From Pixels to Pictures: Understanding the Internal Representation of Latent Diffusion Models

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

Abstract

Many AI image generators use diffusion models to create images from text prompts, including Stable Diffusion, which uses a latent diffusion model (LDM). Even though these LDMs are trained on 2D images, they can generate highly realistic and coherent 3D scenes. Our research explains and visualizes how the LDM encodes 3D information like saliency, depth, and shading. By probing the internal activations of the LDM's U-Net using linear probing classifiers, we conclude that 3D information is encoded as early as step 3 out of 15 of the denoising process. We can further explore how to modify the 3D layout of the generated image without changing the prompt or seed by editing the foreground mask or depth map.

Website:
https://ester-tsai.github.io/diffusion-model-internal-representation
Code:
https://github.com/karinaechen/diffusion-model-internal-representation

1 Introduction

The remarkable advancements of Latent Diffusion Models (LDMs) enable the generation of realistic images from textual descriptions (Rombach et al. 2022). LDMs generate images that contain coherent 3D scene and shading representations, even when trained solely on images lacking explicit depth or shading information. Our project investigates how the LDMs encode depth and shading information in their internal activations. This exploration is pivotal for understanding AI's interpretive capabilities and advancing image synthesis.

We successfully demonstrate that the internal representation of scene geometry is captured much earlier than the human eye can recognize during the diffusion process. As seen in Figure 1, each row starting with "Probe" starts to display relevant information much earlier in the diffusion process than its corresponding non-probe row. Each pair of rows (Mask, Depth, Shading) uses an algorithm to visualize either salient object detection, depth map, or shading map respectively. The top row in each pair of rows is the output of linear probes (Alain and Bengio 2018) trained on the image's internal representation (a tensor), while the bottom row is the output from an off-the-shelf model with the intermediate diffusion images as inputs. This finding concludes that object detection, depth map, and shading qualities are being encoded in the internal activations of the LDMs far earlier than the human eye can see in the intermediate diffusion images.

Additionally, we utilize VGG-16, a 16-layer-deep convolutional neural network trained on the ImageNet database, to explore when an image classification model would recognize an image. We found that the model is not able to detect the diffused image any earlier than the human eye, and in some cases, is slower.

Implications of this finding include ways to fine-tune the diffusion training process. If most of the information is being encoded in the first 80% of time steps, we could potentially speed up the remaining 20% of time steps. Furthermore, by understanding when information is being encoded in the denoising process, we can prevent potential adversarial attacks on the LDM.

1.1 Literature Review

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.

Our paper is heavily derived from the work done by Chen, Viégas and Wattenberg (2023), which found that linear representations of depth and saliency is indeed encoded within the internal activations of the LDM and appears early on in the denoising process. However, we also investigate the encoding of shading information, and show that this is similarly represented as depth and saliency.

