

Expressive Image Generation and Editing with Rich Text

Songwei Ge, Taesung Park, Jun-Yan Zhu, Jia-Bin Huang

Abstract

Plain text has become a prevalent interface for text-based image synthesis and editing. Its limited customization options, however, hinder users from accurately describing desired outputs. For example, plain text makes it hard to specify continuous quantities, such as the precise RGB color value or importance of each word. Creating detailed text prompts for complex scenes is tedious for humans to write and challenging for text encoders to interpret. Furthermore, describing a reference concept or texture in plain text is non-trivial. To address these challenges, we propose using a rich-text editor supporting formats such as font style, size, color, texture fill, footnote, and embedded image. We extract each word’s attributes from rich text to enable local style control, explicit token reweighting, precise color rendering, and detailed region synthesis with reference concepts or texture. We achieve these capabilities through a region-based diffusion process. We first obtain each word’s mask that characterizes the region guided by the word. For each region, we enforce its text attributes by creating customized prompts, applying guidance within the region, and maintaining its fidelity against plain-text generations or input images through region-based injections. We present various examples of image generation and editing from rich text and demonstrate that our method outperforms strong baselines with quantitative evaluations.

Keywords: Text-to-image generation, diffusion model, image editing, generative models

1 Introduction

The development of large-scale text-to-image generative models (Ramesh et al., 2021; Saharia et al., 2022; Rombach et al., 2022; Kang et al., 2023) has propelled image generation to an unprecedented era. The great flexibility of these large-scale models further offers users powerful control of the generation through visual cues (Balaji et al., 2022; Gafni et al., 2022; Zhang and Agrawala, 2023) and textual inputs (Brooks et al., 2023; Hertz et al., 2023). Without exception, existing studies use *plain text* encoded by a pretrained language model to guide the generation. However, in our daily lives, it is rare to use only plain text when working on text-based tasks such as writing blogs or editing essays. Instead, a *rich text* editor (Colorado State University, 2012; Witten et al., 2009) is the more popular choice, providing versatile formatting options for writing and editing text. In this paper, we seek to introduce accessible and precise textual control from rich text editors to text-to-image synthesis.

Rich text editors offer unique solutions for incorporating conditional information separate from the text. For example, using the font color, one can indicate an *arbitrary* color. In contrast, describing the precise color with plain text proves more challenging as general text encoders do not understand RGB or Hex triplets, and many color names, such as ‘olive’ and ‘orange’, have ambiguous meanings. This font color information can be used to define the color of generated objects. For example, in Figure 1, a specific **yellow** can be selected to instruct the generation of a marble statue with that exact color.

Beyond providing precise color information, various font formats make it simple to augment the word-level information. For example, reweighting token influence (Hertz et al., 2023) can be implemented using the font size, which is challenging to achieve with existing visual or textual interfaces. Nevertheless, rich text editors offer more options than font size – similar to how font style distinguishes the styles of individual text elements, we propose using it to capture the artistic style of specific regions. Furthermore,

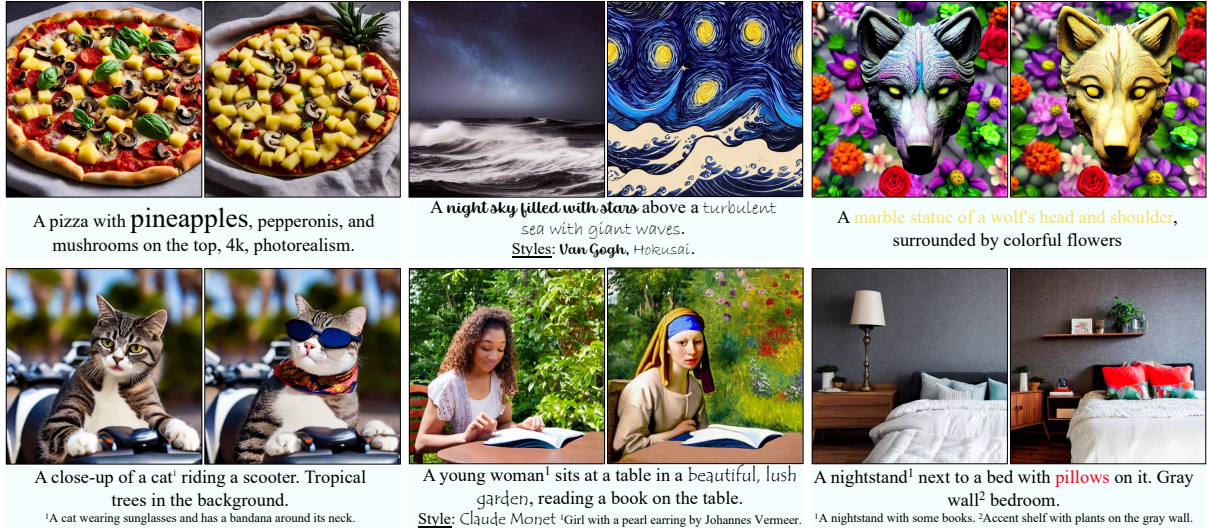


Fig. 1: Plain text (left image) vs. Rich text (right image) Our method allows a user to describe an image using a rich text editor that supports various text attributes like font family, size, color, and footnote. Given these text attributes extracted from rich-text prompts, our method enables precise control of text-to-image synthesis regarding colors, styles, and object details compared to plain text.

embedded images provide extra information in *visual* format. We utilize such images as reference concepts to guide the generated objects. This intuitive design enables personalized/customized generation.

But how can we use rich text? A straightforward implementation is to convert a rich-text prompt with detailed attributes into lengthy plain text and feed it directly into existing methods (Rombach et al., 2022; Hertz et al., 2023; Brooks et al., 2023). Unfortunately, these methods struggle to synthesize images corresponding to lengthy text prompts involving multiple objects with distinct visual attributes, as noted in a recent study (Chefer et al., 2023). They often mix styles and colors, applying a uniform style to the entire image. Furthermore, the lengthy prompt introduces extra difficulty for text encoders to interpret accurate information, making generating intricate details more demanding.

To address these challenges, our insight is to decompose a rich-text prompt into two components (1) a short plain-text prompt (without formatting) and (2) multiple region-specific prompts that include text attributes, as shown in Figure 2. First, we obtain the self- and cross-attention maps using a vanilla denoising process with the short *plain-text* prompt to associate each word with a specific region. Second, we create a prompt for each region using the attributes derived from *rich-text* prompt. For example, we use

“mountain in the style of Ukiyo-e” as the prompt for the region corresponding to the word “mountain” with the attribute “font style: Ukiyo-e”. For RGB font colors that cannot be converted to the prompts, we iteratively update the region with region-based guidance to match the target color. We apply a separate denoising process for each region and fuse the predicted noises to get the final update. During this process, regions associated with the tokens that do not have any formats are supposed to look the same as the plain-text results. Also, the overall shape of the objects should stay unchanged in cases such as only the color is changed. To this end, we propose to use region-based injection approaches.

We demonstrate qualitatively and quantitatively that our method generates more precise color, distinct styles, and accurate details compared to plain text-based methods. We conduct a thorough quantitative evaluation by building a rich-text benchmark by collecting a diverse set of rich-text prompts with font color, style, and footnotes.

A preliminary version of this work was published earlier in (Ge et al., 2023). In this paper, we extend our work and summarize the core differences below.

- We build a benchmark for evaluating the task of rich text-to-image generation. This includes a quantitative evaluation of image generation with

complex prompts. The benchmark presents new challenges for future research.

- We develop two novel applications, leveraging additional rich-text font attributes: 1) using embedded images to guide the concept being generated in the image and 2) enabling texture fill to guide the texture being rendered in the object.
- We apply rich texts for editing *real* images. By leveraging diffusion inversion techniques and segmentation methods, we show rich text-based editing allows precise control of the editing results.

2 Related Work

Text-to-image models. Text-to-image systems aim to synthesize realistic images according to descriptions (Zhu et al., 2007; Mansimov et al., 2016). Fueled by the large-scale text-image datasets (Schuhmann et al., 2022; Byeon et al., 2022), various training and inference techniques (Ho et al., 2020; Song et al., 2021; Ho et al., 2022; Ho and Salimans, 2021), and scalability (Ramesh et al., 2022), significant progress has been made in text-to-image generation using diffusion models (Balaji et al., 2022; Ramesh et al., 2022; Nichol et al., 2022; Saharia et al., 2022; Gafni et al., 2022), autoregressive models (Ramesh et al., 2021; Yu et al., 2022; Chang et al., 2023; Ding et al., 2022), GANs (Sauer et al., 2023; Kang et al., 2023), and their hybrids (Rombach et al., 2022). Our work focuses on making these models more accessible and providing precise controls. In contrast to existing work that uses *plain text*, we use a *rich text* editor with various formatting options.

Controllable image synthesis with diffusion models. A wide range of image generation and editing applications are achieved through either fine-tuning pre-trained diffusion models (Ruiz et al., 2023; Kumari et al., 2023; Zhang and Agrawala, 2023; Avrahami et al., 2023; Wu et al., 2023; Kawar et al., 2023; Ma et al., 2023; Li et al., 2023) or modifying the denoising process (Meng et al., 2022; Choi et al., 2021; Hertz et al., 2023; Parmar et al., 2023; Bansal et al., 2023; Chefer et al., 2023; Avrahami et al., 2022; Balaji et al., 2022; Jiménez, 2023; Bar-Tal et al., 2023; Sarukkai et al., 2023; Zhang et al., 2023; Cao et al., 2023; Phung et al., 2024; Xiao et al., 2023; Feng et al., 2023). For example, Prompt-to-prompt (Hertz et al., 2023) uses attention maps from the original prompt

to guide the spatial structure of the target prompt. Although these methods can be applied to some rich-text-to-image applications, the results often fall short, as shown in Section 4. Concurrent with our work, Mixture-of-diffusion (Jiménez, 2023) and MultiDiffusion (Bar-Tal et al., 2023) propose merging multiple diffusion-denoising processes in different image regions through linear blending. Instead of relying on user-provided regions, we automatically compute regions of selected tokens using attention maps. Gradient (Ho et al., 2022) and Universal (Bansal et al., 2023) guidance control the generation by optimizing the denoised generation at each time step. We apply them to precise color generation by designing an objective on the target region to be optimized.

Layout-controlled image generation. Spatial layouts serve as a meaningful intermediate representation, bridging the gap between abstract scene descriptions and pixel-level image synthesis. Generating images according to layout has been a long-standing and compelling problem (Park et al., 2019; Johnson et al., 2018; Isola et al., 2017; Couairon et al., 2023; Park et al., 2023; Feng et al., 2024; Qu et al., 2023). Park et al. (2023) proposed Gaussian-categorical diffusion process to extend these frameworks by jointly generating images and semantic layouts, further enriching the generative model landscape. Couairon et al. (2023) proposed ZestGuide, a zero-shot method that uses cross-attention-derived segmentation maps for layout-guided image generation. Similarly, Qu et al. (2023) introduced LayoutLLM-T2I to use Large Language Models (LLMs) for layout planning. They propose feedback-based sampler and relation-aware object interaction modules to enhance the quality of complex-scene image generation. Feng et al. (2024) presented LayoutGPT, leveraging in-context learning with structured CSS-style prompts to generate layouts for both 2D images and 3D scenes. A key component in our rich-text-to-image generation framework is the *region-based diffusion process*. Unlike these prior works, where the goal is to produce layout-guided image generation, we aim to *extract the layout* from the diffusion model generation process and synthesize images with rich-text.

Attention in diffusion models. The attention mechanism has been used in various diffusion-based applications such as view synthesis (Liu et al., 2023; Tseng et al., 2023; Watson et al., 2022), image editing (Hertz et al., 2023; Chefer et al., 2023; Patashnik et al., 2023; Parmar et al., 2023; Kumari et al., 2023), and video

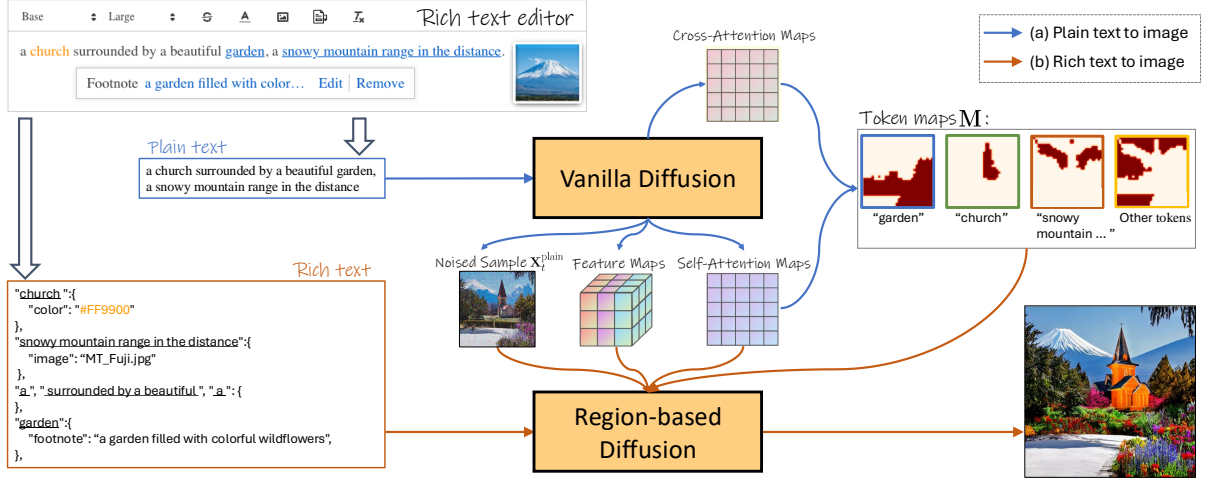


Fig. 2: Rich-text-to-image framework. First, the plain-text prompt is processed by a diffusion model to collect self- and cross-attention maps, noised generation, and residual feature maps at certain steps. The token maps of the input prompt are constructed by first creating a segmentation using the self-attention maps and then labeling each segment using the cross-attention maps. Then the rich texts are processed as JSON to provide attributes for each token span. The resulting token maps and attributes are used to guide our region-based control. We inject the self-attention maps, noised generation, and feature maps to improve fidelity to the plain-text generation.

editing (Liu et al., 2023; QI et al., 2023; Ceylan et al., 2023; Ma et al., 2023). We also leverage the spatial structure in self-attention maps and alignment information between texts and regions in cross-attention maps for rich-text-to-image generation.

Rich text modeling and application. Exploiting information beyond the intrinsic meanings of the texts has been previously studied (Meng et al., 2019; Sun et al., 2021; Xu et al., 2020; Li et al., 2022). For example, visual information, such as underlining and bold type, have also been extracted for various document understanding tasks (Xu et al., 2020; Li et al., 2022). To our knowledge, we are the first to leverage rich text information for text-to-image synthesis.

Image stylization and colorization. Style transfer (Gatys et al., 2016; Zhu et al., 2017; Luan et al., 2017) and Colorization (Reinhard et al., 2001; Tai et al., 2005; Xu et al., 2013; Levin et al., 2004; Zhang et al., 2016, 2017) for *editing real images* have also been extensively studied. In contrast, our work focuses on local style and precise color control for *generating images* from text-to-image models.

Image generation with complex prompts. To accurately generate the image the users expect, one option is to provide more detailed prompts. Several studies

have thus focused on generating images based on complex prompts (Betker et al., 2023; Wang et al., 2024; Wu et al., 2023). DALL-E 3 (Betker et al., 2023) finds that training on highly descriptive synthetic captions reliably improves the alignment between text prompts and generation results. ParaDiffusion (Wu et al., 2023) utilizes pretrained LLM with a larger context window to process complex prompts. Instead, we decouple the complex prompts into multiple detailed prompts that describe local regions. Similar to ours, InstanceDiffusion (Wang et al., 2024) also studies detailed prompts for individual regions, while ours does *not* require layout as part of the user input.

Text-to-image generation benchmark. As text-to-image models develop rapidly, many works have paid attention to the evaluation of these models (Bakr et al., 2023; Hu et al., 2023; Huang et al., 2024; Patel et al., 2024; Zhao et al., 2024). While these benchmarks focus on different aspects of text-to-image generation such as the text-image alignment (Hu et al., 2023; Bakr et al., 2023) and concept learning (Kumari et al., 2023; Patel et al., 2024), we aim at building a benchmark for evaluating rich text to image generation on several applications, including precise color rendering, local style control, and complex prompt alignment.

