Progressive Text-to-Image Diffusion with Soft Latent Direction

Yuteng Ye, Jiale Cai, Hang Zhou, Guanwen Li, Youjia Zhang, Zikai Song, Chenxing Gao, Junqing Yu, Wei Yang*
Huazhong University of Science and Technology, Wuhan, China
{yuteng_ye, jaile_cai, henryryzh, jugglee, youjiazhang, skyesong, cxg, yjqing, weiyangcs}@hust.edu.cn

Abstract

In spite of the rapidly evolving landscape of text-to-image generation, the synthesis and manipulation of multiple entities while adhering to specific relational constraints pose enduring challenges. This paper introduces an innovative progressive synthesis and editing operation that systematically incorporates entities into the target image, ensuring their adhesion to spatial and relational constraints at each sequential step. Our key insight stems from the observation that while a pre-trained text-to-image diffusion model adeptly handles one or two entities, it often falters when dealing with a greater number. To address this limitation, we propose harnessing the capabilities of a Large Language Model (LLM) to decompose intricate and protracted text descriptions into coherent directives adhering to stringent formats. To facilitate the execution of directives involving distinct semantic operations—namely insertion, editing, and erasing—we formulate the Stimulus, Response, and Fusion (SRF) framework. Within this framework, latent regions are gently stimulated in alignment with each operation, followed by the fusion of the responsive latent components to achieve cohesive entity manipulation. Our proposed framework yields notable advancements in object synthesis, particularly when confronted with intricate and lengthy textual inputs. Consequently, it establishes a new benchmark for text-to-image generation tasks, further elevating the field’s performance standards.

Introduction

Text-to-image generation is a vital and rapidly evolving field in computer vision that has attracted unprecedented attention from both researchers and the general public. The remarkable advances in this area are driven by the application of state-of-the-art image-generative models, such as auto-regressive (Ramesh et al. 2021; Wang et al. 2022) and diffusion models (Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022), as well as the availability of large-scale language-image datasets (Sharma et al. 2018; Schuhmann et al. 2022). However, existing methods face challenges in synthesizing or editing multiple subjects with specific relational and attributive constraints from textual prompts (Chefer et al. 2023). The typical defects that occur in the synthesis results are missing entities, and inaccurate inter-object relations, as shown in ???. Existing work improves the compositional skills of text-to-image synthesis models by incorporating linguistic structures (Feng et al. 2022), and attention controls (Hertz et al. 2022; Chefer et al. 2023) within the diffusion guidance process. Notably, Structured Diffusion (Feng et al. 2022) parse a text to extract numerous noun phrases, Attend-and-Excite (Chefer et al. 2023) strengthen attention activations associated with the most marginalized subject token. Yet, these remedies still face difficulties when the text description is long and complex, especially when it involves two and more subjects. Furthermore, users may find it necessary to perform subtle modifications to the unsatisfactory regions of the generated image, while preserving the remaining areas.

In this paper, we propose a novel progressive synthesizing/editing operation that successively incorporates entities, that conform to the spatial and relational constraint defined in the text prompt, while preserving the structure and aesthetics in each step. Our intuition is based on the observation that text-to-image models tend to better handle short-sentence prompts with a limited number of entities (1 or 2) than long descriptions with more entities. Therefore, we can parse the long descriptions into short-sentence prompts and craft the image progressively via a diffusion model to prevent the leakage and missing of semantics.

However, applying such a progressive operation to diffusion models faces two major challenges:

• The absence of a unified method for converting the integrated text-to-image process into a progressive procedure that can handle both synthesis and editing simultaneously. Current strategies can either synthesize (Chefer et al. 2023; Ma et al. 2023) or edit (Kawar et al. 2023; Goel et al. 2023; Xie et al. 2022; Avrahami, Fried, and Lischinski 2022; Yang et al. 2023), leaving a gap in the collective integration of these functions.

• The need for precise positioning and relational entity placement. Existing solutions either rely on user-supplied masks for entity insertion, necessitating manual intervention (Avrahami, Fried, and Lischinski 2022; Nichol et al. 2021), or introduce supplementary phrases to determine the entity editing direction (Hertz et al. 2022; Brooks, Holynski, and Efros 2023), which inadequately addressing spatial and relational dynamics.

