Self-Attention-Driven Refinements for Training-Free Segmentation in Diffusion Transformers

Rento Yamaguchi[1][0009−0000−9163−8031] and Keiji Yanai[1][0000−0002−0431−183X]

The University of Electro-Communications, Tokyo, Japan
{yamaguchi-r, yanai}@mm.inf.uec.ac.jp

Abstract. Diffusion models have made remarkable advancements in the field of image and video generation, and the generation capabilities of recent Diffusion Transformer models have significantly improved. This paper proposes a method for high-quality, training-free semantic segmentation using a single inference step from a pretrained Multi-Modal Diffusion Transformer model. In this approach, a Self-Attention mechanism for image features is applied to the Cross-Attention Map obtained from the inference of a denoising model. Unsupervised clustering extracts object masks without labels, later linked to Cross Attention labels. This method enables segmentation for targets with flexible vocabulary, demonstrating its effectiveness using the PASCAL VOC, a standard segmentation benchmark dataset. While existing image generation models have certain limitations, we investigate the latent capability of Diffusion Transformer models in capturing object-text relationships through attention mechanisms.

Keywords: Training-Free Semantic Segmentation · Multi-modal Diffusion Transformer · Stable Diffusion · Attention Map

1 Introduction

Semantic segmentation, the process of assigning precise class labels to every individual pixel within an image, has emerged as a cornerstone in cutting-edge technology with the potential to revolutionize numerous fields, including healthcare, autonomous driving, and robotics. This burgeoning interest can be attributed to the pivotal role it plays in enhancing machine perception and understanding of complex visual scenes, facilitating advancements in areas like medical image analysis, where accurate segmentation could lead to better diagnostics and treatment planning, and in driver-less vehicles, where it ensures safer navigation and obstacle avoidance. Consequently, semantic segmentation has become a focal point for extensive research and innovation.

Conventional approaches to semantic segmentation predominantly rely on supervised learning methodologies, which necessitate the labor-intensive process of annotating object-mask pair data at a large scale. These methods incur substantial costs and resources for data creation, curation, and subsequent model

training. Despite their widespread use, models developed using traditional supervised learning have limitations, particularly when confronted with classes or object types not present within the training dataset. This often results in a lack of generalization ability, posing significant challenges in deploying such models in dynamic, real-world environments where unforeseen scenarios are common.

To navigate these challenges, the advent of Training-Free Semantic Segmentation techniques has marked a significant breakthrough. These novel techniques leverage the capabilities of large-scale multi-modal foundation models, providing a promising alternative that circumvents the necessity for explicit training data. By capitalizing on the extensive knowledge embedded in foundational models, pre-trained on vast datasets comprising both text and images, these approaches facilitate the application of semantic segmentation across a multitude of downstream tasks. Foundation models such as CLIP (Contrastive Language-Image Pretraining) [8] and Stable Diffusion [9] exemplify this paradigm shift. CLIP successfully learns a unified embedding space for images and text, enabling it to generalize across a diverse range of visual and textual inputs. Similarly, Stable Diffusion demonstrates an unprecedented ability to generate any conceivable image from a textual input, harnessing a rich repository of extensive, real-world knowledge sourced from diverse domains.

In the realm of image generation, diffusion models have garnered attention for their remarkable quality of outputs. The Diffusion Transformer (DiT) architecture [6], in particular, represents a sophisticated evolution in this field. By integrating the powerful capabilities of transformers with diffusion processes, these models have achieved state-of-the-art results in generating realistic images and videos, showcasing an advanced comprehension of both text and image features. This paper builds upon the foundational research efforts [13] that have pioneered the application of Cross-Attention Maps, which are crucial components for the incorporation of text conditions in Diffusion Transformers, in the domain of Training-Free Semantic Segmentation.

In our contribution to this evolving narrative, we introduce a method that expands on the established use of Cross-Attention Maps by incorporating Self-Attention Maps. These maps enable more nuanced and fine-grained object grouping within scenes. Our approach refines the results of semantic segmentation through the application of unsupervised clustering techniques, which enhance the granularity and accuracy of segmentation outputs. This study not only advances the workflow of segmentation tasks but also delves into the internal workings of the DiT architecture, promising insights that could significantly boost its interpretability and adaptability in practical applications.

