Department of Informatics
Athens University of Economics and Business (AUEB)

Generative Image Inpainting for Text-Guided Object Removal

using a Conditioned UNet and LoRA-Tuned Diffusion Model

Athanasopoulou Lefki
f3352404

https://github.com/LefkiAth/Deep-Learning-Project

July 11, 2025

Contents

1. Abstract

The task of removing specific objects from images based on textual instructions is a complex challenge in computer vision, requiring a nuanced understanding of both language and visual context. This paper presents a two-stage pipeline to address this problem. The first stage introduces a novel **Conditioned UNet** architecture for segmentation mask generation. This model, which utilizes a ResNet-18 backbone and a **CLIP text encoder** connected via cross-attention, achieved a Dice score of 0.4386 on the test set, significantly outperforming a pretrained **CLIPSeg baseline**. The second stage employs the **stabilityai/stable-diffusion-2-inpainting** model, fine-tuned using Low-Rank Adaptation **(LoRA)**, to inpaint the regions identified by the segmentation masks. Qualitative analysis shows that our fine-tuned inpainting model produces substantially more context-aware and seamless results than the original baseline model, particularly in non-human object removal scenarios. This work demonstrates a complete and effective end-to-end pipeline for text-guided object removal.

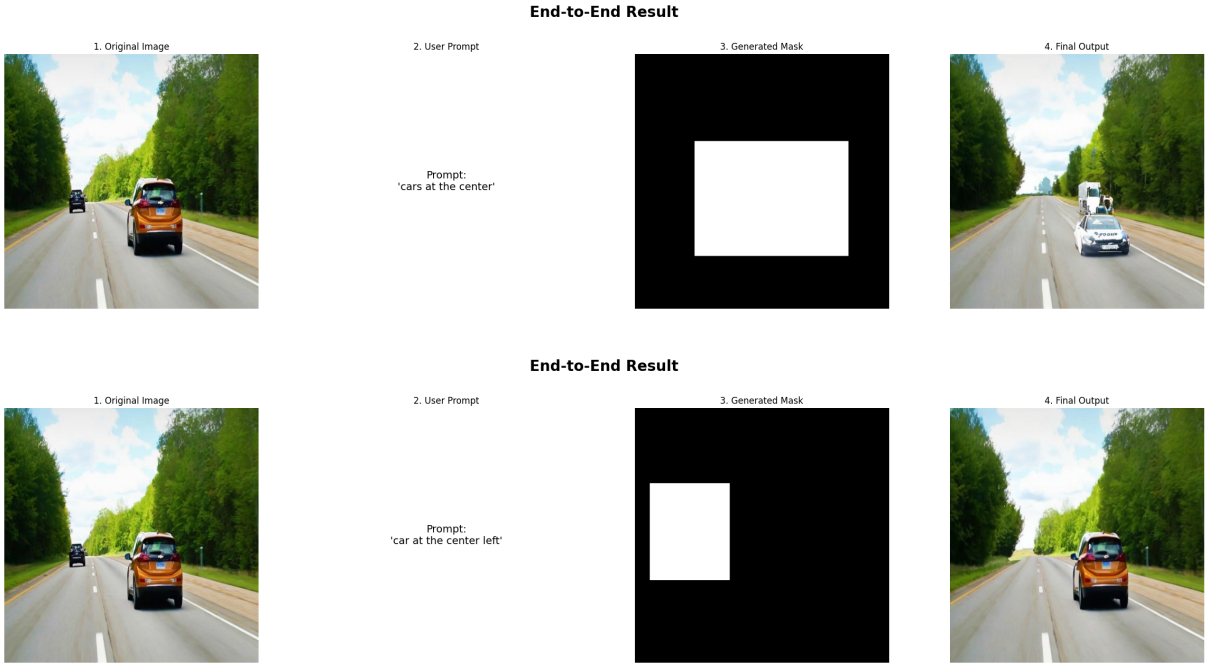


Figure 1: Demonstration of our pipeline's ability to interpret nuanced text prompts. **Top:** Given the plural prompt "cars at the center", the model generates a mask encompassing both vehicles to be removed. **Bottom:** When the prompt is changed to the singular "car at the center left", the pipeline correctly isolates and generates a mask for only the single specified vehicle. This highlights the model's fine-grained contextual understanding, moving beyond simple object detection to precisely follow user instructions.

