Beyond Text-to-Image: Layout-Grounded Stable Diffusion Optimization for Accurate Image Generation

Jiatu Li jil165@ucsd.edu
Wanting Mao wamao@ucsd.edu
Zhengyun Nie znie@ucsd.edu
Jessica Song jwsong@ucsd.edu
Alex Cloninger acloninger@ucsd.edu
Rayan Saab rsaab@ucsd.edu

Abstract

Diffusion models have been shown as the state-of-the-art methods to generate realistic and diverse images based on text prompts, powering models like Stable Diffusion, MidJourney and DALL·E. However, stable diffusion still often make mistakes regarding generating accurately according to numerical information and the object positions. This work proposes a new method to solve the object position problem by letting users draw bounding boxes of objects to indicate their location, which to be used as part of the input, besides providing text prompts only. Our model then utilizes Layout-Grounded Stable Diffusion from Lian et al. (2023) as the backbone to solve the numeracy question so that the actual number of objects in the generated image could match the number specified in the original text prompt. By composing objects onto the positions specified by the user, our model can accurately generate images with numerical and positional information coherent with the text prompt. Through our work, we identify the best coefficients for the above steps using Fréchet Inception Distance. In addition, our model also introduces more diverse and specific user input through integrating a canvas interface that supports specifying image generation subject location and size.

Website: https://dsc180-b11-2.github.io/layout-grounded-optimization/
Code: https://github.com/dsc180-b11-2/layout-grounded-optimization

1 Introduction

In recent years, there has been a significant surge in the popularity of text-prompt image generation models like MidJourney and DALL·E. With the rise of ChatGPT's image generation capabilities, the public is introduced to the magic of turning words into pictures like never before.

The incredible abilities of many state-of-the-art image generation models are made possible by the concept of Stable Diffusion. Diffusion models take an image, convert it into random noise, and then learn how to reverse this process to recreate the original image or sample noise to create new images. Stable Diffusion takes in text prompts and augment the training process in diffusion models to generate images based on the textual description.

However, beneath their remarkable capabilities lies a challenge that needs attention: numerical accuracy in Stable Diffusion. In our experiments, we found that Stable Diffusion often fails the number requirement in its image creation process. A specific example is that when user enters the prompt "Five cats sitting around a table, eating dinner," an image with only three cats is generated. This seemingly simple task is a significant challenge for current state-of-the-art diffusion models. In fact, in Lian et al. (2023)'s empirical research, only 39 out of 100 generated images match the number requirement.

On the other hand, current generate models take plain text as input, which offers users less flexibility in creating images. Previous work, Ge et al. (2023), embeds rich-text editor into models to provided more precise control over the image, however, there still remains limitations. For instance, with plain textual description, users are unable to regulate the specific position of an item other than a vague region. Thus, an interactive canvas with the assistance of text description would be another popular choice for users. Incorporating canvas into the input process enables the users to pass in desired layout, size and object specifications, enhancing their creative freedom in generating the image.

Therefore, improving both Stable Diffusion's numerical accuracy and interactivity will greatly improve the model accuracy, overall reliability, and user experience. In this paper, we set to improve diffusion model's ability to generate images accurately, based on canvas input.

In short, our novel contributions in this project are:

- Developing a canvas interface for user-input,
- Identify the best coefficient of Frozen Step Ration using Fréchet Inception Distance,
- Evaluate the numerical accuracy of our model

