on two types of input specifications:

1. **Segment Layout**: a set of free-form segments \( S = [s^1, s^2, \ldots, s^n] \), and

2. **Segment Content**: a set of natural text descriptions, \( D = [d^1, d^2, \ldots, d^n] \), and optional additional reference images, \( R = [r^1, r^2, \ldots, r^n] \), or control conditions, \( C = [c^1, c^2, \ldots, c^n] \).

Each segment \( s^j \) in \( S \) describes the spatial form of a sub-scene and has a corresponding natural text description \( d^j \) in \( D \), and optionally a corresponding reference image \( r^j \) in \( C \), or a control condition \( c^j \) in \( C \). The segments don’t non-overlap and fully partition the image space of \( x \). Additionally, we convert the segment layout to segment-specific masks, \( M = [m^1, m^2, \ldots, m^n] \), as one-hot encoding vectors.

Our method divides the generative process of a diffusion model into two successive temporal stages: (a) the Scaffolding stage and (b) the Harmonization stage. We define a parameter called the scaffolding factor, denoted by \( \kappa \) (kappa), whose value determines the percentage of the diffusion process that we assign to the scaffolding stage. \( \kappa = \frac{\text{number of scaffolding steps}}{\text{total diffusion steps}} \times 100 \).

The number of harmonization steps is calculated as total diffusion steps minus the scaffolding steps. We explain the two generative stages below:

### 3.1. Scaffolding stage

We introduce the concept of *scaffolding*, which we define as a mechanism for guiding image generation within a segment with some external help. We borrow the term ‘scaffolding’ from the construction industry [26], where it refers to the temporary structures that facilitate the construction of the main building or structure. These scaffolding structures are removed in the building construction once the construction work is complete or has reached a stage where it does not require external help. Similarly, we may drop the scaffolding help after completing the scaffolding stage.

| 1. woman with a scarf |
| 2. man with a hat |
| 3. boy with baseball cap |
| 4. girl with colorful hairband |

The scaffolding factor (\( \kappa \)) (Sec. 3.3) controls the influence of reference images on the final composite image. At low \( \kappa \) values, the reference images are heavily noised and exercise little control; the segments merge drastically. At high \( \kappa \) values, the reference images are lightly noised and the resulting image is nearer to the reference images. A middle \( \kappa \) value balances the influences of reference images and textual descriptions.

Figure 2. Use of reference images for scaffolding: The scaffolding factor (\( \kappa \)) controls the influence of reference images on the final composite image. At low \( \kappa \) values, the reference images are heavily noised and exercise little control; the segments merge drastically. At high \( \kappa \) values, the reference images are lightly noised and the resulting image is nearer to the reference images. A middle \( \kappa \) value balances the influences of reference images and textual descriptions.

**Algorithm 1:** Composite Diffusion: Scaffolding Stage. The input is as defined in the Sec. 3.

```plaintext
if Segment Reference Images then
    for all segments i from 1 to n do
        x_{i,scaff} \leftarrow \text{Noise}(r^i, \kappa); \quad \triangleright Q-sample reference images to last timestep of scaffolding stage.
    end
else if Only Segment Text Descriptions then
    for all t from T to \kappa do
        for all segments i from 1 to n do
            x_t \leftarrow \text{Noise}(x_{i,scaff}, t); \quad \triangleright Q-sample scaffold.
            x_{i,seq} \leftarrow \text{Denoise}(x_t, x_{i,scaff}, m^i, d^i); \quad \triangleright Step-inpaint with the scaffolding image.
        end
    end
else if Text and Segment Control Conditions then
    for all t from T to \kappa do
        for all segments i from 1 to n do
            x_{i,seq} \leftarrow \text{Denoise}(x_t, x_{i,scaff}, m^i, d^i, c^i); \quad \triangleright Scaffolding with the control condition and denoise.
        end
    end
end
x_{comp}^{kappa} \leftarrow \sum_{i=1}^{n} x_{i,seq}^{kappa} \oplus m^i; \quad \triangleright Merge segments.
return x_{comp}^{kappa-1} \oplus m^i; \quad \triangleright Merge segments.
```
The external structural help, in our case, can be provided by any means that help generate or anchor the appropriate image within a segment. We provide this help through either (i) scaffolding reference image - in the case where reference example images are provided for the segments, (ii) a scaffolding image - in the case where only text descriptions are available as conditioning information for the segments, or (iii) a scaffolding control condition - in the case where the base generative model supports conditioning controls and additional control inputs are available for the segments.

