Dense Text-to-Image Generation with Attention Modulation

Yunji Kim\textsuperscript{1}  
Jiyoung Lee\textsuperscript{1}  
Jin-Hwa Kim\textsuperscript{1}  
Jung-Woo Ha\textsuperscript{1}  
Jun-Yan Zhu\textsuperscript{2}

\textsuperscript{1}NAVER AI Lab  
\textsuperscript{2}Carnegie Mellon University

Abstract

Existing text-to-image diffusion models struggle to synthesize realistic images given dense captions, where each text prompt provides a detailed description for a specific image region. To address this, we propose DenseDiffusion, a training-free method that adapts a pre-trained text-to-image model to handle such dense captions while offering control over the scene layout. We first analyze the relationship between generated images’ layouts and the pre-trained model’s intermediate attention maps. Next, we develop an attention modulation method that guides objects to appear in specific regions according to layout guidance. Without requiring additional fine-tuning or datasets, we improve image generation performance given dense captions regarding both automatic and human evaluation scores. In addition, we achieve similar-quality visual results with models specifically trained with layout conditions. Code and data are available at https://github.com/naver-ai/DenseDiffusion.
1. Introduction

Recently developed text-to-image models can synthesize diverse, high-quality images from short scene descriptions [24, 15, 38, 44, 46, 45, 29, 58, 7]. However, as shown in Figure 1, they often encounter challenges when handling dense captions and tend to omit or blend the visual features of different objects. Here the term “dense” is inspired by the work of dense captioning [27] in which each phrase is used to describe a specific image region. Moreover, it is difficult for users to precisely control the scene layout of a generated image with text prompts alone.

Several recent works propose providing users with spatial control by training or fine-tuning layout-conditioned text-to-image models [2, 54, 59, 33, 45, 19, 56, 37]. While Make-a-Scene [19] and Latent Diffusion Models [45] train models from scratch given both text and layout conditions, other recent and concurrent works, such as SpaText [2] and ControlNet [59], introduce additional spatial controls to existing text-to-image models with model fine-tuning. Unfortunately, training or fine-tuning a model can be computationally expensive. Moreover, the model needs to be retrained for every new user condition, every new domain, and every new text-to-image base model.

To address the above issue, we propose a training-free method that supports dense captions and offers layout control. We first analyze the intermediate features of a pre-trained text-to-image diffusion model to show that the layout of a generated image is significantly related to self-attention and cross-attention maps. Based on this observation, we modulate intermediate attention maps according to the layout condition on the fly. We further propose considering the value range of original attention scores and adjusting the degree of modulation according to the area of each segment.

Our experiments show that our method improves the performance of Stable Diffusion [45] and outperforms several compositional diffusion models [35, 6, 17] on dense captions regarding both text conditions, layout conditions, and image quality. We verify them using both automatic metrics and user studies. In addition, our method is on par with existing layout-conditioned models regarding qualitative results. Finally, we include a detailed ablation study regarding our design choices and discuss several failure cases. Our code, models, and data are available at https://github.com/naver-ai/DenseDiffusion.

2. Related work

Text-to-Image Diffusion Models. The development of large-scale diffusion models [38, 44, 46, 45] has enabled us to synthesize diverse photorealistic images with flexible editing capacity [36, 12, 31, 1, 30]. In particular, recent works [21, 52, 14, 40] edit images by analyzing the interaction between intermediate image features and input textual features during the generation process. Prompt-to-Prompt [21] extracts cross-attention maps for each text token and injects them when denoising with an edited caption to preserve the input structure. Plug-and-Play [52] takes a similar approach but reuses self-attention and spatial features. pix2pix-zero [40] proposes cross-attention guidance to encourage the cross-attention maps of original and edited images to stay close. Inspired by the above works, we seek to leverage the close connection between the text features and the image layout inside a pre-trained diffusion model but focus on image synthesis with spatial controls rather than image editing.

Compositional Diffusion Methods. Since text-to-image models are typically trained on datasets with short text captions, they often fail to reflect all the details of dense captions composed of several phrases. A line of work tries to address the problem with a pre-trained model in a training-free manner. For example, Composable Diffusion [35] and MultiDiffusion [6] perform a separate denoising process for each phrase at every timestep. Attend-and-Excite [8] optimizes the noise map to increase the activation of the neglected tokens in cross-attention maps. Structure Diffusion [17] indirectly modulates cross-attention scores by enriching textual features for parsed texts.