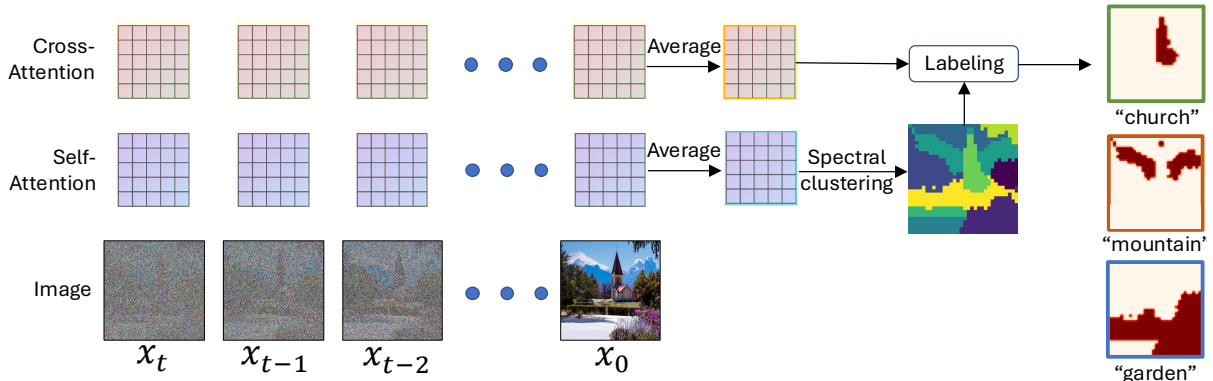


Fig. 3: Token map creation. We average the collected self- and cross-attention maps to create token maps that indicate the layout of the input prompt. The segmentation is first constructed by spectral clustering using the self-attention maps. Then, the averaged cross-attention maps are adopted to label each segment using annotated tokens.

3 Rich Text to Image Generation

From writing messages on communication apps, designing websites (Sahuguet and Azavant, 1999), to collaboratively editing a document (Litt et al., 2022; Ignat et al., 2021), a rich text editor is often the primary interface to edit texts on digital devices. Such rich-text formats allow complementary information to be added to the plain texts. Nonetheless, only plain text has been used in text-to-image generation. To use formatting options in rich-text editors for more precise control over the black-box generation process (Agrawala, 2023), we first introduce a problem setting called *rich-text-to-image generation and editing*. We then discuss our approach to this task.

3.1 Problem Setting

As shown in Figure 2, a rich text editor supports various formatting options, such as font styles, font size, color, and more. We leverage these text attributes as extra information to increase control of text-to-image generation. We interpret the rich-text prompt as JSON, where each text element consists of a span of tokens e_i (e.g., ‘church’) and attributes \mathbf{a}_i describing the span (e.g., ‘color:#FF9900’). Note that some tokens e_U may not have any attributes, and some other tokens may have multiple attributes. Using these annotated prompts, we explore several applications: 1) local style control using *font style*, 2) precise color control using *font color*, 3) texture rendering with *texture fill*, 4) detailed region description using *footnotes*, 5) concept guidance through *embedded images*, and 6) explicit token reweighting with *font sizes*.

Font style is used to apply a specific artistic style \mathbf{a}_i^s , e.g., $\mathbf{a}_i^s = \text{‘Ukiyo-e’}$, to the synthesis of the span of tokens e_i . For instance, in Figure 1, we apply the Ukiyo-e painting style to the ocean waves and the style of Van Gogh to the sky, enabling the application of localized artistic styles. This task presents a unique challenge for existing text-to-image models, as there are limited training images featuring multiple artistic styles. Consequently, existing models tend to generate a *uniform* mixed style across the entire image rather than distinct local styles.

Font color indicates a specific color of the modified text span. Given the prompt “a red toy”, the existing text-to-image models generate toys in various shades of red, such as light red, crimson, or maroon. The color attribute provides a way for specifying a *precise color* in the RGB color space, denoted as \mathbf{a}_i^c . For example, to generate a toy in fire brick red, one can change the font color to “a toy”, where the word “toy” is associated with the attribute $\mathbf{a}_i^c = [178, 34, 34]$. However, as shown in the experiment section, the pretrained text encoder cannot interpret the RGB values and have difficulty understanding obscure color names, such as lime and orange.

Texture fill is another way to stylize text, similar to font color, which is often known as solid fill in the rich-text editor. Such flexible text-fill formatting further supports finer-grained control of the object’s appearance using more concise information beyond text prompts. Specifically, we use \mathbf{a}_i^t to denote the target texture that one wants to render for certain objects. In practice, we use a texture image as the reference.

Footnote provides supplementary explanations of the target span without hindering readability with

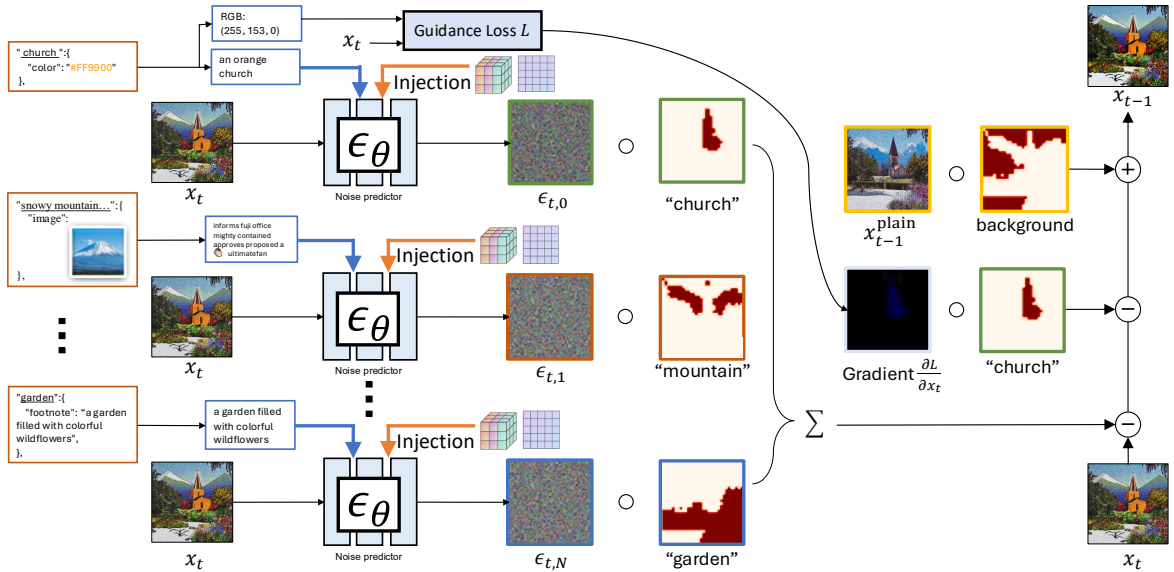


Fig. 4: Region-based diffusion. We fulfill the guidance specified by the rich-text attributes through separate diffusion processes. Depending on the functionality, the attributes are either interpreted as a region-based guidance target (e.g. re-coloring the church), or as a textual input to the diffusion UNet (e.g. handling the embedded image describing the snowy mountain). The self-attention maps and feature maps extracted from the plain-text generation process are injected to help preserve the structure. The predicted noise ϵ_{t,e_i} , weighted by the token map, and the guidance gradient $\frac{\partial L}{\partial x_t}$ are used to denoise and update the previous generation x_t to x_{t-1} . The noised plain text generation x_t^{plain} is blended with the current generation to preserve the exact content in those regions of the unformatted tokens.

lengthy sentences. Writing detailed descriptions of complex scenes is tedious work, and it inevitably creates lengthy prompts (Karpathy and Fei-Fei, 2015; Johnson et al., 2016). Additionally, existing text-to-image models are prone to ignoring some objects when multiple objects are present (Chefer et al., 2023), especially with long prompts. Moreover, excess tokens are discarded when the prompt’s length surpasses the text encoder’s maximum length, e.g., 77 tokens for CLIP models (Radford et al., 2021). We aim to mitigate these issues using a footnote string a_i^f .

Font size can indicate an object’s importance, quantity, or size. We use a scalar a_i^w to denote the weight of each token.

Embedded images provide visual cues complementary to the textual information. Such visual information is more accurate than texts when describing the objects on the identity and low-level details. For example, by connecting the word “cat” in “a cat chasing a butterfly” to a specific cat image, one can generate an image of their cat chasing a butterfly. Generating customized images of the reference concepts has been an active research question (Kumari et al., 2023; Gal

et al., 2023; Chen et al., 2023; Li et al., 2023; Jia et al., 2023; Chen et al., 2023; Gal et al., 2023; Shi et al., 2023; Chen et al., 2023; Xiao et al., 2023). In contrast to free-form generation in prior work, we aim to preserve the original image structure in plain-text generation while synthesizing customized concepts.

3.2 Method

To utilize rich text annotations, our method consists of two steps, as shown in Figure 2. First, we compute the spatial layouts of individual token spans. Second, we use a new region-based diffusion to render each region’s attributes into a globally coherent image.

Step 1. Token maps for spatial layout. Several works (Tang et al., 2022; Ma et al., 2023; Balaji et al., 2022; Hertz et al., 2023; Chefer et al., 2023; Patashnik et al., 2023; Tumanyan et al., 2023) have discovered that the attention maps in the self- and cross-attention layers of the diffusion UNet characterize the spatial layout of the generation. As shown in Figure 3, we develop a method to precisely extract the layout of the diffusion model generation. Our intuition is that the

overall structure of image is determined by the plain text prompt while the rich-text formatting offers more informative descriptions of the local regions.

We use the plain text as the input to the diffusion model and collect self-attention maps of size $32 \times 32 \times 32 \times 32$ across different heads, layers, and time steps. We take the average across all the extracted maps and reshape the result into 1024×1024 . Note that the value at i^{th} row and j^{th} column of the map indicates the probability of pixel i attending to pixel j . We average the map with its transpose to convert it to a symmetric matrix. It is used as a similarity map to perform spectral clustering (Shi and Malik, 2000; Von Luxburg, 2007) and obtain the binary segmentation maps $\widehat{\mathbf{M}}$ of size $K \times 32 \times 32$, where K is the number of segments.

To associate each segment with a textual span, we also extract cross-attention maps for each token w_j :

$$\mathbf{m}_j = \frac{\exp(\mathbf{s}_j)}{\sum_k \exp(\mathbf{s}_k)}, \quad (1)$$

where \mathbf{s}_j is the attention score. We first interpolate each cross-attention map \mathbf{m}_j to the same resolution as $\widehat{\mathbf{M}}$ of 32×32 . Similar to the processing steps of the self-attention maps, we compute the mean across heads, layers, and time steps to get the averaged map $\widehat{\mathbf{m}}_j$. We associate each segment with a texture span e_i following Patashnik et al. (2023):

$$\mathbb{M}_{e_i} = \{\widehat{\mathbf{M}}_k \mid \left| \widehat{\mathbf{M}}_k \cdot \frac{\widehat{\mathbf{m}}_j - \min(\widehat{\mathbf{m}}_j)}{\max(\widehat{\mathbf{m}}_j) - \min(\widehat{\mathbf{m}}_j)} \right|_1 > \epsilon, \quad (2)$$

$$\forall j \text{ s.t. } w_j \in e_i\}, \quad (3)$$

where ϵ is a hyperparameter that controls the labeling threshold, that is, the segment $\widehat{\mathbf{M}}_k$ is assigned to the span e_i if the normalized attention score of any tokens in this span is higher than ϵ . We associate the segments unassigned to any formatted spans with the unformatted tokens e_U . Finally, we obtain the *token map* in Figure 2 as below:

$$\mathbb{M}_{e_i} = \frac{\sum_{\widehat{\mathbf{M}}_j \in \mathbb{M}_{e_i}} \widehat{\mathbf{M}}_j}{\sum_i \sum_{\widehat{\mathbf{M}}_j \in \mathbb{M}_{e_i}} \widehat{\mathbf{M}}_j} \quad (4)$$

Step 2. Region-based denoising and guidance. As shown in Figure 2, given the text attributes and *token maps*, we divide the overall image synthesis into several region-based denoising and guidance processes to

incorporate each attribute, similar to an ensemble of diffusion models (Kumari et al., 2023; Bar-Tal et al., 2023). More specifically, given the span e_i , the region defined by its *token map* \mathbb{M}_{e_i} , and the attribute \mathbf{a}_i , the predicted noise ϵ_t for noised generation \mathbf{x}_t at time step t is

$$\epsilon_t = \sum_i \mathbb{M}_{e_i} \cdot \epsilon_{t, e_i} = \sum_i \mathbb{M}_{e_i} \cdot D(\mathbf{x}_t, f(e_i, \mathbf{a}_i), t), \quad (5)$$

where D is the pretrained diffusion model, and $f(e_i, \mathbf{a}_i)$ is a plain text representation derived from text span e_i and attributes \mathbf{a}_i using the following process:

1. Initially, we set $f(e_i, \mathbf{a}_i) = e_i$.
2. If an embedded image is available, we convert it into a footnote \mathbf{a}_i^f using the gradient-based discrete optimization (Wen et al., 2023).
3. If footnote \mathbf{a}_i^f is available, we set $f(e_i, \mathbf{a}_i) = \mathbf{a}_i^f$.
4. The style \mathbf{a}_i^s is appended if it exists. $f(e_i, \mathbf{a}_i) = f(e_i, \mathbf{a}_i) + \text{'in the style of'} + \mathbf{a}_i^s$.
5. The closest color name (string) of font color $\hat{\mathbf{a}}_i^c$ from a predefined set \mathcal{C} is prepended. $f(e_i, \mathbf{a}_i) = \hat{\mathbf{a}}_i^c + f(e_i, \mathbf{a}_i)$. For example, $\hat{\mathbf{a}}_i^c = \text{'brown'}$ for RGB color $\mathbf{a}_i^c = [136, 68, 20]$.
6. If a texture description $\hat{\mathbf{a}}_i^t$ is provided together with the texture image \mathbf{a}_i^c , we prepend the texture description: $f(e_i, \mathbf{a}_i) = \hat{\mathbf{a}}_i^t + f(e_i, \mathbf{a}_i)$.

We use $f(e_i, \mathbf{a}_i)$ as the original plain text prompt of Step 1 for the unformatted tokens e_U . This helps us generate a coherent image, especially around region boundaries.

Guidance. By default, we use classifier-free guidance (Ho and Salimans, 2022) for each region to better match the prompt $f(e_i, \mathbf{a}_i)$. In addition, if the font color or texture fill is specified, to exploit the appearance information further, we apply gradient guidance (Ho et al., 2022; Dhariwal and Nichol; Bansal et al., 2023) on the current clean image prediction:

$$\hat{\mathbf{x}}_0 = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_t}{\sqrt{\bar{\alpha}_t}}, \quad (6)$$

where \mathbf{x}_t is the noisy image at time step t , and $\bar{\alpha}_t$ is the coefficient defined by noise scheduling strategy (Ho et al., 2020). When font color is specified, we compute an MSE loss $\mathcal{L}_{\text{color}}$ between the average color of $\hat{\mathbf{x}}$ weighted by the *token map* \mathbb{M}_{e_i} and the RGB triplet



Fig. 5: Qualitative comparison on precise color generation. We show images generated by Prompt-to-Prompt (Hertz et al., 2023), InstructPix2Pix (Brooks et al., 2023), and our method using prompts with font colors. Our method generates precise colors according to either color names or RGB values. Both baselines generate plausible but inaccurate colors given color names, while neither understands the color defined by RGB values. InstructPix2Pix tends to apply the color globally, even outside the target object.

\mathbf{a}_i^c . The gradient is calculated below,

$$\frac{d\mathcal{L}_{\text{color}}}{d\mathbf{x}_t} = \frac{d\|\sum_p (\mathbf{M}_{e_i} \cdot \hat{\mathbf{x}}_0) / \sum_p \mathbf{M}_{e_i} - \mathbf{a}_i^c\|_2^2}{\sqrt{\alpha_t} d\hat{\mathbf{x}}_0}, \quad (7)$$

where the summation is over all pixels p .

