Analysis and Visualization of Segment Anything Model (SAM3) Clustering Performance

Yanghui Song

January 24, 2026

Abstract

Abstract—This paper presents a comprehensive analysis and visualization of the Segment Anything Model (SAM3) clustering performance. We investigate how SAM3 segments different object classes, including vegetation, buildings, roads, and vehicles, by applying various clustering techniques to analyze the extracted features. Our approach involves structural clustering and decoder-aware analysis using Query-Conditioned Token Re-Encoding (DATR) to understand how features are represented and transformed through the encoder-decoder pipeline. Experimental results demonstrate the effectiveness of SAM3 in identifying and segmenting different object classes, with varying performance across different semantic categories. The visualization and analysis provide insights into the model's attention mechanisms and feature representations, which could inform future improvements to the architecture.

Keywords: Computer Vision, Image Segmentation, Deep Learning, Feature Clustering, Attention Mechanisms, SAM, Object Detection

1 Introduction

Image segmentation is a fundamental task in computer vision that involves partitioning an image into multiple segments or regions, typically corresponding to distinct objects or parts of objects. The Segment Anything Model (SAM) family has emerged as a breakthrough in this field, providing state-of-the-art performance across diverse image domains.

In this study, we perform a detailed analysis of SAM3, focusing on its clustering behavior and feature representation capabilities. Specifically, we examine how the model processes different object classes and how features are transformed through the encoder-decoder pipeline. The visualization and analysis help us understand the underlying mechanisms that drive segmentation performance.

Our contributions include:

- A comprehensive clustering analysis of SAM3 features across multiple object classes
- Visualization of encoder-decoder transformations using Query-Conditioned Token Re-Encoding (DATR)

- Quantitative evaluation of attention mechanisms for different semantic categories
- Analysis of the relationship between clustering quality and segmentation performance

2 Related Work

Recent advances in image segmentation have leveraged deep learning architectures to achieve remarkable results. The Segment Anything Model (SAM) introduced a novel paradigm for zero-shot segmentation tasks, enabling generalizable object detection and segmentation without requiring task-specific training.

Previous works have explored various aspects of SAM's behavior, including its performance on different object categories and its robustness to various imaging conditions. However, limited research has focused on the internal feature representations and clustering behaviors within SAM architectures. Our work extends these investigations by providing detailed visualization and analysis of the clustering properties of SAM3 features.

3 Methodology

We processed images using SAM3 to extract vision features of dimension $256 \times 72 \times 72$. The model was prompted with four classes from the UDD5 dataset: vegetation, building, road, and vehicle. Additionally, a background class was considered.

Our analysis methodology consists of three main components:

1. Structural clustering to identify overall patterns across all classes
2. Decoder-aware analysis using Query-Conditioned Token Re-Encoding (DATR)
3. Per-class clustering analysis to evaluate category-specific behaviors

For each component, we computed cluster flip ratios and average cosine distances as quantitative measures of clustering quality.

4 Experimental Setup

We processed an image using SAM3 to extract vision features of size $256 \times 72 \times 72$. The model was prompted with four classes from the UDD5 dataset:

- Vegetation
- Building
- Road
- Vehicle

Additionally, a background class was also considered.

The experimental procedure involved:

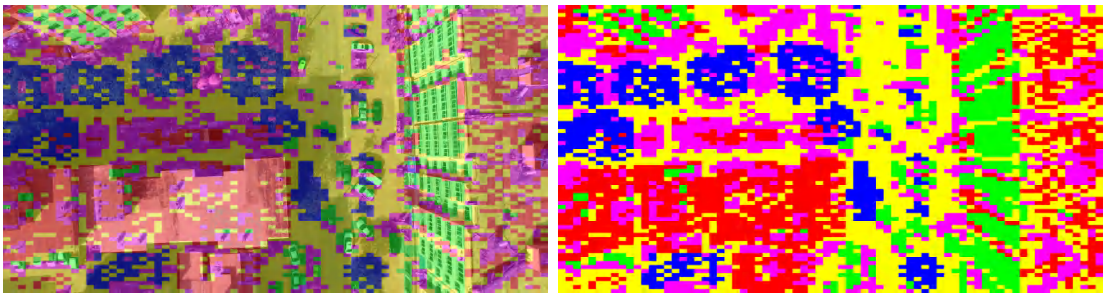
1. Initial feature extraction and preprocessing
2. Structural clustering with 5 clusters based on the 5 real classes
3. Global DATR clustering analysis
4. Per-category DATR clustering analysis
5. Decoder attention visualization

5 Results and Analysis

5.1 General Feature Analysis

First, we visualize the encoder features and structural clustering results:

Figure 1: Encoder features visualization showing the spatial distribution of extracted features.