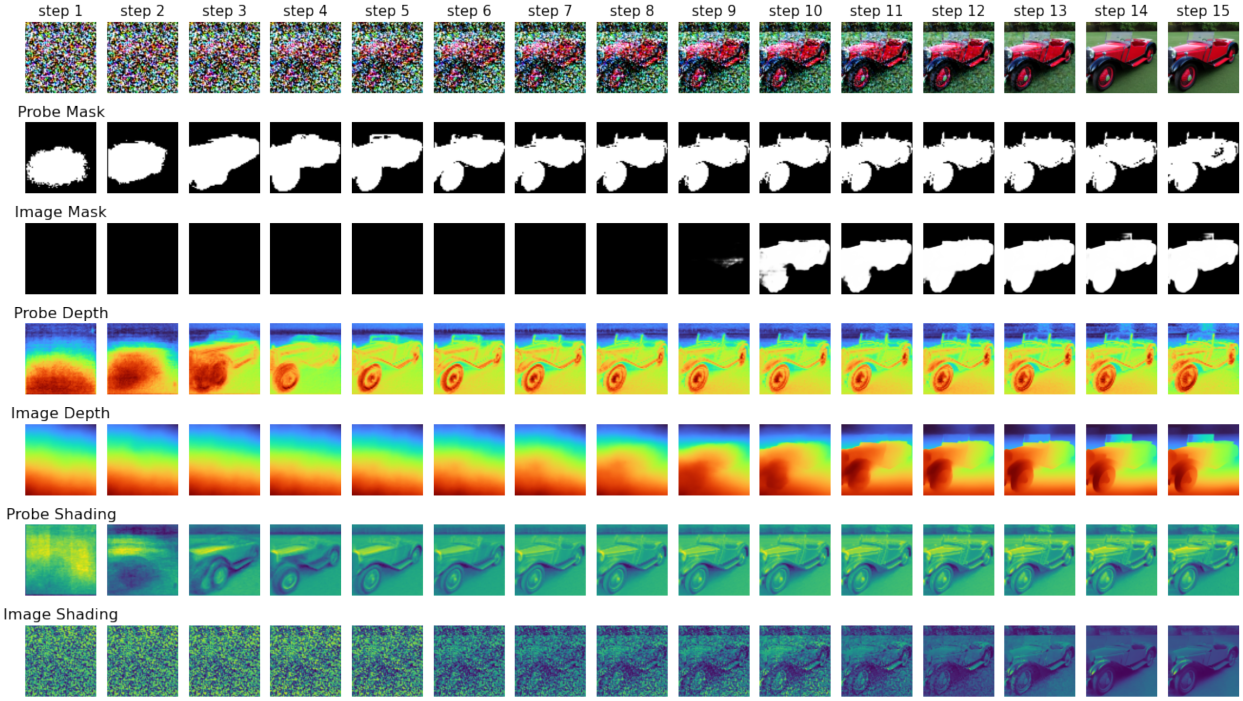

Figure 1: Intermediate steps for the generated image, probe, and model results.

1.2 Data Description

1.2.1 Generating the Image Dataset

Stable Diffusion is an open-source diffusion model that generates images from text prompts. Stable Diffusion is a two-stage framework that consists of (1) an LDM and (2) a variational autoencoder (VAE). The LDM learns to predict and remove noise by reversing a forward diffusion process. The VAE converts data between latent and image space. After the LDM synthesizes a denoised latent z, the decoder of VAE converts the denoised latent z to the image space.

Starting at v1.5, the Stable Diffusion model has a built-in depth model that can predict a depth map. We generate training images using the older Stable Diffusion v1.4 because it was trained without explicit depth information. Our result shows that even if the LDM is trained without explicit depth information, its internal representation of 3D information still exhibits an impressive ability to capture an underlying model of scene geometry.

Our diffusion image dataset consists of 617 images (512 pixels x 512 pixels) generated using Stable Diffusion v1.4. We have a CSV file that contains the prompt index, text prompt, and seed for each image. For example, the image with the prompt index 5246271, the text prompt "ZIGGY - EASY ARMCHAIR", and the seed 64140790 generated the 512 by 512 in Figure 2.

1.2.2 Generating the Ground Truth Images

The diffusion images we generate using Stable Diffusion v1.4 do not have ground truth labels for salient object detection, depth, or shading, so we apply off-the-shelf models to those images to synthesize the ground truth images. The labels are the same size as the diffusion images.