*indicates corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
"...There is a gentle river flowing down the forest. On the left side of the river stands a bear wearing a shirt and a hat, while on the right is a mushroom..."

To overcome these hurdles, we present the Stimulus, Response, and Fusion (SRF) framework, assimilating a stimulus-response generation mechanism along with a latent fusion module into the diffusion process. Our methodology involves employing a fine-tuned GPT model to deconstruct complex texts into structured prompts, including synthesis, editing, and erasing operations governed by a unified SRF framework. Our progressive process begins with a real image or synthesized background, accompanied by the text prompt, and applies the SRF method in a step-by-step approach. Unlike previous strategies that aggressively manipulate the cross-attention map (Wu et al. 2023; Ma et al. 2023), our operation guides the attention map via a soft direction, avoiding brusque modifications that may lead to discordant synthesis. Additionally, when addressing relationships like "wearing" and "playing with", we begin by parsing the relative positions of the objects with the help of GPT, after which we incorporate the relational description and relative positions into the diffusion model to enable object interactions.

In summary, we unveil a novel, progressive text-to-image diffusion framework that leverages the capabilities of a Language Model (LLM) to simplify language description, offering a unified solution for handling synthesis and editing patterns concurrently. This represents an advancement in text-to-image generation and provides a new platform for future research.

Related Work

Our method is closely related to image manipulation and cross-attention control within diffusion models.

Image manipulating refers to the process of digitally manipulating images to modify or enhance their visual appearance. Various techniques can be employed to achieve this end, such as the use of spatial masks or natural language descriptions to guide the editing process towards specific goals. One promising line of inquiry involves the application of generative adversarial networks (GANs) for image domain transfer (Isola et al. 2017; Sangkloy et al. 2017; Zhu et al. 2017; Choi et al. 2018; Wang et al. 2018; Huang et al. 2018; Park et al. 2019; Liu, Breuel, and Kautz 2017; Baek et al. 2021) or the manipulation of latent space (Zhu et al. 2016; Huh et al. 2020; Richardson et al. 2021; Zhu et al. 2020; Wulff and Torralba 2020; Bau et al. 2021). Recently, diffusion models have emerged as the mainstream. GLIDE (Nichol et al. 2021), Blended diffusion (Avrahami, Fried, and Lischinski 2022) and SmartBrush (Xie et al. 2022) replace masked image regions with predefined objects while preserving the inherent image structure. Additionally, techniques such as prompt-to-prompt (Hertz et al. 2022) and instructpix2pix (Brooks, Holynski, and Efros 2023) enable the modification of image-level objects through text alterations. Contrasting previous methods that solely cater to either synthesis or editing, we construct a unified framework that accommodates both.

Objects and positional relationships are manifested within
Text Decomposition

"...There is a dog and a cat playing together on the right side of the yard... These apples are then replaced with oranges, which are subsequently removed..."

Step i  
"[a dog] [plays with] [a cat] [on the right side of] [the yard]."

Step i+1  
"change [apples] to [oranges]."

Step i+2  
"delete [oranges]."

Figure 2: We employ a fine-tuned GPT model to deconstruct a comprehensive text into structured prompts, each classified under synthesis, editing, and erasing operations.

Figure 3: For the synthesis operation, we generate the layout indicated in the prompt from a frozen GPT-4 model, which subsequently yields the new bounding box coordinates for object insertion.

The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)
6695

Method

Problem Formulation

we elaborate upon our innovative progressive text-to-image framework. Given a multifaceted text description \( P \) and a real or generated background \( I \), our primary goal is to synthesize an image that meticulously adheres to the modifications delineated by \( P \) in alignment with \( I \). The principal challenge emerges from the necessity to decode the intricacy of \( P \), manifesting across three complex dimensions:

- The presence of multiple entities and attributes escalates the complexity of the scene, imposing stringent demands on the model to generate representations that are not only accurate but also internally coherent and contextually aligned.

- The integration of diverse positional and relational descriptions calls for the model to exhibit an advanced level of understanding and to employ sophisticated techniques to ascertain precise spatial configuration, reflecting both explicit commands and implied semantic relations.