3.1.1 Scaffolding with reference image
An individual segment may be provided with an example image called scaffolding reference image to gain specific control over the segment generation. This conditioning is akin to using image-to-image translation [19] to guide the production of images in a particular segment. Algorithmically, we directly noise the reference image (refer to Q-sampling in Supp sec 2) to the time-stamp \( t = \kappa \) that depicts the last time-step of the scaffolding stage in the generative diffusion process (Algo. 1, 1-4, and Supp Fig. 5, A). The generated segment can be made more or less in the likeness of the reference image by varying the initializing noising levels of the reference images. Refer to Fig. 2 for an example of scaffolding using segment-specific reference images.

3.1.2 Scaffolding with scaffolding image
This case is applicable when we have only text descriptions for each segment. The essence of this method is the use of a predefined image called scaffolding image \( x_s\text{c}a\text{f} \), to help with the segment generation process. Refer to Algo. 1, 5-11 and Supp Fig. 5, B. Algorithmically, to generate one segment at any timestep \( t \): (i) we apply the segment mask \( m \) to the noisy image latent \( x_t \) to isolate the area \( x_t \odot m \) where we want generation, (ii) we apply a complementary mask \( (1 - m) \) to an appropriately noised (q-sampled to timestep \( t \)) version of scaffold image \( x_t^{s\text{c}a\text{f}} \) to isolate a complementary area \( x_t^{s\text{c}a\text{f}} \odot (1 - m) \), and (iii) we merge these two complementary isolated areas and denoise the composite directly through the denoiser along with the corresponding textual description for the segment. Refer to Supp Fig. 6(a) for an illustration of the single-step generation. We then replicate this process for all the segments.

These steps are akin to performing an inpainting [1] step on each segment but in the context of a scaffolding image. Please note that our method step (Algo. 1, 9) is generic and flexible to allow the use of any inpainting method, including the use of a specially trained model (e.g., RunwayML Stable Diffusion inpainting 1.5 [19]) that can directly generate inpainted segments. We repeat this generative process for successive time steps till the time step \( t = \kappa \). The choice of scaffolding image can be arbitrary. Although convenient, we do not restrict keeping the same scaffolding image for every segment.

3.1.3 Scaffolding with control image
This case is applicable where the base generative model supports conditioning controls, and, besides the text-conditioning, additional control inputs are available for the segment. In this method, we do away with the need for a scaffolding image. Instead of a scaffolding image, an artist provides a scaffolding control input for the segment. The control conditioning input can be a line art, an open pose model, a scribble, a canny image, or any other supported control input that can guide image generation in a generative diffusion process.

Algorithmically, we proceed as follows: (i) We use
a control input specifically tailored to the segment’s dimensions, or we apply the segment mask $m$ to the control condition input $c^i$ to restrict the control condition only to the segment where we want generation, (ii) The image latent $x_t$ is directly denoised through a suitable control-denoiser along with conditioning inputs of natural text and control inputs for the particular segment. We then repeat the process for all segments and for all the timesteps till $t = \kappa$. Refer to Algo.1, 12-17, and Supp Fig. 5, C.

Note that since each segment is denoised independently, the algorithm supports the use of different specialized denoisers for different segments. For example, refer to Fig. 1 where we use three distinct control inputs, viz., scribble, lineart, and openpose. Combining control conditions into Composite Diffusion enables capabilities more powerful than both - the text-to-image diffusion models [19] and the control-conditioned models [27]. Fig. 4 refers to two example cases where we accomplish image generation tasks that are not feasible through either of these two models.