When texture fill is selected, we adopt a region-based perceptual loss (Johnson et al., 2016), $\mathcal{L}_{\text{perceptual}}$, to optimize the region specified by *token map* \mathbf{M}_{e_i} to align with the guided texture:

$$\mathcal{L}_{\text{perceptual}}(\mathbf{x}, \mathbf{y}) = \sum_j \|G_j^{\text{VGG}}(\mathbf{x}) - G_j^{\text{VGG}}(\mathbf{y})\|_2^2, \quad (8)$$

where $G_j^{\text{VGG}}(\cdot)$ computes the Gram matrix of neural features extracted from a VGG model at layer j . Specifically, we mask the image with \mathbf{M}_{e_i} and fill the hole with the texture image to ensure that the loss is only computed on the token region. The guidance gradient is then computed as:

$$\frac{d\mathcal{L}_{\text{texture}}}{d\mathbf{x}_t} = \frac{d\mathcal{L}_{\text{perceptual}}(\mathbf{M}_{e_i} \cdot \mathbf{x}_t + (1 - \mathbf{M}_{e_i}) \cdot \mathbf{a}_i^t, \mathbf{a}_i^t)}{\sqrt{\alpha_t} d\hat{\mathbf{x}}_0}, \quad (9)$$

We denote the total guidance loss as $\mathcal{L} = \lambda_{\text{texture}}\mathcal{L}_{\text{texture}} + \lambda_{\text{color}}\mathcal{L}_{\text{color}}$, where λ_{texture} and λ_{color} are hyperparameters to control the strength of the guidance. We use $\lambda_{\text{color}} = 1$ and $\lambda_{\text{texture}} = 0.2$ unless denoted otherwise. We then update \mathbf{x}_t with the

following equation:

$$\mathbf{x}_t \leftarrow \mathbf{x}_t - \lambda \cdot \mathbf{M}_{e_i} \cdot \frac{d\mathcal{L}}{d\mathbf{x}_t}, \quad (10)$$

Token reweighting with font size. Last, to re-weight the impact of the token w_j according to the font size \mathbf{a}_j^w , we modify its cross-attention maps \mathbf{m}_j . However, instead of applying direct multiplication as in Prompt-to-Prompt (Hertz et al., 2023) where $\sum_j \mathbf{a}_j^w \mathbf{m}_j \neq 1$, we find that it is critical to preserve the probability property of \mathbf{m}_j . We thus propose the following reweighting approach:

$$\hat{\mathbf{m}}_j = \frac{\mathbf{a}_j^w \exp(\mathbf{s}_j)}{\sum_k \mathbf{a}_k^w \exp(\mathbf{s}_k)}. \quad (11)$$

We can compute the token map (Equation 4) and predict the noise (Equation 5) with the reweighted attention map.

Preserve the fidelity against plain-text generation. Although our region-based method naturally maintains the layout, there is no guarantee that the details and shape of the objects are retained when no rich-text attributes or only the color is specified, as shown in Figure 15. To this end, we follow Plug-and-Play (Tumanyan et al., 2023) to inject the self-attention maps and the residual features extracted from the plain-text generation process when $t > T_{\text{pnp}}$ to improve the structure fidelity. In addition, for the regions associated with the unformatted tokens e_U ,



Fig. 6: Qualitative comparison on style control. We show images generated by Prompt-to-Prompt, InstructPix2Pix, and our method using prompts with multiple styles. Only our method can generate distinct styles for both regions.

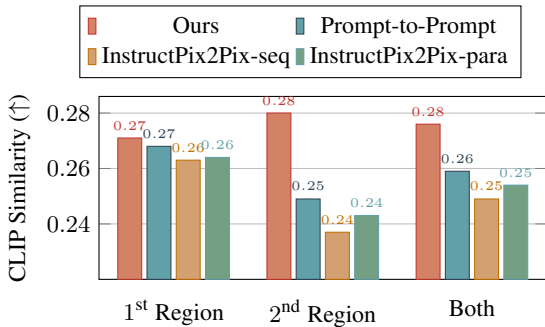


Fig. 7: Quantitative evaluation of local style control. We report the CLIP similarity between each stylized region and its region prompt. Our method achieves the best stylization.

stronger content preservation is desired. Therefore, at certain $t = T_{\text{blend}}$, we blend the noised sample $\mathbf{x}_t^{\text{plain}}$ based on the plain text into those regions:

$$\mathbf{x}_t \leftarrow \mathbf{M}_{e_U} \cdot \mathbf{x}_t^{\text{plain}} + (1 - \mathbf{M}_{e_U}) \cdot \mathbf{x}_t \quad (12)$$

4 Experimental Results

4.1 Experimental Setups

Implementation details. We use Stable Diffusion V1-5 (Rombach et al., 2022) for our main comparisons, as most baselines are built on this model. We also demonstrate the generalizability of the method using SDXL (Podell et al., 2024), which achieves better-quality and higher-resolution generation results over SD. To create the token maps, we use the cross-attention layers in all blocks, excluding the first encoder and last decoder blocks, as the attention maps in these high-resolution layers are often noisy. We discard the maps at the initial denoising steps with $T > 750$. We use $K = 15$, $\epsilon = 0.3$, $T_{\text{pnp}} = 0.3$, $T_{\text{blend}} = 0.3$, and report the results averaged from three random seeds for all quantitative experiments. More details, such as the running time, can be found in Appendix B. We notice that most hyperparameters we use in SD work well for SDXL. Notably, SDXL additionally conditions the generation on the pooled CLIP feature (Radford et al., 2021) by using adaptive normalization. For rich-text generation, we use region-specific prompts to compute these features for



Fig. 8: Qualitative results on texture transfer. We show images our method generates with certain textures specified in the rich-text prompt by texture infill. We present the plain-text generation, reference texture image, and rich-text generation for each example. Note that we resize texture images to a similar size for display. Our method can faithfully align with the texture guidance while organically combining it with the rest of the generated image.

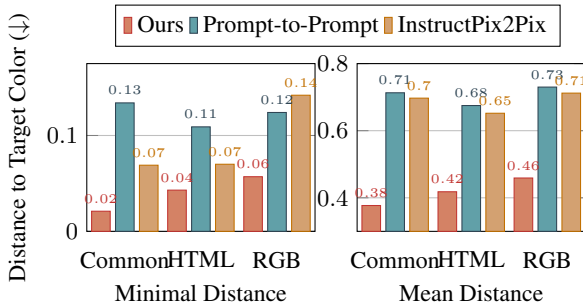


Fig. 9: Quantitative evaluation on precise color generation. Distance against target color is reported (lower is better). Our method consistently outperforms baselines.

region-based diffusion process. In addition, we use the residual features from the first layer of the decoder for the Plug-and-Play method.

Baselines. For font color and style, we quantitatively compare our method with two strong baselines, Prompt-to-Prompt (Hertz et al., 2023) and InstructPix2Pix (Brooks et al., 2023). When two instructions exist for each image in our font style experiments, we apply them in parallel (InstructPix2Pix-para) and sequential manners (InstructPix2Pix-seq). More details can be found in Appendix B. We also perform a human evaluation on these two methods in Appendix Table 1. For re-weighting token importance, we visually compare with Prompt-to-Prompt (Hertz et al., 2023) and two heuristic methods, repeating

and adding parentheses. For complex scene generation with footnotes, we also compare with Attend-and-Excite (Chefer et al., 2023).

4.2 Rich-Text Benchmark

Most existing studies on text-to-image generation (Ramesh et al., 2021; Saharia et al., 2022; Hertz et al., 2023) are only evaluated on the relatively short prompts (Bakr et al., 2023; Hu et al., 2023) with a single image style or simple colors. To evaluate the model capacity on the challenging rich-text applications, including local style rendering, precise color synthesis, and complex prompt alignment, we build a benchmark with pairs of rich-text prompts and their equivalent plain-text prompts focusing on these tasks. For each task, we also design the pipeline to automatically assess the critical aspects of generation quality. More details about the construction process is explained in Appendix B.

Footnote. To understand how well a model performs when a lengthy text prompt is given, we collect a set of long prompts using the GPT-4 model (OpenAI, 2023). We follow a two-step procedure for prompting the language model. First, we ask the model to generate global descriptions for a scene and then generate detailed captions for the objects that appear in the scene. Such a format also makes it easy to convert to rich-text prompts with footnotes. Second, we create the full plain and rich text prompts based on these global and local descriptions.

A coffee table¹ sits in front of a sofa² on a cozy carpet. A painting³ on the wall. cinematic lighting, trending on artstation, 4k, hyperrealistic, focused, extreme details.

¹A rustic wooden coffee table adorned with scented candles and many books.

²A plush sofa with a soft blanket and colorful pillows on it.

³A painting of wheat field with a cottage in the distance, close up shot, trending on artstation, HD, calm, complimentary color, realistic lighting, by Albert Bierstadt, Frederic Church.

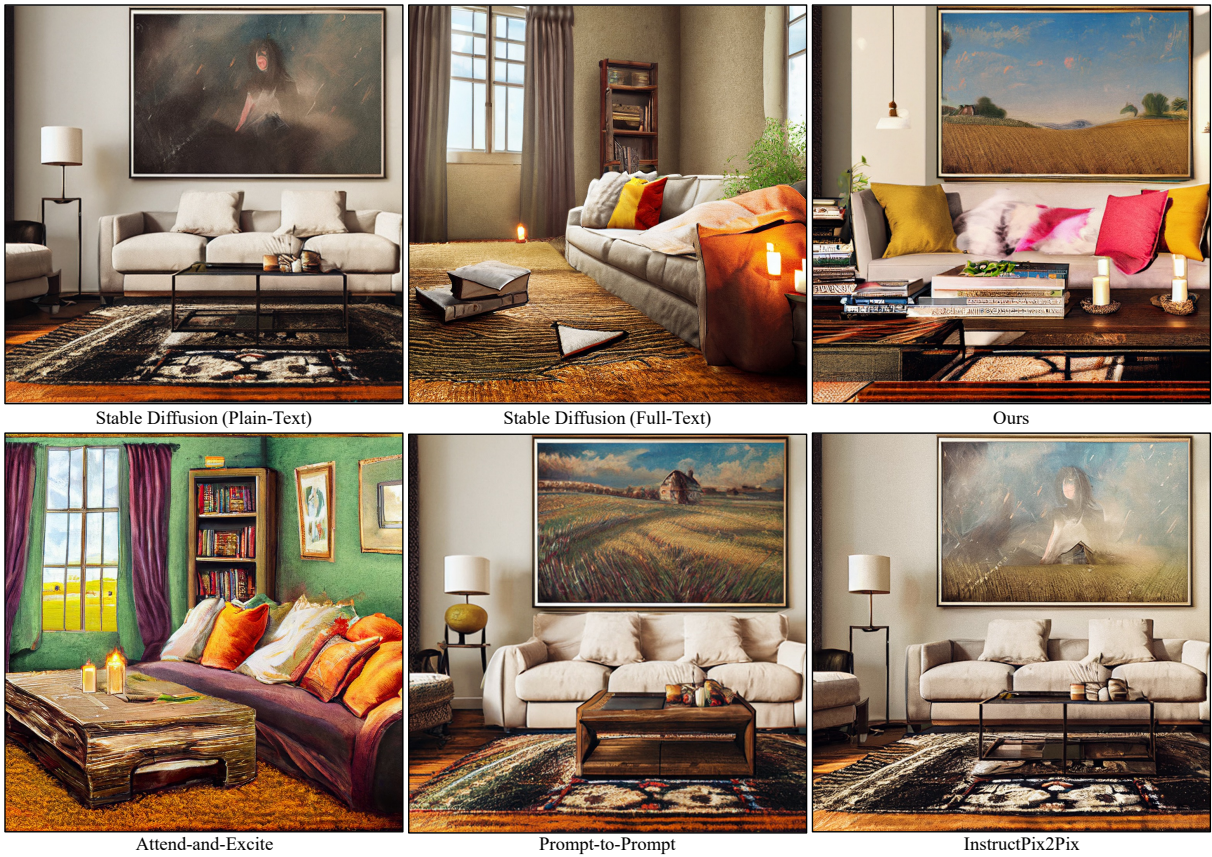


Fig. 10: Qualitative comparison on detailed description generation. We show images generated by Attend-and-Excite, Prompt-to-Prompt, InstructPix2Pix, and our method using complex prompts. Our method is the only one that can generate all the details faithfully.

In the first step, to control the quality of the collected prompts, we manually create a few hierarchical descriptions as the in-context examples. In addition, we find that the model can hallucinate in the local object descriptions by describing the objects not shown in the scene. Also, some descriptions can include non-visual features like the flavor of the food. Therefore, after collecting the results from GPT-4, we manually filter and edit the low-quality text prompts.

We collect 100 prompts in total. In the second step, we use GPT-4 to produce single, long, plain-text prompts. We also create scripts to produce rich-text prompts from these hierarchical prompts. We provide the prompt template we used for both stages in Appendix B. We summarize the statistics of the plain-text prompts in Table 1. Note that the average length of the collected prompts is around $4\times$ longer than the

prior arts and closer to the concurrent work (Wu et al., 2023).

To evaluate the alignment between the generated images and the lengthy text prompts, we follow the process in TIFA (Hu et al., 2023) to sample question-answer pairs according to the scene and object prompts using language models. For example, given the scene prompt “a nightstand next to a bed in the bedroom,” the question-and-answer pair could be “Q: Is there a bed? A: Yes.” Through this process, we collect 2921 such pairs. To evaluate a generated image, either a VQA model or human annotators can answer the question based on the generated image and check whether it is consistent with the ground truth answer.

We manually modify the questions based on the object prompts to reflect that the object is part of the



Fig. 11: Qualitative comparison on token reweighting. We show images generated by our method and Prompt-to-Prompt using token weight of 13 for ‘mushrooms’. Prompt-to-Prompt suffers from artifacts due to the large weight. Heuristic methods like repeating and parenthesis do not work well.



Fig. 12: Qualitative results on customized concept generation. We underline the modified concepts in the prompts and display the reference image on the top right of each rich-text generation result. Our method is able to synthesize the image with the object according to the reference image without changing the overall structure.

scene. For example, for the object prompt “a nightstand with some books,” we modify the question “Are there books?” to “Are there books on the nightstand?” We manually review and drop low-quality, irrelevant, or repeated questions. For example, we drop the pair “Q: What is in the bedroom? A: A bed.” since “a nightstand” or other reasonable objects could be the correct answer. Instead, we prefer to ask the more concrete question, “Is there a bed in the bedroom?” After the manual processing, we obtain 1974 question-answer pairs in total.

Font color. We divide colors into three categories to evaluate a method’s capacity to understand and generate a specific color. The *Common color* category contains 17 standard names, such as “red”, “yellow”, and “pink”. The *HTML color* names are selected from the web color names¹ used for website design, such as “sky blue”, “lime green”, and “violet purple”. The *RGB color* category contains 50 randomly sampled RGB triplets to be used as “color of RGB

values [128, 128, 128]”. To create a complete prompt, we use 12 objects exhibiting different colors, such as “flower”, “gem”, and “house”. This gives us a total of 1,200 prompts. We evaluate color accuracy by computing the mean L2 distance between the region and target RGB values. We also compute the minimal L2 distance as sometimes the object should contain other colors for fidelity, e.g., the “black tires” of a “yellow car”.

Font style. We come up with prompts featuring two objects and styles to evaluate the ability to generate accurate local styles. We create combinations using 7 popular styles and 10 objects, resulting in 420 prompts. To automatically evaluate the performance, we compute CLIP scores (Radford et al., 2021) for each local region to evaluate the stylization quality. Specifically, for each generated image, we mask it by the token maps of each object and attach the masked output to a black background. Then, we compute the CLIP score using the region-specific prompt. For example, for the prompt “a lighthouse (Cyberpunk) among the turbulent waves (Ukiyo-e)”, the local

¹https://simple.wikipedia.org/wiki/Web_color

Table 1: The statistics of text prompts collected in the previous benchmarks or adopted in the previous studies.

Benchmark or paper	Number of words			Number of tokens		
	Avg	Max	Min	Avg	Max	Min
DALLE (Ramesh et al., 2021)	15.8	36	6	18.35	42	9
Prompt-to-Prompt (Hertz et al., 2023)	7.43	12	2	8.56	14	3
DrawBench (Saharia et al., 2022)	11.37	51	1	14.03	57	3
HRS Bench (Bakr et al., 2023)	12.85	36	1	14.94	42	3
TIFA (Hu et al., 2023)	11.46	67	3	11.63	82	3
ParaImage-Big (Wu et al., 2023)	132.9	-	-	-	-	-
ParaImage-Small (Wu et al., 2023)	70.6	-	-	-	-	-
Ours	46.56	68	19	57.68	90	22

Table 2: Quantitative evaluation of long prompt generation. We report the percentage of the time that the VQA output is aligned with the answer to a question regarding the generation. Our method consistently improves Stable Diffusion models.