Meanwhile, similar to the work of Dong et al. [16] that directly increases attention scores of certain words to increase the likelihood of them being generated for language modeling tasks, Paint-with-words (eDiff-Pww) [4] increases the cross-attention scores between image and text tokens that describe the same object given segmentation masks. However, they sometimes fail to reflect text and layout conditions faithfully. In this work, we propose a more extensive modulation method by devising multiple regularisation terms.

Image Synthesis with Spatial Control. Seminal works [11, 28] have proposed data-driven systems for synthesizing images given text descriptions and spatial controls based on image matching and blending. Later, neural networks such as conditional GANs have been adopted to directly synthesize images [26, 10, 57, 39, 51, 60, 25] but only work on limited domains such as natural landscapes or streets. More recently, several works propose to add spatial modulations to a pre-trained text-to-image diffusion model [45, 2, 54, 56, 59, 33]. For example, SpaText [2] fine-tunes a pre-trained model using a paired dataset of images and segmentation maps. ControlNet [59] and GLIGEN [33] develop more effective fine-tuning methods with adapters. Meanwhile, Universal-Guided-Diffusion [5] and Attention Refocusing [41] encourage intermediate outputs to align with the layout condition using guidance functions. Although these models offer control over image layout, they often require retraining each time for a new
We formally define our condition as a set of conditional information $c$, such as a text caption or semantic segmentation map, the condition can be provided via cross-attention layers. In this work, we experiment with Stable Diffusion [45] and use 50 DDIM denoising steps.

The objects of interest are outlined with blue and yellow bounding boxes. In (b), we show the attention maps for token keys marked in the red boxes in self-attention layers. As the timestep $t$ approaches zero, tokens affiliated with the same object communicate more closely, influencing the image layout.

Several concurrent works also propose training-free spatial control for diffusion models [20, 42, 9]. We encourage the readers to check out their works for more details.

### 3. Method

Our goal is to improve the text-to-image model’s ability to reflect textual and spatial conditions without fine-tuning. We formally define our condition as a set of $N$ segments $\{(c_n, m_n)\}_{n=1}^{N}$, where each segment $(c_n, m_n)$ describes a single region, as shown in Figure 3. Here $c_n$ is a non-overlapping part of the full-text caption $c$, and $m_n$ denotes a binary map representing each region. Given the input conditions, we modulate attention maps of all attention layers on the fly so that the object described by $c_n$ can be generated in the corresponding region $m_n$. To maintain the pre-trained model’s generation capacity, we design the modulation to consider original value range and each segment’s area.

#### 3.1. Preliminaries

**Diffusion Models** learn to generate an image by taking successive denoising steps starting from a random noise map $z_0$ [24, 50, 48]. A predefined number of denoising steps determines the degree of noise at each step, and a timestep-dependent noise prediction network $\theta_t$ is trained to predict the noise $\hat{e}_t$ added to a given input $z_t$. Although the earlier models, such as DDPM [24], are computationally-expensive, the non-Markovian diffusion method, DDIM [49], has improved the inference speed by drastically reducing the number of denoising steps. More recently, latent diffusion models enable even faster generation with improved image quality [45]. If the task is to generate an image using conditional information $c$, such as a text caption or semantic segmentation map, the condition can be provided via cross-attention layers. In this work, we experiment with Stable Diffusion [45] and use 50 DDIM denoising steps.

### Attention Score Analysis

Previous studies show that text-conditioned diffusion models tend to form the position and appearance of each object in early stages and refine details such as color or texture variations in relatively later stages [13, 4, 18, 21]. We show a similar trend by analyzing $16 \times 16$ attention maps produced from Stable Diffusion [45]. We first synthesize images using 250 captions from the MS-COCO validation set [34]. As shown in Figure 2, attention maps tend to resemble the image layout as the generation proceeds. Table 1 shows a quantitative analysis of attention scores, revealing distinct differences based on whether query-key pairs belong to the same object for both cross-attention and self-attention layers. This result suggests that query-key pairs belonging to the same object tend to have larger scores during the generation process.