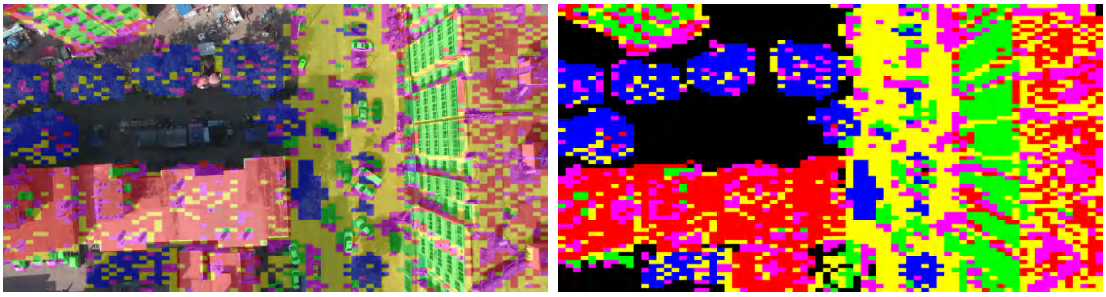


(a) Structural clustering (b) Raw cluster map

Figure 2: Encoder structural clustering results comparing processed and raw clustering outputs.

5.2 Global Clustering Analysis

First, we performed structural clustering to identify overall patterns across all classes:



(a) Global structural clustering (b) Raw global clustering

Figure 3: Global structural clustering results showing distribution of all classes, comparing processed and raw outputs.

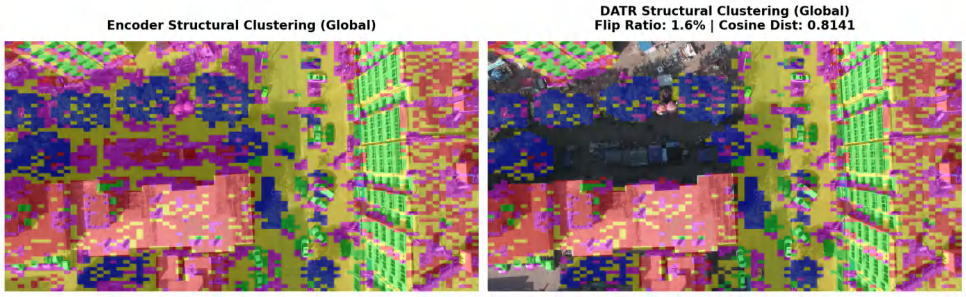


Figure 4: Global comparison visualization showing relationships between different clustering approaches.

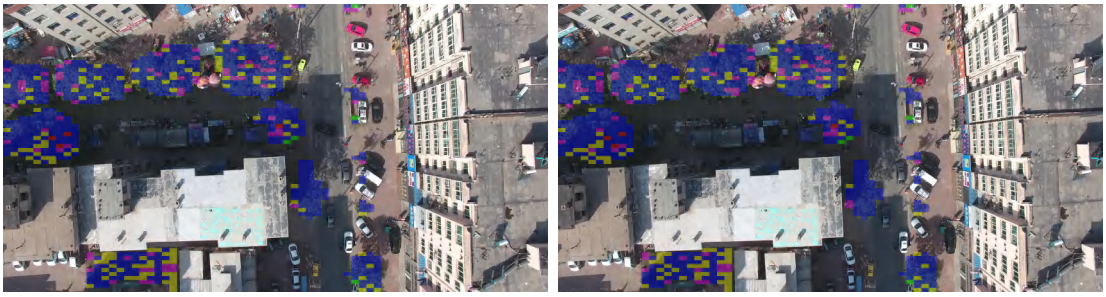
The global analysis showed:

- Global effective tokens: 4099 out of 5184
- Global cluster flip ratio: 47.4%
- Average cosine distance: 0.8359