2. Introduction

The ability to manipulate digital images with natural language commands represents a significant frontier in human-computer interaction. One of the most practical applications in this domain is text-guided object removal, where a user can specify an object to be deleted from a photo using a simple text prompt. This technology has wide-ranging applications, from professional photo editing and content creation to data augmentation and privacy protection.

However, this task presents considerable challenges. A successful system must first correctly interpret the user's textual prompt to identify the target object. It then needs to perform precise spatial segmentation of that object within the image. Finally, it must inpaint the resulting void with a background that is not only visually realistic but also semantically consistent with the rest of the scene.

This paper proposes a two-stage deep learning pipeline to tackle this problem. Our contributions are:

1. The design and implementation of a **Conditioned UNet**, which integrates text embeddings directly into an image segmentation pipeline using cross-attention, for precise mask generation.
2. A rigorous comparison of our model against a CLIPSeg baseline, demonstrating a **17.4% improvement** in Dice score on the test set.
3. The application of **Low-Rank Adaptation (LoRA)** to efficiently fine-tune a state-of-the-art diffusion model for the inpainting task.
4. A qualitative analysis showing our fine-tuned inpainting model provides more contextually aware results than the original, pre-trained model.

3. Related Work

Our project builds upon several key areas in deep learning.

3.1. Image Segmentation

Semantic segmentation, the task of classifying each pixel in an image, is foundational to our work. The **UNet** architecture [1] is a highly successful model for biomedical image segmentation, and its encoder-decoder structure with skip connections has been widely adapted. For text-guided segmentation, models like **CLIPSeg** [4] leverage the joint text-image embedding space of CLIP to perform segmentation from arbitrary text prompts.

3.2. Generative Inpainting

Historically, inpainting was addressed with statistical methods. Modern approaches utilize generative models, with **Diffusion Models** [6] now representing the state of the art. Models like **Stable Diffusion** [5] can generate highly realistic images and have specialized versions for tasks like inpainting.

3.3. Parameter-Efficient Fine-Tuning (PEFT)

Fine-tuning large-scale models like Stable Diffusion is computationally expensive. **Low-Rank Adaptation (LoRA)** [2] is a PEFT technique that freezes the pretrained model weights and injects trainable rank-decomposition matrices into the Transformer layers, dramatically reducing the number of trainable parameters while achieving comparable performance to full fine-tuning.

4. Methodology

Our pipeline consists of two primary stages: Mask Generation and Image Inpainting.

4.1. Dataset and Preprocessing

The foundation of this project is the `paint-by-inpaint/PIPE` dataset [3], sourced from the Hugging Face repository. To balance data diversity with available computational resources, a curated subset of 15,000 samples was selected for our experiments. For our object removal task, we strategically inverted the dataset's intended use. The original 'target' images, which contained the added objects, were repurposed as our source inputs, while the original 'source' images became our ground truth targets for the inpainting process.

A critical preprocessing step involved refining the original prompts, which were generated by a VLM and often proved to be verbose and noisy. For example, a descriptive prompt such as "Add a man in his late 20s to early 30s, with short dark hair, wearing a black jacket" was simplified by leveraging the dataset's existing class and location attributes. This resulted in a more direct and structured format, like **'person at the left'**, which provided clearer guidance for our segmentation model. Following this refinement, which also included replacing invalid samples and converting all text to lowercase, the final dataset was partitioned into training (75%), validation (15%), and test (10%) sets.

4.2. Stage 1: Mask Generation with a Conditioned UNet

This section details the architecture and training regimen of our proposed model for text-guided object segmentation. Our approach is centered around a custom U-Net architecture that is deeply conditioned on textual prompts, trained with a hybrid loss function, and optimized using a carefully selected set of modern training techniques.

4.2.1. Model Architecture. The core of our segmentation model is a custom U-Net style architecture, designed to effectively integrate textual guidance with image features. For the image encoder, we selected a **ResNet-18** model [7], pretrained on ImageNet. This choice was driven by its strong balance of robust feature extraction capabilities and computational efficiency. The decoder was constructed symmetrically, using a series of custom `UpBlock` modules which contain `ResNetBlocks` to process features and restore spatial resolution.

The key innovation of our model is the use of a **Transformer Cross-Attention** module to fuse text embeddings with image features. The text embeddings are generated

from the instructional prompts using a pretrained `CLIP` text encoder. Unlike simpler conditioning methods, this text conditioning is injected at multiple stages of the U-Net. This multi-stage design, applied to the deeper layers of the encoder and within each block of the decoder, allows the textual prompt to guide the feature extraction and image reconstruction processes at various levels of semantic and spatial resolution.