2 Methods

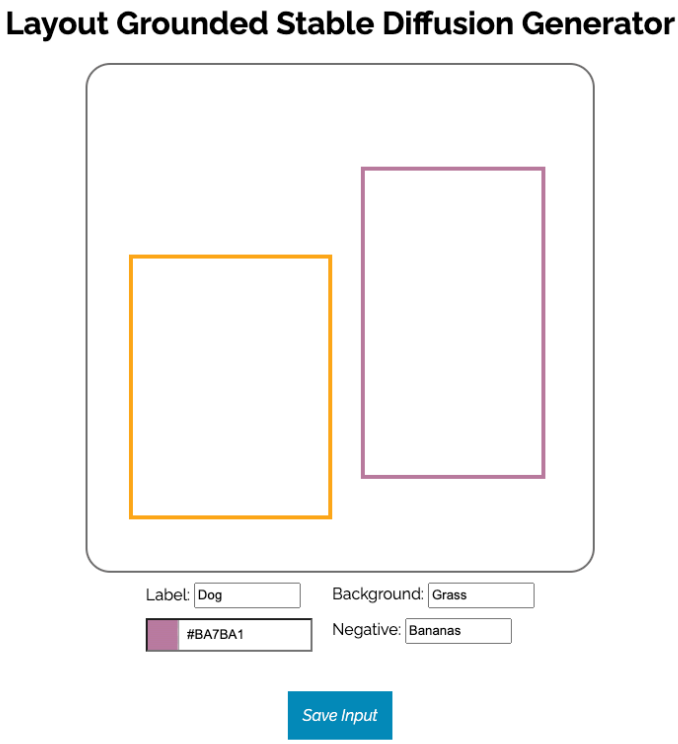
A sketch is usually the first stage of an image. An interactive canvas that enables users to operate freely would be an intuitive alternative option than plain text editor. In this section, we first introduce our approach to this canvas interface, then follows our image generation model: layout-grounded Stable Diffusion. This image generation model reaches

high accuracy in object numeracy with pixel-wise and RGB-level control over objects, while maintaining a coherent style among objects and background. We further evaluate the influence of a critical parameter, frozen step ratio, with Fréchet Inception Distance (FID) which measures the similarity of generated images to real images (Heusel et al. (2017)).

2.1 Front-end Canvas

In order to introduce a more diverse and specific user input, we intend to build a canvas interface that supports specifying image generation subject location, size and color through a front-end website.

The website features several elements: a 512x512 sized canvas, a label section, a color picker, a background prompt text input box, a negative prompt text input box, and a button that allows the user to save all inputs.



The website requires the user to input a label and a color to a bounding box before the user starts drawing. Once label and color are inputted, the user can draw a box on the canvas to specify an object's location and size. Our diffusion model will then use this user input to generate latent masks at this specific input location for the specified text. After drawing one bounding box, the website defaults that the user is using the same input. If the user wants to generate a different object or a different color, they would have to re-pick.

The website requires the user to enter a background prompt, but not a negative prompt.

The background input specifies what environment should the generated objects be in. The negative prompt denotes what should not appear in the image.

After drawing all intended objects and a background, the user can click the `Save Input` button, and a `.json` file will be automatically downloaded. This file contains all user inputted information.

As of now, we are unable to host our model on a live server. Therefore, this section of the process requires human intervention. I will move the `.json` file into the `canvas input` directory in LMD.

Once imported into LMD, the `parse_input_from_canvas` function, called in `prompt_batch.py`, processes the user inputs into appropriate format for our model. Then, the pipeline detects any overly large or out of bounds bounding boxes and resize them into scale. It generates a layout that contains user inputted information, in preparation for future image generation in LMD. We explain how LMD generates the image and our experimentation with fine-tuning the model parameters in the next section.

Through implementing the front-end canvas, we hope to give the user the power to design realistic bounding boxes of objects. This enhances the user's artistic creativity control with stable diffusion. Moreover, we hope that, given more realistic location of objects, the model have more guidance over generating the correct number of objects. We evaluate DUDE's capability in producing numerically accurate images in the Results section.

2.2 Layout-grounded Stable Diffusion

2.2.1 Related Work

In the exploration of utilizing stable diffusion for image generation from prompts containing numerical information, a notable challenge is the model's frequent inaccuracies in visually representing the numerical details specified in the prompts. A pioneering approach to address this issue is presented in "LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models" by Lian et al. (2023). This methodology introduces an innovative strategy to overcome the numeracy challenges encountered by contemporary state-of-the-art diffusion models.