2 Related Work

2.1 Training-Free Semantic Segmentation

There exists a variety of innovative approaches that employ large-scale models for open-vocabulary segmentation [4, 5, 12]. These efforts signify substantial progress

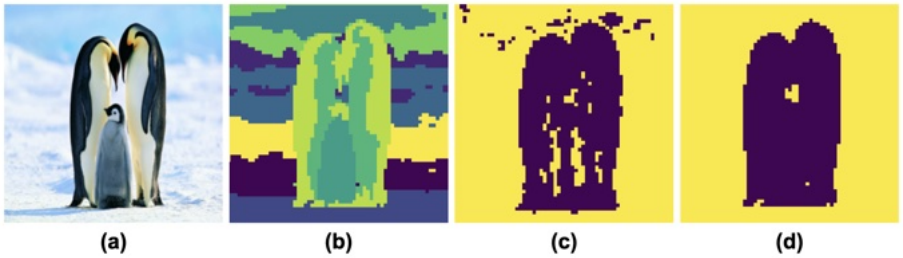

Fig. 1. Example of our method's result. (a) Input image (b) Self Attention Map extracted from DiT (c) Aggregated Cross Attention Map for "penguin" and "background" extracted from DiT (d) Refined mask generated by our method.

in extending segmentation capabilities beyond fixed vocabulary limits. Nevertheless, this section is dedicated to discussing the specific domain of Training-Free Semantic Segmentation methods, which distinctly eschew the training of dedicated decoders for segmentation tasks. Training-Free Semantic Segmentation stands as a paradigm shift in the field, offering a method to execute semantic segmentation by leveraging pre-trained models. This technique substantially lowers the cost and effort involved in segmentation by eliminating the need for extensive training data and the traditional training process itself. By capitalizing on models capable of extracting extensive text and image features at scale, it provides unmatched flexibility, allowing seamless adaptation to novel classes and datasets without the need for retraining.

A particularly notable model that has been commonly utilized in these Training-Free Semantic Segmentation initiatives is the CLIP model. The versatility of CLIP in handling diverse datasets and its proven efficacy in various segmentation tasks underline its significance. Advances such as MaskCLIP [14] exemplify the robust potential of CLIP for segmentation tasks; this technique ingeniously transforms the linear layer of CLIP into a convolutional layer, thus enabling open-vocabulary segmentation while maintaining fixed weights. Such modifications exhibit the potential for mining rich semantic information without intensive investment in computational resources.

In addition to the advances enabled by CLIP, recent years have witnessed the emergence of various methods employing diffusion models, presenting further improvements in the realm of semantic segmentation. One such pioneering effort is embodied in StableSeg [3], which exploits the U-Net [10] based Stable Diffusion framework to synthesize highly accurate segmentation masks derived from both Cross-Attention Map and Self-Attention Map. These maps serve as essential elements in elevating segmentation precision, illustrating the efficacy of diffusion processes in extracting detailed semantic information.

Additionally, there exists an exploratory approach that utilizes the Diffusion Transformer-based Stable Diffusion 3. This technique conducts segmentation operations utilizing the Cross-Attention Map within the intricate Multi-Modal

Diffusion Transformer [13]. Such an approach indicates a significant stride in combining diverse modalities and harnessing the synergistic potential of textual and visual features, enriching the segmentation landscape.

Further enriching this domain, research exemplified by DiffSeg [11] investigates unsupervised methodologies. Using a novel approach involving iterative merging and synthesizing of Self-Attention Maps, DiffSeg excels in accurately extracting object candidate masks. This high-precision method attests to the power of unsupervised strategies in achieving top-tier segmentation outcomes, once again highlighting the transformative potential of self-attention mechanisms in modern semantic segmentation tasks.