At the end of the scaffolding stage, we construct an intermediate composite image by composing from the segment-specific latents. For each segment-specific latent, we retain the region corresponding to the segment masks and discard the complementary region (Refer to Supp Fig. 5 and Algo. 1, 20-21). The essence of the scaffolding stage is that each segment develops independently and has no influence on the development of the other segments. We next proceed to the ‘harmonization’ stage, where the intermediate composite serves as the starting point for further diffusion steps.

3.2. Harmonizing stage

The above method, if applied to all diffusion steps, can produce good multi-segment inpainted images. However, because the segments are being constructed independently, the composite tends to be less harmonized and less well-blended at the segment edges. To alleviate this problem, we introduce a new succeeding stage called the ‘harmonization stage’. The essential difference from the preceding scaffolding stage is that in this stage each segment develops in the context of the other segments. We also drop any help through scaffolding images in this stage.

We can further develop the intermediate composite from the previous stage in the following ways: (i) by direct denoising the composite image latent via a global prompt (Algo. 2, 2-3, and Supp Fig. 7, A), or (ii) by denoising the intermediate composite latent separately with each segment specific conditioning and
Algorithm 2: Composite Diffusion: Harmonization
Stage. Input same as Algo. 1, plus $x_{comp}^{\kappa=1}$

1. for all $t$ from $\kappa-1$ to 0 do
2. if Global Text Conditioning then
3. $x_{t-1} \leftarrow$ Denoise($x_t, D$) ; $\triangleright$ Base Denoiser
4. else if Segment Text Conditioning then
5. for all segments $i$ from 1 to $n$ do
6. $\hat{x}_{t-1}^{\kappa_0}$ \leftarrow Denoise($x_t, d^i, c^i$) ; $\triangleright$ Controlled Denoiser
7. end
8. $x_{t-1}^{comp} \leftarrow \sum_{i=1}^{n} x_{t-1}^{\kappa_0} \odot m^i$ ; $\triangleright$ Merge segments
9. else if Segment Control+Text Conditioning then
10. for all segments $i$ from 1 to $n$ do
11. $\hat{x}_{t-1}^{\kappa_0}$ \leftarrow Denoise($x_t, d^i, c^i$) ; $\triangleright$ Controlled Denoiser
12. end
13. $x_{t-1}^{comp} \leftarrow \sum_{i=1}^{n} x_{t-1}^{\kappa_0} \odot m^i$ ; $\triangleright$ Merge segments
14. end
15. return $x_{comp} \leftarrow (x_{comp}^{\kappa=1})$ ; $\triangleright$ Final Composite

then composing the denoised segment-specific latents. The segment-specific conditions can be either pure natural text descriptions or may include additional control conditions (Refer to Algo. 2, 4-8 and 9-13, and Supp Fig. 7, B and C). While using global prompts, the output of each diffusion step is a single latent and we do not need any compositional step. For harmonization using segment-specific conditions, the compositional step of merging different segment latents at every time step (Algo. 2, 8 and 13) ensures that the context of all the segments is available for the next diffusion step. This leads to better blending and harmony among segments after each denoising iteration. Our observation is that both these methods lead to a natural coherence and convergence among the segments of the composite image (Supp Fig. 8 provides an example illustration).

3.3. Impact of Scaffolding factor $\kappa$:

Increasing the $\kappa$ value allows the segments to develop independently longer. This gives better conformance with the segment boundaries while reducing the blending and harmony of the composite image. Our experience has shown that the appropriate value of $\kappa$ depends upon the domain and the creative needs of an artist. Typically, we find that values of kappa around 20-50 are sufficient to anchor an image in the segments. Figure 3 illustrates the impact of $\kappa$ on image generation that gives artists an interesting creative control on segment blending. Supp Tab. 4 provides a quantitative evaluation of the impact of the scaffolding factor on the various parameters of image quality.

