ECE 285 - FINAL PROJECT REPORT

Pushkal Mishra
Electrical and Computer Engineering
A69033424

ABSTRACT

This project aims to re-implement the recent EmerDiff [1] framework for unsupervised semantic segmentation and apply it to a custom dataset of realistic driving scenes collected using the CARLA simulator. EmerDiff [1] leverages the internal semantic understanding of diffusion models to produce fine-grained segmentation maps without any labels or additional training. The challenges of annotation cost of data from autonomous driving systems are extremly high and so this project aims to evaluate the effectiveness of EmerDiff [1] on our custom dataset [2], which consists of camera images collected alongside Radar and LiDAR. Furthermore, EmerDiff's [1] segmentation quality is compared against other segmentation methods as mentioned in the paper. The goal is to explore the transferability and robustness of unsupervised segmentation models in real-world autonomous driving environments to simulated scenarios.

1 Introduction

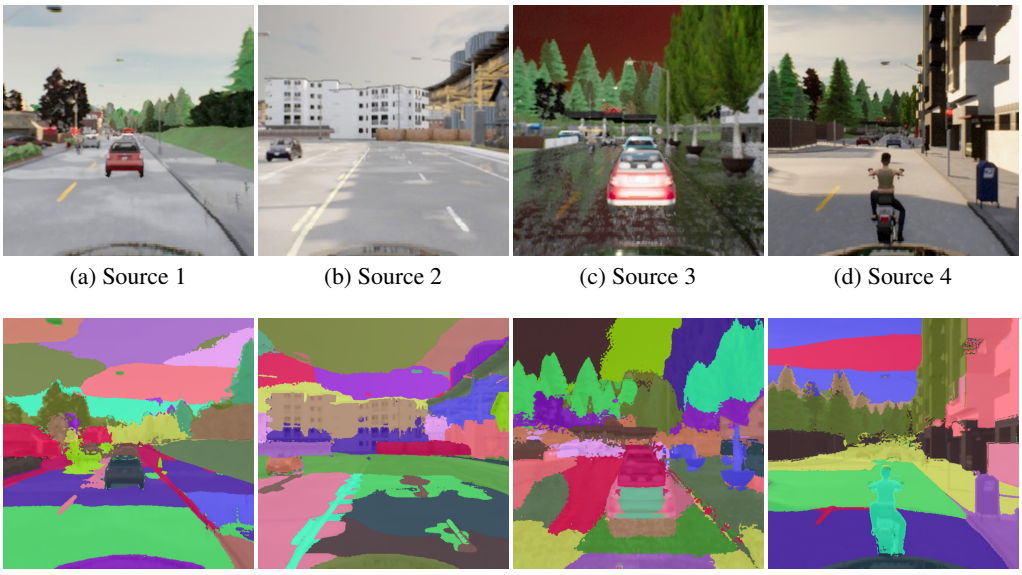
Pixel-wise semantic segmentation is the task of assigning a semantic category to each pixel in an image, is a foundational problem in computer vision with broad applications across autonomous driving and scene understanding. However, traditional semantic segmentation models heavily rely on large-scale, finely annotated datasets, which are labor-intensive, expensive to curate and not easy to find. This challenge is particularly observable in the automotive domain, where diverse driving scenarios and complex scene layouts make manual annotation process time-consuming and error-prone. Many real-world automotive datasets either lack segmentation labels entirely or provide only coarse annotations, which limits supervised learning models in such settings.

One way out of this problem is to go with unsupervised semantic segmentation, which has emerged as a promising alternative due to the availablity of large-scale trained models. These methods aim to discover semantically meaningful groupings of pixels directly from raw images without the use of ground truth labels. Some methods currently use self-supervised vision transformers (e.g., DINO-based methods) which are often incapable of producing fine-grained segmentations which map entire input pixel domain to segmented groups.

Recent advances in text-conditioned diffusion models, particularly Stable Diffusion [3], have revealed that these generative models encode rich semantic knowledge in their internal representations, particularly in the lower levels of representation space. Prior works on extracting segmentations from diffusion models rely on extra supervision or handcrafted layout priors, and are also incapable of producing fine-grained segmentation maps.