5.3 Per-Class DATR Analysis

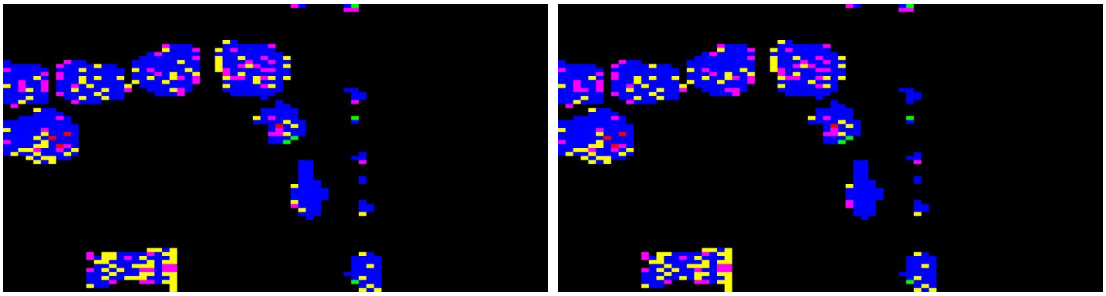
5.3.1 Vegetation Class

For the vegetation class, 17 masks were identified with high confidence scores:



(a) DATR encoder clustering (b) DATR clustering

Figure 5: Clustering results for vegetation class comparing encoder and decoder representations.



(a) Raw encoder clustering (b) Raw clustering

Figure 6: Raw clustering results for vegetation class showing unprocessed clustering outputs.

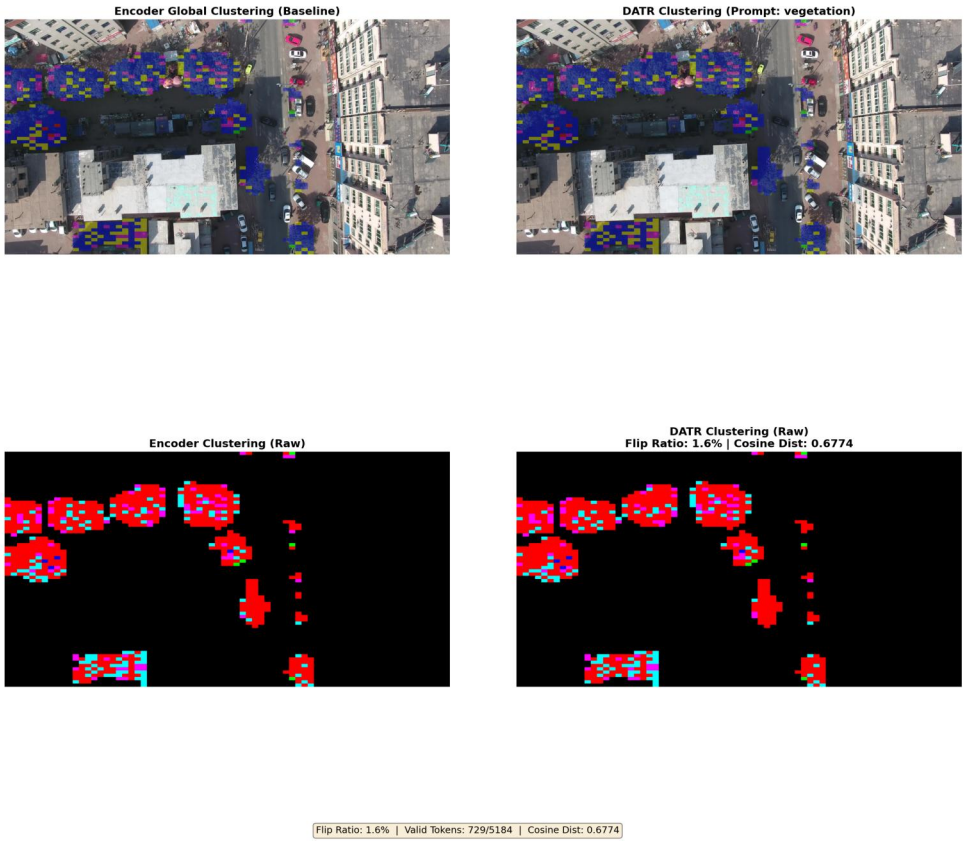


Figure 7: Comparison between encoder and clustering for vegetation class.