#### 3.2. Layout-guided Attention Modulation

The analysis results on attention maps motivate us to intervene in the generation process and modulate original scores to better reflect textual and layout conditions. Specifically,
we modulate the attention maps as below,

\[
A' = \text{softmax} \left( \frac{QK^T + M}{\sqrt{d}} \right),
\]

\[
M = \lambda_t \cdot R \odot M_{\text{pos}} \odot (1 - S)
- \lambda_t \cdot (1 - R) \odot M_{\text{neg}} \odot (1 - S),
\]  

where the query-key pair condition map \( R \in \mathbb{R}^{[\text{queries}] \times [\text{keys}]} \) defines whether to increase or decrease the attention score for a particular pair. If two tokens belong to the same segment, they form a positive pair, and their attention score will be increased. If not, they form a negative pair, with their attention decreased.

Due to the difference between cross-attention and self-attention layers, \( R \) is defined separately for the two layers. We introduce the matrices \( M_{\text{pos}}, M_{\text{neg}} \in \mathbb{R}^{[\text{queries}] \times [\text{keys}]} \) to consider the original value range, which aims to preserve the pre-trained model’s generation capacity. To further adjust the degree of modulation according to the size of each object, we calculate the matrix \( S \in \mathbb{R}^{[\text{queries}] \times [\text{keys}]} \) that indicates the segment area for each image query token.

Finally, we observe that a large modification may deteriorate the image quality as the timestep \( t \) approaches zero. Therefore, we use a scalar \( \lambda_t \) to adjust the degree of modulation using the power function as below,

\[
\lambda_t := w \cdot t^p,
\]

where the timestep \( t \in [0, 1] \) has been normalized. We use different co-efficients \( w^c \) and \( w^s \) for cross-attention and self-attention layers.

**Attention Modulation at Cross-attention Layers.** In cross-attention layers, intermediate image features are updated according to textual features, which construct objects’ appearance and layouts. The degree and location to which they are reflected are determined by the attention scores between image tokens and text tokens. Hence, we modify cross-attention maps to induce certain textual features congregate in a specific region according to its corresponding layout condition \( m_n \), with the query-key pair condition map \( R_{\text{cross}} \) defined as below,

\[
R_{ij}^{\text{cross}} := \begin{cases} 
\bar{k}, & \text{if } \bar{k}[j] = 0 \\
\bar{m}_{k[j]}, & \text{otherwise}
\end{cases}
\]

where \( \bar{m}_n \) is a binary map flattened into a vector and resized to match the spatial resolution of each layer. Here \( \bar{k} \in \mathbb{R}^{[\text{keys}]} \) is a vector that maps a token index to the segment index. If the \( j \)-th text token is not part of any of the \( N \) segments, \( \bar{k}[j] \) is set to zero.

**Attention Modulation at Self-attention Layers.** Self-attention layers allow intermediate features to interact with each other to create a globally coherent result. Our attention modulation is designed to restrict communication between tokens of different segments, thereby preventing the mixing of features of distinct objects. Specifically, we increase the attention scores for tokens in the same segment and decrease it for those in different segments. The query-key pair condition map \( R_{\text{self}} \) for this objective is defined as below,

\[
R_{ij}^{\text{self}} := \begin{cases} 
1, & \text{if } \exists n \text{ s.t. } \bar{m}_n[i] = 1 \text{ and } \bar{m}_n[j] = 1 \\
0, & \text{otherwise}
\end{cases}
\]

**Value-range Adaptive Attention Modulation.** Since our method alters the original denoising process, it may potentially compromise the image quality of the pre-trained model. To mitigate this risk, we modulate values according to the range of original attention scores. We calculate the following matrices that identify each query’s maximum and minimum values, ensuring the modulated values stay close to the original range. Therefore, the adjustment is proportional to the difference between the original values and either the maximum value (for positive pairs) or the minimum value (for
negative pairs).

\[ M_{\text{pos}} = \max(QK^\top) - QK^\top, \]
\[ M_{\text{neg}} = QK^\top - \min(QK^\top). \] (6)

The max and min operations return the maximum and minimum values for each query, initially producing a vector with dimensions \( \mathbb{R}^{|\text{queries}| \times 1} \). To restore the original dimensions \( \mathbb{R}^{|\text{queries}| \times |\text{keys}|} \), we replicate values along the key-axis.

