Report: Text Controlled Object Relocation and Relighting

From: Muhammad Faheem

Table of Contents

## Objectives

## Pipeline Overview

## Strengths

## Existing Issues and Limitations

## Future Work

# Objectives

To develop a vision-language pipeline capable of detecting, segmenting, removing, and relocating objects in a scene based on a natural language prompt, while maintaining visual consistency through inpainting and relighting.

# Pipeline Overview

1. **Input**:

* Scene image (e.g., scene\_resized.png)
* Text prompt (e.g., *"move the red chair to the right"*)

1. **Parsing**:

* Extract object of interest and action (relocation direction) from prompt using rule-based NLP.

1. **Object Detection**:

* Used **Grounding DINO** (via Replicate API) for text-guided detection.
* Switched to **DETR** from Hugging Face Transformers for local, open-source detection.
* **Limitation**: DETR lacks text grounding, which slightly reduced precision without manual filtering.

1. **Object Segmentation**:

* Used **Segment Anything Model (SAM)** to extract precise binary masks.
* Initial issue: Fully black masks due to incorrect normalization; resolved by verifying post-processing and mask scaling.

1. **Object Removal**:

* Performed **inpainting** using **Stable Diffusion Inpainting** via Replicate.
* Prompt used: *“Remove the object {parsed['object']} and fill in the background naturally.”*

1. **Object Relocation**:

* Object extracted using mask and composited into new location.

# Strengths

1. Fully modular and testable pipeline.
2. Leverages SOTA models (SAM, DETR, Stable Diffusion).
3. Respects visual realism using alpha blending and bounds checking.
4. Localizable and scalable (Hugging Face models can run offline).

# Existing Issues and Limitations

1. **Detection** via DETR lacks grounding capability — object class filtering needed.
2. **Relighting** is currently unimplemented or handled implicitly via inpainting; could be improved with neural relighting or depth-aware models.
3. **Perspective matching** is not supported in relocation (warp can be added).
4. **Edge artifacts** may occur if mask is coarse or misaligned.

# Future Work

1. Integrate grounding-aware detectors like **Grounding DINO** locally (via TorchServe or Hugging Face).
2. Add **depth estimation** and **neural relighting** modules (e.g., RelightNet, Stable Zero-DCE).
3. Allow **multi-object prompts** and more complex spatial reasoning.
4. Optimize for real-time or interactive use cases.