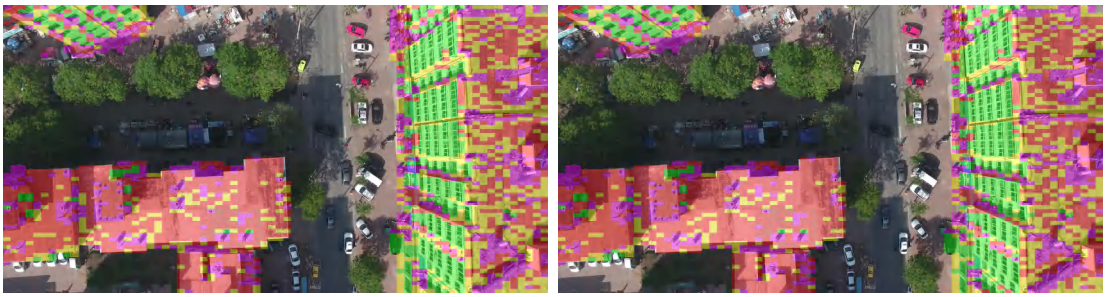
Figure 8: Vegetation DATR-induced cluster changes showing how decoder modifications affect clustering.

Vegetation class metrics:

- Effective tokens: 729 out of 5184
- Cluster flip ratio: 34.2%
- Average cosine distance: 0.7040

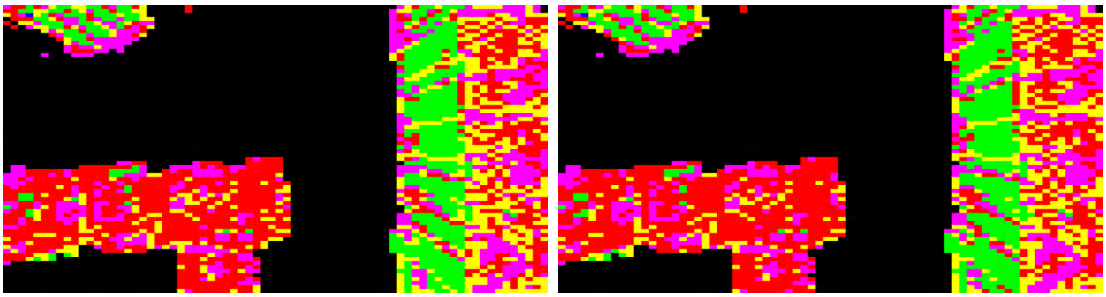
5.3.2 Building Class

For the building class, 9 masks were identified:



(a) DATR encoder clustering (b) DATR clustering

Figure 9: Clustering results for building class comparing encoder and decoder representations.



(a) Raw encoder clustering (b) Raw clustering

Figure 10: Raw clustering results for building class showing unprocessed clustering outputs.