To address this challenges, EmerDiff [1] proposes a novel framework that extracts pixel-level semantic knowledge from pre-trained diffusion models without any additional training or annotations. The generative process of Stable Diffusion is used to identify semantic correspondences between image pixels and low-resolution feature map regions through targeted perturbations. This allows the construction of fine-grained, image-resolution segmentation maps using only the intrinsic semantic structure embedded in the diffusion model. The result is an unsupervised image segmentor that operates label-free and demonstrates strong generalization across complex scenes.

In this work, we reimplement and evaluate the EmerDiff framework with a particular focus on custom, unlabeled automotive datasets. This study aims to validate its potential for real-world deployment in scenarios where segmentation annotations are either unavailable or infeasible to obtain. Through a series of ablation studies and qualitative analyses, we explore the internal mechanisms of the model and its performance under various hyperparameter configurations,



(a) Source 1 (b) Source 2 (c) Source 3 (d) Source 4

(e) Segmented 1 (f) Segmented 2 (g) Segmented 3 (h) Segmented 4

Figure 1: Example of segmented images from the CARLA dataset. The top row shows source images from the dataset, while the bottom row displays the corresponding segmentation outputs generated by EmerDiff.

showing its robustness and versatility of this approach for unsupervised scene understanding. Some results of the segmentation can be seen in Figure 1.

2 Problem

In this project, we reimplement the EmerDiff framework from scratch and apply it to a custom driving dataset collected in the CARLA simulator. Our goal is to evaluate whether EmerDiff can generalize to novel, unlabeled driving scenarios. We conduct ablation studies on key hyperparameters such as the number of masks, perturbation strength, and injection timestep. Without relying on ground truth, we assess segmentation quality qualitatively and explore how semantic structure emerges across varied scenes. This work highlights EmerDiff's potential as a practical unsupervised segmentation method for automotive perception tasks.

3 Approach

This section describes the EmerDiff methodology, which enables pixel-level semantic segmentation using a frozen diffusion model, specifically Stable Diffusion [3].

The pipeline consists of three major stages: (1) Latent Inversion and Feature Extraction, (2) Low-Resolution Clustering for Segmentation, and (3) Semantic Correspondence via Modulated Reverse Diffusion.

3.1 Latent Inversion and Feature Extraction

EmerDiff operates on **Stable Diffusion**, a text-to-image latent diffusion model (LDM) where the denoising process is performed not on raw image pixels but in a compressed *latent space*. The original image $\mathbf{x} \in \mathbb{R}^{512\times512\times3}$ is first encoded into a low-dimensional latent tensor $\mathbf{z}_0 \in \mathbb{R}^{64\times64\times4}$ using a VAE encoder.

The latent $\mathbf{z}_0$ is then inverted into a noise trajectory using DDPM-based inversion, resulting in a set of latents varying across time and increasing in noise levels. To extract semantically rich features, EmerDiff leverages *query vectors* from the *first cross-attention layer* in the 16×16 upsampling block of the U-Net architecture at a fixed timestep t_f. These vectors are trained to interact with text tokens and have been empirically shown to encode meaningful object-level representations even in the absence of prompts.

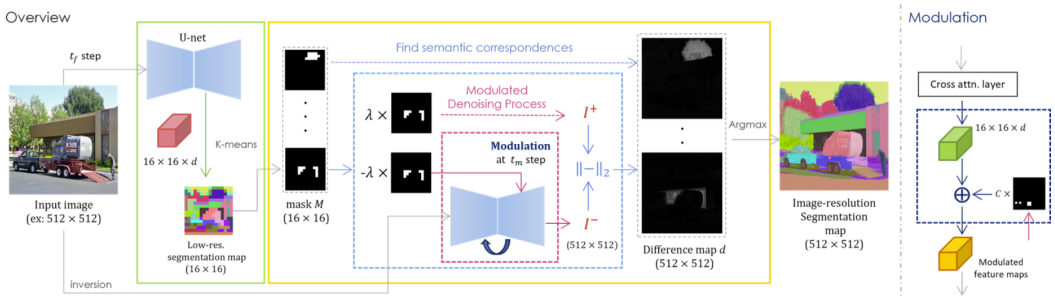


Figure 2: Overview of the EmerDiff framework. Green: Low-resolution segmentation. Orange: Image-resolution segmentation maps.

