DreamPLACE : Person Scene Composition

Harsh Gupta,Rishubh Parihar, R. Venkatesh Babu

Indian Institute of Science, Bangalore

Introduction

- Our project introduces an innovative two-phase method for seamlessly integrating subjects into background scenes using text-guided image processing.
- **Phase 1: Mask Preparation** Involves processing a text prompt, background image to generate an inpainting mask. This mask identifies the optimal subject placement within the scene.
- **Phase 2: Scene Synthesis.** Utilizes the Stable DiffusionXL pipeline with Textual Inversion, taking the initial inputs and the generated mask to synthesize the final scene, ensuring natural and coherent integration of the subject.

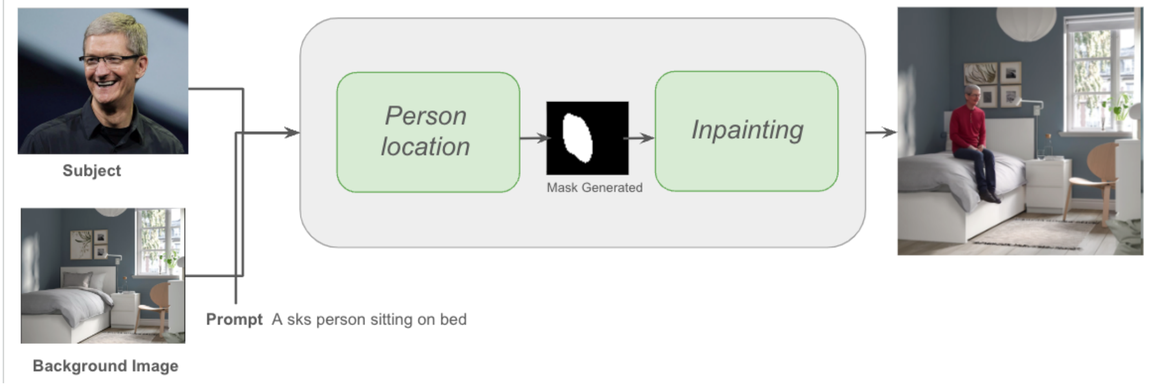

Figure 1: Framework Overview: Three inputs - subject, background image, and prompt ("a person sitting on a bed") ,where background image and prompt are processed to generate a mask. This mask, along with other inputs, guides the inpainting pipeline for subject-scene integration.

Methodology

Phase 1: Inpainting Mask Creation and Person Embedding

- Description of Phase 1: This phase focuses on generating an inpainting mask and embedding the person into the scene using the mask encoder. It takes the background image and textual description to ascertain the optimal placement location for the person.

Human location is represented as binary mask in the image space

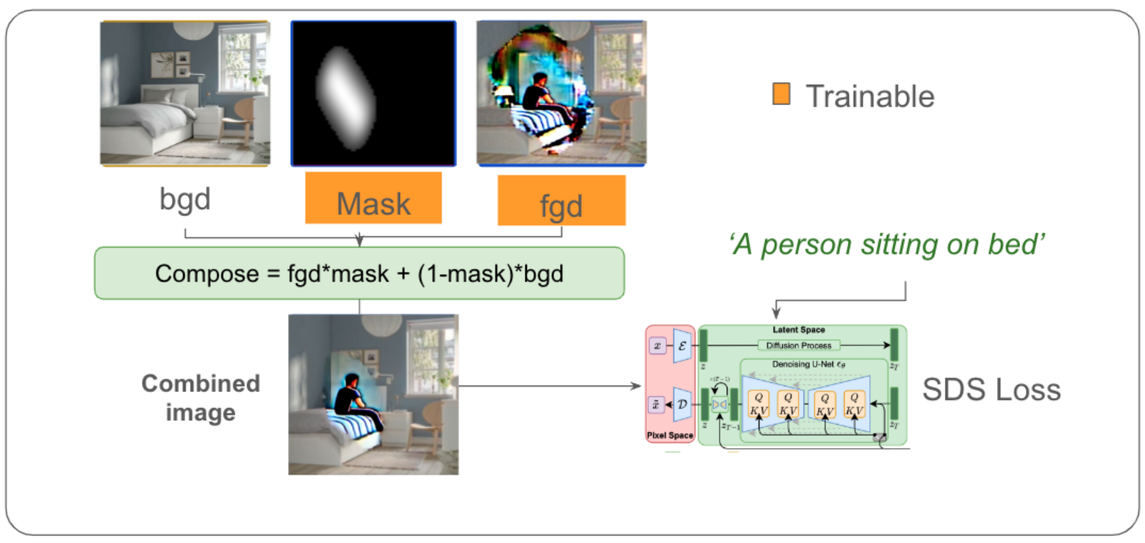

figure 2: Placement Module,which identifies the optimal location within the scene for subject placement

Phase 2: Scene Synthesis via Stable Diffusion DDPM

- Description of Phase 2: This phase involves scene synthesis using the Stable Diffusion DDPM pipeline. The model processes the background image, the mask from Phase 1, and the textual description to produce a final image that blends the person naturally with the background.