Specifically, the LLM-Grounded Diffusion method employs a two-stage process to generate images from text prompts. In the first stage, a pre-trained large language model (LLM) is leveraged to create a scene layout that includes captioned bounding boxes based on the given prompt, which describes the desired image, and a caption for the background. For instance, given the prompt "A watercolor painting with three cats and two dogs playing on the grass," the LLM first segments the prompt into components delineating foreground objects and the background scene. The foreground is identified as three cats and two dogs, for which positions are determined and represented as bounding boxes generated by the LLM. Concurrently, the background is conceptualized with a caption, in this case "A realistic photo of a grassy outdoor scene," enabling the subsequent generation of a corresponding background image by the diffusion model.

The second stage is an layout-grounded generative model, where a novel controller is introduced to guide an existing diffusion model, such as Stable Diffusion, which originally lacks this specific training objective. This guidance ensures adherence to the layout generated in the initial phase. The process of generating the image in this stage involves two main steps: generating masked latents for each object specified in the layout (e.g., cats and dogs) with attention control to ensure accurate placement within the designated boxes, followed by the coherent composition of these masked latents onto the background noisy image. This results in a final image that aligns with the specified foreground and background based on the text prompt.

2.2.2 Evaluation of Frozen Step Ratio

In the layout-grounded model, we discovered a critical parameter: frozen step ratio, or r_T in Lian et al. (2023)'s paper. It represents the percentage of the total time steps T during which LMD denoise from the masked latents and background image distribution together. This parameter ranges from 0 to 1.

(a) r=0.1 (b) r=0.2 (c) r=0.6

(d) r=0.7 (e) r=0.8 (f) r=0.9

Figure 1: Vary r with prompt: A realistic image of a white deer and a gray bear in an empty factory scene

A small r value indicates a minimal number of time steps dedicated to the standalone creation process of the masked latents, potentially leading to a mismatch between the prompt

and final product, especially the color. This occurs because CLIP and its word embedding matching with image pixels guide the reverse diffusion process. Therefore, the color may be influenced by its training data instead of the user prompt. Moreover, any mismatch between CLIP's embeddings from language to pixels can result unwanted objects or backgrounds in the final image.

Conversely, a higher r value (approaching 1) allocates a substantial proportion of the denoising process time steps to generating the masked latents and background image separately. In other words, the model spends more time generating the latent mask of specified objects. While this can enhance the quality of the objects, it may also result in an unnatural appearance, as if the objects were merely superimposed onto the background, lacking seamless integration. For this reason, it is crucial to find the best frozen step ratio, r. The metric we are using to identify the best values is Fréchet Inception Distance.

Fréchet Inception Distance is a metric for measuring the quality and diversity of image created by generative model. It is improved upon Inception Score and computes the similarity between the distribution of generated images and the distribution of a set of real images (Heusel et al. (2017)). In this project, we evaluate images generated with five different frozen step ratio, r: 0, 0.25, 0.5, 0.75, and 1. For each r, we generate $1,000$ images for five prompts using the same bounding box image with the same label.

We first found two datasets of real images as our ground truth. The first is the Stanford Dogs Dataset from Khosla et al. (2011), which includes 20580 dog images from 120 different breeds. The second is a Kaggle Flowers Dataset from flo (2199), which contains 15740 flower images of 16 different species.

3 Results

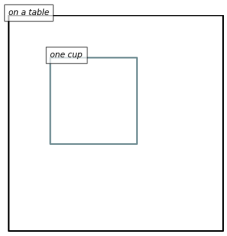
3.1 Numerical Accuracy

For each of the following prompts, we first used our front-end canvas to specify the corresponding number of bounding boxes.

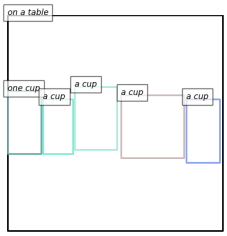
1. A realistic image of one cup
2. A realistic image of five cups
3. A realistic image of ten cups

The layouts and coordinates of the bounding boxes are as follows. For each prompt, a sample image is included. Then we generated 500 images for each prompt based on the coordinates of the bounding boxes we drew.

