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

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

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

Project Background

What is Stable Diffusion?

- Stable Diffusion is an open-source diffusion model that generates images from text prompts.
- Stable Diffusion is a two-stage framework that consists of:
  - A latent diffusion model (LDM)
    - The LDM learns to predict and remove noise in the latent space by reversing a forward diffusion process.
  - A variational autoencoder (VAE)
    - 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.

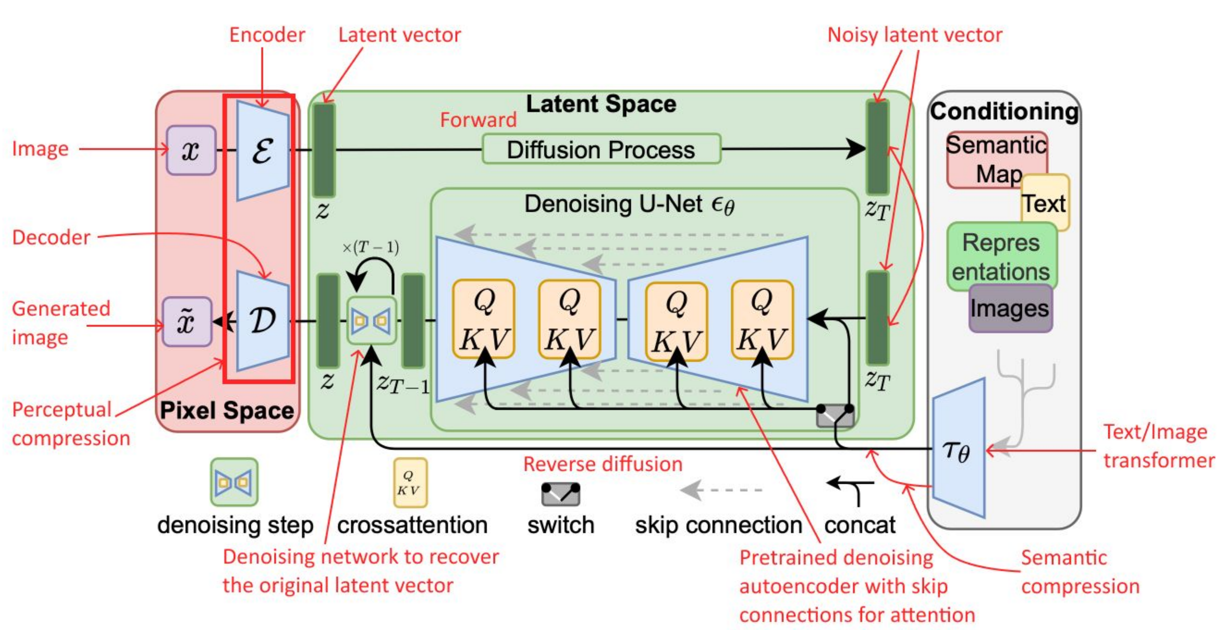


Figure 1. Architecture of an LDM (Rombach, Blattmann 2022)

Problem Statement

- Does an LDM create an internal 3D representation of the object it portrays?
- How early in the denoising process do depth, saliency, and shading information develop in the internal representation?
- At what time step does an image classifier correctly detect the object?

Data

617 images (512 pixels x 512 pixels) generated using Stable Diffusion v1.4

Image generated by Stable Diffusion v1.4 using the text prompt "ZIGGY - EASY ARMCHAIR" and seed 64140790.

Salient object detection mask generated by TRACER.

Shading and illumination map generated by Intrinsic.

Depth map generated by MiDaS.