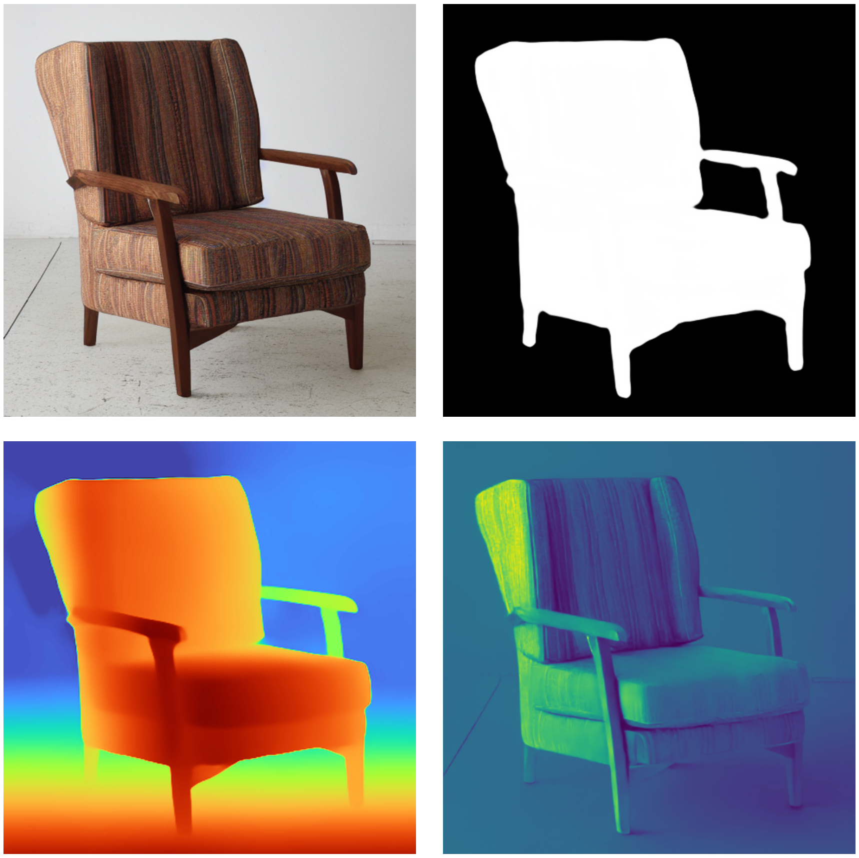

Figure 2: (top left) 512 x 512 image generated by Stable Diffusion v1.4 using the text prompt "ZIGGY - EASY ARMCHAIR" and seed 64140790.
(top right) Salient object detection mask generated by TRACER.
(bottom left) Depth map generated by MiDaS.
(bottom right) Shading and illumination map generated by Intrinsic.

For salient object detection, we apply the salient object tracing model TRACER by Lee, Shin and Han (2022) to generate a mask for each image. The masks are black and white, where white indicates the salient object, or foreground, and black indicates the background.

For depth labels, we apply the pre-trained MiDaS model designed by Ranftl et al. (2020) to the diffusion images to estimate their relative inverse depth maps.

For shading labels, we apply the pre-trained Intrinsic model designed by Careaga and Aksoy (2023) to the diffusion images to generate highly accurate intrinsic decompositions and estimate the shading maps.

2 Methods

Our research method draws inspiration from Chen, Viégas and Wattenberg (2023) and moves beyond their focus on depth to explore other image information such as shading and illumination. We also explore at what point image recognition models like VGG-16 can correctly identify the image subject in the reverse diffusion process.

2.1 Extracting the Internal Representation

Our research aims to explain and visualize the changes in the LDM's latent space as the LDM generates an original image from pure noise. We extract the self-attention layer's intermediate outputs from the U-Net denoising block within the LDM at each denoising step. We use a hook to save the features of a U-Net module every time it runs. Each *features* Tensor has the shape of torch.Size([2, 4096, 320]). We select the *features* from a specific time step, block type, block index, and layer as the input for the linear probing classifier. The name of the module we select from the U-Net is "up_blocks.3.attentions.0.transformer_blocks.0.attn1.to_out.0," which stands for the block "up,", block index 3, layer index 0, and layer name "transformer_blocks.0.attn1.to_out.0."