3.2 Low-Resolution Clustering for Segmentation

The extracted feature vectors are spatially arranged on a 16×16 grid and serve as a semantic representation of the image at reduced resolution. EmerDiff applies k-means clustering to group these feature vectors into K clusters, producing a *low-resolution segmentation map* as $M_{\text{low}} \in \{1, \ldots, K\}^{16\times16}$:

$$\mathcal{C} = \text{kmeans}(\{\mathbf{q}_i\}, K)$$

Each cluster corresponds to a semantically distinct region in the image. These clusters serve as modulation targets in the reverse diffusion process.

3.3 Semantic Correspondence via Modulated Reverse Diffusion

To recover full-resolution segmentations, EmerDiff measures the *semantic influence* of each low-resolution cluster on the final image by performing *modulated denoising* during reverse diffusion.

During this reverse process, EmerDiff injects perturbations into the value vector $\mathbf{V}$ in the cross-attention layer computation:

$$\text{Attn}_{\text{mod}} = f\left(\sigma\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} + c \cdot \mathbf{M}\right)$$

Here, $\mathbf{M} \in \{0, 1\}^{hw}$ is a binary mask corresponding to the spatial region of a cluster, and $c = \pm\lambda$ is the perturbation strength. Two denoising runs are performed: one with $+\lambda$ and one with $-\lambda$. The resulting images $\mathbf{I}^+$ and $\mathbf{I}^-$ are then used to compute the *difference map*:

$$\mathbf{d}_i = \left\|\mathbf{I}^+ - \mathbf{I}^-\right\|_2$$

This map encodes the *semantic correspondence strength* between image pixels and the target cluster i.

Finally, the full-resolution segmentation map $M_{\text{high}} \in \{1, \ldots, K\}^{512\times512}$ is constructed by assigning each pixel (x, y) to the cluster with the maximum response in the difference maps:

$$M_{\text{high}}(x, y) = \arg\max_i \mathbf{d}_i(x, y)$$

Additional Enhancements: To preserve structural consistency, EmerDiff optionally injects original attention maps (i.e., uses fixed $\mathbf{Q}\mathbf{K}^\top$) during reverse diffusion and applies Gaussian smoothing to the final segmentation maps to reduce pixelation artifacts.

4 Dataset Description

For this project, we utilize RGB camera images from the C-Shenron dataset [2] I collected in an earlier work—a high-fidelity, multimodal dataset generated within the CARLA simulator. Notably, this dataset does not include any

ground-truth segmentation labels, making it ideal for evaluating unsupervised segmentation methods like EmerDiff. The RGB images represent diverse urban driving environments and are collected using an expert agent capable of executing realistic driving behaviors such as obstacle avoidance, compliance with traffic rules, and high-level route planning. This ensures the collected scenes are visually complex and accurate, capturing realistic vehicle scenes and dynamics.

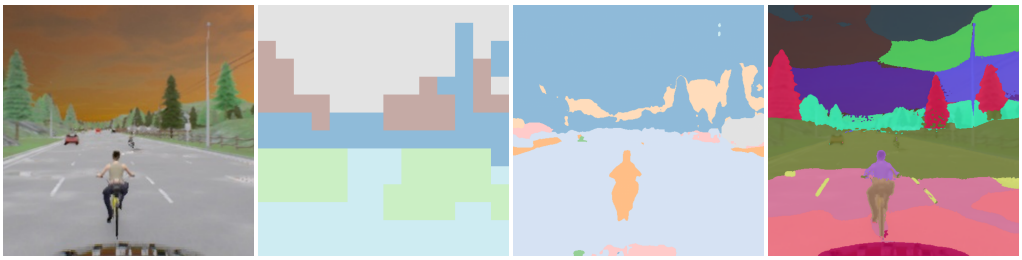
The dataset spans over 185,000 unique frames extracted from eight different simulated towns (Town01–Town07 and Town10), covering a rich spectrum of driving scenarios, road layouts, and weather conditions. Scenes include highways, residential neighborhoods, and dense urban intersections populated with dynamic actors such as vehicles and pedestrians. This diversity makes the dataset a strong testbed for evaluating the generalization of unsupervised semantic segmentation methods in novel, unlabelled, and visually varied real-world-inspired conditions. Due to size constraints, we will only provide a subset of the dataset for this project.