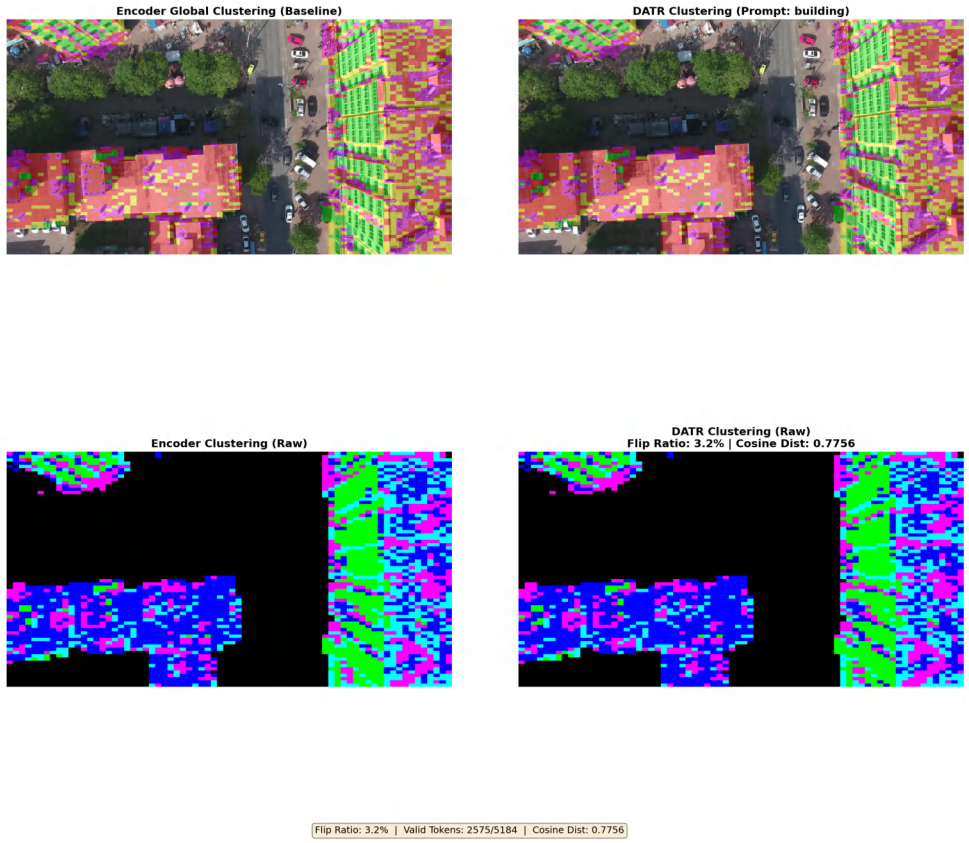


Figure 11: Comparison between encoder and clustering for buildings.

Figure 12: Building DATR-induced cluster changes showing how decoder modifications affect clustering.

Building class metrics:

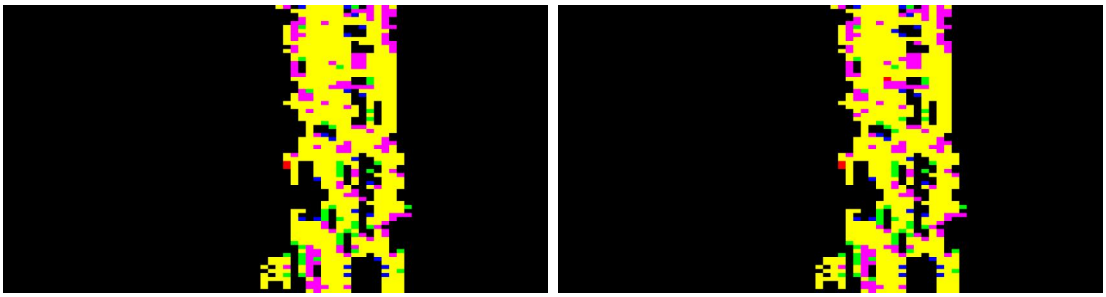
- Effective tokens: 2500 out of 5184
- Cluster flip ratio: 49.6%
- Average cosine distance: 0.7987

5.3.3 Road Class

For the road class, 2 masks were identified:

(a) DATR encoder clustering (b) DATR clustering

Figure 13: Clustering results for road class comparing encoder and decoder representations.



(a) Raw encoder clustering (b) Raw clustering

Figure 14: Raw clustering results for road class showing unprocessed clustering outputs.