4.2.2. Loss Function. To train the model, we employed a hybrid loss function that combines the strengths of both pixel-level and region-based objectives. This approach provides stable gradients while directly optimizing for segmentation overlap, which is crucial for handling the class imbalance often present in this task. The final loss is a weighted sum of Binary Cross-Entropy (BCE) and Dice Loss [7] :

$$L_{\text{total}} = \lambda_{\text{BCE}} \cdot L_{\text{BCEWithLogits}} + \lambda_{\text{Dice}} \cdot L_{\text{Dice}}$$

Based on empirical results, we set the weights to $\lambda_{\text{BCE}} = 0.3$ and $\lambda_{\text{Dice}} = 0.7$. This places a greater emphasis on the Dice term, prioritizing the model's ability to produce spatially accurate and cohesive masks.

4.2.3. Training Details. The model was trained for a total of 40 epochs using the **AdamW** optimizer with a base learning rate of 1×10^{-4}. A differential learning rate was used, allowing for a different, potentially smaller rate for fine-tuning the pre-trained text encoder. To facilitate convergence, a `ReduceLROnPlateau` learning rate scheduler was implemented, which reduces the learning rate by a factor of 0.5 if the validation loss fails to improve for 3 consecutive epochs.

To prevent overfitting and ensure we retain the best-performing model, we utilized two key techniques. First, **Best Model Checkpointing** was used to save the model's weights each time a new minimum in validation loss was achieved. Second, **Early Stopping** with a patience of 10 epochs was enabled to halt the training process if the validation loss no longer showed improvement.

4.3. Stage 2: Image Inpainting with a LoRA-Tuned Diffusion Model

The second stage of our pipeline addresses the inpainting task, for which we fine-tuned a large-scale diffusion model. The selected base model was `stabilityai / stable-diffusion-2-inpainting`, a powerful pre-trained model designed for this purpose. To ensure a consistent and high-quality denoising process, we replaced the default scheduler with a `DDIMScheduler`.

Given that fully fine-tuning a model of this scale is computationally prohibitive, we employed the parameter-efficient fine-tuning technique (PEFT) of **low-rank Adaptation (LoRA)** [2]. We injected trainable LoRA adapters into the attention layers of the model's U-Net, specifically targeting the `to_q`, `to_k`, `to_v`, and `to_out.0` modules. A small rank of $r = 4$ and an alpha of $\alpha = 4$ were used, which drastically reduced the number of trainable parameters while still allowing the model to effectively adapt to our specific dataset.

The model was fine-tuned using a standard latent diffusion objective. For each training step, the ground truth target image was first encoded into the latent space using the model's Variational Autoencoder (VAE). A random amount of noise was then added to these latents. The model's task was to predict the original noise that was

added, conditioned on the noisy latents, the resized mask, and the masked image latents. The loss was calculated as the Mean Squared Error (MSE) between the predicted noise and the original noise. Notably, the model was trained unconditionally, using an empty text prompt, to specialize it for the task of background reconstruction based solely on visual context. The training was conducted for 2 epochs with a learning rate of 1×10^{-4} and a batch size of 1. [6]

5. Results

5.1. Mask Generation Performance

To evaluate the performance of our segmentation model, we conducted a rigorous analysis on the validation set. We first compared two distinct output strategies: a precise pixel-level mask ("Raw") and a simplified bounding box mask ("Box"). The evaluation, illustrated by the score distributions in our violin plots, revealed that the "Box vs. Box" comparison consistently yielded higher and more stable Dice and IoU scores. This indicates that our model excels at robust object localization, even when it struggles with pixel-perfect segmentation. Consequently, the bounding box strategy was selected for the final model comparison and for use in the downstream inpainting task.

We then compared our `ConditionedUNet` against the `CLIPSeg` baseline using this "Box vs. Box" method on the validation set. Our model demonstrated a significant performance improvement, achieving a mean Dice score of **0.4292** compared to CLIPSeg's 0.3663.

Having selected our model and evaluation strategy, a final assessment was performed on the held-out test set to report an unbiased measure of performance. The quantitative results are summarized in Table 1. Our final model achieved a Dice score of **0.4386** and an IoU score of **0.3118**, confirming its effectiveness and ability to generalize to unseen data.

Table 1: Final "Box vs. Box" performance on the test set. Our model shows a 17.4% improvement in Dice score over the baseline.