	Stable Diffusion	Stable Diffusion XL
Plain-text generation	72.86 \pm 1.13	79.05 \pm 0.41
Rich-text generation	73.24 \pm 1.00	81.08 \pm 0.66

CLIP score of the lighthouse region is measured by comparing its similarity with the prompt “lighthouse in the style of cyberpunk.” In this example, we refer to “lighthouse” as the first region and “waves” as the second region.

4.3 Quantitative Comparison

We report the local CLIP scores computed by a ViT-B/32 model in Figure 7. Our method achieves the best overall CLIP score compared to the two baselines. This demonstrates the advantage of our region-based diffusion method for localized stylization. To further understand the each model’s capacity for generating multiple styles, we report the metric on each region. Prompt-to-Prompt and InstructPix2Pix-para achieve a decent score on the 1st Region, i.e., the region first occurs in the sentence. However, they often fail to fulfill the style in the 2nd Region. We conjecture that the Stable Diffusion model tends to generate a uniform style for the entire image, which can be attributed to single-style training images. Furthermore, InstructPix2Pix-seq performs the worst in 2nd Region. This is because the first instruction contains

no information about the second region, and the second region’s content could be compromised when we apply the first instruction.

We show quantitative results of precise color generation in Figure 9. The distance of *HTML color* is generally the lowest for baseline methods, as they provide the most interpretable textual information for text encoders. This aligns with our expectation that the diffusion model can handle simple color names, whereas they struggle to handle the RGB triplet. Our rich-text-to-image generation method consistently improves on the three categories and two metrics over the baselines.

In Table 2, we follow TIFA (Hu et al., 2023) to evaluate the long prompt generation using either Stable Diffusion or Stable Diffusion XL as the base models. Specifically, we generate 5 images for each text prompt using different random seeds and use the *mplug-large* model (Li et al., 2022) to generate the answers to each benchmark question. We report how often the generated answers are consistent with the benchmark answers by averaging across different seeds. We also report the standard deviation. Our rich-text method consistently improves the generation quality with different base models when a lengthy prompt is given. Moreover, we find that the Stable Diffusion XL significantly outperforms the Stable Diffusion model.

4.4 Visual Comparison

Precise color generation. We show qualitative comparison on precise color generation in Figure 5. InstructPix2Pix (Brooks et al., 2023) is prone to create global color effects rather than accurate local control. For example, in the flower results, both the vase

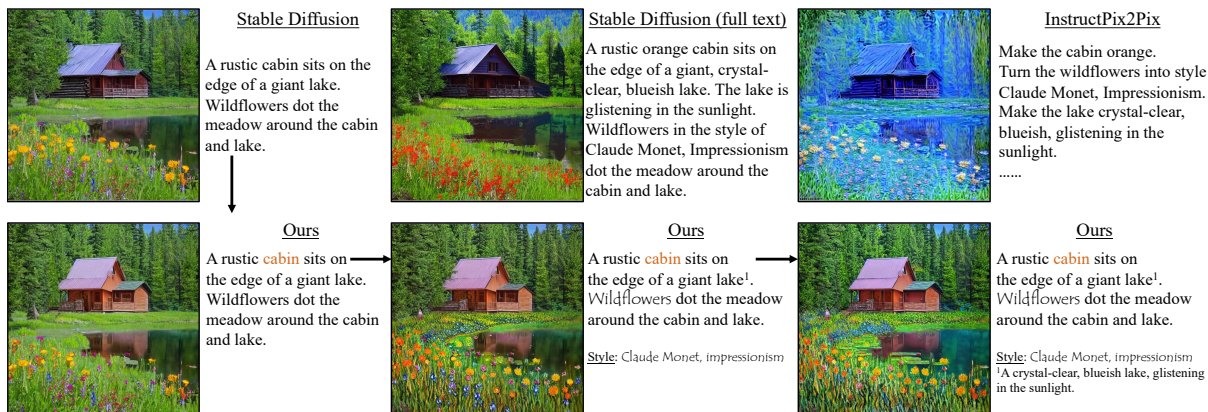


Fig. 13: Our workflow. (top left) A user begins with an initial plain-text prompt and wishes to refine the scene by specifying the color, details, and styles. (top center) Naively inputting the whole description in plain text does not work. (top right) InstructPix2Pix (Brooks et al., 2023) fails to make accurate editing. (bottom) Our method supports precise refinement with region-constrained diffusion processes. Moreover, our framework can naturally be integrated into a rich text editor, enabling a tight, streamlined UI.

and background change to the target colors. Prompt-to-Prompt (Hertz et al., 2023) provides more precise control over the target region. However, neither Prompt-to-Prompt nor InstructPix2Pix can generate precise colors. In contrast, our method can generate precise colors for all categories and prompts.

Local style generation. Figure 6 visually compares local style generation. When applying InstructPix2Pix-seq, the style in the first instruction dominates the entire image and undermines the second region. Figure 13 in the Appendix shows this cannot be fully resolved using different hyperparameters of classifier-free guidance. Similar to our observation in the quantitative evaluation, our baselines tend to generate the image in a globally uniform style instead of distinct local styles for each region. In contrast, our method synthesizes the correct styles for both regions. One may suggest independently applying baselines with two stylization processes and composing the results using token maps. However, Figure 12 (Appendix) shows that such methods generate artifacts on the region boundaries.

Texture-guided generation. The results in Figure 8 demonstrate our method’s ability to seamlessly incorporate texture information specified by a reference image into image generation results while preserving the other information of the textual descriptions, such as the overall shape and geometry. For instance, the generated cake aligns with the visual style of the salmon in the texture guidance while

preserving the overall shape and structure of the original cake. The other objects, such as the cat, remain unchanged. Also, the chef’s apron exhibits a clear artistic style inspired by the reference while retaining the size and shape of the original apron in plain-text generation.

Complex scene generation. Figure 10 shows comparisons on complex scene generation. Attend-and-Excite (Chefer et al., 2023) uses the tokens missing in the full-text generation result as input to fix the missing objects, like the coffee table and carpet in the living room example. However, it still fails to generate all the details correctly, e.g., the books, the painting, and the blanket. Prompt-to-Prompt (Hertz et al., 2023) and InstructPix2Pix (Brooks et al., 2023) can edit the painting accordingly, but many objects, like the colorful pillows and stuff on the table, are still missing. In contrast, our method faithfully synthesizes all these details described in the target region.

Token importance control. Figure 11 shows the qualitative comparison on token reweighting. When using a large weight for ‘mushroom,’ Prompt-to-Prompt generates clear artifacts as it modifies the attention probabilities to be unbounded and creates out-of-distribution intermediate features. Heuristic methods fail when adding more mushrooms, while our method generates more mushrooms and preserves the quality. More results of different font sizes and target tokens are shown in Figures 23 - 25 in the Appendix.

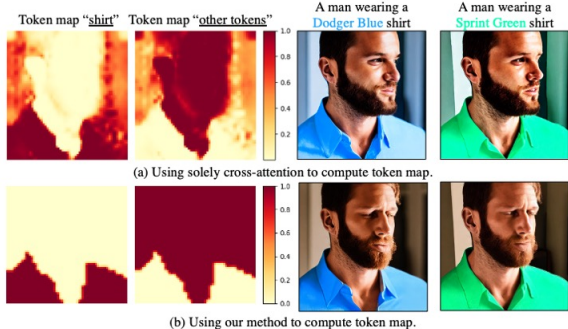


Fig. 14: Ablation of token maps. Using solely cross-attention maps to create token maps leads to inaccurate segmentations, causing the background to be colored in an undesired way.

Customized concept generation. We show reference-driven image generation in Figure 12. Our method can generate the customized concepts specified by the embedded images without making undesired changes to the plain-text generation. For example, only the backpack design is modified for the example on the left, and the dog’s breeds are changed (instead of the pose) on the right.

More visual generation results. We show more rich-text generation results using Stable Diffusion XL model (Podell et al., 2024) with a focus on multiple rich-text attributes in Figure 16. We show that these rich-text formatting options provide great flexibility in customizing the text-to-image generation process using our region-based diffusion and guidance. Our region-based injection also helps preserve the fidelity with respect to plain-text results. For example, the font color supports accurate controlling the color of dressing and looking, font style allows freely rendering artistic styles in a certain region, embedded image insert personal reference image into the generation seamlessly, and so on. In addition, combining these options can further boost the controllability. For example, with the font color specifying the color of the hat, a footnote can be further used to describe the type of hat. Also, powered by the improved generation capacity of SDXL, we also show the generation of large and complex scenes, such as a snowy forest and a city square, with many visual elements indicated by the rich-text attributes. In conclusion, these results highlight that our method and design choices are agnostic to the base model and can be readily generalized to different pre-trained diffusion models.

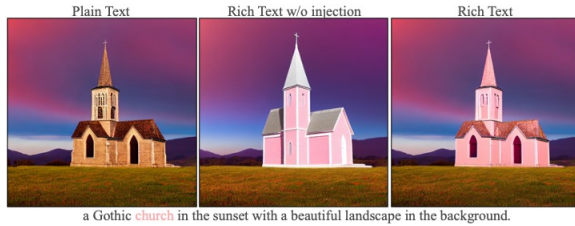


Fig. 15: Ablation of injection method. We show images generated based on plain text and rich text with or without injection methods. Injecting features and noised samples help preserve the structure of the church and unformatted token regions.

Real image editing. We also explore using rich text as an alternative to plain text in real image editing. Given a real image, we manually caption the image with text descriptions and use the off-the-shelf diffusion inversion methods to obtain the noise latent that reconstructs the image (Huberman-Spiegelglas et al., 2023; Wu and De la Torre, 2022; Ju et al., 2024). We find that our method is generally robust to different inversion methods. In the experiments, we use the edit-friendly DDPM (Huberman-Spiegelglas et al., 2023) as our default inversion method. Unlike the image generation setting, the source image is already available here. We thus use the state-of-the-art grounded segmentation methods (Kirillov et al., 2023; Liu et al., 2023; Ren et al., 2024) to produce the token maps. This is generally more robust than the cross-attention-based methods since the manually created text prompts for inversion may not accurately describe the image as the model expects.

We adopt the same configuration we used in image generation, including the region-based guidance denoising and injection methods. We compare with existing editing methods, including inversion and editing with plain text (Huberman-Spiegelglas et al., 2023), InstructPix2Pix (Brooks et al., 2023), and Plug-and-Play (Tumanyan et al., 2023).

As shown in Figure 17, the rich text allows editing the regions’ color, style, and content with accurate controllability. Specifically, all existing methods struggle with preserving the details in the non-edited regions. For example, in the Van Gogh portrait example, all the baselines change the styles of areas other than the hat. Instead, our injection method and the region information in the token masks greatly improve the fidelity in these regions. In addition, as we allow direct usage of the color RGB information, we edit the dress color (in the example of the second row) more accurately.



A cat eating food¹ in a bowl on the table in the kitchen.
¹The round dry cat food.



A wizard holding a wand¹ and wearing a robe and wizard hat.
¹A wooden wand made of tree branch and a gem.



A girl with long hair sitting in a cafe, by a table with coffee¹ on it.
¹Ceramic cup with intricate design, a dance of earthy browns and delicate gold. Dark, velvety latte in it.



A close-up photo of a corgi wearing a hat¹, beach and ocean in the background.
Style: Impressionism.
¹A lady's hat.



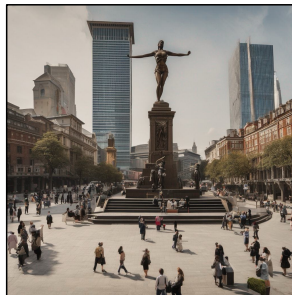
A watercolor painting of the detective duck wearing a sheriff uniform and holding a vintage handgun.
¹A dark green, washed jacket.
²A beautiful flower bouquet made of pink roses.



A male Lycan wolf wearing jet trench coat playing a guitar in a dark forest¹. high fantasy, digital painting, concept art.
¹A forest with dead woods and flying ravens
Style: Steam Punk.



A cozy cabin¹ nestled in a snowy forest², and a snowman³ stands in the yard.
¹A charming wooden cabin with Christmas decoration, warm light coming from windows.
²Towering evergreen trees covered in a thick layer of pristine snow.
³Cute snowman wearing a carrot nose, coal eyes, and a colorful scarf.



A bustling city square¹ with a large bronze statue². Skyscrapers³ tower above.
¹City square paved with cobblestones, hosting a farmers' market with colorful stalls.
²A majestic statue of a city founder holding a scroll, pigeons perched on his outstretched arm.
³Modern glass and steel skyscrapers reflecting the midday sun, bustling with corporate activity.

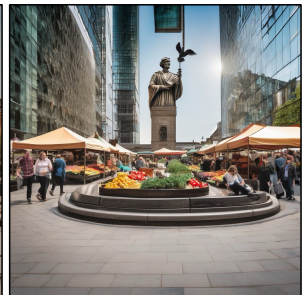
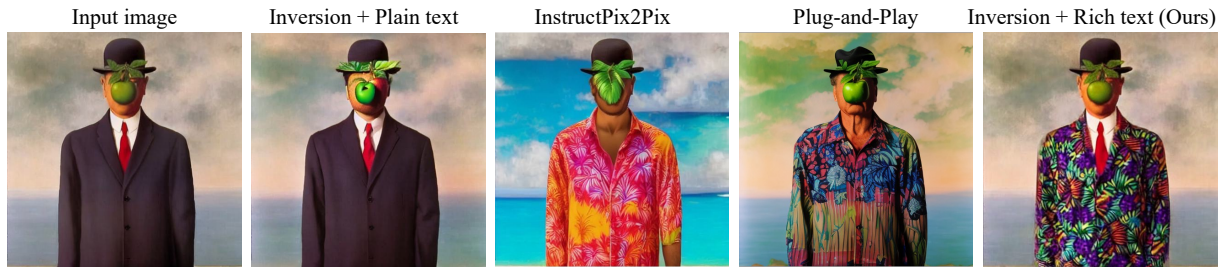


Fig. 16: More rich-text guided image generation results. We show more visual generation results by SDXL using rich texts with *multiple* text attributes, with a highlight on generating complex scenes.



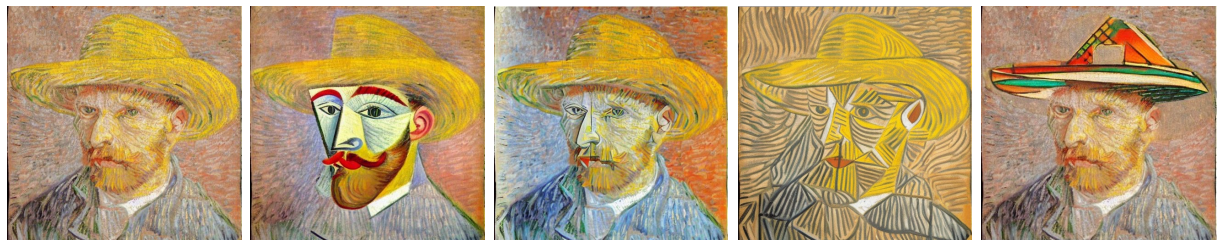
A man in suit¹ with a green apple on his face.
¹A colorful Hawaiian shirt



A stylish woman in suit and a dress walking down a Tokyo street, high quality.



Close-up of a man wearing a helmet. Snow¹ in the background.
¹A wheat field



A portrait of a man with a beard wearing a hat
 Styles: Picasso, cubism

Fig. 17: Image editing with rich text. We show that with the off-the-shelf inversion techniques, our methods easily slot in as an alternative to plain-text editing. Using rich text with our region-based guidance and diffusion allows more precise control in editing real images than existing methods.

Interactive editing. In Figure 13, we showcase a sample workflow to illustrate our method’s interactive strength and editing capacity over Instruct-Pix2Pix (Brooks et al., 2023).