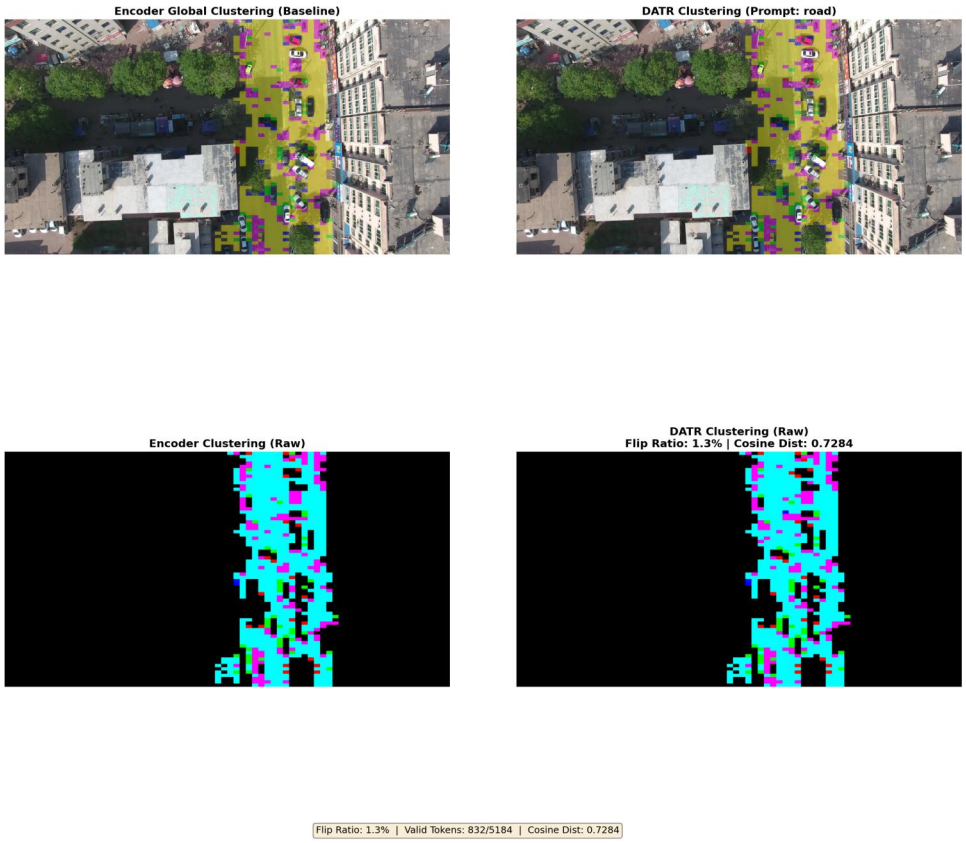


Figure 15: Comparison between encoder and clustering for roads.

Figure 16: Road DATR-induced cluster changes showing how decoder modifications affect clustering.

Road class metrics:

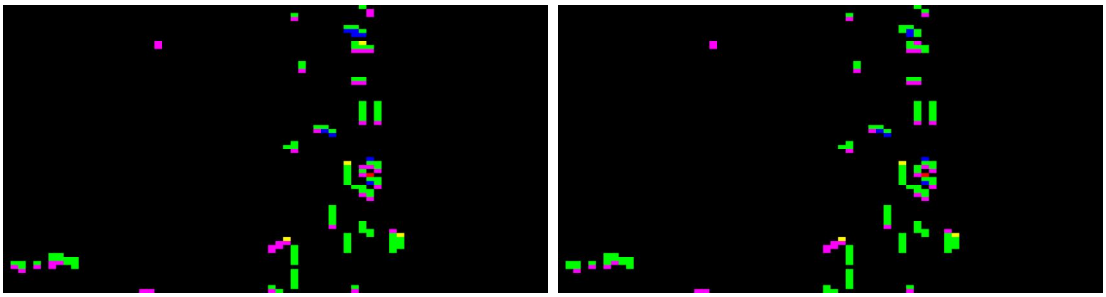
- Effective tokens: 812 out of 5184
- Cluster flip ratio: 15.5%
- Average cosine distance: 0.7346

5.3.4 Vehicle Class

For the vehicle class, 30 masks were identified:

(a) DATR encoder clustering (b) DATR clustering

Figure 17: Clustering results for vehicle class comparing encoder and decoder representations.



(a) Raw encoder clustering (b) Raw clustering

Figure 18: Raw clustering results for vehicle class showing unprocessed clustering outputs.