2.2 Developing the Probing Classifier

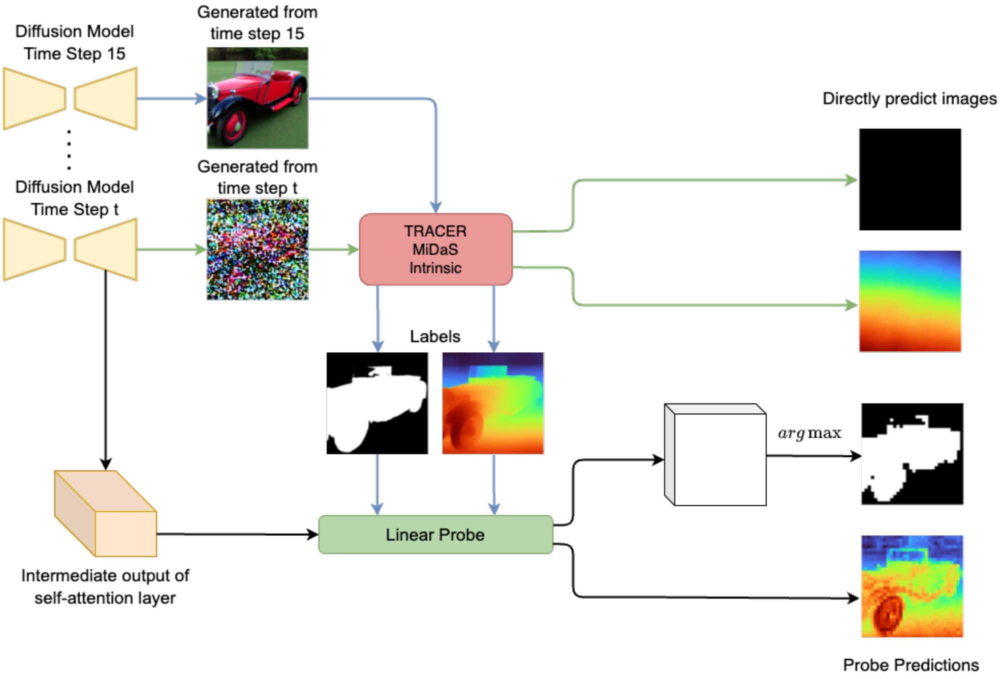

Figure 3: Probing Workflow

The linear probe is a neural network with a linear layer. We train the probe for 30 epochs using the Adam optimizer and cross entropy loss.

The linear probe takes the internal representation of images as its input, with the outputs from the TRACER, MiDaS, and Intrinsic models serving as labels. These labels represent the accurate salient, depth, and shading information derived from the final image.

The probing binary classifier trained on the TRACER image outputs distinguishes foreground from background at the pixel level.

For continuous attributes such as depth and shading, we retrieve the outputs from the self-attention layers and employ a linear regressor trained on these outputs to predict the nuanced variations in depth and shading. The linear regressor trained on the MiDaS image outputs predicts a depth map that shows the 3D scene depicted in the Stable Diffusion image. The linear regressor trained on the Intrinsic image outputs predicts a shading map that shows the illumination depicted in the Stable Diffusion image.

2.3 Choosing the Best Performing Probe

Table 1: Spatial and Feature Dimensions of Self-Attention Layers in the LDM

Blocks	Number of Self-Attn Layers	Spatial $h \times w$	Feature c
Encoder 1	2	64 × 64	320
Encoder 2	2	32 × 32	640
Encoder 3	2	16 × 16	1280
Encoder 4	0	-	-
Bottleneck	1	8 × 8	1280
Decoder 1	0	-	-
Decoder 2	3	16 × 16	1280
Decoder 3	3	32 × 32	640
Decoder 4	3	64 × 64	320

The probes are trained on different U-Net blocks and layers and have varying performances because of their difference in location, number of features, and output size (see Table 1). We choose the probe trained on Decoder 4 layer 1 because it produces the biggest outputs and achieves the highest performance metrics. Here are the performance metrics we consider:

Dice coefficient is a similarity measure that quantifies the similarity between two binary images. We use the Dice coefficient to compare the probe performances for salient object detection. The Dice coefficient is defined as:

$$\text{Dice} = \frac{2 \times \text{area of overlap}}{\text{total area}}$$

The Dice coefficient ranges from 0 to 1, where:

- 0 indicates no overlap between the binary images (complete dissimilarity).
- 1 indicates complete overlap between the binary images (complete similarity).

Spearman's rank correlation coefficient measures the strength and direction of association between the ranks of pixel intensities in two images. We use rank correlation to compare the probe performances for depth and shading prediction.

The value of Spearman's rank correlation coefficient (ρ) ranges from -1 to 1:

- ρ = 1 indicates a perfect positive monotonic relationship (as pixel intensities increase in one image, they also increase in the other image). We want our probing results to show high rank correlation with the labels.

- ρ = -1 indicates a perfect negative monotonic relationship (as pixel intensities increase in one image, they decrease in the other image).

- ρ = 0 indicates no monotonic relationship between the pixel intensities of the two images.

Pearson's linear correlation coefficient measures the strength and direction of the linear relationship between two images. We use linear correlation in addition to rank correlation to compare the probe performances for depth and shading prediction. We want our probing results to show a high linear correlation with the labels.

Using the output from Decoder 4 layer 1, our saliency, depth, and shading probes produce the following test metrics at the last time step, which are higher than the results from all other blocks and layers (Table 2).