Figure 2. Ground truth images

Internal Representation

Methods

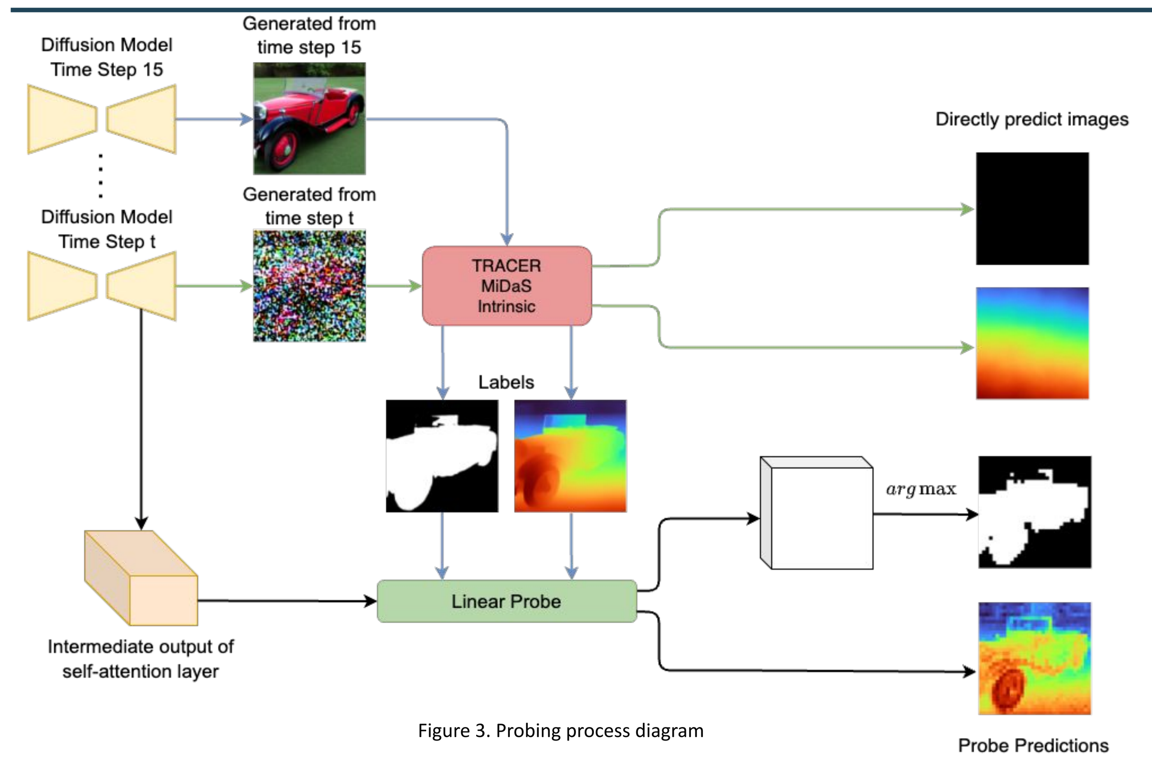


Figure 3. Probing process diagram

- **Internal representation** = the neural network's self-attention layer's intermediate activation output.
- **The linear probe model**
  - **Input:** internal representation of a LDM.
    - A tensor of shape [2, 4096, 320].
    - At a specific time step, for a specific block and layer of the U-Net.
  - **Output:** predicted image showing a certain property.
    - e.g. depth, salient-object detection, shading.

Results: Probing the LDM

Probe performance at the last step	Score between -1 and 1
Foreground Segmentation Dice Coefficient	0.85
Depth Estimation Rank Correlation	0.71
Shading Estimation Rank Correlation	0.62

- Using intermediate activations of noisy input images, linear probes can accurately predict the foreground, depth, and shading.
  - Shown by high Dice Coefficient and Rank Correlation in the table.
- All three properties emerge early in the denoising process (around step 3 out of 15), suggesting that the spatial layout of the generated image is determined at the very beginning of the generative process.

3D properties in LDM emerge at step 3 **models detect 3D properties in the image at step 10**