Mask-area Adaptive Attention Modulation. We observe a noticeable quality degradation when there is a large area difference between segments. Specifically, if one segment’s area is much smaller than others, our method may fail to generate realistic images. To resolve this, we use the matrix \( S \) in Equation 2 to automatically adjust the modulation degree according to the area of each segment: increase the degree for small segments and decrease for large segments. To calculate the matrix \( S \), we first compute the area percentage of the mask that each query token belongs to, and then replicate the values along the key axis.

Implementation Details. We use Stable Diffusion [45], trained on the LAION dataset [47]. In our experiments, we only apply the attention modulation to the initial denoising steps \( (t \approx 0.7) \), as we observe no clear improvement beyond that. The hyperparameters in Equation 3 are set as follows: \( p = 5 \), \( w^c = 1 \) for cross-attention layers, and \( w^s = 0.3 \) for self-attention layers. To further increase the effectiveness of our method, we replace parts of textual features with separately encoded ones for each text segment \( c_n \). This strategy is particularly useful when a text caption contains multiple closely related objects, such as a microwave and an oven.

4. Experiments

4.1. Evaluation Setting

Baselines. We compare our method with various training-free methods, which aim to improve the fidelity of the pre-trained Stable Diffusion [45] when using dense captions. Composable Diffusion [35] predicts noises for each phrase separately and merges them at every timestep, leading to a longer inference time proportional to the number of phrases. Structure Diffusion [17] employs a language parser to enhance the value features of all parsed phrases at every cross-attention layer. eDiffi-Pww [4] manipulates the attention scores for specific query-key pairs according to an additional layout condition. However, since the eDiffi model is not publicly available, we conduct experiments with the Pww implemented on Stable Diffusion (SD-Pww).

<table>
<thead>
<tr>
<th></th>
<th>SOA-I ↑</th>
<th>CLIP-Score ↑</th>
<th>Human Preference ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Images</td>
<td>93.07</td>
<td>0.2821</td>
<td>-</td>
</tr>
<tr>
<td>SD-Pww [4]</td>
<td>73.92 ± 1.84</td>
<td>0.2800 ± 0.0005</td>
<td>53%</td>
</tr>
<tr>
<td>DenseDiffusion (Ours)</td>
<td>77.61 ± 1.75</td>
<td>0.2814 ± 0.0005</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: Quantitative evaluation results on the fidelity to textual condition. We compare our method with various automatic metrics and user study. The human preference percentage shows how often AMT participants prefer our results over the baselines.

<table>
<thead>
<tr>
<th></th>
<th>IoU ↑</th>
<th>Local CLIP-Score ↑</th>
<th>Human Preference ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Images</td>
<td>59.13</td>
<td>0.2163</td>
<td>-</td>
</tr>
<tr>
<td>SD-Pww [4]</td>
<td>23.76 ± 0.50</td>
<td>0.2123 ± 0.0003</td>
<td>63%</td>
</tr>
<tr>
<td>DenseDiffusion (Ours)</td>
<td>34.99 ± 1.13</td>
<td>0.2176 ± 0.0007</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: Quantitative evaluation results on the fidelity to layout condition. We compare only with SD-Pww [4] since it is the only baseline that uses segmentation maps. 63% of AMT participants prefer our results over those from SD-Pww.

Metrics. We evaluate each method regarding two criteria: fidelity to the text prompt and alignment with the layout condition. For text prompt, we calculate the CLIP-Score [22], which measures the distance between input textual features and generated image features, and the SOA-I score [23] that uses YOLOv7 [55] to check if the described objects are present in the generated image. In terms of layout alignment, we compare IoU scores of a segmentation map predicted by YOLOv7 [55], with respect to the given layout condition. We further evaluate CLIP-scores on cropped object images (Local CLIP-score [2]) to check whether generated objects followed detailed descriptions. As Composable Diffusion [35] and Structure Diffusion [17] do not take layout conditions, they are excluded for a fair comparison.