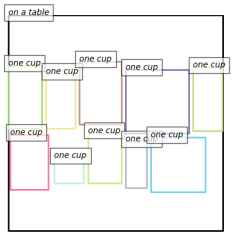
For each prompt, we manually evaluated the numerical accuracy of our model for each prompt among 500 images. In addition, we used Stable Diffusion XL to generate 200 images for each of the aforementioned prompt for comparison and calculated their corresponding numerical accuracy. The results are shown in the following table.



(a) one cup



(b) five cups

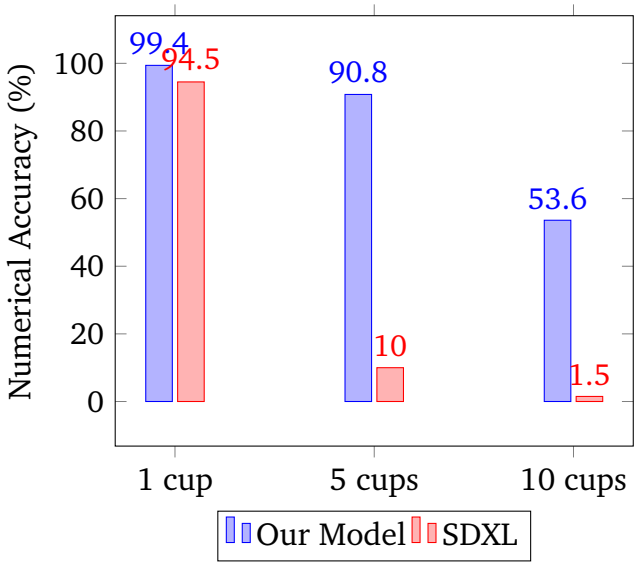


(c) ten cups

(d) sample image: one cup

(e) sample image: five cups

(f) sample image: ten cups



From 3.1, we can see that our model performed very well for prompts with one cup and five cups, achieving over 90% for both prompts. However, numerical accuracy dropped significantly for the prompt with ten cups. Examining the generated pictures, we found that the reason for such low accuracy is because sometimes there will be cups outside the specified positions, or there are two cups appear within the same bounding box that merged together.

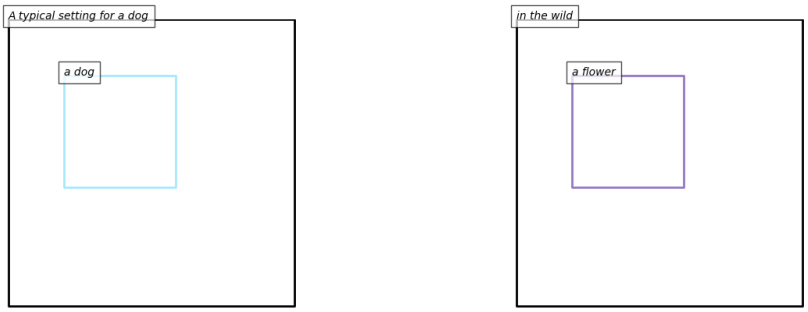
Yet, even with the decrease in accuracy, our model still significantly outperformed SDXL, the baseline model we used behind our optimized layout-grounded stable diffusion model. This

shows that employing bounding boxes was truly effective in improving numerical accuracy in generated images.

3.2 Evaluation of Frozen Step Ratio

To evaluate the quality of generated images, we compare the distribution of generated images with the distribution of a set of real images (ground truth) in order to assess the quality of images created by a generative model.

We aimed to find the frozen-step ratio r_T that leads to the best generated images, and we tested 0, 0.25, 0.5, 0.75, and 1, to perform experiments. For each r_T, we generated 1000 images using the prompt "A realistic image of a dog" and the following layout.