5 Experiments

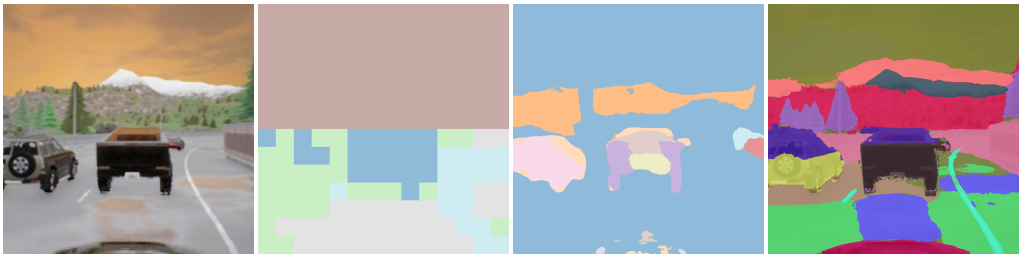
To evaluate the performance of EmerDiff, we conduct a series of qualitative experiments across multiple configurations. Unless otherwise specified, all experiments are performed using EmerDiff's default parameters which have been found optimal in prior studies: number of masks = 25, modulation strength $\lambda = 10$, and injection timestep $t_s = 281$. All results are qualitative due to the lack of segmentation annotations in our dataset.

5.1 Comparison of Segmentation Methods

We compare EmerDiff with two leading unsupervised segmentation approaches: DINO [4] and SegFormer [5]. DINO performs patch-level segmentation by clustering features from self-supervised Vision Transformers (ViTs), typically yielding coarse segmentation maps with little object boundary precision. In contrast, SegFormer performs pixel-level classification and produces significantly finer boundaries due to its hierarchical transformer-based encoder and lightweight decoder design.



(a) Original 1 (b) DINO 1 (c) SegFormer 1 (d) EmerDiff 1



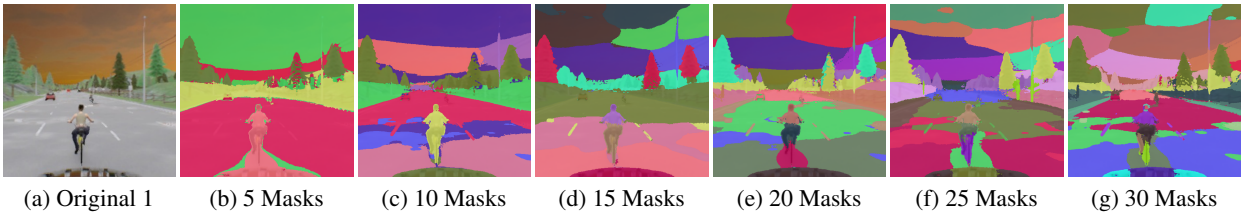
(e) Original 2 (f) DINO 2 (g) SegFormer 2 (h) EmerDiff 2

Figure 3: This figure shows examples of segmented images from the CARLA dataset using DINO-v2 [4] with ViT-B/16 backbone, SegFormer [5] and the implemented EmerDiff with standard hyperparameters.

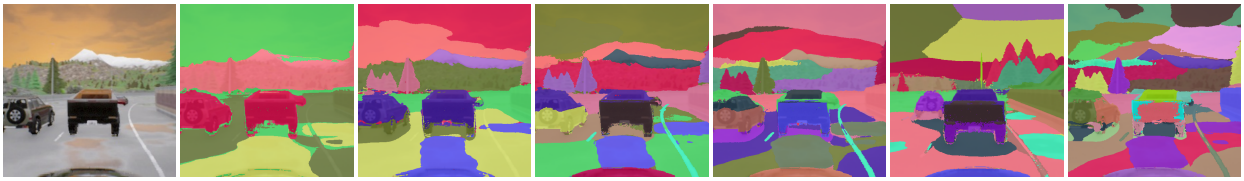
From the Figure 3, we can observe that DINO ViT-B/16 provides only rough object segmentation. While useful for foreground extraction, it lacks clear semantic differentiation across objects. SegFormer, on the other hand, identifies boundaries more clearly but suffers from poor class separation, likely due to the limited and fixed number of learned classes. EmerDiff on the other hand consistently identifies roads, vehicles, sky, and even subtle features such as wet ground patches. This improvement stems from its use of the latent representations of a pre-trained Stable Diffusion model, which captures pixel-level semantic information even in unsupervised settings.