4.5 Ablation Study

Generating token maps solely from cross-attention. The other straightforward way to create token maps is

to use cross-attention maps directly. To ablate this, we first take the average of cross-attention maps across heads, layers, and time steps and then take the maximum across tokens. Finally, we apply softmax across all the spans to normalize the token maps. However, as shown by the example in Figure 14, since the prompt has no correspondence with the background, the token map of “shirt” also covers partial background regions. Note that simple thresholding is ineffective as some

regions still have high values, e.g., the right shoulder. As a result, the target color *bleeds* into the background. Our methods obtain more accurate token maps and, consequently, more precise colorization.

Ablation of the injection methods. To demonstrate the effectiveness of our injection method, we compare image generation with and without it in Figure 15. In the font color example, we show that applying the injection effectively preserves the shape and details of the target church and the structure of the sunset in the background. In the footnote example, we show that the injection keeps the look of the black door and the color of the floor.

5 Discussion and Limitations

In this paper, we have expanded the controllability of text-to-image models by incorporating rich-text attributes as the input. We have demonstrated the potential for generating and editing images with local styles, precise colors, texture guidance, different token importance, reference images, and complex descriptions. Nevertheless, numerous formatting options remain unexplored, such as bold/italic, hyperlinks, spacing, and bullets/numbering. Also, there are multiple ways to use the same formatting options. For example, one can use font style to characterize the shape of the objects. We hope this paper encourages further exploration of integrating accessible user interfaces into text-based content creation tasks, even beyond images.

References

- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML, pp. 8821–8831 (2021)
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
- Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: CVPR (2023)
- Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
- Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: ECCV (2022)
- Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
- Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
- Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: ICLR (2023)
- Colorado State University, T.A.P.: tutorial: Rich Text Format (RTF) from Microsoft Word - The ACCESS Project (2012). http://accessproject.colostate.edu/udl/modules/word/tut_rtf.cfm
- Witten, I.H., Bainbridge, D., Nichols, D.M.: How to Build a Digital Library, (2009)
- Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
- Ge, S., Park, T., Zhu, J.-Y., Huang, J.-B.: Expressive text-to-image generation with rich text. In: ICCV (2023)
- Zhu, X., Goldberg, A.B., Eldawy, M., Dyer, C.R., Strock, B.: A text-to-picture synthesis system for augmenting communication. In: AAAI (2007)
- Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R.: Generating images from captions with attention. In: ICLR (2016)
- Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
- Byeon, M., Park, B., Kim, H., Lee, S., Baek, W., Kim, S.: COYO-700M: Image-Text Pair Dataset. <https://github.com/kakaobrain/coyo-dataset> (2022)
- Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
- Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
- Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi,

- M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research* **23**(47), 1–33 (2022)
- Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications* (2021)
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022)
- Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: *ICML* (2022)
- Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research* (2022)
- Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704* (2023)
- Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217* (2022)
- Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515* (2023)
- Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: *CVPR* (2023)
- Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.-Y.: Multi-concept customization of text-to-image diffusion. In: *CVPR* (2023)
- Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: Spatext: Spatio-textual representation for controllable image generation. In: *CVPR* (2023)
- Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation, 7623–7633 (2023)
- Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: *CVPR* (2023)
- Ma, Y., He, Y., Cun, X., Wang, X., Shan, Y., Li, X., Chen, Q.: Follow Your Pose: Pose-Guided Text-to-Video Generation using Pose-Free Videos (2023)
- Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: *CVPR* (2023)
- Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., Ermon, S.: Sdedit: Image synthesis and editing with stochastic differential equations. In: *ICLR* (2022)
- Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. In: *ICCV* (2021)
- Parmar, G., Singh, K.K., Zhang, R., Li, Y., Lu, J., Zhu, J.-Y.: Zero-shot image-to-image translation. In: *SIGGRAPH* (2023)
- Bansal, A., Chu, H.-M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. *arXiv preprint arXiv:2302.07121* (2023)
- Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. *arXiv preprint arXiv:2206.02779* (2022)
- Jiménez, Á.B.: Mixture of diffusers for scene composition and high resolution image generation. *arXiv preprint arXiv:2302.02412* (2023)
- Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. In: *ICML* (2023)
- Sarukkai, V., Li, L., Ma, A., R’e, C., Fatahalian, K.: Collage diffusion. *ArXiv abs/2303.00262* (2023)
- Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.-Y.: Diffcollage: Parallel generation of large content with diffusion models. (2023)
- Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: *ICCV* (2023)
- Phung, Q., Ge, S., Huang, J.-B.: Grounded text-to-image synthesis with attention refocusing. In: *CVPR* (2024)
- Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431* (2023)
- Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: *ICLR* (2023)
- Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: *NeurIPS* (2023)

- (2022)
- Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)
- Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: CVPR (2018)
- Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
- Couairon, G., Careil, M., Cord, M., Lathuiliere, S., Verbeek, J.: Zero-shot spatial layout conditioning for text-to-image diffusion models. In: ICCV (2023)
- Park, M., Yun, J., Choi, S., Choo, J.: Learning to generate semantic layouts for higher text-image correspondence in text-to-image synthesis. In: ICCV (2023)
- Feng, W., Zhu, W., Fu, T.-j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. (2024)
- Qu, L., Wu, S., Fei, H., Nie, L., Chua, T.-S.: Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In: ACM MM (2023)
- Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV (2023)
- Tseng, H.-Y., Li, Q., Kim, C., Alsisan, S., Huang, J.-B., Kopf, J.: Consistent view synthesis with pose-guided diffusion models. In: CVPR (2023)
- Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel View Synthesis with Diffusion Models (2022)
- Patashnik, O., Garibi, D., Azuri, I., Averbuch-Elor, H., Cohen-Or, D.: Localizing object-level shape variations with text-to-image diffusion models. In: ICCV (2023)
- Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. arXiv:2303.04761 (2023)
- QI, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. In: ICCV (2023)
- Ceylan, D., Huang, C.-H., Mitra, N.J.: Pix2video: Video editing using image diffusion. arXiv:2303.12688 (2023)
- Ma, W.-D.K., Lewis, J., Kleijn, W.B., Leung, T.: Directed diffusion: Direct control of object placement through attention guidance. arXiv preprint arXiv:2302.13153 (2023)
- Meng, Y., Wu, W., Wang, F., Li, X., Nie, P., Yin, F., Li, M., Han, Q., Sun, X., Li, J.: Glyce: Glyph-vectors for chinese character representations. NeurIPS **32** (2019)
- Sun, Z., Li, X., Sun, X., Meng, Y., Ao, X., He, Q., Wu, F., Li, J.: Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. ACL (2021)
- Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: SIGKDD (2020)
- Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.: Dit: Self-supervised pre-training for document image transformer. In: MM (2022)
- Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016)
- Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
- Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: CVPR (2017)
- Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications **21**(5), 34–41 (2001) <https://doi.org/10.1109/38.946629>
- Tai, Y.-W., Jia, J., Tang, C.-K.: Local color transfer via probabilistic segmentation by expectation-maximization. In: CVPR (2005)
- Xu, L., Yan, Q., Jia, J.: A sparse control model for image and video editing. ACM Transactions on Graphics (TOG) **32**, 1–10 (2013)
- Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. SIG (2004)
- Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
- Zhang, R., Zhu, J.-Y., Isola, P., Geng, X., Lin, A.S., Yu, T., Efros, A.A.: Real-time user-guided image colorization with learned deep priors. ACM Transactions on Graphics (TOG) **9**(4) (2017)
- Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., *et al.*: Improving image generation with better captions. Computer Science. <https://cdn.openai.com/papers/dall-e-3.pdf> **2**(3), 8 (2023)
- Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. arXiv preprint arXiv:2402.03290 (2024)
- Wu, W., Li, Z., He, Y., Shou, M.Z., Shen, C., Cheng,

- L., Li, Y., Gao, T., Zhang, D., Wang, Z.: Paragraph-to-image generation with information-enriched diffusion model. arXiv preprint arXiv:2311.14284 (2023)
- Bakr, E.M., Sun, P., Shen, X., Khan, F.F., Li, L.E., Elhoseiny, M.: Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In: ICCV (2023)
- Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: ICCV (2023)
- Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In: NeurIPS (2024)
- Patel, M., Gokhale, T., Baral, C., Yang, Y.: Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models. In: AAAI (2024)
- Zhao, L., Zhao, T., Lin, Z., Ning, X., Dai, G., Yang, H., Wang, Y.: Flasheval: Towards fast and accurate evaluation of text-to-image diffusion generative models. arXiv preprint arXiv:2403.16379 (2024)
- Sahuguet, A., Azavant, F.: Wysiwyg web wrapper factory (w4f) (1999)
- Litt, G., Lim, S., Kleppmann, M., Hardenberg, P.: Peritext: A crdt for collaborative rich text editing. Proceedings of the ACM on Human-Computer Interaction (PACMHCI) (2022)
- Ignat, C.-L., André, L., Oster, G.: Enhancing rich content wikis with real-time collaboration. Concurrency and Computation: Practice and Experience **33**(8), 4110 (2021)
- Agrawala, M.: Unpredictable Black Boxes are Terrible Interfaces. <https://magrawala.substack.com/p/unpredictable-black-boxes-are-terrible> (2023)
- Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
- Johnson, J., Karpathy, A., Fei-Fei, L.: Denscap: Fully convolutional localization networks for dense captioning. In: CVPR (2016)
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763 (2021). PMLR
- Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023). <https://openreview.net/forum?id=NAQvF08TcyG>
- Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
- Li, D., Li, J., Hoi, S.C.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. arXiv preprint arXiv:2305.14720 (2023)
- Jia, X., Zhao, Y., Chan, K.C., Li, Y., Zhang, H., Gong, B., Hou, T., Wang, H., Su, Y.-C.: Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642 (2023)
- Chen, W., Hu, H., Li, Y., Rui, N., Jia, X., Chang, M.-W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. arXiv preprint arXiv:2304.00186 (2023)
- Gal, R., Arar, M., Atzmon, Y., Bermano, A.H., Chechik, G., Cohen-Or, D.: Encoder-based domain tuning for fast personalization of text-to-image models. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023)
- Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023)
- Tang, R., Pandey, A., Jiang, Z., Yang, G., Kumar, K.V.S.M., Lin, J., Ture, F.: What the daam: Interpreting stable diffusion using cross attention. ArXiv **abs/2210.04885** (2022)
- Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR (2023)
- Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8), 888–905 (2000) <https://doi.org/10.1109/34.868688>
- Von Luxburg, U.: A tutorial on spectral clustering. Statistics and computing **17**, 395–416 (2007)
- Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., Goldstein, T.: Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In: NeurIPS (2023)
- Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
- Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS
- Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses

- for real-time style transfer and super-resolution. In: ECCV (2016)
- Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. In: ICLR (2024)
- OpenAI: GPT-4 Technical Report (2023)
- Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., *et al.*: mplug: Effective and efficient vision-language learning by cross-modal skip-connections. In: EMNLP (2022)
- Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly ddpm noise space: Inversion and manipulations. arXiv preprint arXiv:2304.06140 (2023)
- Wu, C.H., Torre, F.: Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. arXiv preprint arXiv:2210.05559 (2022)
- Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In: ICLR (2024)
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment anything. In: ICCV (2023)
- Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., *et al.*: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
- Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks (2024)

In this appendix, we provide additional experimental results and details. In section A, we show the images generated by our model, Attend-and-Excite (Chefer et al., 2023), Prompt-to-Prompt (Hertz et al., 2023), and InstructPix2Pix (Brooks et al., 2023) with various RGB colors, local styles, and detailed descriptions via footnotes. In section B, we provide additional details on the implementation and evaluation.

A Additional Results

In this section, we first show additional results of rich-text-to-image generation on complex scene synthesis (Figures 15, 16, and 17), precise color rendering (Figures 18, 19, and 20), local style control (Figures 21 and 22), and explicit token re-weighting (Figure 23, 24, and 25). We also show an ablation study of the averaging and maximizing operations across tokens to obtain token maps in Figure 26. We present additional results compared with a composition-based baseline in Figure 27. Last, we show an ablation of the hyperparameters of our baseline method InstructPix2Pix (Brooks et al., 2023) on the local style generation application in Figure 28.

A car¹ driving on the road. A bicycle² nearby a tree³. A cityscape⁴ in the background.

¹A sleek sports car gleams on the road in the sunlight, with its aerodynamic curves and polished finish catching the light. ²A bicycle with rusted frame and worn tires.

³A dead tree with a few red apples on it. ⁴A bustling Hongkong cityscape with towering skyscrapers.



Fig. 18: Additional results of the footnote. We show the generation from a complex description of a garden. Note that all the methods except for ours fail to generate accurate details of the mansion and fountain as described.

A lush garden¹ with a fountain². A grand mansion³ in the background.

¹A garden is full of vibrant colors with a variety of flowers.

²A fountain made of white marble with multiple tiers. The tiers are intricately carved with various designs.

³An impressive two-story mansion with a royal exterior, white columns, and tile-made roof. The mansion has numerous windows, each adorned with white curtains.



Stable Diffusion (Plain-Text)



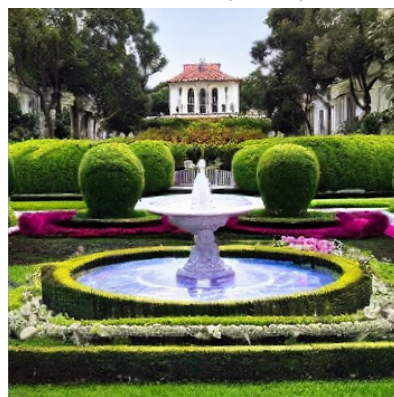
Stable Diffusion (Full-Text)



Ours



Attend-and-Excite



Prompt-to-Prompt



InstructPix2Pix

Fig. 19: Additional results of the footnote. We show the generation from a complex description of a garden. Note that all the methods except for ours fail to generate accurate details of the mansion and fountain as described.

A small chair¹ sits in front of a table² on the wooden floor. There is a bookshelf³ nearby the window⁴.

¹A black leather office chair with a high backrest and adjustable arms.

²A large wooden desk with a stack of books on top of it.

³A bookshelf filled with colorful books and binders.

⁴A window overlooks a stunning natural landscape of snow mountains.



Stable Diffusion (Plain-Text)



Stable Diffusion (Full-Text)



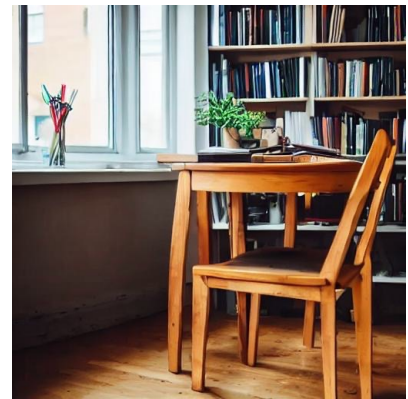
Ours



Attend-and-Excite



Prompt-to-Prompt



InstructPix2Pix

Fig. 20: Additional results of the footnote. We show the generation from a complex description of an office. Note that all the methods except ours fail to generate accurate window overlooks and colorful binders as described.



Fig. 21: Additional results of the font color. We show the generation of different objects with colors from the *Common category*. Prompt-to-Prompt has a large failure rate of respecting the given color name, while InstructPix2Pix tends to color the background and irrelevant objects.