Model	Dice Score (Test Set)	IoU Score (Test Set)
CLIPSeg (Baseline)	0.3736	0.2821
ConditionedUNet (Ours)	**0.4386**	**0.3118**

The qualitative difference is shown in Figure 2 below, where our model's mask is significantly more accurate.

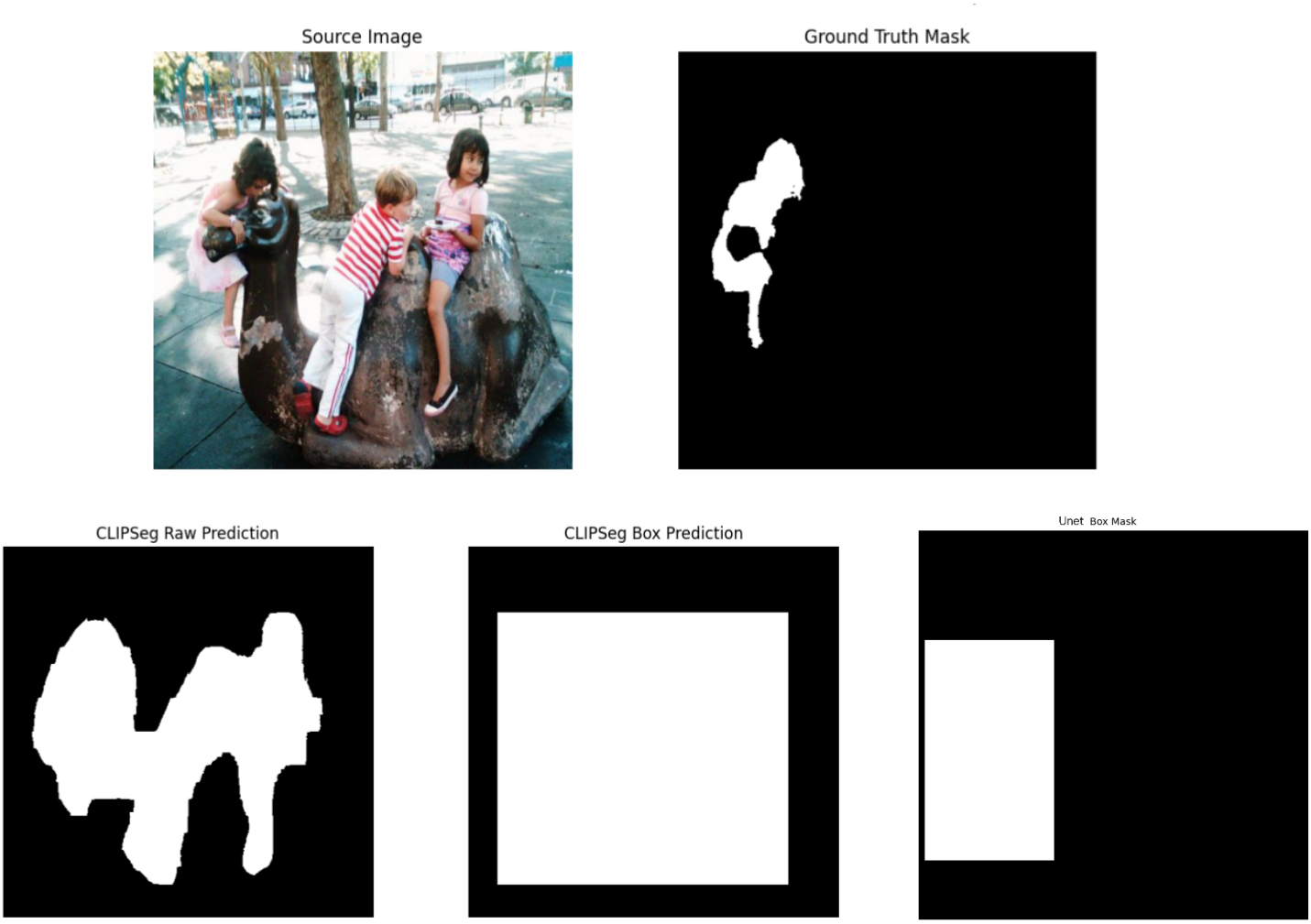
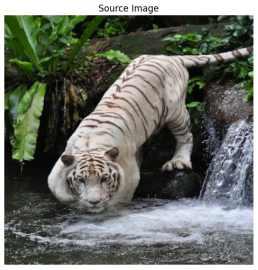


Figure 2: Comparison of different mask generation methods. Top: The original source image and the ground truth mask. Bottom-left: The raw and box prediction from the CLIPSeg model. Bottom-right: The box mask from our UNet model.

