Avataar Assignment

Naman Jaswani

Abstract. This project addresses the development of user-friendly technology for manipulating objects within a scene. It involves two key tasks: the segmentation of a specified object in a given scene based on a user-provided class prompt, and the subsequent repositioning of the segmented object within the scene to create a composite image. These tasks contribute to the overall goal of enhancing the ease of interaction with objects in a digital environment. I propose a solution by combining GroundingDINO, Segment-Anything and Stable Diffusion models to achieve this objective.
Project Page[cpu inference]: https://huggingface.co/spaces/noni27/Avataar$_{a}ssignment$.

1 Task

Recent advancements in generative artificial intelligence (AI) have introduced innovative workflows, particularly in the domain of post-production editing for product photographs. This includes utilizing generative AI techniques to refine and modify studio-captured product images, optimizing them for presentation on e-commerce platforms. The primary focus of this study encompasses two interrelated tasks with the overarching objective of developing technology to enable a user-friendly functionality for manipulating objects within a given scene.

The first task involves the segmentation of a designated object within the provided scene, guided by a user-specified class prompt. This segmentation process is fundamental to establishing a user-friendly interface for subsequent interactions. The second task centers around dynamically repositioning the segmented object within the same scene and seamlessly integrating it to produce a composite image. The ultimate goal is to contribute to the advancement of technology that facilitates intuitive and user-friendly manipulation of objects within a visual context. This research endeavors to enhance creative workflows, particularly in the context of post-production editing for e-commerce product images.

2 Approach

GroundingDINO: The authors of the SOTA model GroundingDINO [2] proposed an open set object detection using text prompts. The idea was to detect objects not specific to training data. I used the bounding box predictions made from groundingDINO model, changed its reference from center to corner and scaled it according to image size. I had to check predictions for various combinations of box-threshold, text-threshold values to reduce false positives. The box threshold specifies the minimum similarity score required for an object box to be

Fig. 1: Results on variety of images. Below each image is the prompt used to select an object class.

considered a positive detection while the text threshold specifies the minimum similarity score required for an image to be considered a positive detection.

Segment Anything Model: This SOTA model from Meta [1] does Zero-Shot segmentation using prompts. Prompts specifying what to segment in an image allow for a wide range of segmentation tasks without the need for additional training. Even though the authors specified text prompt as one of the three types of prompts that SAM takes, I didn't find the text prompt support in its code. Hence I decided to make use of GroundingDINO's ability to use text prompt to generate bounding box(bbox) predictions and used these bbox predictions as prompt to SAM to segment the object of class defined in text prompt.

Stable Diffusion: All the masks obtained from SAM were combined to get a 2D composite mask, which I use to cut out objects from scene. This creates holes in image. Now, I used stable diffusion inpainting model to fill these holes. Stable Diffusion Inpainting [3, 4] is a latent text-to-image diffusion model capable of

generating photo-realistic images given any text input, with the extra capability of inpainting the pictures by using a mask. The prompt I used for all results is "a photo of background", but it can be played around with.

3 Experimental Evaluation

3.1 Dilation effect:

While SAM combined with GroundingDINO was able to give very sharp segmentation masks for objects specified in text prompt, I observed that stable diffusion(SD) was generating artifacts on the masked region. I suspected that this might be due to the fact that there is not much space for SD models to do their magic. Hence, I decided to dilate the mask (extrapolate the masked region) before inpainting the masked region. In (Fig. 2), we can clearly observe that without dilation, SD model is generating a sharp object (since the mask was sharp), while with dilation, I was able to get smooth inpainted image. I use this trick for all results attached below.

3.2 Variety of inputs:

In Fig. 1, I show outputs from my models on different types of images and prompts. A few interesting observations from each of the five images : (a) It is able to differentiate between white dog and black dog. (b) It detected right most flower pot as stool, which can be rectified by adjusting box-threshold, (c) It generated an object which fits well in the background(with some artifacts though), (d) It generated (and hence extended) the shadow from table, (e) It was able to differentiate between balck and green grapes!

3.3 Complex Image:

The scene looks quite complex with variety of objects. The model is able to segment prompted carrots and inpaint according to background Fig. 3. It shows, model can work well in this complex domain. One observation though, it distorts human faces and text present in the scene. We can use bounding boxes from groundingDINO, cutout from inpainted image the pixel locations present within bbox coordinates and make a composite with original image. For better text and faces within mask, we can increase num-inference-steps of SD models or try out some face restoration models.