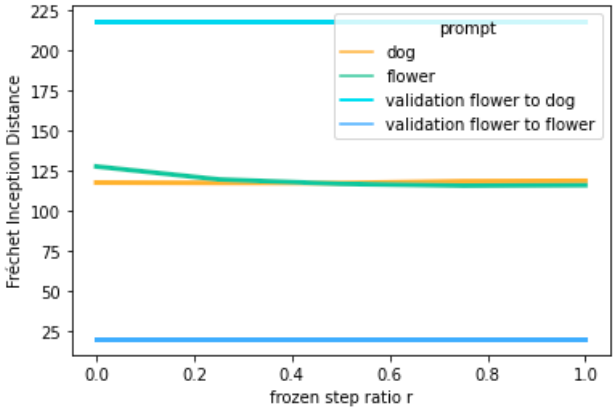
(a) dog (b) flower

For each r, we compared the distance between the distribution of true dog images and the distribution of generated dog images. To create a benchmark for FID score, we first sampled 1000 images from the Stanford Dogs dataset and used them as a validation set. We used the rest 19580 dog images at the ground truth. Comparing the distribution of the validation images and ground truth images gave a FID score of 19.950 (our benchmark 1), indicating the noise levle for this score (difference between images from the same distribution). Similarly, we sampled 1000 images from the Kaggle Flowers Dataset and used them as another validation set. Comparing flower validation images and dog ground truth images gave a FID score of 218.161 (our benchmark 2). This number shows how large the difference is for images from two completely different distributions.

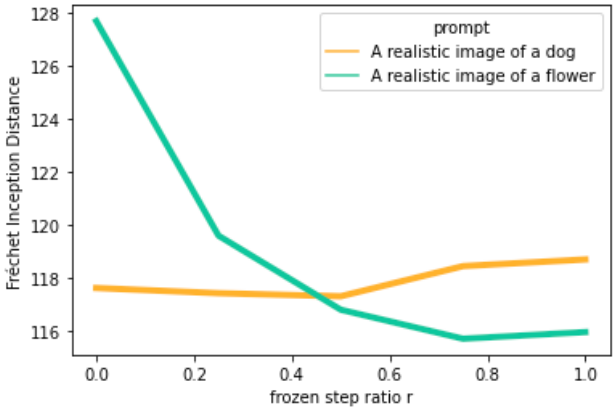
Comparing the generated dog images to their ground truth produced very similar FID scores for different r_T. Results are shown in 4a.

Similarly, we calculate FID between realistic and generated flower images. For benchmark 1, we compare 1000 flower images with 14740 ground truth images, finding an FID score of 19.665. We also compare 1000 dog images with the same ground truth imgaes, getting an FID of 217.383. The results are shown in 4b.

From these plots, we can see that $r_T = 0.75$ results in the most realistic images.



(a) dog



(b) flower

3.2.1 Frozen Ratio FIDs Interpretation

It turned out that there is no significant difference among the scores obtained for different r_T for the dogs prompt. However, for flowers, the FID scores vary more significantly with different r_T, with 127.68 being the highest for r_T=0 and 115.70 being the lowest for r_T=0.75.

We have two educated guesses for the reason behind this result. First, dogs may not vary much in terms of color, shape, or details as much as flowers do. While the dogs can be a few colors, such as black, white or brown, flowers have a much wider range of realistic colors. Secondly, a typical setting for a dog would also not differ too much, since the number of typical settings associated with a dog is limited. Meanwhile, for the flower FID test, we decided to use the background prompt "in the wild," in order to maintain a relatively similar background with the ground truth dataset. However, the words "the wild" also give more variability of the background environment. Therefore, for FID test on dog images, the differences between images with the same seed but different r_T are very small.

The generated images **??** demonstrate why the flower image generated using r_T=0.75 is the most realistic (thus having lowest FID on average).

Figure 5: Flower images with different r_T

First, it is apparent that the resulting image of $r_T=0$ is not realistic. Stable diffusion wrongly incorporated a scene with blue sea and skies with a flower and a rock. For $r_T=0.25$, the generated image is more realistic but still look generated and lacks details. For example, the curvature of the rock is unnaturally round. Furthermore, the pedals in the middle of the flower doesn't match with the rest of them. The generated image for $r_T=1$ is realistic for both the flower and the wall. However, the flower seems to grow in the wall rather than somewhere in the wild, indicating the model's lack of capability in integrating the boxed object and the background together, in a natural way. The two images for $r_T=0.5$ and $r_T=0.75$ have the best quality compared to other images and thus have the lowest FID scores. The image for $r_T=0.5$ is fairly realistic with fine-grained rock details and a protruding round rock surface. Meanwhile, we found the generated image for $r_T=0.75$ to be the most realistic. The rock has a protruding structure and reasonable crack patterns. The flower has accurate pedal sizes and shapes.