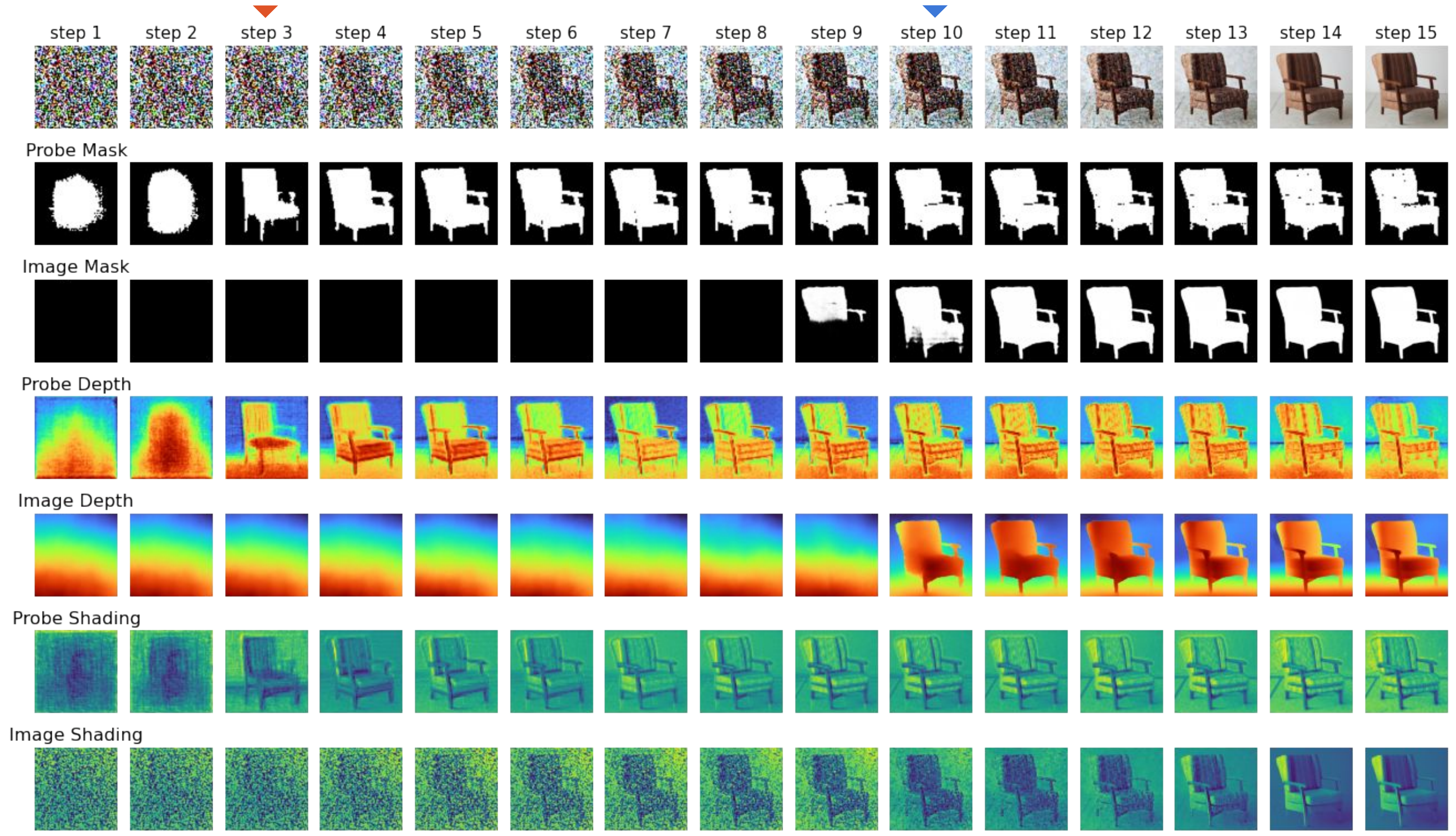


Figure 4. Intermediate steps for the generated image, probe, and model results

Image Classification

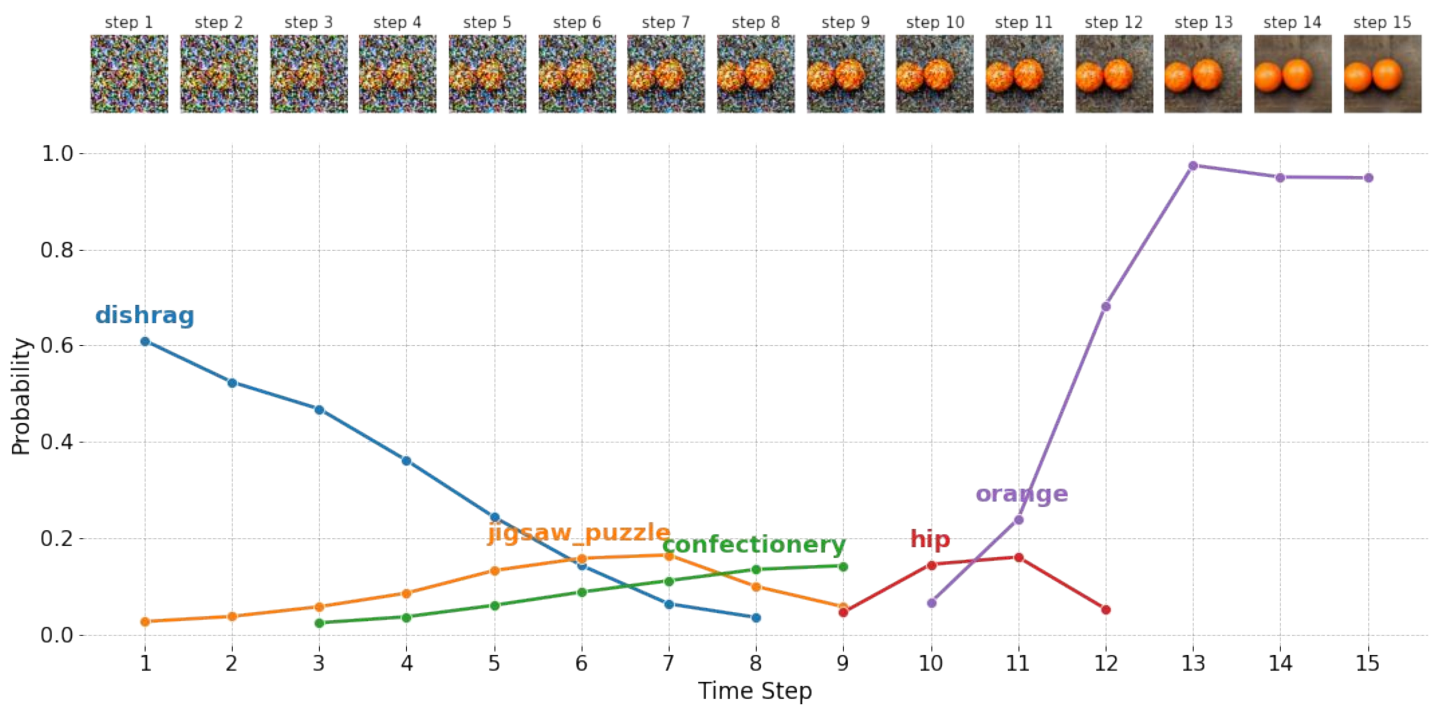
Methods

1. Generate images using Stable Diffusion with prompts that match ImageNet categories.
   - For example, prompt = "lemon".
2. Run each intermediate image through VGG-16, an image classification model trained on ImageNet.
3. Visualize predictions results.