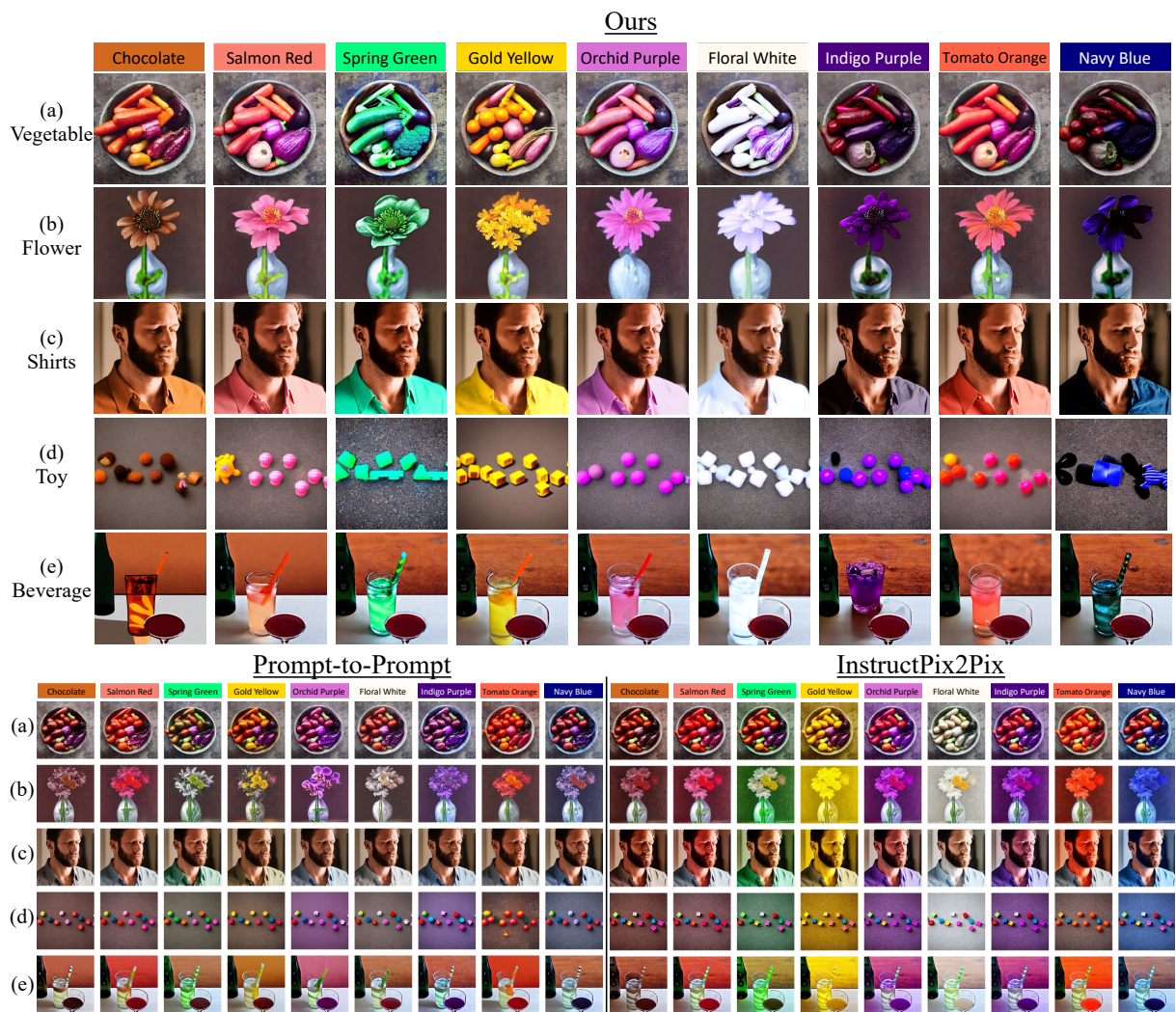


Fig. 22: Additional results of the font color. We show the generation of different objects with colors from the *HTML* category. Both methods fail to generate the precise color, and InstructPix2Pix tends to color the background and irrelevant objects.



Fig. 23: Additional results of the font color. We show the generation of different objects with colors from the *RGB* category. Both baseline methods cannot interpret the RGB values correctly.

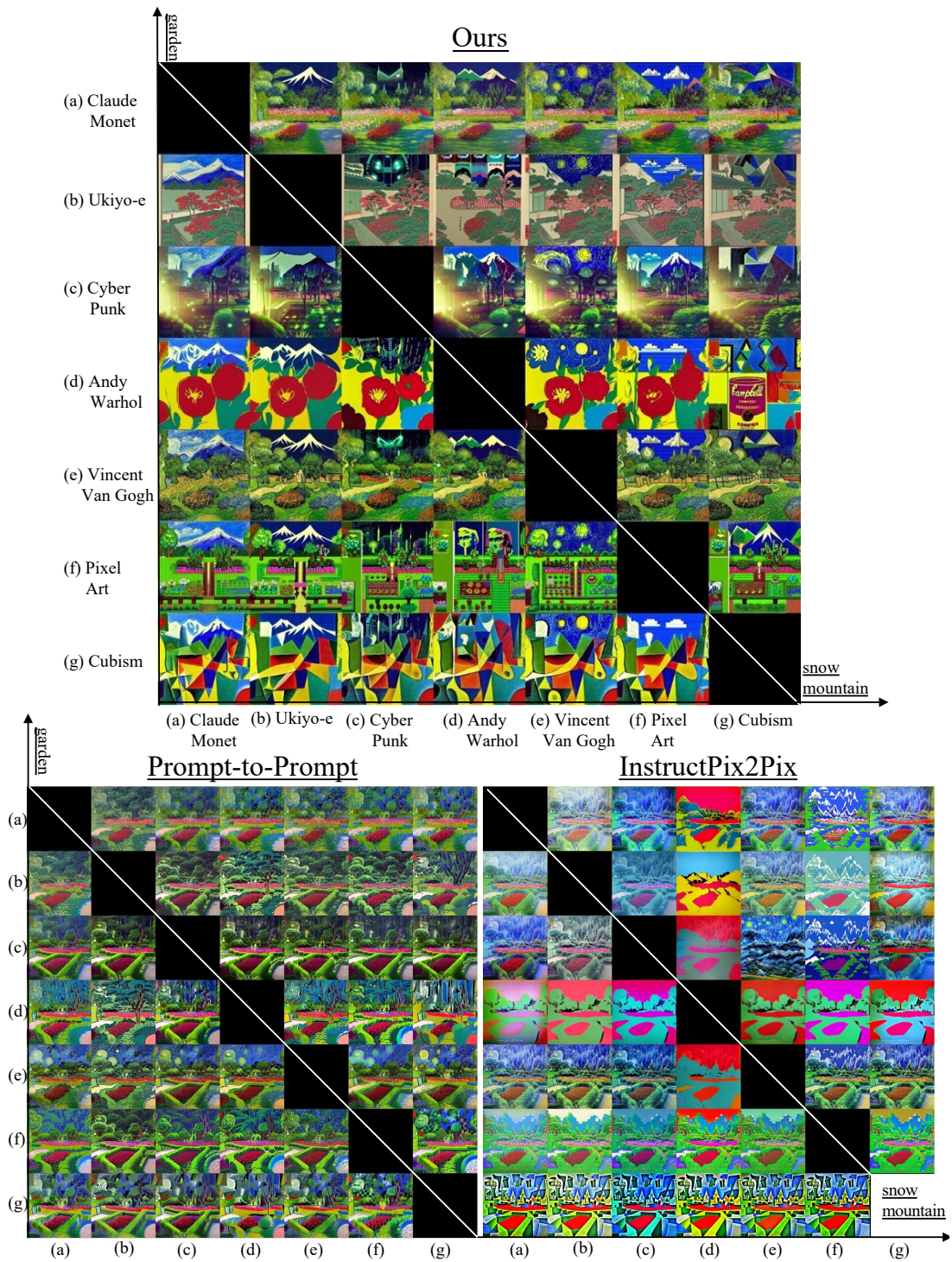


Fig. 24: Additional results of the font style. We show images generated with different style combinations and prompt “a beautiful garden in front of a snow mountain”. Each row contains “snow mountain” in 7 styles, and each column contains “garden” in 7 styles. Only our method can generate distinct styles for both objects.

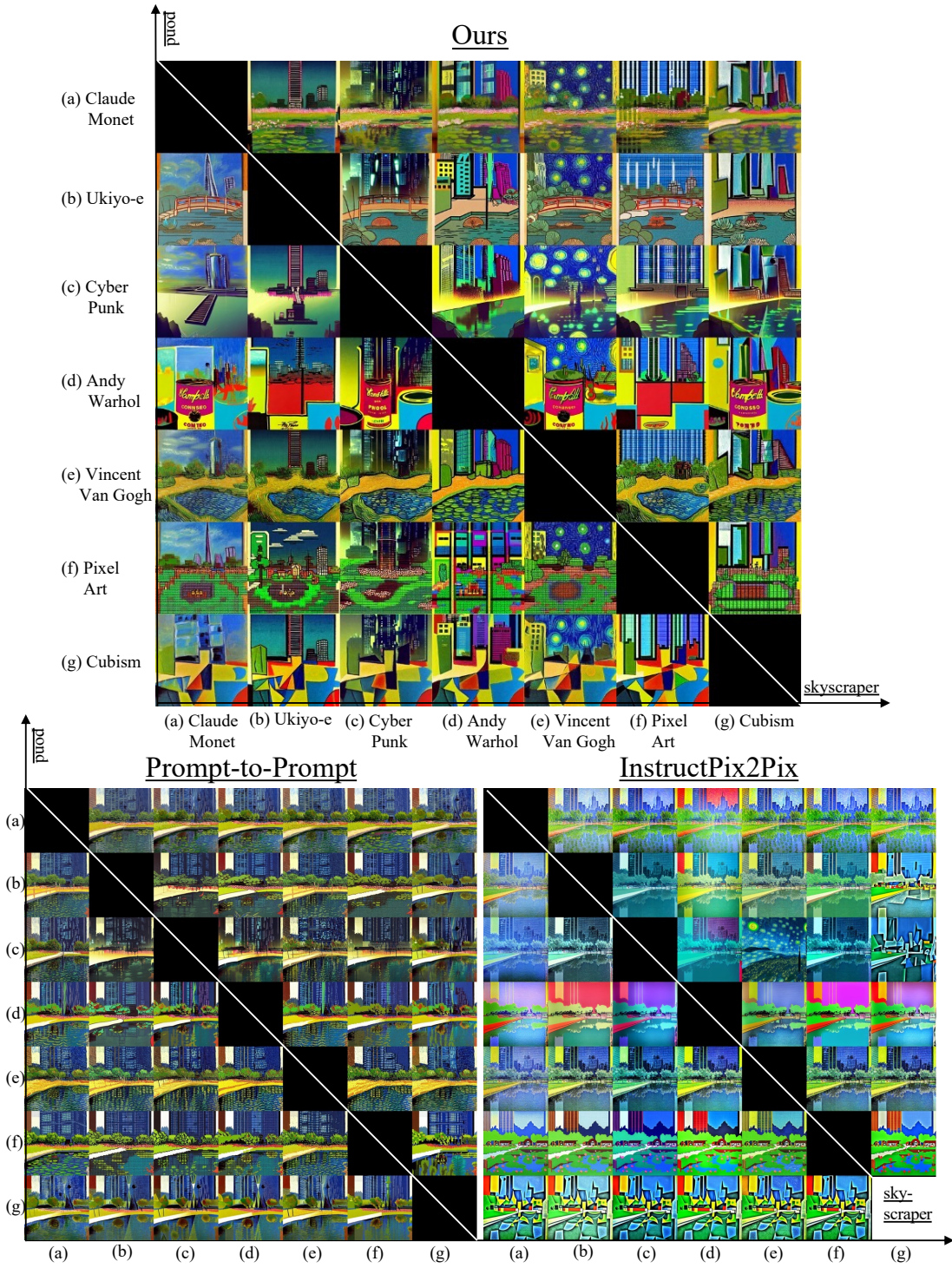


Fig. 25: Additional results of the font style. We show images generated with different style combinations and prompt “a small pond surrounded by skyscraper”. Each row contains “skyscraper” in 7 styles, and each column contains “pond” in 7 styles. Only our method can generate distinct styles for both objects.

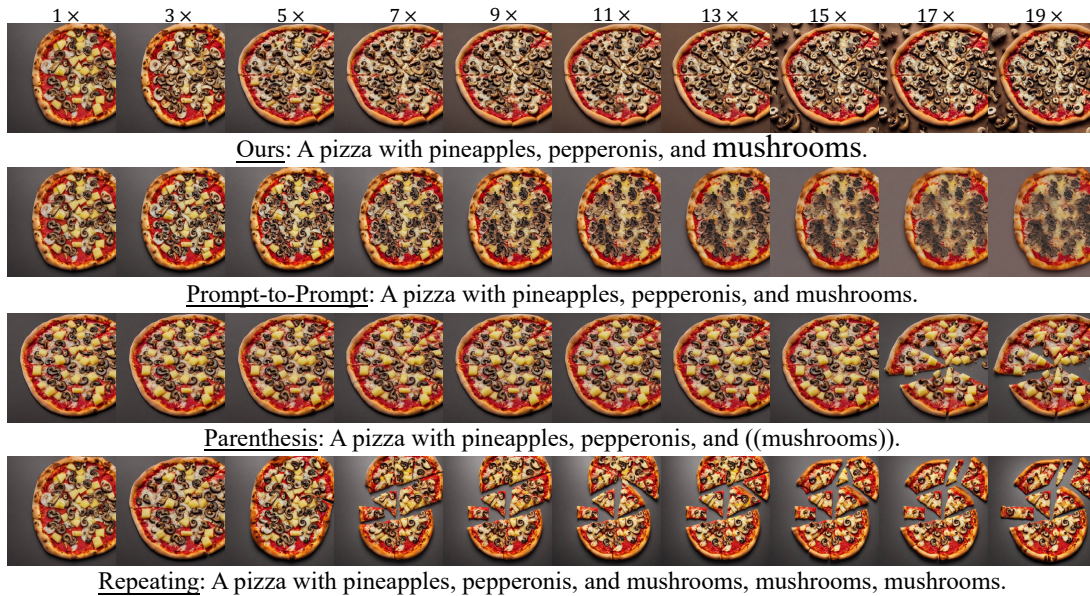


Fig. 26: Additional results of font sizes. We use a token weight evenly sampled from 1 to 20 for the word ‘mushrooms’ with our method and Prompt-to-Prompt. For parenthesis and repeating, we show results by repeating the word ‘mushrooms’ and adding parentheses to the word ‘mushrooms’ for 1 to 10 times. Prompt-to-Prompt suffers from generating artifacts. Heuristic methods are not effective.

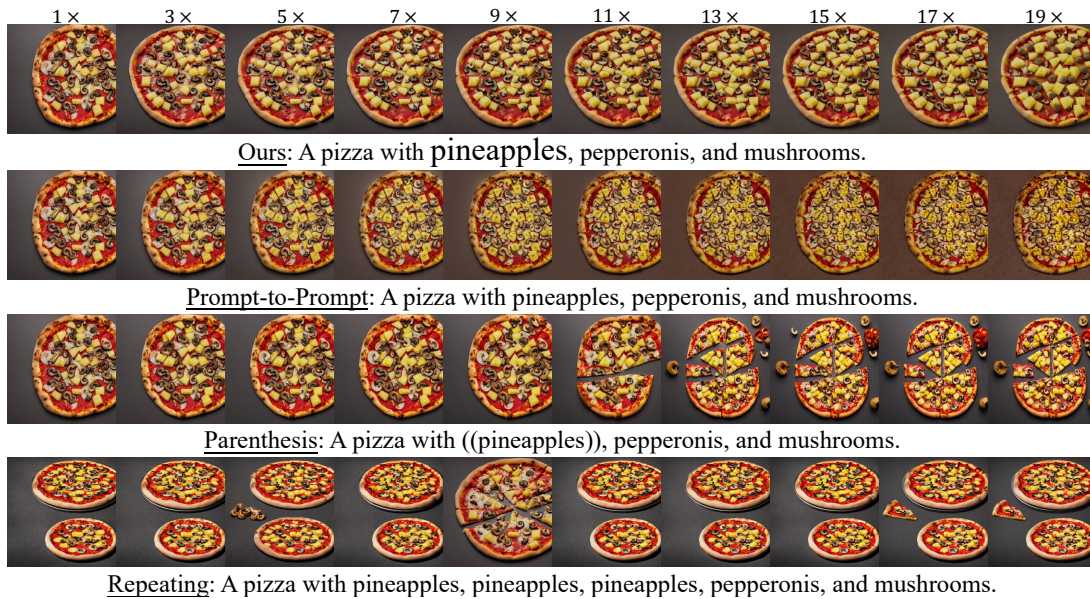


Fig. 27: Additional results of font sizes. We use a token weight evenly sampled from 1 to 20 for the word ‘pineapples’ with our method and Prompt-to-Prompt. For parenthesis and repeating, we show results by repeating the word ‘pineapples’ and adding parentheses to the word ‘pineapples’ for 1 to 10 times. Prompt-to-Prompt suffers from generating artifacts. Heuristic methods are not effective.