3 Methodology

In our approach, we expand on Yamaguchi's method [13] by using the pre-trained Stable Diffusion 3 model [1]. This model facilitates the extraction of the Joint Attention Map from a single inference step of the noise removal system MM-DiT. This process is key in acquiring two distinct Cross Attention Maps that encapsulate both image and text features, along with a Self-Attention Map dedicated to image features. By adhering to the framework established in prior research [13], we implement semantic segmentation using the Cross-Attention Maps. Subsequently, we employ k-Means clustering on the Self-Attention Map to create refined object candidate masks, which are then matched with the labels identified in the Cross-Attention Maps.

3.1 Preliminary

Stable Diffusion 3 is composed of several sophisticated components: a Variational Auto Encoder for embedding images into a Latent Space, three advanced text encoders—namely CLIP-L, CLIP-G, and T5-XXL—for crafting text embeddings, and a Multi-Modal Diffusion Transformer tasked with noise reduction. Each component works synergistically to enhance the accuracy and efficiency of the model. Within the attention block, the concept of Joint Attention is innovatively employed. This mechanism operates similarly to the self-attention process, achieved by concatenating query, key, and value vectors projected from both image and text features. This joint attention construct calculates a probability, known as Joint Attention Probability, through the interactions between extended query and key vectors derived from the feature sets. This probability incorporates the Image Self-Attention Probability, originating from the image features, the Text Self-Attention Probability from text features, and a Cross-Attention Probability that intertwines both image and text features.

Following the methodology outlined in previous works [13], we commence by fixing a single timestep and executing a step of inference using MM-DiT on the input image to procure the Joint Attention Map. Subsequently, from this Joint Attention Map, we extract two Cross-Attention Maps that define relationships between image and text features. These maps are then averaged individually

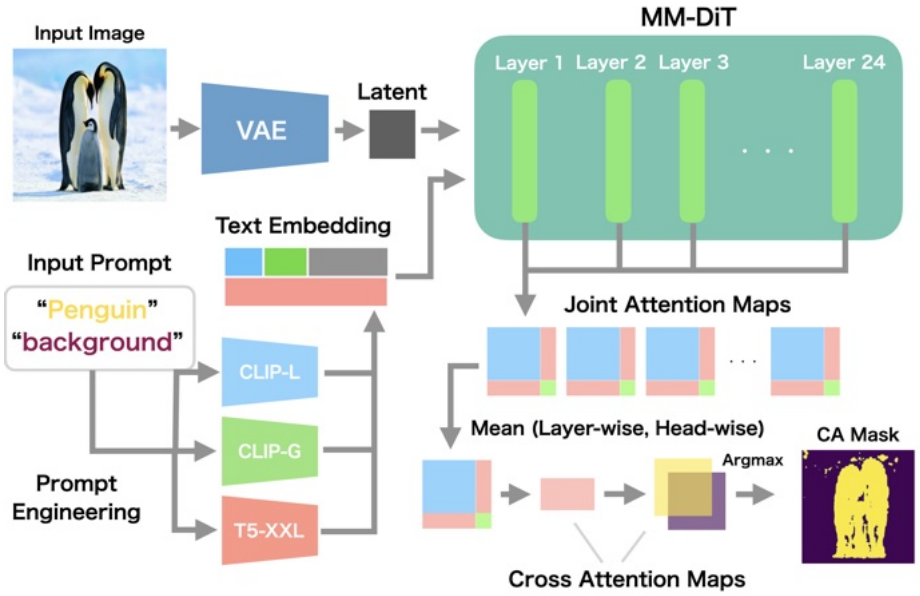

Fig. 2. An overview of extracting the Joint Attention Map from a single step of MM-DiT and deriving the Cross-Attention Map that describes the relationship between image and text features. The extracted Cross-Attention Maps are averaged across both the multi-head dimension and the layer dimension of MM-DiT to form a single cross-attention map. Subsequently, using the argmax operation on the map of tokens belonging to the segmentation class of interest, it results in a single segmentation map. (Partially adapted from [13])

and collectively with multi-head attention averages. We further integrate them with predefined Cross Attention Maps from MM-DiT layers. Finally, by assigning each pixel the class with the most significant response, we derive the final segmentation mask (see Figure 2).