4. Quality criteria and evaluation

As stated earlier, one of the objectives of this research is to ask the question: Is the quality of the composite greater than or equal to the sum total of the quality of the individual segments? In this section, we first lay out our quality criteria and the evaluation approach, and then discuss the results of our implementations.

4.1. Quality criteria

We argue that the present methods of evaluating the image quality of image generation models are not sufficient for our purposes. For example, methods such as FID, Inception Score, Precision, and Recall [6,11,22,23] that are traditionally used for measuring the quality and diversity of generated images, do so only with respect to the set of reference images used in training. Further, they do not evaluate some key properties of concern to us such as conformity of the generated images to the provided inputs, the harmonization achieved when forming images from sub-scenes, and the overall aesthetic and technical quality of the generated images. These properties are key to holistically evaluating the Composite Diffusion approach. Hence, we propose the following set of quality criteria:

1. **CF: Content Fidelity**: The purpose of the text prompts is to provide a natural language description of what needs to be generated in a particular region of the image. The purpose of the control conditions is to specify objects or visual elements within a sub-scene. This parameter measures how well the generated image represents the textual prompts (and/or the control conditions) used to describe the sub-scene.

2. **SF: Spatial Layout Fidelity**: The purpose of the spatial layout is to provide spatial location guidance to various elements of the image. This parameter measures how well the parts of the generated image conform to the boundaries of specified segments or sub-scenes.

3. **BH: Blending and Harmony**: When we compose an image out of its parts, it is important that the different regions blend together well and we do not get abrupt transitions between any two regions. Also, it is important that the image as a whole appears harmonious, i.e., the contents, textures, colors, etc. of different regions form a unified whole. This parameter measures the smoothness of the transitions between the boundaries of the segments, and the harmony among different segments of the image.

4. **QT: Technical Quality**: The presence of noise and unwanted artifacts that can appear in the image generations can be distracting and may reduce the visual quality of the generated image. This parameter measures how clean the image is from the unwanted noise, color degradation, and other unpleasant artifacts.
4.2. Evaluation baselines

We measure the performance of our Composite Diffusion approaches, using text-only conditioning (Ours:CD-T) and text+control conditioning (Ours:CD-TC). To measure the performance of our approaches using the above quality criteria, we deploy the following three baselines:

1. Baseline B1: Base models This baseline has two forms: (i) B1-T: [19] This is the Text to Image base diffusion model that takes only text prompts as the input. Since this input is unimodal, the spatial information is provided solely through natural language descriptions. (ii)B1-TC: [27] This is the Text+Control to Image base diffusion model that takes additional control condition inputs besides the text prompts.

2. Baseline B2: Serial inpainting As indicated in the Sec. 2, we should be able to achieve a composite generation by serially applying inpainting to an appropriate background image and generating one segment at a time. We use two serial inpainting methods: one based on blended latent diffusion (B2:BLD) [1], and another based on a specially trained inpainting method for stable diffusion (B2:RSD) [19].

3. Baseline B3: Related approaches For this, we consider two publicly available implementations of the related methods - ediff-I paint-by-word (B3:ediff-I) [4] and Multidiffusion (B3:MD) [5].
Table 1. Automated evaluation results. The best-performing algorithm in a category is marked in bold.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Fidelity</td>
<td>0.23</td>
<td>0.24</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>Spatial Fidelity</td>
<td>0.24</td>
<td>0.26</td>
<td>0.26</td>
<td>0.27</td>
<td>0.26</td>
<td>0.27</td>
<td>0.27</td>
<td>0.28</td>
</tr>
<tr>
<td>Technical Quality</td>
<td>1.34</td>
<td>1.89</td>
<td>2.69</td>
<td>1.15</td>
<td>0.49</td>
<td>1.02</td>
<td>1.24</td>
<td>2.26</td>
</tr>
<tr>
<td>Aesthetic Quality</td>
<td>6.3</td>
<td>6.5</td>
<td>5.5</td>
<td>5.6</td>
<td>5.0</td>
<td>6.4</td>
<td>6.4</td>
<td>6.6</td>
</tr>
<tr>
<td>Blend&amp;Harmony</td>
<td>6903</td>
<td>2999</td>
<td>725</td>
<td>1112</td>
<td>2696</td>
<td>5239</td>
<td>7404</td>
<td>5302</td>
</tr>
<tr>
<td>Human Preference</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>CreateTime(s)/Art</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>7</td>
<td>10</td>
<td>13</td>
<td>13</td>
<td>19</td>
</tr>
</tbody>
</table>