- The concurrent introduction of synthesis, editing, and erasing operations introduces additional layers of complexity to the task. Managing these intricate operations within a unified model presents a formidable challenge, requiring a robust and carefully designed approach to ensure seamless integration and execution.

We address these challenges through a unified progressive text-to-image framework that: (1) employs a fine-tuned GPT model to distill complex texts into short prompts, categorizing each as synthesis, editing, or erasing mode, and accordingly generating the object mask; (2) sequentially processes these prompts within the same framework, utilizing attention-guided generation to capture position-aware features with soft latent direction, and subsequently integrates them with the previous stage’s outcomes in a subtle manner. This approach synthesizes the intricacies of text-to-image transformation into a coherent, positionally aware procedure.

Text Decomposition

\( P \) may involve multiple objects and relations, we decompose \( P \) into a set of short prompts, which produces an image accurately representing \( P \) when executed sequentially. As illustrated in fig. 2, we fine-tune a GPT with OpenAI API (OpenAI 2023) to decompose \( P \) into multiple structured prompts, denoted as \( \{P_1, P_2, \ldots, P_n\} \). Each \( P_i \) falls into one of the three distinct modes: Synthesis mode: "[object 1] [relation] [object 2] [position] [object 3]", Editing mode: "change [object 1] to [object 2]", and Erasing mode: "delete [object]". In pursuit of this aim, we start by collecting full texts using ChatGPT (Brown et al. 2020) and then manually deconstruct them into atomic prompts. Each prompt has a minimal number of relations and is labeled with synthesis/editing/erasing mode. Using these prompts and their corresponding modes for model supervision, we fine-tune the GPT model to enhance its decomposition and generalization ability.
Operational Layouts. For the synthesis operation, as shown in fig. 3, we feed both the prompt and a reference bounding box into a frozen GPT-4 API. This procedure produces bounding boxes for the target entity that will be used in the subsequent phase. We exploit GPT-4’s ability to extract information from positional and relational text descriptors. For example, the phrase “cat and dog play together” indicates a close spatial relationship between the “cat” and “dog”. Meanwhile, “on the right side” suggests that both animals are positioned to the right of the “yard”. For the editing and erasing operations, we employ Diffusion Inversion (Mokady et al. 2023) to obtain the cross-attention map of the target object, which serves as the layout mask. For example, when changing “apples” to “oranges”, we draw upon the attention corresponding to “apples”. On the other hand, to “delete the oranges”, we focus on the attention related to “oranges”. Notably, this approach avoids the need to retrain the diffusion model and is proficient in managing open vocabularies. We denote generated layout mask as \( \mathcal{M} \) for all operations in following sections for convention.

In the following section, we provide a complete introduction to the synthesis operation. At last, we exhibit that the editing and erasing operations only differ from the synthesis operation in parameter settings.

Stimulus & Response

With the synthesis prompt \( \mathcal{P}_i \) to be executed and its mask configuration \( \mathcal{M}_i \). The goal of Latent Stimulus & Response is to enhance the positional feature representation on \( \mathcal{M} \). As illustrated in fig. 4, this is achieved by guided cross-attention generation. Differing from the approaches (Ma et al. 2023; Wu et al. 2023), which manipulate attention through numerical replacement, we modulate the attention within mask regions associated with the entity in \( \mathcal{P}_i \) via a soft manner. Rather than directly altering the attention, we introduce a stimulus to ensure that the object attention converges to the desired scores. Specifically, we formulate a stimulus loss function between the object mask \( \mathcal{M} \) and the corresponding attention \( A \) as:

\[
\mathcal{L}_s = \sum_{i=1}^{n} (\text{softmax}(A_i^t) - \delta \cdot \mathcal{M}^t)
\]

where \( A_i^t \) signifies the cross-attention map of the \( i \)-th object at the \( t \)-th timestep, \( \mathcal{M}^t \) denotes the mask of the \( i \)-th object, \( \delta \) represents the stimulus weights. The intent of stimulus attention leans towards a spatial-wise generation process. This is achieved by backpropagating the gradient of the stimulus loss function, as defined in Eq. 1, to update the latent code. This process serves as a latent response to the stimulated attention, which can be formally expressed as:

\[
\delta z_i^t \leftarrow \delta z_i - \alpha_t \cdot \nabla_{z_i} \mathcal{L}_s
\]