3.2 Self-Attention Refinements

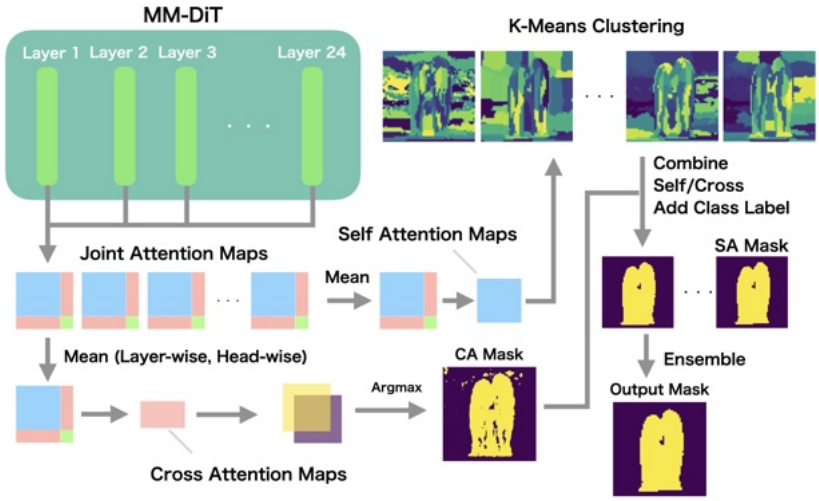

Fig. 3. An overview of the method to refine MM-DiT's Cross Attention Map using the Self Attention Map. The Self-Attention Map obtained from a single inference step of MM-DiT is clustered using K-Means to create multiple unlabeled clusters. Each cluster is then checked against the regions of labels in the Cross-Attention Map to determine which label region it overlaps with the most. By assigning the most frequent label to each cluster, labels are applied to the Self Attention Map.

This segment elaborates on the methodology of refining segmentation masks derived from Cross-Attention using the MM-DiT model. Illustrated in Figure 3, the refinement process initiates with the extraction of the Self-Attention Map from the Joint Attention Map generated at one processing step within MM-DiT. The Self-Attention Map is tailored to image features analyzed through Self Attention principles. It is structured as a tensor with dimensions $(L_w L_h, L_w L_h)$, where L_w and L_h indicate the width and height of the Latent space, representing the strength of relationships between each pixel in the Latent space and surrounding pixels. This configuration highlights zones of increased affinity and segment candidate areas of interest.

To further segment these intricacies, we apply K-Means clustering to the Self-Attention Map, divided into n clusters as determined by a hyperparameter. This clustering effectively segments each pixel into specific regions, leading to

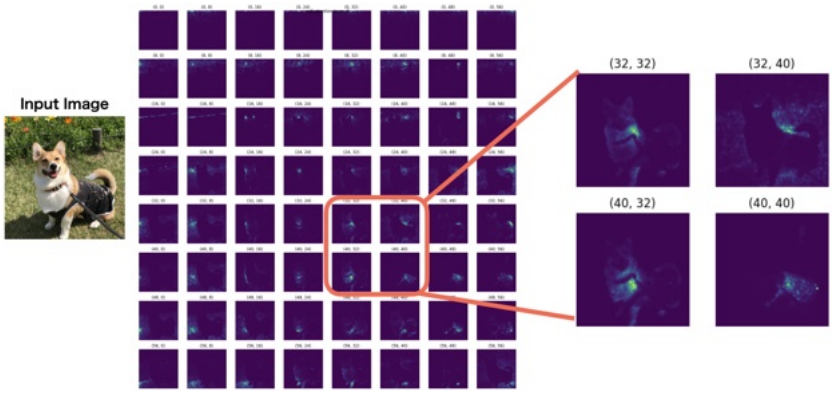

Fig. 4. Visualization of the Self-Attention Map in MM-DiT. The Self-Attention Map extracts a probability map that highlights coordinates similar to a given coordinate. Pixels belonging to a dog strongly respond with other pixels in the dog's region, while pixels belonging to the background strongly respond with other pixels in the background region.

