Quarter 2 Project Proposal - Diffusion Model Scene Representation

Atharva Kulkarni apkulkarni@ucsd.edu
Ester Tsai etsai@ucsd.edu
Karina Chen kac009@ucsd.edu
Zelong Wang zew013@ucsd.edu
Alex Cloninger acloninger@ucsd.edu
Rayan Saab rsaab@ucsd.edu

1 Broad Problem Statement

The ability of latent diffusion models (LDMs) to generate realistic images from textual descriptions has seen remarkable advancements. Even when trained purely on images without explicit depth information, they typically output coherent pictures of 3D scenes. These models have the astonishing capacity to create detailed, coherent scene representations.

However, their ability to represent depth and saliency within generated images remains unclear. Our project aims to delve into the diffusion process of LDMs, unraveling how they internally represent and process scene geometry.

This investigation is crucial as it not only enhances our understanding of AI's interpretive capabilities but also paves the way for further advancements in image synthesis. Existing research primarily focuses on the output capabilities of these models, leaving a gap in comprehending their internal processing mechanics.

2 Narrow Problem Statement

Previous work mostly focused on using diffusion models to produce higher-quality results, particularly in image generation. In our Quarter 1 Project, we focused on simplifying the diffusion process, looking at the process of how DDPMs work on two-dimensional distributions, specifically how the Gaussian noise is removed in the reverse diffusion process.

In our Quarter 2 Project, we will similarly focus on the diffusion process, but we are now looking at their internal representations of scene geometry as noise is being removed. Even though diffusion models are trained on 2D images, we want to show that they encode some internal representations of depth and saliency even in early denoising steps before the human eye can detect anything other than random noise.

Previous work has attempted to answer this question and has found that there is depth information that emerges in early denoising. Baranchuk et al. (2021) extrapolated the intermediate activations of a pre-trained diffusion model for semantic segmentation. Their high segmentation performance reveals that the diffusion model encodes the rich semantic representations during training for generative tasks. Our work shows that the internal representation of LDM also captures the geometric properties of its synthesized images.

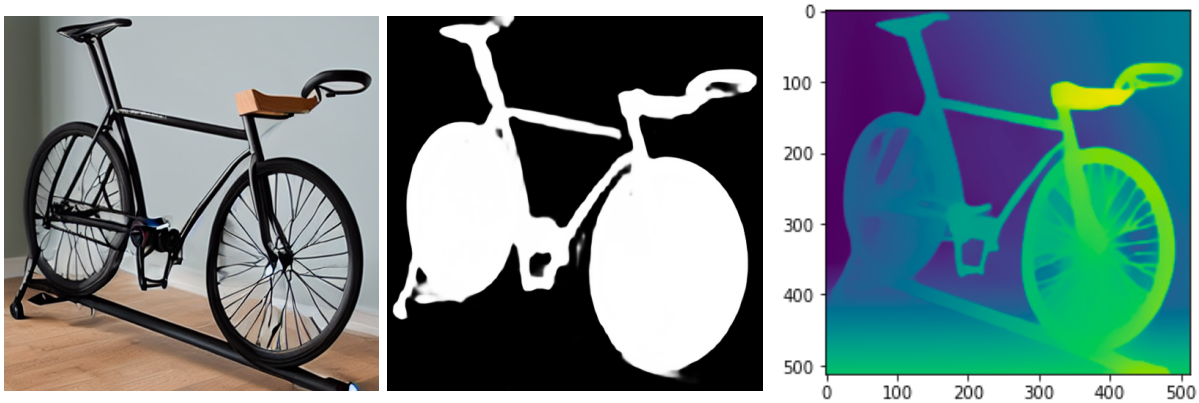
3 Primary Output Statement

Our primary output will be a research paper and a website to showcase our results. In our research paper, we will explain the motivation, methods, results, and discussion of our project. We should include images that demonstrate the LDM's ability to separate foreground from background in the early denoising stages. In contrast, image segmentation and salient-object detection models perform poorly on noisy images. We will compare the results from the LDM and the other methods to show that diffusion models can capture the

internal 3D representation of a scene. On our website, we will display the most important information from our research paper with an emphasis on visualizations. To take advantage of the website format, we can include a tab for additional comparison examples and other interesting visualizations. As a stretch goal, we can implement an interactive widget that generates an image according to the object mask inputted by the user.

4 Data

For this project, our training dataset consists of 617 images (512 pixels x 512 pixels) generated from Stable Diffusion v1.4. If we want to show the internal representation of our diffusion model through salient object detection, then our testing dataset consists of the salient object mask outputs from applying TRACER (`https://github.com/Karel911/TRACER`) to our generated images. If our goal is depth estimation, then our testing dataset consists of the monocular depth estimation outputs from applying MiDaS (`https://github.com/isl-org/MiDaS`) to our generated images. We have successfully obtained all training and testing datasets, but we plan to generate more images using different prompts according to the directions we hope to explore.



(a) (b) (c)

Figure 1: Three images of a bike: (a) *bike.png* - An image of a bike generated by Stable Diffusion v1.4 using the prompt "Lapierre Pulsium 600 FDJ Road Bike 2017" (b) Salient object detection mask output of *bike.png* using TRACER (c) Depth output of *bike.png* using MiDaS

The data that we have obtained includes the diffusion model generated images and their corresponding salient object mask output and depth information output. This is all the information that we need.

We obtained the same data as Chen, Viégas and Wattenberg (2023), and this data was sufficient for their similar experiments. Therefore, for our purposes, the data should also be of sufficient quality.

5 Backup Plan

In the case that we face insurmountable technical issues with exploring our own questions, our backup plan is to recreate the paper by Chen, Viégas and Wattenberg (2023) and make additional observations about the results, such as whether the output is consistent in quality at the same time steps across different prompts. This works well as a backup plan because they explored a similar direction as us and paved the road for our project. In our schedule, we have allotted 1-2 weeks to recreate the probing result using MiDAS and TRACER models, and the rest of the quarter is for exploration beyond existing work.

References

Baranchuk, Dmitry, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. 2021. "Label-Efficient Semantic Segmentation with Diffusion Models." *CoRR* abs/2112.03126. [Link]

Chen, Yida, Fernanda Viégas, and Martin Wattenberg. 2023. "Beyond Surface Statistics: Scene Representations in a Latent Diffusion Model."