Datasets. We curate a new evaluation dataset containing detailed descriptions for each segment. Specifically, we select 250 samples with two or more unique objects from the MS-COCO validation set [34]. Then we manually replace the class label for each segmentation map with a phrase extracted from the caption; for example, “dog” to “a black and white dog”. We generate four random images per caption, leading to 1,000 images for each baseline used in the evaluation.

User study. We conduct a user study using Amazon Mechanical Turk. For each task, we present users with two sets of 4 images along with the same input conditions. They are
Figure 4: Comparisons with other training-free methods based on Stable Diffusion [45]. Each image is generated with the same dense caption for all methods. However, only SD-Pww [4] and our method DenseDiffusion support segmentation maps for layout control. Our method aligns more closely with the input mask than SD-Pww [4].

4.2. Results

Evaluation on the fidelity to textual condition. In Figure 4, we compare DenseDiffusion with all baselines for images generated with dense captions. While baseline methods sometimes omit one or more objects described in the text captions, our results are more faithful to both textual and layout conditions. In particular, the comparison with SD-Pww [4] highlights the effectiveness of our training-free modulation method.

The quantitative evaluation results in Table 2 reveal a consistent trend. DenseDiffusion outperforms others in terms of both automatic and human evaluation. However, SOA-I seems to be loosely correlated with the human evaluation results due to the domain gap between LAION [47] and MS-COCO [34], used for training Stable Diffusion [45] and YOLOv7 [55] respectively. Interestingly, the performance tends to suffer significantly when the inference method is altered too much from the original one, as seen in the case of Composable Diffusion [35].

Evaluation on the fidelity to layout condition. To evaluate the fidelity to layout conditions, we only compare with the results of SD-Pww [4], the only baseline that can control an image layout. Figures 4, 5 and Table 3 show that DenseDiffusion outperforms SD-Pww by a large margin. SD-Pww not only fails to reflect layout conditions faithfully but also tends to mix different object features or omits key objects. In particular, the substantial difference in IoU scores shows that DenseDiffusion is more effective in reflecting layout conditions. Figure 6 shows that our method responds well to various conditions created by changing part of a given textual condition, such as object types or image styles while maintaining the original layout condition.
Dense Diffusion (Ours)  

SD-Pww [4]  

MAS [19]  

SpaText [2]  

Stable Diffusion [45]

Figure 5: Comparison with different layout-guided text-to-image methods. While MAS [19] and SpaText [2] are specifically trained for layout control, SD-Pww [4] and DenseDiffusion (ours) are training-free methods based on pre-trained Stable Diffusion models [45]. Nevertheless, our results adhere to layout conditions comparably to SpaText [2], and even outperform MAS [19] in many cases.

Comparison to layout-conditioned models. To highlight that DenseDiffusion is effective even as a training-free method, we further compare with MAS [19] and SpaText [2], both of which are text-to-image models trained with layout conditions. MAS [19] uses tokenized semantic segmentation maps as an additional condition, and SpaText [2] fine-tunes Stable Diffusion [45] with spatially splatted CLIP image features according to a layout condition. Since these models are not publicly available, we use the examples presented in the original SpaText paper. Figure 5 shows that DenseDiffusion can reflect layout conditions comparably and even outperforms MAS [19] for diverse concepts.

Ablation Study. Below we evaluate each component used in our DenseDiffusion: (a) Attention Modulation at Cross-attention Layers, (b) Attention Modulation at Self-attention Layers, (c) Value-range Adaptive Attention Modulation, and (d) Mask-area Adaptive Attention Modulation.

We first present visual results from our ablation study in Figure 7. All images in the same row are generated from the same initial noise map. As shown in columns w/o (a) and w/o (b), attention modulations in both cross-attention and self-attention layers are crucial for satisfying both textual and layout conditions. The images in column w/o (c) show that value-range adaptive modulation improves the fidelity of the method to given conditions further. Finally, according
A painting of a dog riding a flying bicycle, over a big city with a yellowish full moon in the night sky.

A painting of a black horse under a red full moon, in the style of The Starry Night.