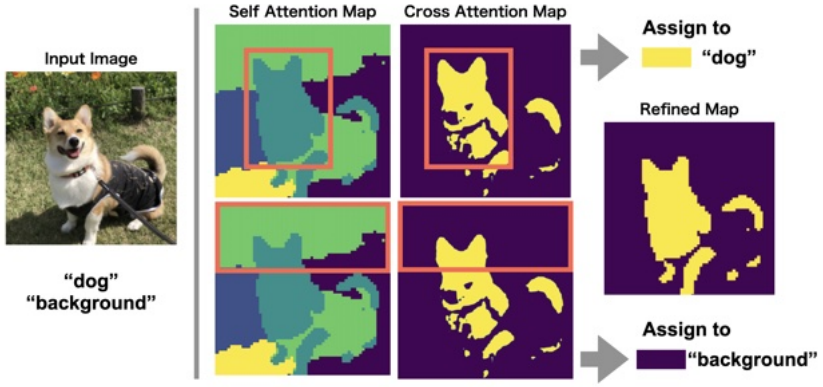

Fig. 5. Overview of assigning class labels to clusters within the Self-Attention Map using Cross-Attention guidance. For each cluster, determine the prevalent class label by overlaying it with the corresponding region from the Cross-Attention Map. This ensures consistency in labeling and associates each cluster with the most representative class.

the generation of abstraction masks for potential object candidates that lack predefined class labels.

In following steps, as showcased in Figure 5, we assign class labels to each of the delineated k Self-Attention Maps. For each cluster discerned through K-Means within the Self-Attention Map, a Cross-Attention Map overlay is employed to ascertain the most frequently appearing class label within the overlap region. This frequently appearing label is assigned to the entire cluster. Clusters that share identical labels are classified collectively as part of the same object class. By consolidating the class labels most frequently utilized across each of the k Self-Attention Masks, we are able to ensemble these outcomes to yield a final segmentation mask with significantly enhanced precision and accuracy.

4 Experiments

4.1 Experimental Setup

In the following experiments, we utilize the weights of Stable Diffusion 3, specifically 'stabilityai/stable-diffusion-3-medium-diffusers' on diffusers [7]. Unless otherwise specified, we adopt a timestep of $t = 150$ and extract attention from layers 9 to 14 in the 24-layer MM-DiT, following the existing research setup [13]. The number of clusters for K-Means clustering is set to $n = 50$.

4.2 Layer-wise Self-Attention Analysis

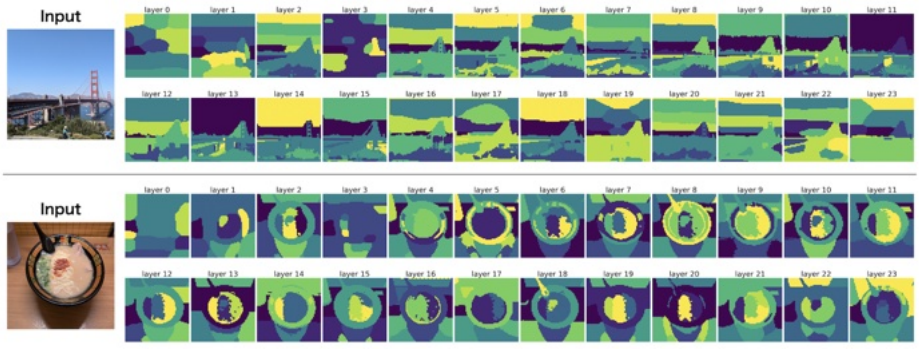

Fig. 6. Visualization of the differences in Self-Attention Maps obtained from each layer of MM-DiT.

To investigate the characteristics of Self-Attention Maps from each layer within the 24-layer MM-DiT, we extracted the Self-Attention Map from each layer and applied K-Means clustering to each. The differences in these maps are visualized in Figure 6. For clearer illustration of the differences, the number of

clusters was set to $n = 10$. The figure shows that in the early stages, such as layers 0 to 3, the Self Attention Maps are clustered in a manner that differs from the object's boundaries within the image. Additionally, in the intermediate layers, it was observed that the maps do not necessarily split exclusive objects into one or more distinct clusters, and some clusters may include two different objects. This can be attributed to the nature of self-attention, which not only maintains the probability map of the relationships between pixels through also considers the appearance and structure for image generation, thus reflecting overall image features regardless of object boundaries or structures.