3.3 Text-to-Image Alignment

We also performed experiments to show that our optimized layout-grounded stable diffusion model is much better at adhering to the text prompts in image generation than the baseline Stable Diffusion XL and the original LMD we are improving. We used the following prompts to demonstrate this result.

For the first prompt, we can see that SDXL only showed a man with one hand holding a blue bag. The red bag did not seem to be held by his hand. Therefore, the image did not capture the correct action specified in the prompt. The reason for this is that current state-of-the-art stable diffusion still could not depict the interaction between different objects within an image correctly. The relative position of one object in terms of the other is very hard for SDXL to understand. For the original LMD model, we can see that the bounding boxes whose coordinates are generated by GPT-4 seem to be in the right position, with one bag box in the bottom left of the human box and the other bag box in the bottom right of the human box. However, this is not accurate enough, as the resulting image showed that the man in the middle of the picture is not holding the bags but rather dragging. In addition, the colors of the bags also did not match with the prompt. Therefore, the generated picture still did not align with the prompt well. Finally, by manually drawing the bounding boxes, our model not only captures the holding action accurately but also the color information

Figure 6: Comparison of our model with SDXL, LMD, with Vermeer (1657) and Jon (2021)

correctly. Note that the bounding boxes of the bags we drew for this prompt overlap with the bounding box of the man, which is reasonable because the hands of the man from the generated image are in the middle bounding box, meaning that some portion of the bags must also be in this bounding box to ensure the contact between bags and hands (interaction).

For the second and third prompts, we showcase our model's ability to superior ability in giving users creative capabilities with stable diffusion. Although the quality of images generated by SDXL is wonderful, the actual content did not align with the prompt well. Without indication of location of objects, stable diffusion purely tries to match token embedding from the prompt with image embedding. Therefore, we see keywords present in the images, such as red bag and blue bag, milk maid, girl, and mountains. Yet, SDXL fails to capture the association and interactions between objects.

In the LMD model, Lian et al. (2023) relies on Chat-GPT to automatically generate the bounding boxes based on the given prompt. The specified locations of these bounding boxes give relatively more accurate location and relationship between each object. However, GPT has a great possibility of interpreting the relationship between bounding boxes wrongly, or assigning the attributes wrongly. Hence, it did generate red and blue bags relative to the person's left and right hands, but wrongly assigned the attributes to each bounding box. In the milkmaid case, because of the complexity of the prompt, GPT was unable to create bounding boxes that's accurate to the relationship described in the text prompt.

Lastly, through human-intervened artistic placement of objects in the scene, we created layouts that's more realistic to the given prompt. Therefore, our model correctly assigned col-

ors to each object. For the milkmaid, the relative location of milkmaid, jar and table helped us to recreated the ”pouring” relationship. Moreover, the layout successfully guided stable diffusion to create more artistic images, recreating “The Milkmaid” by Johannes Vermeer and modern singer Zeph’s album cover for the song ”miss me.” in similar styles.

4 Discussion & Future Work

In the results section, we gave an abundant amount of analysis of our observations of our results. In this section, we discuss existing areas of improvement, our analysis, and future work.

4.1 Numerical Accuracy

We found that there are two main reasons why the numerical accuracy decreased drastically when the number of cups changed from five to ten. First and foremost, objects sometimes interact or are interconnected, causing a single masked latent to encapsulate multiple objects. 7 generated from the prompt “A realistic image of a dog” shows that within the same bounding box, there are two dogs sitting together. Thus, this image does not capture the correct numerical information as specified in the prompt. The reason for this is in reality, dogs might play with each other or stay together, causing stable diffusion could not differentiate one from the other. Moreover, we observed that the larger the number of bounding boxes in the layout, the higher the probability that one or more bounding box might contain more than one object. Therefore, the generated images from the prompt “A realistic images of ten cups” have much higher chance to encounter this issue and thus have much lower numerical accuracy (53.6%) compared with the accuracy of images containing five cups (90.8%). Another reason that decreased numerical accuracy is that there might be