2.4 Image Classification Using VGG-16

We utilized VGG-16, a 16-layer-deep convolutional neural network trained on the ImageNet database, to explore when an image classification model would recognize an image.

First, we generate images using Stable Diffusion using prompts that match ImageNet categories (e.g. "lemon"). For each generated image, we save a total of 15 images, one for each time step. We then put each intermediate image into VGG-16 to perform object classification. After taking the top-five predictions of VGG-16, we plot how the probability for each class changes over the time steps to examine when the model's prediction starts to become confident. See Figures 5 and 6 for the results.

3 Results

The probes are given the LDM's internal representation as inputs while the off-the-shelf models are given intermediate images as inputs. Figure 4 illustrates the LDM's internal encoding of 3D information by comparing the results from the probes with the image models (TRACER, MiDaS, and Intrinsic). For the first 9 time steps, due to the "noisy" nature of the intermediate images, these models struggle to generate accurate predictions. However, as the denoising time steps progress, the models gradually improve in accuracy, coinciding with the point at which human observers start to discern relevant information more clearly.

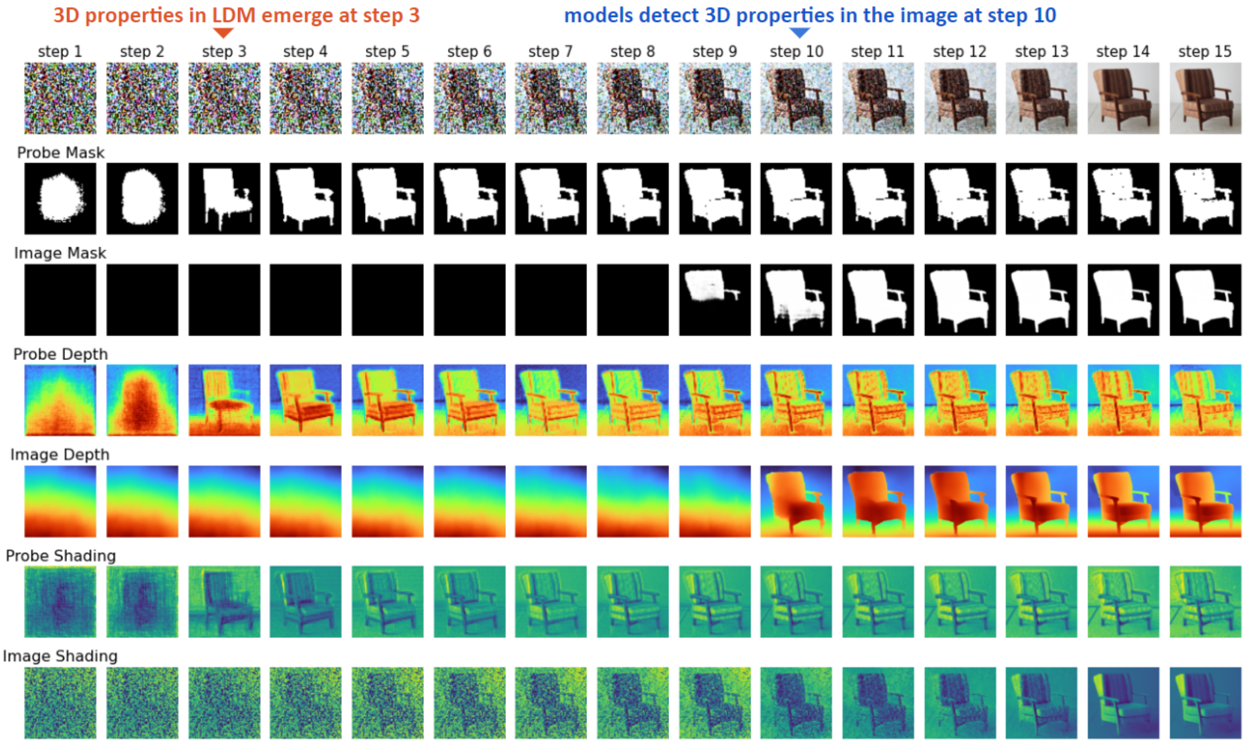

Figure 4: Intermediate steps for the generated image, probe, and model results. 3D properties like saliency, depth, and shading develop in the LDM's internal representation as early as time step 3. However, TRACER (for predicting salient object detection mask), MiDaS (for predicting depth), and Intrinsic (for predicting shading) detect 3D properties much later at time step 10, showing that the LDM encodes 3D information much earlier than the models can detect them in the noisy images.