4.3 Comparison on Number of Clusters

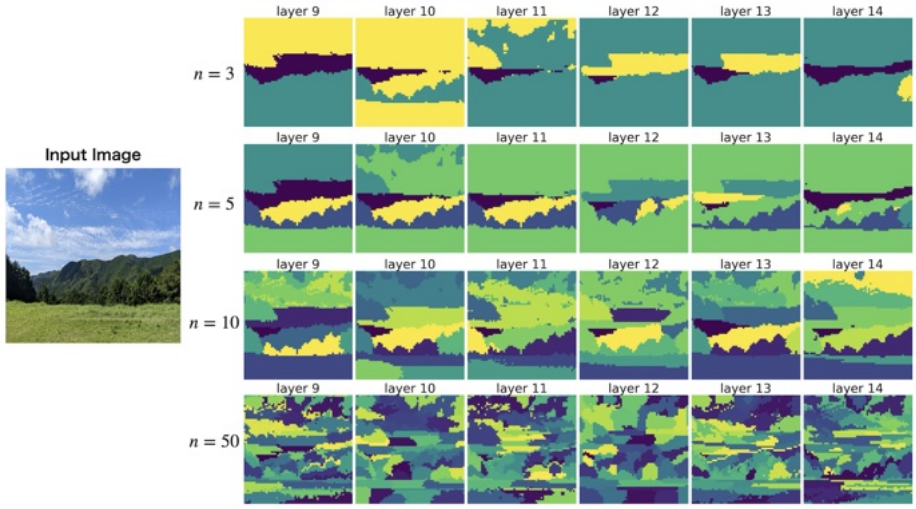

Fig. 7. Comparison of segmentation masks generated with different numbers of clusters.

Subsequently, we conducted a comparison of segmentation masks by varying the number of clusters n. As shown in Figure 7, it was observed that as the number of clusters increases, the boundaries of the objects are segmented more finely, allowing more accurate extraction of the shapes of objects. However, if the number of clusters becomes too large, over-clustering may occur. When there are too many clusters, noise regions in Cross Attention may also be assigned to inappropriate class labels. Therefore, selecting an appropriate number of clusters is crucial. Table 1 presents the change in mIoU on the PASCAL VOC dataset[2] as the number of clusters varies. From these results, it can be concluded that Self Attention Refinement generally improves the segmentation accuracy, However with fewer clusters, the accuracy of the segmentation masks generated by Cross-Attention decreases.

Table 1. Relationship between number of clusters and mIoU.

	10	20	30	50	100	200	w/o SA
mIoU	0.448	0.461	0.462	0.462	0.461	0.459	0.452

4.4 Qualitative Evaluation

To demonstrate the effectiveness of the proposed method, we qualitatively compared the segmentation masks generated from the Cross-Attention Map (CA Map) with those refined using the Self-Attention Map. As shown in Figure 8, it was observed that the segmentation masks generated by the proposed method extract object shapes more accurately than the masks generated solely by Cross-Attention. In addition, the proposed method helps reduce artifacts within masks, resulting in smoother boundaries.