5.2 Varying the Number of Segmentation Masks

We investigate the effect of varying the number of segmentation masks from 5 to 30 under two different modulation strengths: $\lambda = 10$ and $\lambda = 100$. At $\lambda = 10$, we observe that segmentation quality improves initially with the number of masks, peaking around 10–15 masks, and then degrades beyond that. This degradation is attributed to over-segmentation in the CARLA driving scenes, which contain a limited number of meaningful semantic classes. At higher mask counts, noise or subtle appearance differences are misinterpreted as new classes, leading to fragmented and inconsistent segmentations. These results can be seen in Figure 4.



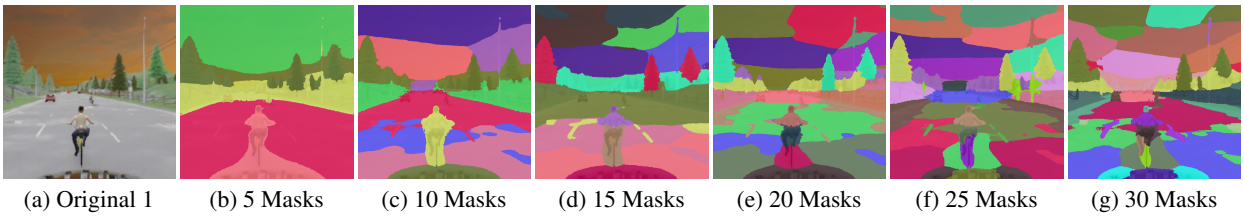
(a) Original 1 (b) 5 Masks (c) 10 Masks (d) 15 Masks (e) 20 Masks (f) 25 Masks (g) 30 Masks



(h) Original 2 (i) 5 Masks (j) 10 Masks (k) 15 Masks (l) 20 Masks (m) 25 Masks (n) 30 Masks

Figure 4: This figure shows segmented images generated by EmerDiff by varying the number of masks to be generated from the pipeline. Here the value of λ is set to 10. The number of masks is varied from 5 to 30.

At $\lambda = 100$, the same trend is observed, although the segmentation outputs appear more coarse. This is because a larger λ amplifies the influence of the modulation, exaggerating semantic boundaries and smoothing over fine details. The difference effect becomes more spatially pronounced, making the model more sensitive to modulation but at the cost of spatial precision. These results can be seen clearly in Figure 5. Also these observations align with the behavior reported in the original EmerDiff paper



(a) Original 1 (b) 5 Masks (c) 10 Masks (d) 15 Masks (e) 20 Masks (f) 25 Masks (g) 30 Masks

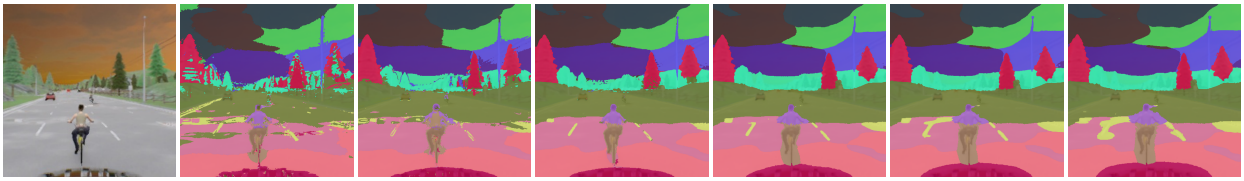
(h) Original 2 (i) 5 Masks (j) 10 Masks (k) 15 Masks (l) 20 Masks (m) 25 Masks (n) 30 Masks

Figure 5: This figure shows segmented images generated by EmerDiff by varying the number of masks to be generated from the pipeline. Here the value of λ is set to 100.