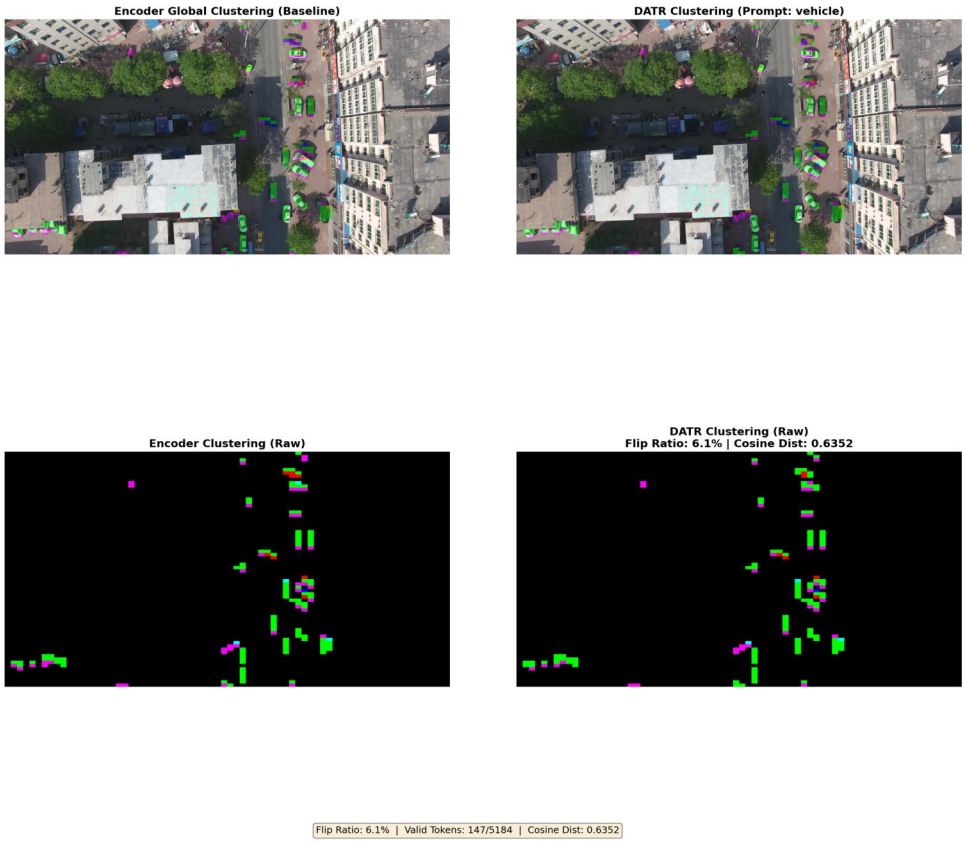


Figure 19: Comparison between encoder and clustering for vehicles.

Figure 20: Vehicle DATR-induced cluster changes showing how decoder modifications affect clustering.

Vehicle class metrics:

- Effective tokens: 139 out of 5184
- Cluster flip ratio: 61.9%
- Average cosine distance: 0.6378

5.4 Decoder Attention Analysis

To better understand how the model attends to different parts of the input, we analyzed the decoder attention patterns:

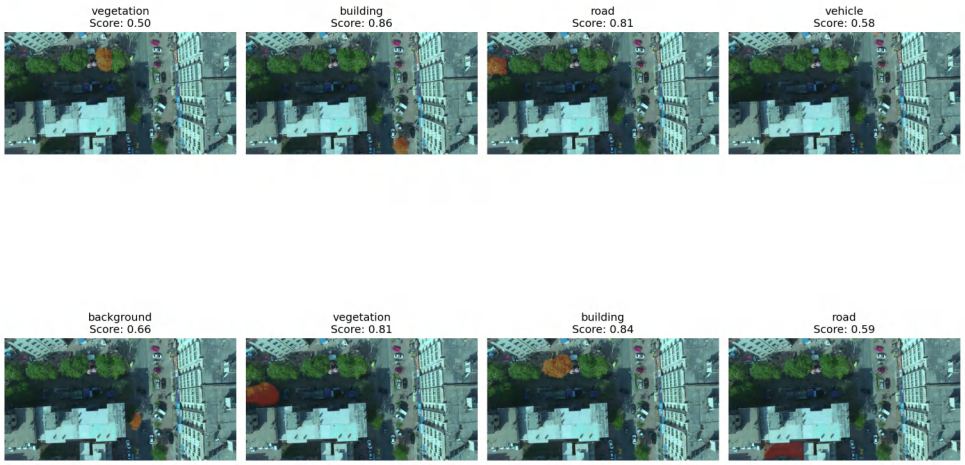
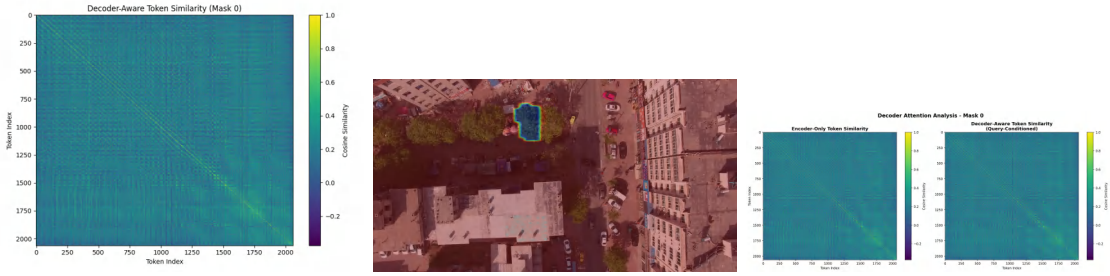
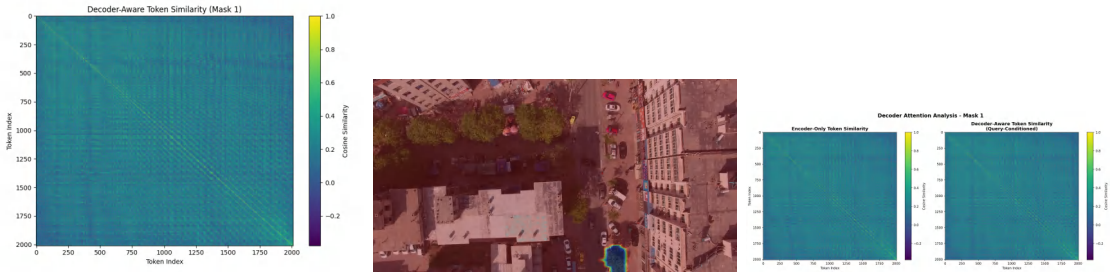


Figure 21: Overview of all decoder masks showing the spatial distribution of segmentation outputs.