Learn the embedding of a given subject then use SD-XL inpainting

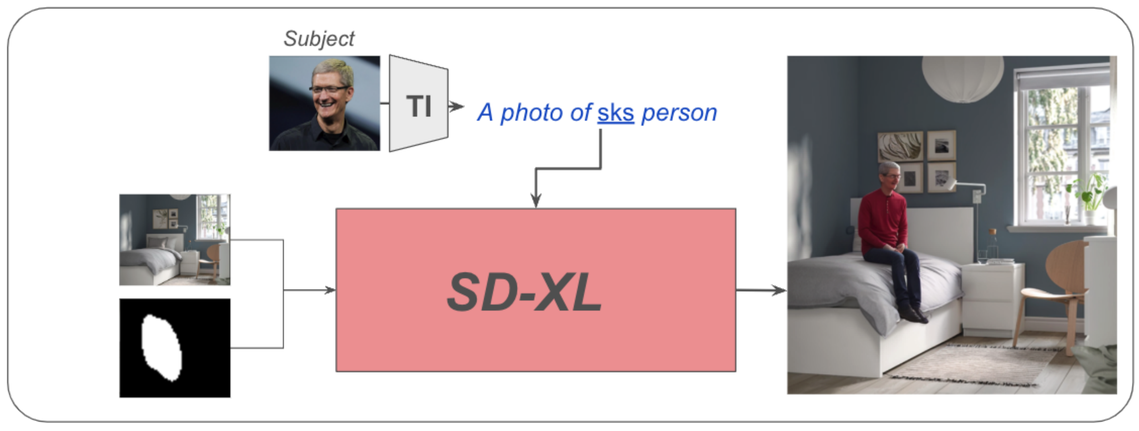

figure 3: Personalization Module, responsible for embedding the subject into the scene in a seamless and contextually appropriate manner

Mask Initialization Process

- The algorithm begins by initializing 'n' Gaussian blobs per subject in the image. It starts with one central blob and adds others based on relative distances, each with a radius of 0.1.
- The centers of these blobs are tied together using relative angles to form a coherent shape. The equation for calculating the center of each subsequent blob is as follows:

$$\text{center}_{i+1} = \text{center}_i + \begin{bmatrix} \text{radius} \cdot \cos(\text{relative\_angles}[i+1]) \\ \text{radius} \cdot \sin(\text{relative\_angles}[i+1]) \end{bmatrix} \quad (1)$$

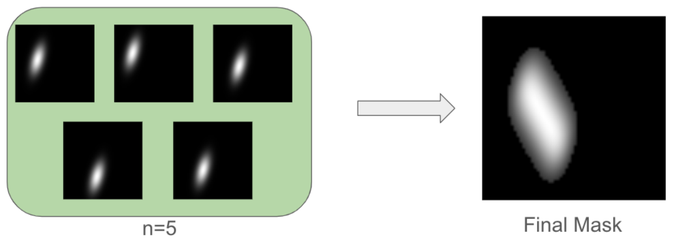

figure 4: Visualization of the Gaussian Blob Mask Generation Process.

Results

figure 5: Diverse Scene Compositions with Subject Tim Cook. This image showcases the versatility of our methodology in placing the subject across various scenarios.

figure 6: Comprehensive Results with Various Subjects. The first row details the subjects used in the experiment, while the first column displays the masks tailored for each background.

Conclusions and Future Work

Our project marks a significant advancement in the field of inpainting mask generation, specifically in aligning subjects with scenes based on textual prompts.

Future Work: We are enhancing the realism of embedded subjects, particularly in facial features and body details based on Custom Diffusion + TI. A major ongoing challenge is optimizing the scale of Gaussian blobs in our process of mask generation.

References

- Ben Poole, Ajay Jain, Jonathan T. Barron, Ben Mildenhall, (2023) DreamFusion: Text-to-3D using 2D Diffusion
- Amir Hertz, Kfir Aberman, Daniel Cohen-Or(2023), Delta Denoising Score