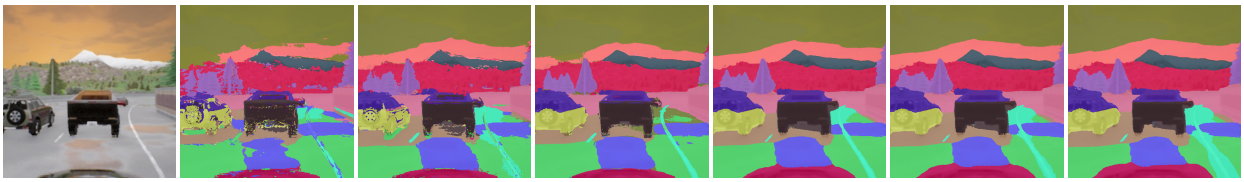
5.3 Varying the Modulation Strength

Based on the previous experiment, we fix the number of masks to 15 and 25, and vary the modulation strength λ across the range 1, 5, 10, 50, 100, 500, 1000. When number of masks are 15, we observe that as λ increases, segmentation

becomes progressively smoother and more semantically consistent. At low values less than 5, the segmentations appear noisy and under-modulated, while at high values greater than 500, semantic boundaries become overly smoothed and classes begin to bleed into one another. This can be seen in Figure 6.



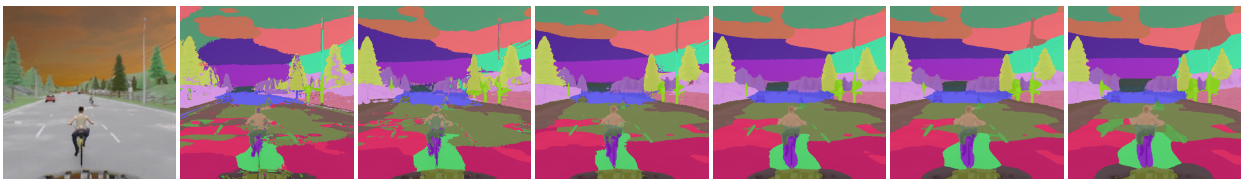
(a) Original 1 (b) $\lambda = 1$ (c) $\lambda = 5$ (d) $\lambda = 10$ (e) $\lambda = 50$ (f) $\lambda = 100$ (g) $\lambda = 1000$



(h) Original 2 (i) $\lambda = 1$ (j) $\lambda = 5$ (k) $\lambda = 10$ (l) $\lambda = 50$ (m) $\lambda = 100$ (n) $\lambda = 1000$

Figure 6: This figure shows segmented images generated by EmerDiff by varying the perturbation strength λ. Here the number of masks generated are fixed to 15 which were found to be optimal from the previous experiment. λ is varied from 1 to 1000 in log scale.

When the number of masks are 25, the segmentation quality is overall poorer due to class confusion caused by excessive segmentation masks. However, the trend with λ remains the same. The best segmentation quality across all conditions is achieved when number of masks is 15 and λ is 10, confirming the default settings proposed in the EmerDiff framework. This can be seen in Figure 7.



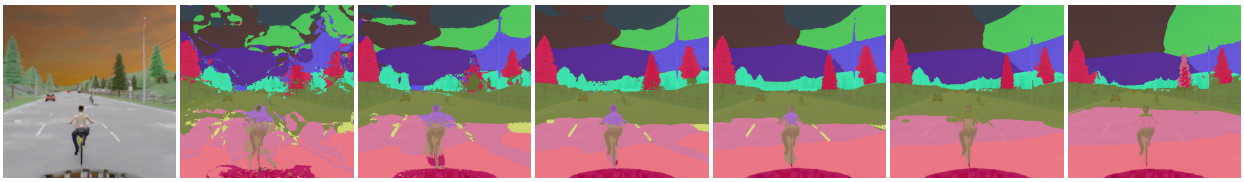
(a) Original 1 (b) $\lambda = 1$ (c) $\lambda = 5$ (d) $\lambda = 10$ (e) $\lambda = 50$ (f) $\lambda = 100$ (g) $\lambda = 1000$