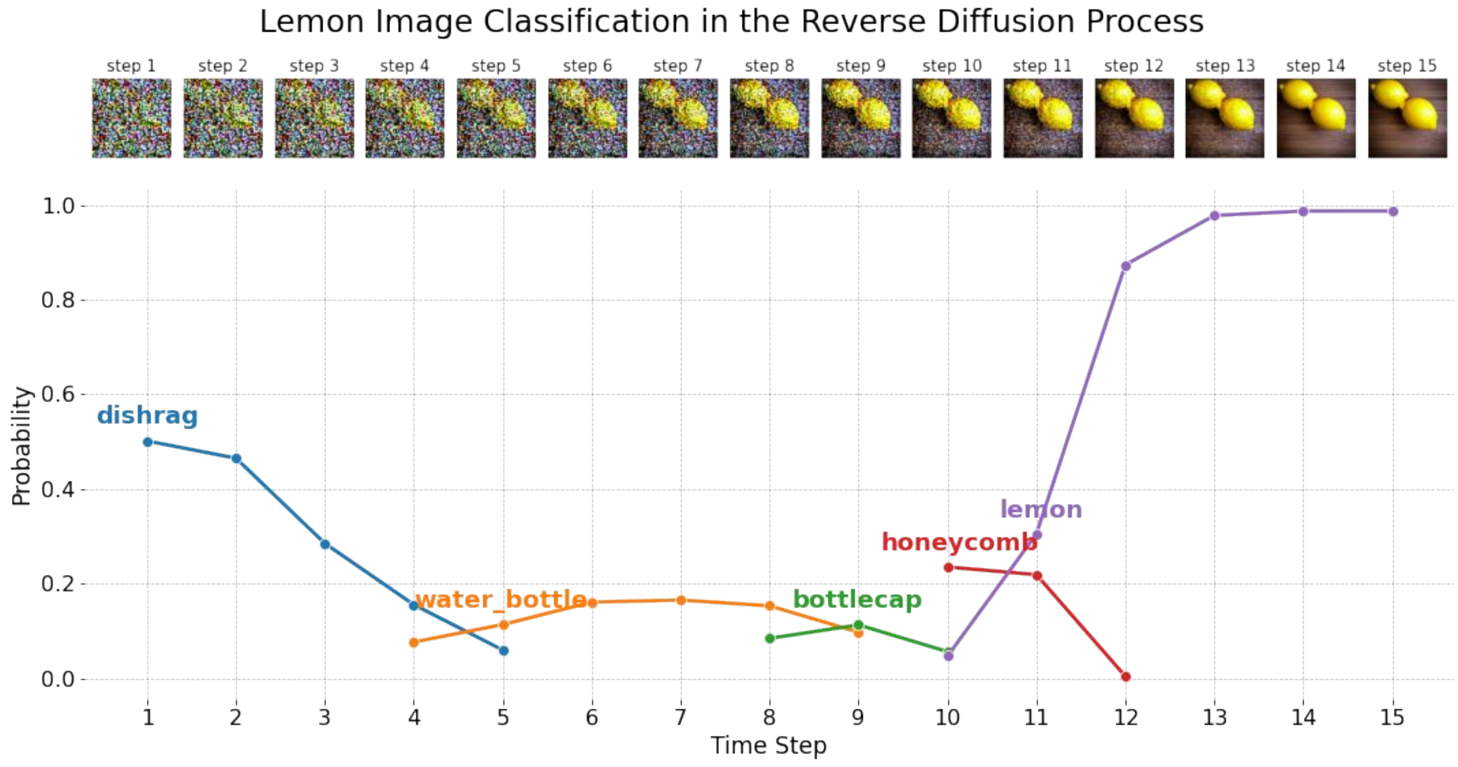
Results

- Comparing classification confidence for generated vs. real images.
  - Generated images: two lemons (98.75%), two oranges (94.8%).
  - Real images: two lemons (99.4%), singular lemon (87.7%), singular orange (87.0%).
- The correct classification has high confidence (> 90%) towards the end of the diffusion process for the majority of generated images.
  - This means that the generated images are fairly good representations of the object prompted.
- VGG-16 correctly identifies the object after step 11.

Orange Fruit Image Classification in the Reverse Diffusion Process



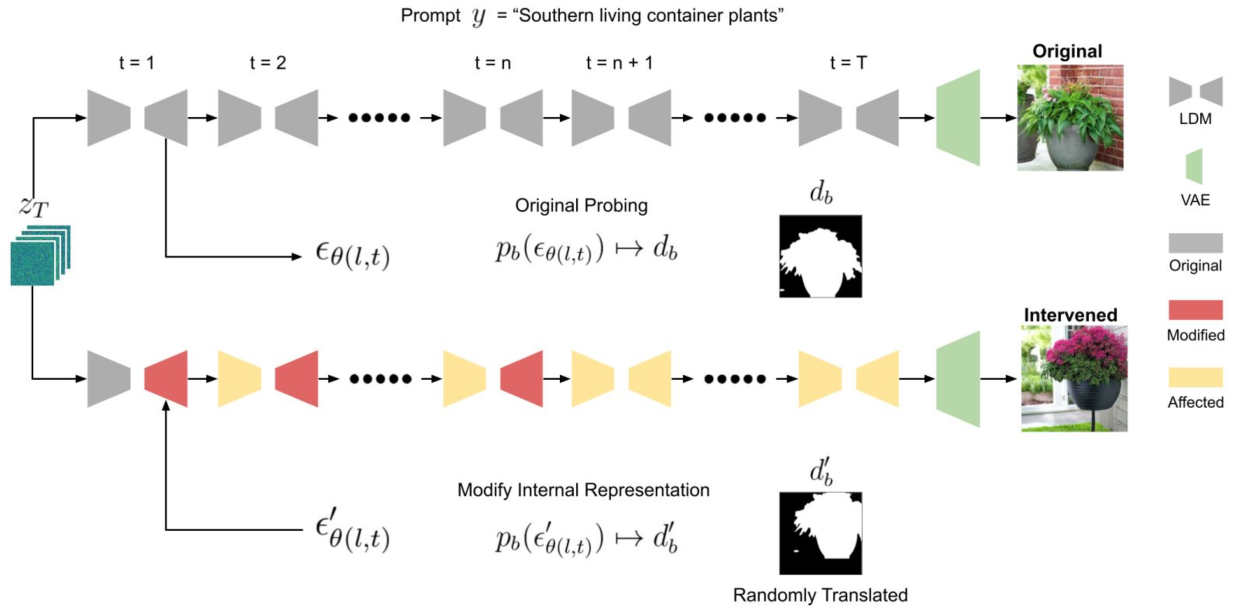
Lemon Image Classification in the Reverse Diffusion Process