Table 2: Probe Test Metrics

Task	Metric	Score Between -1 and 1
Foreground Segmentation	Dice Coefficient	0.85
Pixel-Wise Depth Estimation	Rank Correlation	0.71
Pixel-Wise Depth Estimation	Linear Correlation	0.74
Pixel-Wise Shading Estimation	Rank Correlation	0.62
Pixel-Wise Shading Estimation	Linear Correlation	0.64

Table 2 displays the test metrics for the probe trained on the activation output of Decoder 4 layer 1 (Table 1). The metrics are explained in Section 2.3. All of the metrics are collected from the last time step. The Dice Coefficient for the salient object detection (i.e. foreground segmentation) probe is 0.85. The pixel-wise depth estimation probe has a Rank Correlation of 0.71 and a Linear Correlation of 0.84. The pixel-wise shading estimation probe has a Rank Correlation of 0.62 and a Linear Correlation of 0.64.

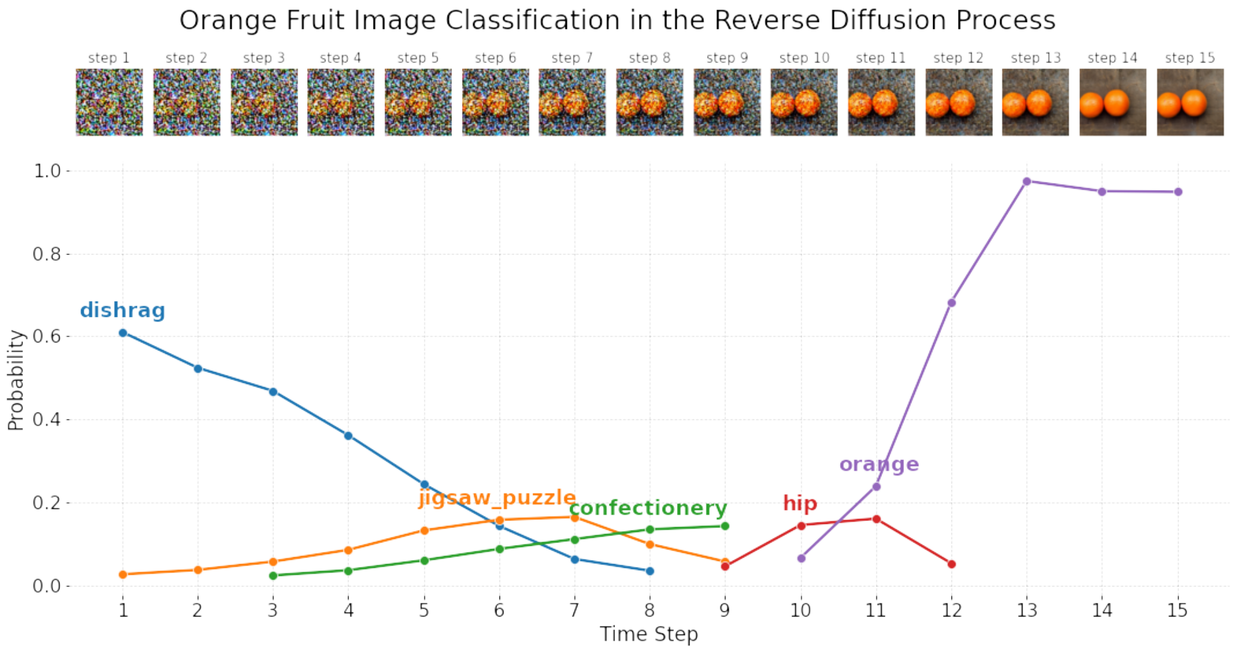

Figure 5: The top VGG-16 predictions and its confidence level at each time step for a Stable Diffusion image generated using the prompt “orange fruit”.