Figure 6. Human evaluation results from the set - General Population(GP) for Composite Diffusion with Text and Segment Layout input. Refer to Supp Fig. 22 for additional human evaluation results from a more specific set of population, viz., Artists and Designers(AD).

implementation details for our algorithms and baselines are available in Supp Secs. 3, 4, and 5.

4.4. Results Summary

In this section, we highlight the main trends we observed in our evaluation. For a detailed analysis and further insights, please refer to Supp Sec. 9. We present the automated evaluation results in Tab. 1.

For content and spatial fidelity, Ours:CD-TC, followed by Ours:CD-T, gets the highest score. This result reinforces our claim that our approach allows for better textual and spatial conformity. In particular, baseline B1:TC gets a lower spatial fidelity score compared to Ours:CD-TC and Ours:CD-T, indicating that using conditional control through, for example, ControlNet [27] alone may not be sufficient to get spatial control. Ours:CD-T gets the highest aesthetic score, and Ours:CD-T follows closely along with the baseline B3:MD. While B1:TC secures the second-highest aesthetic score, it compromises the critical parameters of spatial and textual conformity. B3:ediff-I exhibits minimal noise, although it sacrifices aesthetics and input conformity. Ours:CD-T has comparable noise levels to baseline B1:T, while Ours:CD-TC has slightly higher noise than the baseline B1:TC: both of our approaches result in much higher input conformity and aesthetics.

B2:RSD gets the best blending and harmonization score. We can attribute this result to B2:RSD generating composite images directly in a single diffusion process. However, in our case, we could tune the generation towards blending by using the explicit harmonization stage to generate holistic images. B2:BLD, although it shows better blending scores, performs the worst in a visual examination, which might indicate the shortcomings of our automated method for blending and harmonization. Our approach’s run-time efficiency is similar to the related methods but is higher than the baselines B1 and B2. The human preference ranking, which gives an overall measure of human preference, text alignment, and image fidelity, Ours:CD-T ranks the highest, followed by B3:MD and Ours:CD-T.

In Fig. 6, Supp Fig. 22, and Supp Tabs. 2 and 3, we share the outcomes of human evaluation surveys regarding these quality parameters. The overall patterns in these surveys mostly align with those seen in automated results, thus also providing validation for our automated evaluation methods.

5. Conclusion

From the artists’ affordance perspective, we proposed a sub-scene-based composition of a generated image as it provides an intuitive and easy method for art creation. The finer level of control is best served by segment-specific control conditions. We showed that dividing the composite generation process into two stages - scaffolding and harmonizing modularizes the development of algorithms for composite diffusion. The researchers can, in the future, independently improve the respective stages. We also find that diffusion processes are inherently harmonizers and it is best to exploit the inherent composition of diffusion models rather than blending through external means as tried in the past with image blending methods.
References

[9] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 1
[24] Viktoria Solidarnykh. This artist combines real photos and turns them into amazing digital art. DIY Photography, 2023. 1
[25] Jiasheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation, 2023. 7
[26] Zhe Yin and Carlos Caldas. Scaffolding in industrial construction projects: current practices,

[27] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. 
https://github.com/lllyasviel/ControlNet-v1-1-nightly, 2023. 1, 2, 5, 7, 8