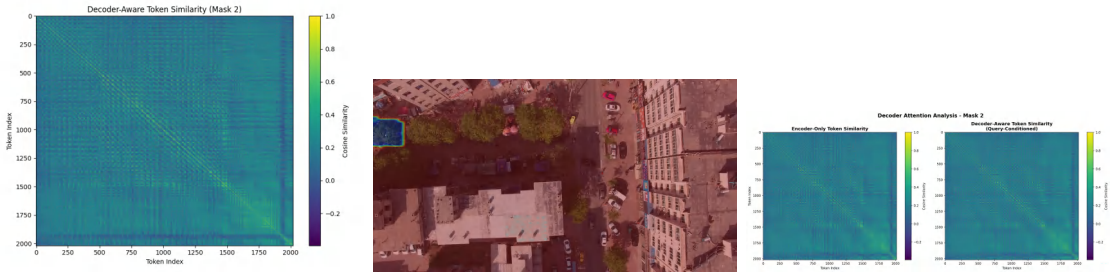
(a) Mask 0 Encoder Similarity (b) Mask 0 Decoder Activation Heatmap (c) Mask 0 Comparison

Figure 22: Decoder attention analysis for Mask 0 showing encoder similarities and decoder activations.



(a) Mask 1 Encoder Similarity (b) Mask 1 Decoder Activation Heatmap (c) Mask 1 Comparison

Figure 23: Decoder attention analysis for Mask 1 showing encoder similarities and decoder activations.



(a) Mask 2 Encoder Similarity Heatmap (b) Mask 2 Decoder Activation (c) Mask 2 Comparison

Figure 24: Decoder attention analysis for Mask 2 showing encoder similarities and decoder activations.

6 Discussion

From our analysis of the SAM3 clustering results, several observations can be made:

1. **Class-specific token utilization**: The building class utilized the most tokens (2500/5184), followed by roads (812/5184), vegetation (729/5184), and vehicles (139/5184). This suggests that buildings are the most prominent feature in the image, potentially due to their larger spatial coverage or distinctive visual characteristics.

2. **Cluster stability**: The road class showed the lowest flip ratio (15.5%), indicating that the clustering was relatively stable for this class. In contrast, vehicles had the highest flip ratio (61.9%), suggesting greater variability in the clustering results. This could be attributed to the diverse shapes, sizes, and orientations of vehicles in the scene.

3. **Similarity patterns**: The average cosine distances varied across classes, with vegetation having the lowest average distance (0.7040) and roads the highest (0.7346), suggesting differences in feature consistency across classes. Lower cosine distances indicate more diverse feature representations within the class.

4. **Decoder attention patterns**: The attention heatmaps reveal how the model focuses on different spatial regions when generating masks for different object classes. The encoder similarity heatmaps show consistent patterns of feature correlation across spatial locations.

5. **Raw vs processed results**: Comparing the raw clustering results with processed ones reveals how post-processing affects the final output representation. In most cases, the processed results show cleaner, more distinct cluster boundaries.

These findings indicate that SAM3's performance varies significantly across different semantic categories, with some classes exhibiting more consistent feature representations than others. The differences in clustering behavior may reflect the inherent complexity of representing diverse object instances within each class.

7 Conclusion

This visualization analysis demonstrates the effectiveness of SAM3 in identifying and segmenting different object classes in the input image. The clustering results provide insights into how the model represents different semantic categories and how these representations are transformed through the encoder-decoder pipeline. The decoder attention analysis reveals the spatial patterns the model uses for segmentation, which could inform future improvements to the architecture.

Future work could involve comparing these results with ground truth annotations to quantify segmentation accuracy and investigating the relationship between clustering quality and segmentation performance. Additionally, extending this analysis to multiple images and scenes would provide more comprehensive insights into SAM3's generalizability and robustness across diverse visual contexts.