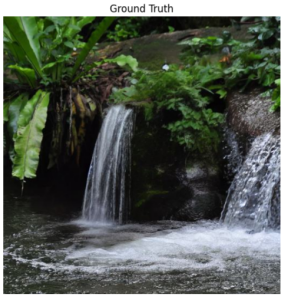
5.2. Inpainting Performance

The performance of the inpainting stage was evaluated qualitatively, as this is the most effective method for assessing the visual fidelity of generative models. This involved a side-by-side visual comparison of outputs from our fine-tuned LoRA model against those from the original, pre-trained Stable Diffusion baseline. A selection of these comparisons is presented in Figure 3.

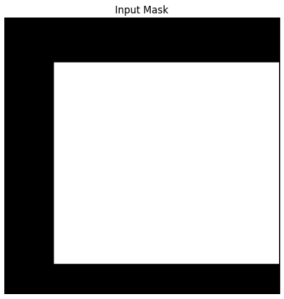
The results demonstrate a substantial and consistent improvement from our LoRA-tuned model. Our model consistently produces more realistic and contextually coherent images, showcasing a much deeper understanding of the required background reconstruction. A clear example of this is the "tiger at the center" task (Figure 3a); the baseline model erroneously inpainted a stone statue, failing conceptually, whereas our model correctly reconstructed the waterfall and rock formations behind the removed tiger. This superior contextual awareness was also observed in other examples as well.



(a) Source Image



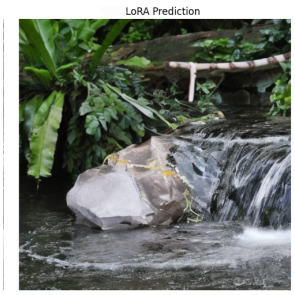
(b) Ground Truth



(c) Inpainting Mask

(d) Text Prompt

(e) Baseline Model Result



(f) Our LoRA Model Result

Figure 3: Qualitative Comparison of Inpainting Models (ID: 1138). The top two rows show the inputs. The bottom row compares the baseline result (e) with our fine-tuned LoRA model (f).

6. Discussion

The results confirm the success of our two-stage approach. The superior quantitative performance of the Conditioned UNet validates our custom architecture. The qualitative analysis of the inpainting stage revealed the clear benefits of LoRA fine-tuning.

However, the project also highlighted key limitations. Both models struggled with conceptual misunderstanding, sometimes replacing an object instead of removing it. More significantly, the models failed on the "human problem," producing distorted imagery when tasked with removing human figures, as shown in Figure 4.

Figure 4: Example of a Critical Model Limitation (The "Human Problem," ID: 8991). When tasked with removing a person's suit, our model hallucinates a complex, nonsensical object, highlighting the difficulty of inpainting human forms.

7. Conclusion

In this project, we successfully developed and validated a complete, end-to-end system that effectively translates a natural language instruction into a high-quality, edited image. By combining a custom-trained segmentation network with a fine-tuned generative inpainting model, we have produced a pipeline that is not only functional but has also been proven superior to established baselines at each critical stage. Future work could focus on improving the model's conceptual understanding of "removal" and exploring more advanced techniques for handling complex subjects like human figures.

References

[1] A.I. S tempData. What is u-net? https://medium.com/analytics-vidhya/what-is-unet-157314c87634, 2019.

[2] HF. Lora. https://huggingface.co/docs/diffusers/tutorials/using_peft_for_inference.

[3] HF. Paint-by-inpaint. https://huggingface.co/datasets/paint-by-inpaint/PIPE_Masks,https://huggingface.co/datasets/paint-by-inpaint/PIPE/viewer/default/train.

[4] Matt Paynne. What is clipseg how we build optimization pipelines for text-guided zero-shot segmentation. https://www.width.ai/post/what-is-clipseg, 2023.

[5] Onkar Mishra. Stable diffusion explained. https://medium.com/@onkarmishra/stable-diffusion-explained-1f101284484d, 2023.

[6] Themos Stafylakis. Diffusion models. https://eclass.aueb.gr/modules/document/file.php/INF401/Lectures%202024-2025%20TStafylakis/DL-Diffusion.pdf, 2025.

[7] Yuzhi Chen. Application of resnet18-unet in separating tumors from brain mri images. https://www.researchgate.net/publication/373944435_Application_of_Resnet18-Unet_in_separating_tumors_from_brain_MRI_images#pf3, 2023.