Future Works: Intervening the LDM

Figure 5. The Intervention workflow (Chen, 2023).

- The foreground object can be repositioned by modifying the activations of the U-Net decoders.
- First, obtain a target mask by translating the original mask.
  - Goal: to find the activations (i.e. probe inputs) that cause the probe to output a mask highly similar to the target mask.
- Perform gradient descent on the activations until the probe can output the desired target mask.
- Replace the original activations with the modified activations, then resume the denoising process.



- Foreground mask has a causal role in image generation.
- Intervention: Without changing the prompt, input latent vector, and model weights, we can modify the scene layout of the generated image by editing the foreground mask (Y. Chen et al.).

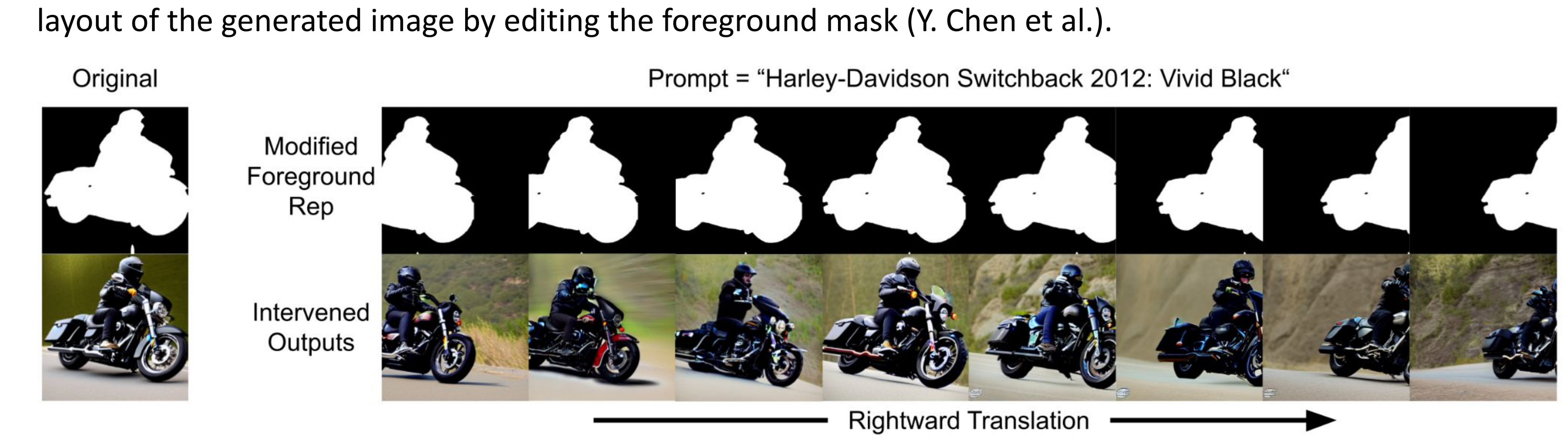


Figure 6. Intervening the LDM to produce different outputs (Chen, 2023)