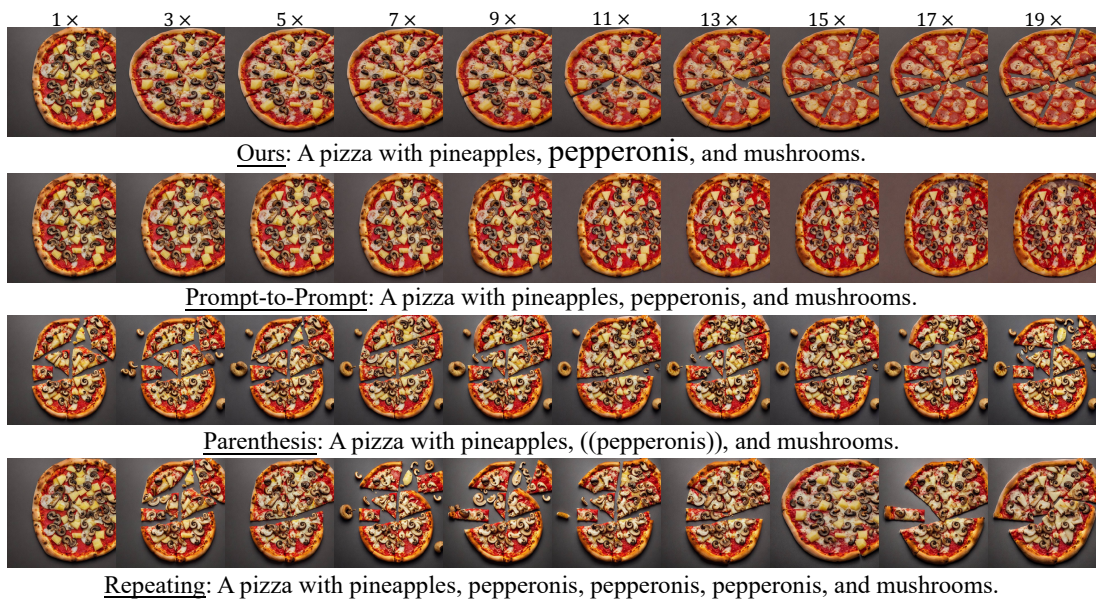
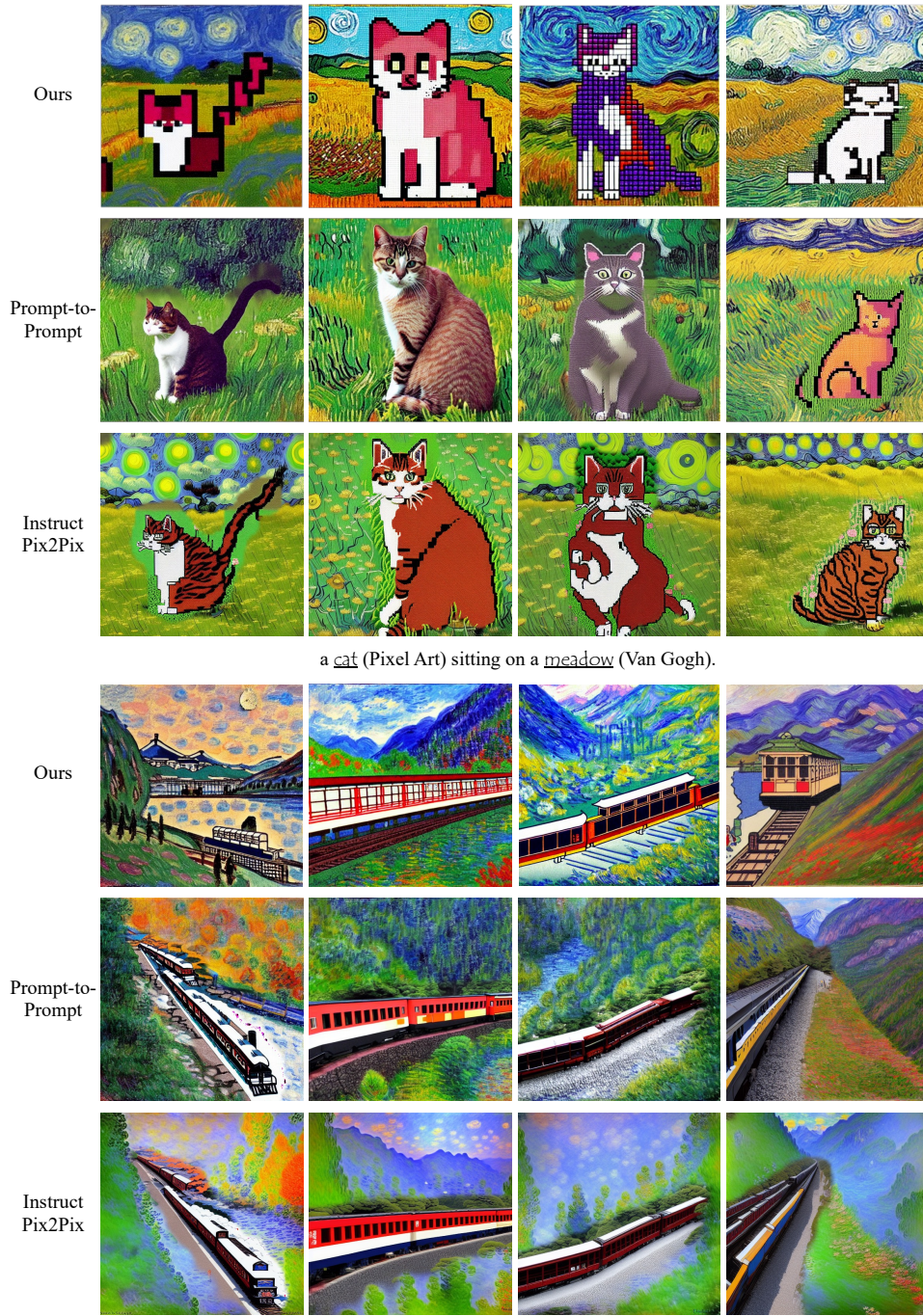


Fig. 28: Additional results of font sizes. We use a token weight evenly sampled from 1 to 20 for the word ‘pepperonis’ with our method and Prompt-to-Prompt. For parenthesis and repeating, we show results by repeating the word ‘pepperonis’ and adding parentheses to the word ‘pepperonis’ for 1 to 10 times. Prompt-to-Prompt suffers from generating artifacts. Heuristic methods are not effective.



A stream train (Ukiyo-e) on the mountain side (Claude Monet).

Fig. 29: Comparison with a simple composed-based method using different random seeds. Since the methods like Prompt-to-Prompt (Hertz et al., 2023) cannot generate multiple styles on a single image, one simple idea to fix this is to apply the methods on two regions separately and compose them using the token maps. However, we show that this leads to sharp changes and artifacts at the boundary areas.

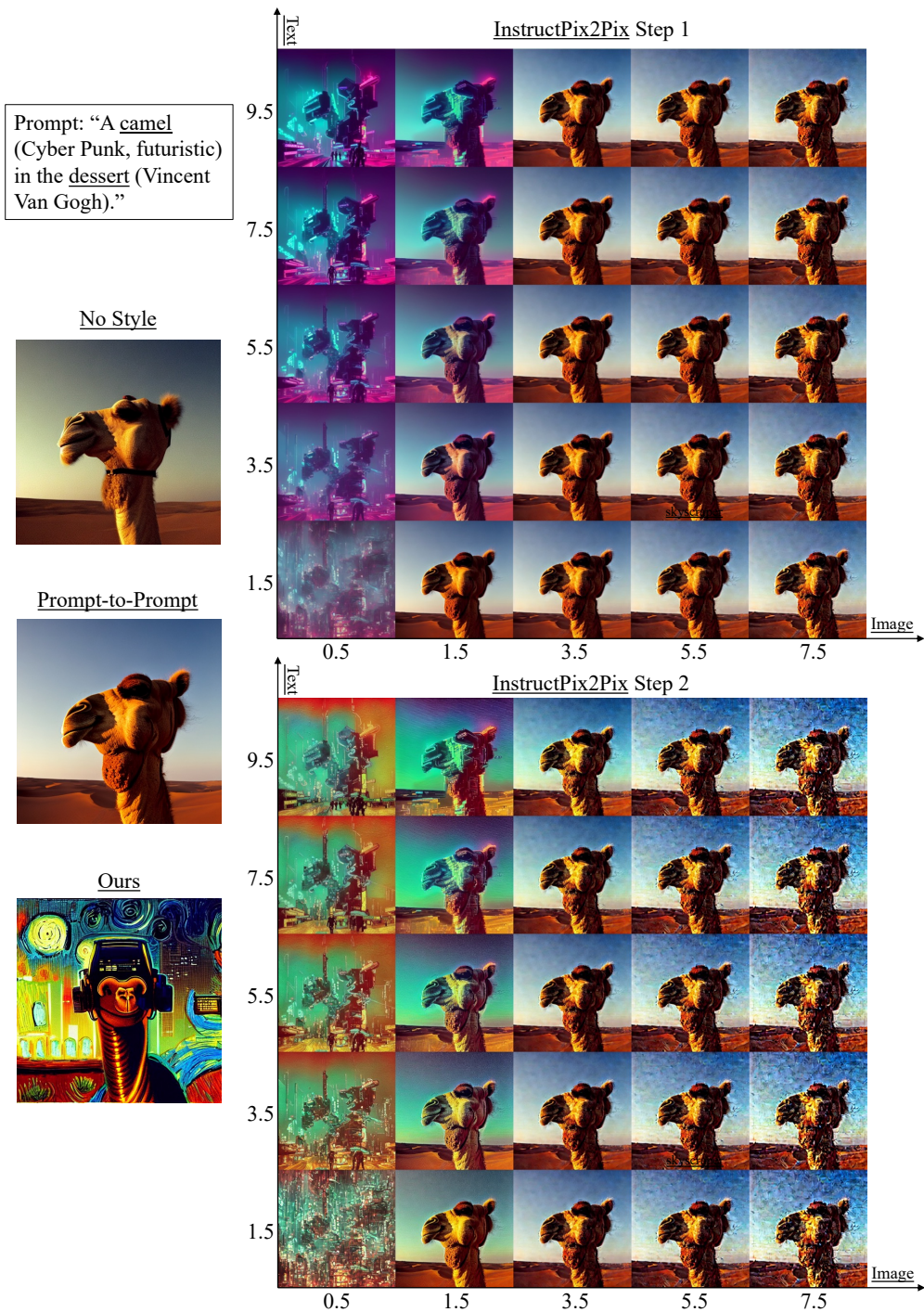


Fig. 30: Ablation of the classifier free guidance of InstructPix2Pix. We show that InstructPix2Pix fails to generate both styles with different image and text classifier-free guidance (cfg) weights. When image-cfg is low, the desert is lost after the first editing. We use image-cfg= 1.5 and text-cfg= 7.5 in our experiment.

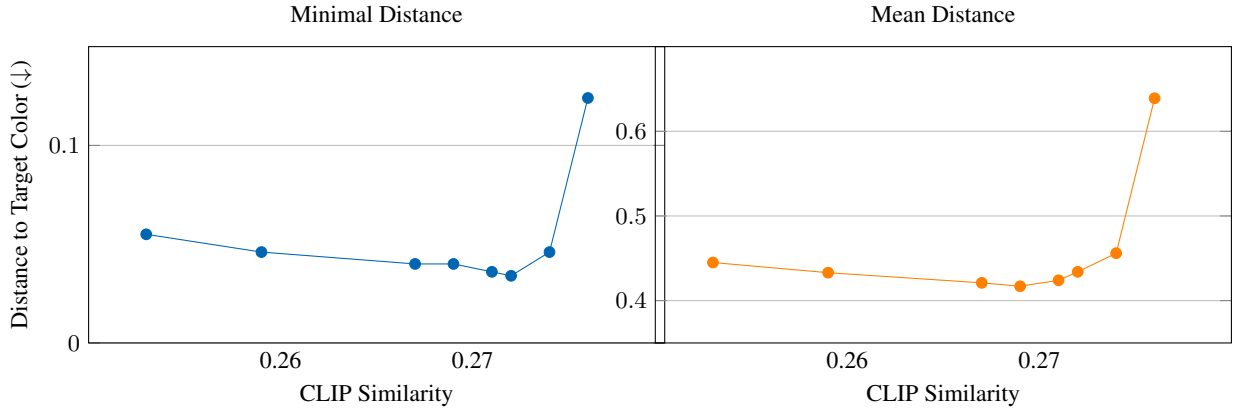


Fig. 31: Ablation on the hyperparameter λ in Equation (7). We report the trade-off of CLIP similarity and color distance achieved by sweeping the strength of color optimization λ .

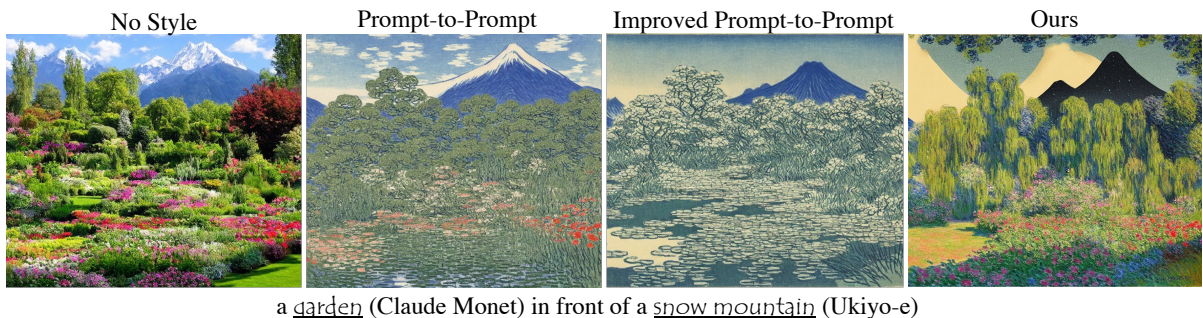


Fig. 32: Improved Prompt-to-Prompt. Further constraining the attention maps for styles does not resolve the mixed style issue.

Ablation of the color guidance weight. Changing the guidance strength λ allows us to control the trade-off between *fidelity* and *color precision*. To evaluate the fidelity of the image, we compute the CLIP score between the generation and the plain text prompt. We plot the CLIP similarity vs. color distance in Figure 31 by sweeping λ from 0 to 20. Increasing the strength always reduces the CLIP similarity as details are removed to satisfy the color objective. We find that larger λ first reduces and then increases the distances due to the optimization divergence.

Constrained Prompt-to-Prompt. The original Attention Refinement proposed in Prompt-to-Prompt (Hertz et al., 2023) does not apply any constraint to newly added tokens’ attention maps, which may be the reason that it fails with generating distinct styles. Therefore, we attempt to improve Prompt-to-Prompt by injecting the cross-attention maps for the newly added style tokens. For example, in Figure 32, we use the cross attention map of “garden” for the style “Claude Monet”. However, the method still produces a uniform style.

Human Evaluation We conduct a user study on crowdsourcing platforms. We show human annotators a pair of generated images and ask them which image more accurately expresses the reference color, artistic styles, or supplementary descriptions. To compare ours with each baseline, we show 135 font color pairs, 167 font style pairs, and 21 footnote pairs to three individuals and receive 1938 responses. As shown in the table below, our method is chosen more than 80% of the time over both baselines for producing more precise color and content given the long prompt and more than 65% of the time for rendering more accurate artistic styles. We will include a similar study at a larger scale in our revision.

Table 3: Human evaluation results.

	Color	Style	Footnote
Ours vs. Prompt-to-Prompt	88.2%	65.2%	84.1%
Ours vs. InstructPix2Pix	80.7%	69.8%	87.3%

B Additional Details

B.1 Rich-text Benchmark

This section details our construction process of the rich-text benchmark.

Footnote evaluation. Here we provide our full prompt template used for GPT-4 to create complex scene prompts with two stages, which typically include three primary parts components:

- **Instruction** describe the goal of task. In the first step, the goal is to generate both global and local description of the scene. In the second step, the goal is to compose the descriptions in an organic manner.
- **In-context exemplars** help enhance the model’s capability for these specific tasks and define the output format by providing manually created examples.
- **User prompt** is added in the end to initiate the task and let the model continue the conversation and generate the complex scene descriptions.