(h) Original 2 (i) $\lambda = 1$ (j) $\lambda = 5$ (k) $\lambda = 10$ (l) $\lambda = 50$ (m) $\lambda = 100$ (n) $\lambda = 1000$

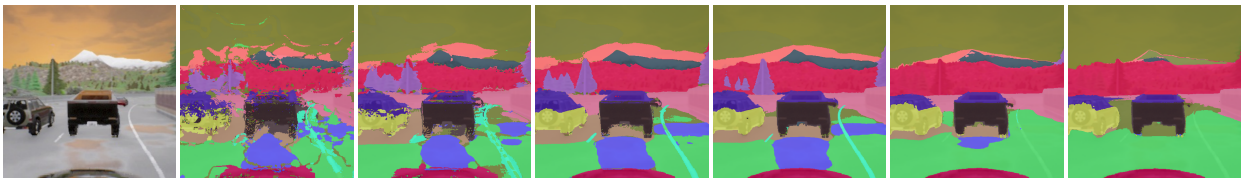
Figure 7: This figure shows segmented images generated by EmerDiff by varying the perturbation strength λ. Here the number of masks generated are fixed to 25. λ is varied from 1 to 1000 in log scale.

5.4 Varying the Modulation Time Step

Finally, we study the impact of varying the modulation timestep t_s in the diffusion denoising process. We test timesteps at 1, 81, 281, 581, 881, 981. At lower timesteps e.g., 1, 81, the segmentation maps are random and fragmented, displaying "island-like" masks with weak semantic coherence. This is expected, as early denoising stages operate on high-noise latent representations, where semantic structure has not yet emerged. As the timestep increases, the segmentations become cleaner and more meaningful. The best results are observed around $t_s = 281$, aligning with the findings in the original paper. At later timesteps e.g., 881, 981, segmentation becomes overly smooth and loses structural specificity. This can be seen in Figure 8.



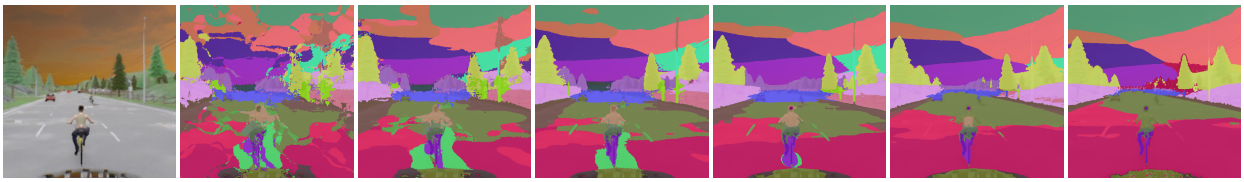
(a) Original 1 (b) $t_s = 1$ (c) $t_s = 81$ (d) $t_s = 281$ (e) $t_s = 581$ (f) $t_s = 8811$ (g) $t_s = 981$



(h) Original 2 (i) $t_s = 1$ (j) $t_s = 81$ (k) $t_s = 281$ (l) $t_s = 581$ (m) $t_s = 881$ (n) $t_s = 981$

Figure 8: This figure shows segmented images generated by EmerDiff by varying the timestep t_s at which the attention maps are perturbed. Here the number of masks generated are fixed to 15 and λ fixed to 10. The timestep t_s is varied from 1 to 981 in log scale.

Additionally, increasing the number of masks from 15 to 25 in this setting worsens the results, as the model struggles to differentiate between too many semantic groups in a dataset with relatively low semantic diversity. This can be seen in Figure 9.



(a) Original 1 (b) $t_s = 1$ (c) $t_s = 81$ (d) $t_s = 281$ (e) $t_s = 581$ (f) $t_s = 8811$ (g) $t_s = 981$

(h) Original 2 (i) $t_s = 1$ (j) $t_s = 81$ (k) $t_s = 281$ (l) $t_s = 581$ (m) $t_s = 881$ (n) $t_s = 981$

Figure 9: This figure shows segmented images generated by EmerDiff by varying the timestep t_s at which the attention maps are perturbed. Here the number of masks generated are fixed to 25 and λ fixed to 10. The timestep t_s is varied from 1 to 981 in log scale.