In the above equation, \( \delta z_i^t \) represents the updated latent code and \( \alpha_t \) denotes the learning rate. Finally, we execute another forward pass of the stable diffusion model using the updated latent code \( \delta z_i^t \) to compute \( \delta z_{i-1}^t \) for the subsequent denoising step. Based on eq. (1) and eq. (2), we observe consistent spatial behavior in both the cross-attention and latent spaces. For a more detailed analysis, we refer to fig. 5 and find this property contributes to producing faithful and position-aware image representations.
Stable Diffusion (SD) V-1.4. During the Stimulus & Response stage, we assign a weight of $\delta$ equals 0.8 in eq. (1), and set $t$ equals 25 and $\alpha_t$ equals 40 in eq. (2). We implement the stimulus procedure over the $16 \times 16$ attention units and integrate the Iterative Latent Refinement design (Chefer et al. 2023). In the latent fusion stage, the parameter $\tau$ is set to a value of 40.

**Qualitative and Quantitative Results**

**Qualitative and Quantitative Comparisons with Single-Generation Baselines.** Fig. 6 reveals that traditional baseline methods often struggle with object omissions and maintaining spatial and interactional relations. In contrast, our progressive generation process offers enhanced image fidelity and controllability. Additionally, we maintain finer details in the generated images, such as the shadows of the

**Experiment**

**Baselines and Evaluation.** Our experimental comparison primarily concentrates on Single-Stage Generation and Progressive Generation baselines. (1) We refer to Single-Stage Generation methods as those that directly generate images from input text in a single step. Current methods include Stable Diffusion (Rombach et al. 2022), Attend-and-excite (Chefer et al. 2023), and Structured Diffusion (Feng et al. 2022). We compare these methods to analyze the efficacy of our progressive synthesis operation. We employ GPT to construct 500 text prompts that contain diverse objects and relationship types. For evaluation, we follow (Wu et al. 2023) to compute Object Recall, which quantifies the percentage of objects successfully synthesized. Moreover, we measure Relation Accuracy as the percentage of spatial or relational text descriptions that are correctly identified, based on 8 human evaluations. (2) We define Progressive Generation as a multi-turn synthesis and editing process that builds on images from preceding rounds. Our comparison encompasses our comprehensive progressive framework against other progressive methods, which includes Instruction-based Diffusion models (Brooks, Holynski, and Efros 2023) and mask-based diffusion models (Rombach et al. 2022; Avrahami, Fried, and Lischinski 2022). To maintain a balanced comparison, we source the same input images from SUN (Xiao et al. 2016) and text descriptions via the GPT API (OpenAI 2023). Specifically, we collate five scenarios totaling 25 images from SUN, a dataset that showcases real-world landscapes. Each image is paired with the text description, which ensures: 1. Integration of synthesis, editing, and easing paradigms; 2. Incorporation of a diverse assortment of synthesized objects; 3. Representation of spatial relations (e.g., top, bottom, left, right) and interactional relations (e.g., “playing with”, “wearing”). For evaluation, we utilize Amazon Mechanical Turk (AMT) to assess image fidelity. Each image is evaluated based on the fidelity of the generated objects, their relationships, the execution of editing instructions, and the alignment of erasures with the text descriptions. Images are rated on a fidelity scale from 0 to 2, where 0 represents the lowest quality and 2 signifies the highest. With two evaluators assessing each generated image, the cumulative score for each aspect can reach a maximum of 100.

**Implementation Details.** Our framework builds upon Stable Diffusion (SD) V-1.4. During the Stimulus & Response stage, we assign a weight of $\delta$ equals 0.8 in eq. (1), and set $t$ equals 25 and $\alpha_t$ equals 40 in eq. (2). We implement the stimulus procedure over the $16 \times 16$ attention units and integrate the Iterative Latent Refinement design (Chefer et al. 2023). In the latent fusion stage, the parameter $\tau$ is set to a value of 40.