Font style evaluation. To compute the local CLIP scores at each local region to evaluate the stylization quality, we need to create test prompts with multiple objects and styles. We use seven popular styles that people use to describe the artistic styles of the generation, as listed below. Note that for each style, to achieve the best quality, we also include complementary information like the name of famous artists in addition to the style.

```
styles = [
    'Claud Monet, impressionism, oil on canvas',
    'Ukiyoe',
    'Cyber Punk, futuristic',
    'Pop Art, masterpiece, Andy Warhol',
    'Vincent Van Gogh',
    'Pixel Art, 8 bits, 16 bits',
    'Abstract Cubism, Pablo Picasso'
]
```

We also manually create a set of prompts, where each contains a combination of two objects, for stylization, resulting in 420 prompts in total. We generally confirm that Stable Diffusion (Rombach et al., 2022) can generate the correct combination, as our goal is not to evaluate the compositionality of the generation as in DrawBench (Saharia et al., 2022). The prompts and the object tokens used for our method are listed below.

```
candidate_prompts = [
    'A garden with a mountain in the distance.': ['garden', 'mountain'],
    'A fountain in front of an castle.': ['fountain', 'castle'],
    'A cat sitting on a meadow.': ['cat', 'meadow'],
    'A lighthouse among the turbulent waves in the night.': ['lighthouse', 'turbulent w'],
    'A stream train on the mountain side.': ['stream train', 'mountain side'],
    'A cactus standing in the desert.': ['cactus', 'desert'],
    'A dog sitting on a beach.': ['dog', 'beach'],
    'A solitary rowboat tethered on a serene pond.': ['rowboat', 'pond'],
    'A house on a rocky mountain.': ['house', 'mountain'],
    'A rustic windmill on a grassy hill.': ['rustic', 'hill'],
]
```

Table 4: The complete prompt to generate scene and object descriptions used in the first step in creating complex image prompt.

Role	Content
Instruction	System: "In this task, you act like visual and captioning experts, write a caption of a complex scene image. The description contains an overall description of the scene and several descriptions of the object details that appear in the scene. Please only describe visual information in the caption. Here are some examples:"
In-context examples	<p>Assistant: "Scene: A cozy cabin nestled in a snowy forest, and a snowman stands in the yard. Cabin: A charming wooden cabin with Christmas decoration, warm light coming out from the windows. Snowy Forest: Towering evergreen trees covered in a thick layer of pristine snow. Snowman: A cute snowman wearing a carrot nose, coal eyes, and a colorful scarf, welcoming visitors with a cheerful vibe."</p> <p>Assistant: "Scene: A coffee table sits in front of a sofa on a cozy carpet. A painting on the wall. Coffee table: A rustic wooden coffee table adorned with scented candles and many books. Sofa: A plush sofa with a soft blanket and colorful pillows on it. Painting: A painting of wheat field with a cottage in the distance, close up shot, trending on artstation, HD, calm, complimentary color, realistic lighting, by Albert Bierstadt, Frederic Church."</p> <p>Assistant: "Scene: A close-up of a cat riding a scooter. Tropical trees in the background. Cat: A cat wearing sunglasses and has a bandana around its neck."</p> <p>Assistant: "Scene: A car driving on the road. A bicycle nearby a tree. A cityscape in the background. Car: A sleek sports car gleams on the road in the sunlight, with its aerodynamic curves and polished finish catching the light. Bicycle: A bicycle with rusted frame and worn tires. Tree: A dead tree with a few red apples on it. Cityscape: A bustling Hongkong cityscape with towering skyscrapers."</p> <p>Assistant: "Scene: A beach bonfire party with people sitting around the fire, surfboards leaning on a dune, and a cooler is nearby. People: People wearing casual beachwear, some wrapped in blankets, some holding cups of beer. Fire: Crackling fire with driftwood logs and glowing embers. Surfboards: Colorful and waxed surfboards are lined up, leaning against a sand dune. Cooler: A white cooler with blue cover filled with an assortment of beverages."</p> <p>Assistant: "Scene: A small chair sits in front of a table on the wooden floor. There is a bookshelf nearby the window. Chair: A black leather office chair with a high backrest and adjustable arms. Table: A large wooden desk with a stack of books on top of it. Bookshelf: A bookshelf filled with colorful books and binders. Window: A window overlooks a stunning natural landscape of snow mountains."</p> <p>Assistant: "Scene: A lush garden with a fountain. A grand mansion in the background. Garden: A garden is full of vibrant colors with a variety of flowers. Fountain: A fountain made of white marble with multiple tiers. The tiers are intricately carved with various designs. Mansion: An impressive two-story mansion with a royal exterior, white columns, and tile-made roof. The mansion has numerous windows, each adorned with white curtains."</p> <p>Assistant: "Scene: A serene beach with crystal-clear turquoise water and palm trees. A couple walking along the shoreline. Beach: The beach has pink sand with seashells, crabs, and starfish on it. Palm Trees: Tall, swaying palm trees with lush green fronds and green coconuts. Couple: A young couple in matching, red beachwear, smiling as they enjoy their romantic stroll."</p> <p>Assistant: "Scene: A cozy cafe with chalkboard menus, a barista behind the counter, and a pastry display. Chalkboard Menus: Chalkboard menus with handwritten lists of coffees and teas, specials highlighted. Barista: A barista wearing black uniform and a tie is steaming milk for a latte. Pastry Display: A glass pastry display filled with croissants, muffins, and scones."</p>
User prompt	User : "Can you provide more examples like those above?"

Table 5: The complete prompt to generate full descriptions used in the second step in creating complex image prompt.

Role	Content
Instruction	System: "In this task you are acting like a text editing expert. You are given an overall description of a scene and several descriptions of the object details that appear in the scene. Your job is to compose the descriptions into a paragraph that describe the scene. Make sure that the grammar is correct and the flow is natural. Here are a few examples:"
In-context examples	<p>User: "Scene: A serene beach with crystal-clear turquoise water and palm trees. A couple walking along the shoreline. Beach: The beach has pink sand with seashells, crabs, and starfish on it. Palm Trees: Tall, swaying palm trees with lush green fronds and green coconuts. Couple: A young couple in matching, red beachwear, smiling as they enjoy their romantic stroll."</p> <p>Assistant: "A serene beach with crystal-clear turquoise water and tall, swaying palm trees with lush green fronds and green coconuts. The beach has pink sand with seashells, crabs, and starfish on it. A couple in matching, red beachwear walking along the shoreline, smiling as they enjoy their romantic stroll."</p> <p>User: "Scene: A cozy cabin nestled in a snowy forest, and a snowman stands in the yard. Cabin: A charming wooden cabin with Christmas decoration, warm light coming out from the windows. Snowy Forest: Towering evergreen trees covered in a thick layer of pristine snow. Snowman: A cute snowman wearing a carrot nose, coal eyes, and a colorful scarf, welcoming visitors with a cheerful vibe."</p> <p>Assistant: "A cozy, charming wooden cabin with Christmas decoration and warm light coming out from the windows nestled in a snowy forest with towering evergreen trees covered in a thick layer of pristine snow. A cute snowman a wearing a carrot nose, coal eyes, and a colorful scarf, welcoming visitors with a cheerful vibe, stands in the yard."</p> <p>User: "Scene: A street vendor surrounded by the shoppers. Buildings in the background. Street Vendor: A friendly street vendor stands behind a cart filled with an assortment of fresh fruits and vegetables. Shoppers: Diverse shoppers wearing Indian traditional clothes, carrying eco-friendly bags. Buildings: Indian-style old buildings in various colors."</p> <p>Assistant: "A friendly street vendor stands behind a cart filled with an assortment of fresh fruits and vegetables surrounded by the diverse shoppers wearing Indian traditional clothes, carrying eco-friendly bags. Indian-style old buildings in various colors in the background."</p> <p>User: "Scene: A vineyard in front of a winery on a sunny day. Winery: An elegant winery building with a rustic charm, set against a backdrop of rolling hills. Oak Wine barrels stacked outside. Vineyard: vineyard with rows of Lush green grapevines heavy with ripe, plump grapes, waiting to be harvested."</p> <p>Assistant: "A vineyard with rows of Lush green grapevines heavy with ripe, plump grapes, waiting to be harvested, in front of a elegant winery building with a rustic charm, set against a backdrop of rolling hills. Oak Wine barrels stacked outside, a sunny day"</p>
User prompt	User : "Can you write the descriptions given the new user input below. Do not add descriptions that are not part of the user input. + [user prompt]"

Font color evaluation. To evaluate precise color generation capacity, we create a set of prompts with colored objects. We divide the potential colors into three levels according to the difficulty of text-to-image generation models to depend on. The *easy-level* color set contains 17 basic color names that these models generally understand. The complete set is as below.

```
COLORS_easy = {
    'brown': [165, 42, 42],
    'red': [255, 0, 0],
    'pink': [253, 108, 158],
    'orange': [255, 165, 0],
    'yellow': [255, 255, 0],
    'purple': [128, 0, 128],
    'green': [0, 128, 0],
    'blue': [0, 0, 255],
    'white': [255, 255, 255],
    'gray': [128, 128, 128],
    'black': [0, 0, 0],
    'crimson': [220, 20, 60],
    'maroon': [128, 0, 0],
    'cyan': [0, 255, 255],
    'azure': [240, 255, 255],
    'turquoise': [64, 224, 208],
    'magenta': [255, 0, 255],
}
```

The *medium-level* set contain color names that are selected from the HTML color names². These colors are also standard to use for website design. However, their names are less often occurring in the image captions, making interpretation by a text-to-image model challenging. To address this issue, we also append the coarse color category when possible, e.g., “Chocolate” to “Chocolate brown”. The complete list is below.

```
COLORS_medium = {
    'Fire Brick red': [178, 34, 34],
    'Salmon red': [250, 128, 114],
    'Coral orange': [255, 127, 80],
    'Tomato orange': [255, 99, 71],
    'Peach Puff orange': [255, 218, 185],
    'Moccasin orange': [255, 228, 181],
    'Goldenrod yellow': [218, 165, 32],
    'Olive yellow': [128, 128, 0],
    'Gold yellow': [255, 215, 0],
    'Lavender purple': [230, 230, 250],
    'Indigo purple': [75, 0, 130],
    'Thistle purple': [216, 191, 216],
    'Plum purple': [221, 160, 221],
    'Violet purple': [238, 130, 238],
    'Orchid purple': [218, 112, 214],
    'Chartreuse green': [127, 255, 0],
    'Lawn green': [124, 252, 0],
    'Lime green': [50, 205, 50],
    'Forest green': [34, 139, 34],
    'Spring green': [0, 255, 127],
    'Sea green': [46, 139, 87],
}
```

²<https://simple.wikipedia.org/wiki/Web.color>

```

'Sky blue': [135, 206, 235],
'Dodger blue': [30, 144, 255],
'Steel blue': [70, 130, 180],
'Navy blue': [0, 0, 128],
'Slate blue': [106, 90, 205],
'Wheat brown': [245, 222, 179],
'Tan brown': [210, 180, 140],
'Peru brown': [205, 133, 63],
'Chocolate brown': [210, 105, 30],
'Sienna brown': [160, 82, 4],
'Floral White': [255, 250, 240],
'Honeydew White': [240, 255, 240],
}

```

The *hard-level* set contains 50 randomly sampled RGB triplets as we aim to generate objects with arbitrary colors indicated in rich texts. For example, the color can be selected by an RGB slider.

```

COLORS_hard = {
  'color of RGB values [68, 17, 237]': [68, 17, 237],
  'color of RGB values [173, 99, 227]': [173, 99, 227],
  'color of RGB values [48, 131, 172]': [48, 131, 172],
  'color of RGB values [198, 234, 45]': [198, 234, 45],
  'color of RGB values [182, 53, 74]': [182, 53, 74],
  'color of RGB values [29, 139, 118]': [29, 139, 118],
  'color of RGB values [105, 96, 172]': [105, 96, 172],
  'color of RGB values [216, 118, 105]': [216, 118, 105],
  'color of RGB values [88, 119, 37]': [88, 119, 37],
  'color of RGB values [189, 132, 98]': [189, 132, 98],
  'color of RGB values [78, 174, 11]': [78, 174, 11],
  'color of RGB values [39, 126, 109]': [39, 126, 109],
  'color of RGB values [236, 81, 34]': [236, 81, 34],
  'color of RGB values [157, 69, 64]': [157, 69, 64],
  'color of RGB values [67, 192, 60]': [67, 192, 60],
  'color of RGB values [181, 57, 181]': [181, 57, 181],
  'color of RGB values [71, 240, 139]': [71, 240, 139],
  'color of RGB values [34, 153, 226]': [34, 153, 226],
  'color of RGB values [47, 221, 120]': [47, 221, 120],
  'color of RGB values [219, 100, 27]': [219, 100, 27],
  'color of RGB values [228, 168, 120]': [228, 168, 120],
  'color of RGB values [195, 31, 8]': [195, 31, 8],
  'color of RGB values [84, 142, 64]': [84, 142, 64],
  'color of RGB values [104, 120, 31]': [104, 120, 31],
  'color of RGB values [240, 209, 78]': [240, 209, 78],
  'color of RGB values [38, 175, 96]': [38, 175, 96],
  'color of RGB values [116, 233, 180]': [116, 233, 180],
  'color of RGB values [205, 196, 126]': [205, 196, 126],
  'color of RGB values [56, 107, 26]': [56, 107, 26],
  'color of RGB values [200, 55, 100]': [200, 55, 100],
  'color of RGB values [35, 21, 185]': [35, 21, 185],
  'color of RGB values [77, 26, 73]': [77, 26, 73],
  'color of RGB values [216, 185, 14]': [216, 185, 14],
  'color of RGB values [53, 21, 50]': [53, 21, 50],
  'color of RGB values [222, 80, 195]': [222, 80, 195],
}

```



```

'color of RGB values [103, 168, 84]': [103, 168, 84],
'color of RGB values [57, 51, 218]': [57, 51, 218],
'color of RGB values [143, 77, 162]': [143, 77, 162],
'color of RGB values [25, 75, 226]': [25, 75, 226],
'color of RGB values [99, 219, 32]': [99, 219, 32],
'color of RGB values [211, 22, 52]': [211, 22, 52],
'color of RGB values [162, 239, 198]': [162, 239, 198],
'color of RGB values [40, 226, 144]': [40, 226, 144],
'color of RGB values [208, 211, 9]': [208, 211, 9],
'color of RGB values [231, 121, 82]': [231, 121, 82],
'color of RGB values [108, 105, 52]': [108, 105, 52],
'color of RGB values [105, 28, 226]': [105, 28, 226],
'color of RGB values [31, 94, 190]': [31, 94, 190],
'color of RGB values [116, 6, 93]': [116, 6, 93],
'color of RGB values [61, 82, 239]': [61, 82, 239],
}

```

To write a complete prompt, we create a list of 12 objects and simple prompts containing them as below. The objects would naturally exhibit different colors in practice, such as “flower”, “gem”, and “house”.

```

candidate_prompts = [
    'a man wearing a shirt': 'shirt',
    'a woman wearing pants': 'pants',
    'a car in the street': 'car',
    'a basket of fruit': 'fruit',
    'a bowl of vegetable': 'vegetable',
    'a flower in a vase': 'flower',
    'a bottle of beverage on the table': 'bottle beverage',
    'a plant in the garden': 'plant',
    'a candy on the table': 'candy',
    'a toy on the floor': 'toy',
    'a gem on the ground': 'gem',
    'a church with beautiful landscape in the background': 'church',
]

```

B.2 Experiment Details

Baseline. We compare our method quantitatively with two strong baselines, Prompt-to-Prompt (Hertz et al., 2023) and InstructPix2Pix (Brooks et al., 2023). The prompt refinement application of Prompt-to-Prompt allows adding new tokens to the prompt. We use plain text as the base prompt and add color or style to create the modified prompt. InstructPix2Pix (Brooks et al., 2023) allows using instructions to edit the image. We use the image generated by the plain text as the input image and create the instructions using templates “turn the *[object]* into the style of *[style]*,” or “make the color of *[object]* to be *[color]*”. For the stylization experiment, we apply two instructions in both parallel (InstructPix2Pix-para) and sequence (InstructPix2Pix-seq). We tune both methods on a separate set of manually created prompts to find the best hyperparameters. In contrast, it is worth noting that our method *does not* require hyperparameter tuning.

Running time. The inference time of our models depends on the number of attributes added to the rich text since we implement each attribute with an independent diffusion process. In practice, we always use a batch size of 1 to make the code compatible with low-resource devices. In our experiments on an NVIDIA RTX A6000 GPU, each sampling based on the plain text takes around 5.06 seconds, while sampling an image with two styles takes around 8.07 seconds, and sampling an image with our color optimization takes around 13.14 seconds.