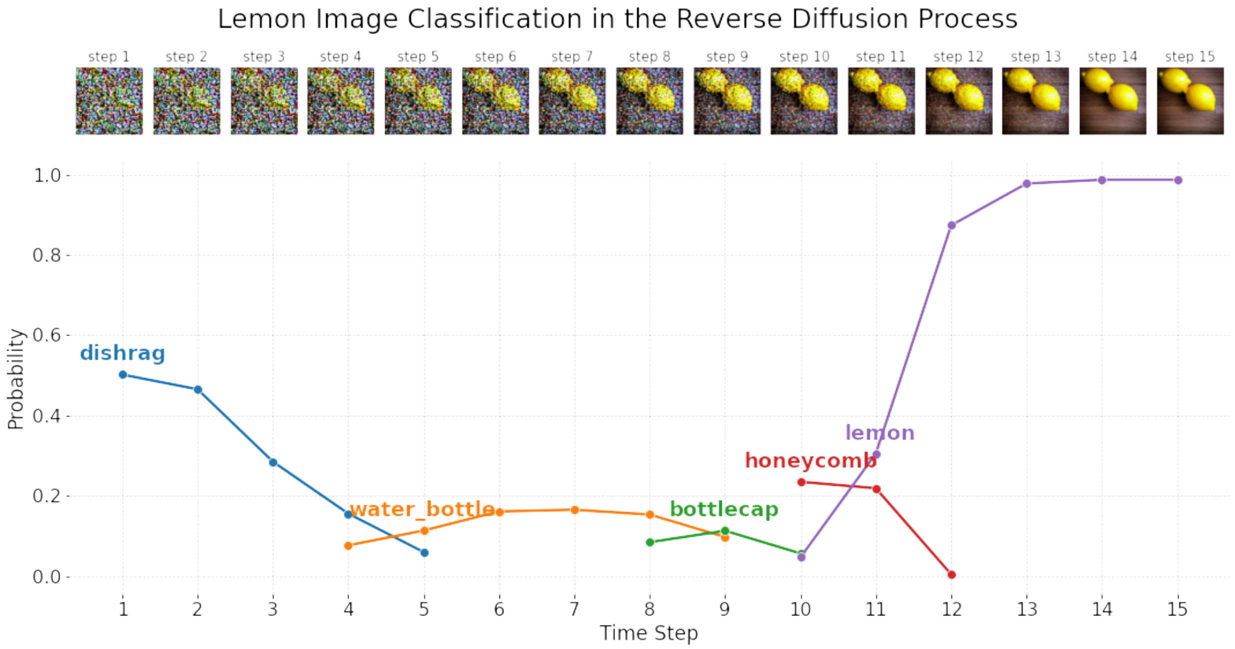

Figure 6: The top VGG-16 predictions and its confidence level at each time step for a Stable Diffusion image generated using the prompt "lemon".

Figure 5 shows the top VGG-16 predictions and its confidence level at each time step for a Stable Diffusion image generated using the prompt "orange fruit". For the first 5 time steps, VGG-16 predicts the noisy image as a dishrag. In the middle of the denoising process, VGG-16 predicts the noisy image as a jigsaw puzzle, confectionery, or hip. Finally, at time step 11, VGG-16 predicts "orange" as the most probable object.

Figure 6 shows the top VGG-16 predictions and its confidence level at each time step for a Stable Diffusion image generated using the prompt "lemon". For the first 4 time steps, VGG-16 predicts the noisy image as a dishrag. In the middle of the denoising process, VGG-16 predicts the noisy image as a water bottle, bottlecap, or honeycomb. Finally, at time step 11, VGG-16 predicts "lemon" as the most probable object.

4 Discussion

4.1 Interpretation of Results

Since all of the test metrics in Table 2 are close to 1, we have evidence that the probing results have a strong positive correlation with the synthetic labels using off-the-shelf models. These test metrics indicate that the probes are generally very good at capturing each of those 3D properties encoded in the LDM's activations. The depth probe archives a higher Rank Correlation and Linear Correlation than the shading probe, providing evidence that the depth probes are better at predicting depth than the shading probes are at predicting shading.