**Qualitative and Quantitative Results**

**Qualitative and Quantitative Comparisons with Single-Generation Baselines.** Fig. 6 reveals that traditional baseline methods often struggle with object omissions and maintaining spatial and interactional relations. In contrast, our progressive generation process offers enhanced image fidelity and controllability. Additionally, we maintain finer details in the generated images, such as the shadows of the
Figure 6: Qualitative comparison with Single-Stage baselines. Common errors in the baselines include missing objects and mismatched relations. Our method demonstrates the progressive generation process.

<table>
<thead>
<tr>
<th>Method</th>
<th>Object Recall ↑</th>
<th>Relation Accuracy ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stable Diffusion</td>
<td>40.7</td>
<td>19.8</td>
</tr>
<tr>
<td>Structured Diffusion</td>
<td>43.5</td>
<td>21.6</td>
</tr>
<tr>
<td>Attend-and-excite</td>
<td>50.3</td>
<td>23.4</td>
</tr>
<tr>
<td>Ours</td>
<td>64.4</td>
<td>50.8</td>
</tr>
</tbody>
</table>

Table 1: Quantitative comparison with Single-Stage Generation baselines.
Figure 7: The analysis of Stimulus & Response in the editing operation. The left side shows a visual comparison between SD (Stable Diffusion) and S&R (Stimulus & Response). The right side presents the convergence curve of cross-attention loss during diffusion sampling steps. The loss is computed as the difference between reference attention and model-generated attention. In the right figure, red, blue, and green colors represent the objects “jaguar”, “cat”, and “monkey” respectively. Solid lines indicate SD loss, while dashed lines represent S&D loss.

Figure 8: Qualitative comparison with Progressive Generation baselines. The first two phases illustrate object synthesis operation, where target objects are color-coded in both the text and layout. Subsequent phases depict object editing and erasing processes, wherein a cat is first transformed into a rabbit and then the rabbit is removed.

Table 2: Quantitative comparison of our method against Progressive Generation baselines, using rating scores.

<table>
<thead>
<tr>
<th>Method</th>
<th>Synthesis</th>
<th>Editing</th>
<th>Erasing</th>
</tr>
</thead>
<tbody>
<tr>
<td>InstructPix2Pix</td>
<td>19</td>
<td>24</td>
<td>32</td>
</tr>
<tr>
<td>Stable-inpainting</td>
<td>64</td>
<td>54</td>
<td>65</td>
</tr>
<tr>
<td>Blended Latent</td>
<td>67</td>
<td>32</td>
<td>67</td>
</tr>
<tr>
<td>Ours</td>
<td>74</td>
<td>60</td>
<td>72</td>
</tr>
</tbody>
</table>

Table 3: Ablation study. LF and S&R represent Latent Fusion and Stimulus & Response respectively.

<table>
<thead>
<tr>
<th>Method Variant</th>
<th>Object Recall ↑</th>
<th>Relation Accuracy ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o LF</td>
<td>38.8</td>
<td>21.8</td>
</tr>
<tr>
<td>w/o S&amp;R</td>
<td>58.3</td>
<td>45.2</td>
</tr>
<tr>
<td>Ours</td>
<td>64.4</td>
<td>50.8</td>
</tr>
</tbody>
</table>

Conclusion

In this study, we addressed the prevailing challenges in the rapidly advancing field of text-to-image generation, particularly the synthesis and manipulation of multiple entities under specific constraints. Our innovative progressive synthesis and editing methodology ensures precise spatial and relational representations. Recognizing the limitations of existing diffusion models with increasing entities, we integrated the capabilities of a Large Language Model (LLM) to dissect complex text into structured directives. Our Stimulus, Response, and Fusion (SRF) framework, which enables seamless entity manipulation, represents a major stride in object synthesis from intricate text inputs.

One major limitation of our approach is that not all text can be decomposed into a sequence of short prompts. For instance, our approach finds it challenging to sequentially parse text such as “a horse under a car and between a cat and a dog.” We plan to gather more training data and labels of this nature to improve the parsing capabilities of GPT.
Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC No. 62272184). The computation is completed in the HPC Platform of Huazhong University of Science and Technology.

References