→ a unicorn
→ a zebra
→ a UFO
→ in the style of Edvard Munch

→ a cat
→ a crescent moon
→ the ocean
→ the pointillism style

Figure 6: We generate images using the same layout condition but with various text prompts, by modifying part of the given textual condition. Our results faithfully adhere to both textual and layout conditions.

Figure 7: Ablation study. We show some example results when various components are ablated from our full method. We define the following components as: (a) Attention Modulation at Cross-attention Layers, (b) Attention Modulation at Self-attention Layers, (c) Value-range Adaptive Attention Modulation, and (d) Mask-area Adaptive Attention Modulation. All images are generated with the same initial noise map. We can infer that all components contribute to improving the fidelity of Stable Diffusion to the given conditions.

to column w/o (d), the method adheres to the conditions but produces images with monotonous textures.

In Table 4, we evaluate various ablated methods using automatic metrics. The results show that, except for component (d), the removal of each component leads to a significant drop in all metric scores. As indicated by the significant drops in IoU scores, the ablated methods fail to achieve high fidelity to layout conditions. The largest performance degradation is observed when attention modulation at cross-attention layers is omitted, as textual features play a crucial role in constructing the image layout. Furthermore, modulating scores without considering the value-range also results in drops across all metrics. These findings collectively confirm that our attention modulation is highly effective in improving the fidelity to given conditions without compromising the generation capability of the pre-trained model.

Regarding component (d), we interpret this exception as a result of the ablated method’s tendency to create a
Table 4: Quantitative evaluations on each component. We evaluate several variants, in which the following components are removed from our full method: (a) Attention Modulation at Cross-attention Layers, (b) Attention Modulation at Self-attention Layers, (c) Value-range Adaptive Attention Modulation, and (d) Mask-area Adaptive Attention Modulation. The results suggest that each component contributes to the final performance of DenseDiffusion.

<table>
<thead>
<tr>
<th>Component</th>
<th>SOA-I ↑</th>
<th>CLIP-Score ↑</th>
<th>IoU ↑</th>
<th>Local CLIP-Score ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o (a)</td>
<td>72.27</td>
<td>0.2785</td>
<td>18.35</td>
<td>0.2050</td>
</tr>
<tr>
<td>w/o (b)</td>
<td>75.60</td>
<td>0.2790</td>
<td>29.57</td>
<td>0.2152</td>
</tr>
<tr>
<td>w/o (c)</td>
<td>76.51</td>
<td>0.2801</td>
<td>29.88</td>
<td>0.2136</td>
</tr>
<tr>
<td>w/o (d)</td>
<td>73.62</td>
<td>0.2808</td>
<td>38.00</td>
<td>0.2191</td>
</tr>
<tr>
<td>DenseDiffusion</td>
<td>77.61</td>
<td>0.2814</td>
<td>34.99</td>
<td>0.2176</td>
</tr>
</tbody>
</table>

Figure 8: Limitations and failure cases. (a) Given captions, “a photo of (a bear / a juggling bear) at the beach”, our method fails in the latter case, since the Stable Diffusion itself cannot generate an image with the correct appearance. (b) As both self-attention and cross-attention layers are fairly coarse, our method fails to follow fine details of the segment condition, such as the shape of leaves.

5. Conclusion

In this work, we have proposed DenseDiffusion, a training-free method that improves the fidelity of a pretrained text-to-image model to dense captions, and enable image layout control. Our finding shows that considering the value range and segment size significantly improves our attention modulation method. Experimental results show that DenseDiffusion outperform other methods on various evaluation metrics. Notably, our training-free method provides comparable layout control to existing models that are specifically trained for this task.

Limitations. DenseDiffusion has several limitations. First, our method is highly dependent on the capacity of its base model, Stable Diffusion [45]. As shown in Figure 8a, our method fails to produce certain objects, such as a juggling bear, if Stable Diffusion is unable to produce them itself.

Second, our method struggles with fine-grained input masks with thin structures, as both self-attention and cross-attention layers are fairly coarse. As examples shown in Figure 8b, our method fails to follow fine details of the segment condition, such as the shape of leaves.

Acknowledgments. We thank Or Patashnik and Songwei Ge for paper proofreading and helpful comments. This work was experimented on the NAVER Smart Machine Learning (NSML) platform [32].

References


[50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In ICLR, 2021. 3