Figure 7: Objects interact, causing a single masked latent to encapsulate multiple objects.

similar objects appearing in the background of the generated image. Although these objects are not specified by the user and thus not inside any bounding box, we still cannot say the image have the correct number objects as specified in the text prompt. 8 generated

from the prompt "A realistic image of a cup" shows that beside the cup we wanted in the bounding box located in the center of the image, there are two more cups in the background, one in the bottom left and one in the upper right. This factor negatively influences all images regardless of the text prompts provided, although it happens with a relatively small probability.

Future work to deal with issues of numerical accuracy as described above can be detect the number of objects in the masked latent first and regenerate masked latents if necessary. In addition, better baseline models can also be incorporated if possible. A method to solve the problem of objects in the background can be specifying the negative prompt (i.e. not include the type of objects we want in the background). To ensure the numerical accuracy of the final generated image, we can also use a detector to count the number of objects first, and then use current state-of-the-art methods to add or remove some objects to align with the text prompt if necessary.

With respect to image quality, we do not know if FID scores around 120 are good enough for the generated images. Therefore, future work can first include some benchmark generated image datasets that are realistic enough and compute the FID score for them to set a standard. In addition, simple prompts such as "A realistic image of xxx" are too general and models cannot really understand what is "realistic" enough. Therefore, more details should be included in the text prompts so that the final images can look realistic and more similar to the real images, thus having a lower FID score potentially. Bounding boxes of different sizes and locations in the layout should also be experimented so that the resulting images can have more variability and thus potentially have a distribution closer to the real image distribution. Better baseline models, such as DALLE if possible, can also be incorporated to our model to improve the quality of generated image.

Figure 8: Diffusion model may use the prompt to produce similar objects in the background.

Lastly, a part of our original project proposal included integrating Ge et al. (2023)'s RGB-level color control into our canvas interface and image generation process. Due to time and technical constraints, we weren't able to reach satisfying results with this integration. Thus, one future direction is to integrate the color of a specific bounding box object into the denoising loop's loss update, to give more control over the image generation to the user.

5 Conclusion

In this paper, we introduced a novel approach to enhance Stable Diffusion's ability to generate realistic and numerically accurate images that assist user-driven creativity, based on Lian et al. (2023)'s LLM-Grounded Diffusion model.

By integrating a canvas interface that allows users to specify the location, size, and color of objects through drawing bounding boxes, we offer a more interactive and precise way of guiding the image generation process. Moreover, we conduct research in our model's effectiveness in creating realistic and numerically accurate images through evaluation of frozen step ratio and manual measurement. We discover that a frozen step ratio of 0.75 (spending 3/4 of the time generating separate masks, and then the rest denoising them together) result in the most realistic images. Furthermore, we compared the numerical accuracy of tour model against Stable Diffusion XL, showing significant improvements, especially for images with a higher number of objects. Although the model's performance declined with an increase in object count, it still substantially outperformed the baseline model, highlighting the effectiveness of using bounding boxes for improved numerical accuracy.

Lastly, we discuss limitations and potential areas for future work, including strategies for addressing the challenges of object interaction and unintended background objects. We suggest that further research could explore methods for detecting and correcting numerical inaccuracies during the image generation process and investigate the impact of more detailed prompts on image realism.

References

,

Ge, Songwei, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. 2023. "Expressive Text-to-Image Generation with Rich Text." In *IEEE International Conference on Computer Vision (ICCV)*.

Heusel, Martin, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 2017. "GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium." *CoRR* abs/1706.08500. [Link]

Jon, Zephani. 2021. "miss me." [Link]

Khosla, Aditya, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. "Novel Dataset for Fine-Grained Image Categorization." In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO

Lian, Long, Boyi Li, Adam Yala, and Trevor Darrell. 2023. "LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models." *arXiv preprint arXiv:2305.13655*

Vermeer, Johannes. 1657. "The Milkmaid." [Link]