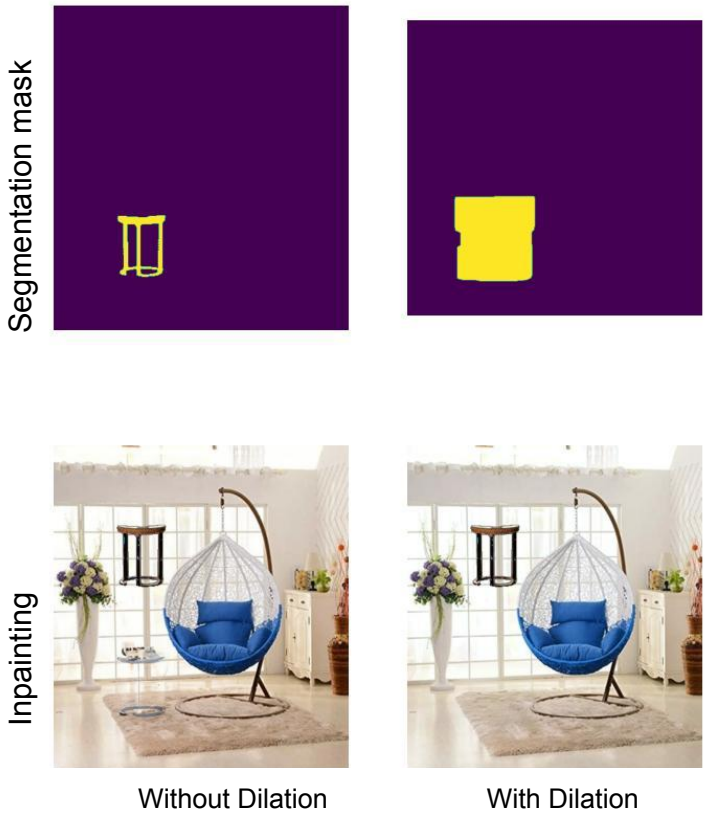

Fig. 2: The effect of dilating the segmentation mask on Stable Diffusion inpainting First column is without dilation while second column is with dilation

Fig. 3: Vegetable market scene Segmented carrots are marked with cyan color. SD model is able to generate plausible images at the masked locations, blending it pretty well with the vegetable stall.

3.4 Complex text prompts:

In Fig. 4, interestingly we can see that groundingDINO + SAM is able to understand relative positioning of objects. Also this combination model is able to segment all body parts quite well. An observation while experimenting different prompts to select objects, I found that groundingDINO is not able to capture prompts for text in image (eg: FIFA or Qatar in this scene). Faces have been distorted by SD model while all other patterns like pattern on Argentine jersey or numbers are intact.

Fig. 4: Complex text prompts: (a) : Guy wearing black robe lifting trophy , (b) : Guy in green tshirt , (c) : Left most guy

3.5 Different stable diffusion models:

I try results from different stable diffusion models. Fig. 5 shows SD2 creates much crispier image than SD1 (just look at the reflection of objects generated on shiny table !). SDXL on the other hand generates a sharp laptop on table(prompt used was "a photo of background").

Fig. 5: Different Stable Diffusion models First row : Stable Diffusion , Second Row : Stable Diffusion 2, Third row : Stable Diffusion XL

4 Conclusion and Future Work

In conclusion, this project aimed at developing user-friendly technology for manipulating objects within a scene through a combination of advanced generative artificial intelligence models. The tasks involved segmentation of specified objects based on user-provided class prompts and the subsequent repositioning of segmented objects to create composite images. Our proposed solution incorporated the synergistic use of GroundingDINO, Segment-Anything, and Stable Diffusion models to achieve these objectives.

Our approach involved adapting the GroundingDINO model for open-set object detection using text prompts. The Segment-Anything model facilitated zero-shot segmentation, utilizing bounding box predictions from GroundingDINO. Stable Diffusion was employed for inpainting and generating photo-realistic images while overcoming artifacts.

The experimental evaluation showcased the effectiveness of our approach. Dilation of masks before inpainting proved to be a crucial step, improving the smoothness of the generated images. Results demonstrated the versatility of the models on various inputs and prompts, successfully handling complex scenes with multiple objects.

Noteworthy observations included the ability of the combined model to understand relative positioning of objects and segmenting various body parts effectively. However, challenges were identified, such as the distortion of faces and text in certain scenarios. Suggestions for further improvement included refining face restoration and increasing inference steps for stable diffusion models. More work for future can include segmenting object shadows along with the object itself.

In summary, this project contributes to the ongoing advancements in generative AI, offering a practical and user-friendly solution for manipulating objects within digital environments. The insights gained from experimentation provide a foundation for future research and enhancements in the field. The models and techniques employed hold promise for a wide range of applications, particularly in the realm of creative content generation and post-production editing.

References

1. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 2
2. 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) 1
3. 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. arXiv preprint arXiv:2307.01952 (2023) 2
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2