The best segmentation quality is achieved at $t_s = 281$ with 15 masks and $\lambda = 10$, confirming the default settings proposed in the EmerDiff framework.

5.5 Summary of Results

- **Segmentation Baseline Comparison:** EmerDiff outperforms both DINO and SegFormer by generating finer, more coherent segmentations that capture meaningful scene elements such as roads, sky, and wet surfaces. This is attributed to its ability to exploit semantic structure from diffusion model latents.
- **Varying Number of Masks:** Increasing the number of masks initially improves segmentation quality, peaking around 10–15 masks. Beyond this, segmentation quality degrades due to over-segmentation, especially given the limited semantic variety in the CARLA dataset.

- **Effect of Modulation Strength (λ):** Moderate values (e.g., $\lambda = 10$) offer the best trade-off between semantic sharpness and spatial coherence. Low values result in noisy masks, while very high values overly smooth semantic boundaries.
- **Varying Injection Timestep (t_s):** Early timesteps (e.g., $t_s = 1$) produce incoherent, island-like segmentations due to high latent noise. Best segmentation results are observed around $t_s = 281$, with later timesteps yielding overly smooth and class-blended masks.

6 Conclusion

In this project, we re-implemented the EmerDiff framework for unsupervised semantic segmentation and evaluated its performance on a novel, unlabeled driving dataset collected using the CARLA simulator. Our study demonstrates that EmerDiff can produce fine-grained, semantically coherent segmentation maps without any ground truth labels or additional training. By leveraging the internal semantic knowledge embedded within the pre-trained Stable Diffusion model, EmerDiff successfully identifies scene elements such as roads, vehicles, sky, and even subtle environmental features like wet patches.

Overall, this work highlights the impact of the EmerDiff architecture in providing high-quality semantic segmentation all while having hyperparameters that can be tuned for different scenarios.

7 Implementation

The EmerDiff framework was implemented entirely from scratch for this project, faithfully following the methodology described in the original paper. Major components such as latent inversion orchestration, k-means-based mask proposal, cross-attention perturbation logic, modulated reverse denoising, and semantic difference map generation were developed independently using PyTorch and HuggingFace's Diffusers library. All visualizations and experimental analyses were also custom-built.

However, a few key utility functions were adapted from existing open-source resources to ensure correctness and reproducibility:

- `sample_xts_from_x0()`: Simulates the forward diffusion process by adding scheduled Gaussian noise to the latent $\mathbf{z}_0$.
- `get_variance()`: Computes the noise variance at a given timestep based on cumulative alpha scheduling.
- `inversion_forward_process()`: Runs the DDPM forward process from $\mathbf{z}_0$ to $\mathbf{z}_T$ and reconstructs noise vectors using U-Net predictions at each step.
- `reverse_step()`: Performs a single DDIM-style reverse step from $\mathbf{z}_t$ to $\mathbf{z}_{t-1}$ using predicted noise and scheduled variance.
- `inject_attention()`: Overrides attention forward passes in the U-Net by injecting custom hooks into specific layers (cross/self, up/down/mid blocks). This function enables recording original Q/K vectors and applying perturbations to feature maps during reverse denoising.

The first four functions were reused from the original DDPM inversion codebase[1], and `inject_attention()` was adapted based on the U-Net implementation in the HuggingFace Diffusers library[2]. All reused code was kept minimal, limited to low-level scheduling and U-Net manipulation logic. The rest of the EmerDiff pipeline was implemented from the ground up.

References

[1] Koichi Namekata, Amirmojtaba Sabour, Sanja Fidler, and Seung Wook Kim. Emerdiff: Emerging pixel-level semantic knowledge in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[2] Pushkal Mishra, Satyam Srivastava, Jerry Li, Kshitiz Bansal, and Dinesh Bharadia. Demo: C-shenron- a realistic radar simulation framework for carla. *In In ACM Conference on Embedded Networked Sensor Systems 2025 (SenSys '25) at Irvine, CA, USA*, 2025.

[1] `https://github.com/inbarhub/DDPM_inversion/blob/main/ddm_inversion/inversion_utils.py`

[2] `https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/unet_2d_condition.py`

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

[5] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: simple and efficient design for semantic segmentation with transformers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.