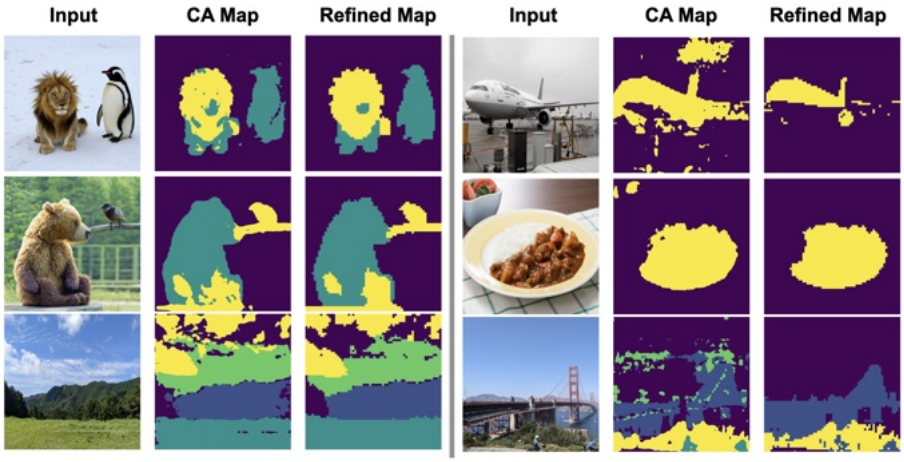

Fig. 8. Qualitative comparison of segmentation masks generated by Cross-Attention Map and Self Attention Refinement Map.

4.5 Quantitative Evaluation

The proposed method was applied to the PASCAL VOC 2012 dataset to assess its semantic segmentation accuracy on this established semantic segmentation benchmark. We calculated the Intersection over Union (IoU) between the segmentation masks generated by the proposed method and the ground truth labels of the PASCAL VOC 2012 dataset, using the mean IoU as a performance metric. Furthermore, we compared the segmentation masks produced by the proposed

method with those generated using Cross Attention to demonstrate the effectiveness of our approach.

For comparison purposes, we selected StableSeg [3], a training-free semantic segmentation method that uses U-Net-based Stable Diffusion, and MaskCLIP [14], which utilizes the CLIP model.

Table 2. Quantitative Comparison on PASCAL VOC

Methods	**MaskCLIP**	**StableSeg**	**Ours (w/o SA)**	**Ours (w/ SA)**
mIoU	0.447	0.472	0.452	0.462

5 Conclusions

In this study, we presented a novel refinement approach for enhancing segmentation results via Self-Attention Maps, specifically tailored for Training-Free Semantic Segmentation using the pretrained Stable Diffusion 3's Multi-Modal Diffusion Transformer model. Our proposed methodology focuses on refining segmentation masks derived from Cross-Attention Maps by utilizing clustering techniques based on Self-Attention Maps to effectively generate candidate masks for objects.

The experimental evaluation, conducted using the well-known PASCAL VOC 2012 dataset, demonstrated the efficacy of our proposed approach. Specifically, it was shown that leveraging Self-Attention Maps significantly improved the accuracy of the segmentation tasks compared to those results obtained merely from Cross-Attention Maps. This highlights the potential of Self-Attention refinement in yielding superior segmentation quality.

However, our study does present certain limitations that merit consideration. A foundational assumption is that the Cross-Attention Map maintains a consistent degree of meaningful accuracy. Furthermore, the pretrained Stable Diffusion model is assumed not to exceed its innate capabilities in text-image grouping. These limitations imply constraints on the generalizability of our method in contexts where these assumptions may not hold.

As we look towards future research endeavors, several key areas for further investigation emerge. Primarily, enhancing the clustering algorithm applied to Self-Attention Maps holds promise for achieving even more refined segmentation results. Furthermore, extending the application of our proposed method across a broader spectrum of segmentation benchmark datasets will help validate its robustness and versatility in varied practical scenarios. Emphasizing these improvements and extensions will enable a deeper understanding and wider applicability of the approach we have proposed.

References

1. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (Jun 2010)
3. Honbu, Y., Yanai, K.: Training-free region prediction with stable diffusion. In: ACM MM (2024)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
5. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: CVPR. pp. 7086–7096 (2022)
6. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
7. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., Wolf, T.: Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers (2022)
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
11. Tian, J., Aggarwal, L., Colaco, A., Kira, Z., Gonzalez-Franco, M.: Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. CVPR (2024)
12. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: ICCV. pp. 1206–1217 (2023)
13. Yamaguchi, R., Yanai, K.: Exploring cross-attention maps in multi-modal diffusion transformers for training-free semantic segmentation. In: ACCV. pp. 260–274 (2024)
14. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. pp. 696–712. Springer Nature Switzerland, Cham (2022)