Figure 4 reinforces the same conclusion as Table 2. 3D properties like saliency, depth, and shading develop in the LDM's internal representation as early as time step 3. However, TRACER (for predicting salient object detection mask), MiDaS (for predicting depth), and Intrinsic (for predicting shading) detect 3D properties much later at time step 10, showing that the LDM encodes 3D information much earlier than the models can detect them in the noisy images.

VGG-16 only started to predict the correct class around step 12 for most of the images we tested (see Figure 5 and Figure 6), contradicting our initial hypothesis that an image classification model could find information that was encoded by the diffusion model early in the denoising process. Contrary to our hypothesis, VGG-16 classifies the object correctly after a human eye would. We knew previously that depth information was being encoded in the internal representation (the tensors), but wanted to explore if the same was happening for the actual image (the pixels). With these results, we've found that the model is not encoding information in the actual image, and that image classification models don't recognize the target class any earlier than a human eye.

4.2 Comparison to Prior Work

Our results provide evidence that shows an encoding of 3D information within the neural network is present early in the generative process. Despite our use of Stable Diffusion v1.4, which is not trained with any depth prior, 3D properties that determine the spatial layout of the image are evident early on in the denoising process. In Stable Diffusion v2.0, they introduced a depth-to-image model that utilizes depth information when generating new images. This model estimates a depth map of the input image using MiDaS, and then uses this map as another condition in addition to the original text prompt and original image conditions to generate an image.

We come to the same conclusion as Chen, Viégas and Wattenberg (2023), as depth and saliency both develop at a stage where the image still appears noisy to the human eye. We expand on their results by also investigating shading and illumination information, and come to the conclusion that this type of 3D information is similarly encoded and appears at the same time step as saliency and depth.

Baranchuk et al. (2021) find that the intermediate activations of the network in the reverse diffusion process capture high-level semantic information. We similarly find that the internal representations encode depth information.

Kim et al. (2023) wanted to generate a depth-aware guidance system for diffusion models using estimated depth information derived from the intermediate representations of the U-Net. This implies a similar result to our work, that there is an encoding of depth within the internal representation of the diffusion model.

Wang et al. (2022) were able to convert the pretrained 2D Stable Diffusion model on images into a 3D generative model of radiance fields without extra 3D information given. We investigate the encoded 3D information already present in the 2D diffusion model instead of extending it to the 3D realm.

5 Future Works

5.1 Intervention

In the paper by Chen, Viégas and Wattenberg (2023), they investigate intervening and modifying the internal representations in order to reposition the salient object. This would be an interesting result to replicate, build upon, and also compare to the depth-to-image capability of Stable Diffusion v2.0. It has implications for making image generation even more customizable and realistic, as the generated image can be tailored to users' needs.

5.2 Speeding up Diffusion

If the bulk of the information is already encoded by very early steps in the denoising process, we can potentially speed up the rest of the steps without sacrificing quality. This brings in enormous cost savings, as training top-of-the line diffusion models like Stable Diffusion can cost hundreds of thousands of dollars, if not more.

5.3 Augmenting Datasets

When it comes to autonomous vehicles, where safety and accuracy are paramount, the need for diverse and comprehensive datasets is critical. One way to enhance these datasets is by leveraging encoded depth information through diffusion models.

Usually, depth information is captured using a LiDAR sensor or depth cameras. However, collecting such information can be resource intensive, and existing images may not have these pieces of information.

Diffusion models can help here in two ways. Synthetic depth maps that closely resemble real-world scenarios can be generated, and potentially filling in depth information for existing images or images without complete depth maps.

References

Alain, Guillaume, and Yoshua Bengio. 2018. "Understanding intermediate layers using linear classifier probes."

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

Careaga, Chris, and Yağız Aksoy. 2023. "Intrinsic Image Decomposition via Ordinal Shading." *ACM Trans. Graph.*

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

Kim, Gyeongnyeon, Wooseok Jang, Gyuseong Lee, Susung Hong, Junyoung Seo, and Seungryong Kim. 2023. "DAG: Depth-Aware Guidance with Denoising Diffusion Probabilistic Models."

Lee, Min Seok, Wooseok Shin, and Sung Won Han. 2022. "TRACER: Extreme Attention Guided Salient Object Tracing Network."

Ranftl, René, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. "Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer."

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. "High-Resolution Image Synthesis with Latent Diffusion Models."

Wang, Haochen, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. 2022. "Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation."