Exploring Cross-Attention Maps in Multi-modal Diffusion Transformers for Training-Free Semantic Segmentation

Rento Yamaguchi[1] and Keiji Yanai[1]

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

Abstract. This paper presents a novel training-free semantic segmentation method that leverages a pre-trained large-scale image generation model incorporating the Multi-modal Diffusion Transformer (MM-DiT) architecture. Inspired by training-free segmentation techniques using the U-Net-based noise removal model in the Stable Diffusion framework, our approach extracts cross-attention maps between textual and visual features during the inference stages of the MM-DiT to generate mask images. Experimental results demonstrate that our method achieves segmentation accuracy comparable to CLIP-based and U-Net-based stable diffusion methods. While the direct segmentation scores are relatively modest, the significance of our work lies in the exploration of cross-attention maps within the DiT. This investigation provides critical insights that could advance training-free segmentation methodologies and enhance the interpretability of diffusion-based models.

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

1 Introduction

Semantic segmentation in computer vision involves assigning a class label to each pixel in an image, which is a task that holds significant importance across a myriad of domains such as image editing, autonomous driving, and medical image analysis. Conventional supervised learning approaches to semantic segmentation demand extensive labeled annotation datasets, which are costly and labor-intensive to generate. Additionally, these models typically exhibit poor generalization to unseen classes. This limitation hinders their practical applicability.

To alleviate these challenges, unsupervised and zero-shot segmentation techniques have surfaced as promising alternatives. Zero-shot semantic segmentation, in particular, aims to generalize to new categories without requiring explicit training on those categories, thereby addressing the data scarcity and labeling cost issues. Noteworthy advancements in zero-shot segmentation leverage large pre-trained models like CLIP (Contrastive Language-Image Pretraining) [12] and

U-Net-based architectures integrated within the Stable Diffusion framework [14], which exploit the synergy of textual and visual features.

Despite the efficacy of these approaches, limitations remain, especially when adapting to the latest architectures. For instance, while U-Net-based noise reduction models within Stable Diffusion have demonstrated some success, the emergence of the Multi-modal Diffusion Transformer (MM-DiT) in Stable Diffusion 3 [4] introduces a new paradigm that existing methods cannot directly apply to. This necessitates innovative methodologies to harness the potential of MM-DiT effectively.

In this paper, we propose a novel zero-shot training-free semantic segmentation method that utilizes the MM-DiT architecture from the pre-trained large-scale image generation model Stable Diffusion. Drawing inspiration from U-Net-based training-free segmentation techniques, our approach extracts cross-attention maps during the inference stages of MM-DiT to produce segmentation masks. Our experiments reveal that this method attains segmentation accuracy on par with established CLIP-based and U-Net-based approaches, although direct segmentation metrics are still moderate. Crucially, our work emphasizes the exploration of cross-attention mechanisms within the DiT architecture, offering vital insights that could drive the future development of training-free segmentation techniques and enhance the interpretability of diffusion-based models.

2 Related Work

2.1 Training-Free Semantic Segmentation

Zero-shot learning is a paradigm wherein models classify and segment data into categories that are not explicitly encountered during their training phase. This methodology utilizes knowledge from known categories to infer and segment unseen categories, significantly alleviating the challenges associated with collecting and annotating large volumes of labeled data. Specifically, in the realm of semantic segmentation, zero-shot learning is invaluable, offering practical solutions in areas like autonomous driving, healthcare, and remote sensing.

One significant advancement in this domain is the Segment Anything model by Kirillov *et al.* [7]. The model serves as a foundation for segmentation, being pre-trained on extensive annotated datasets. It provides exceptional performance across diverse images and annotations, establishing a benchmark in the use of large-scale pre-trained models for semantic segmentation.

CLIP models have also been paramount in zero-shot semantic segmentation. Works by Rao *et al.* [13] and Luddecke *et al.* [9] have combined textual descriptions with visual features. The key idea lies in mapping both text and images into a unified embedding space, aligning similar concepts closely. This shared space allows models to infer segmentation masks for unseen classes based on textual descriptions, circumventing the need for annotated examples for every possible class. Recently, diffusion models have drawn attention for their precision in image generation. Studies by Wu *et al.* [16] and Tian *et al.* [15] have proposed leveraging these models to generate accurate mask images for segmentation tasks.

Moreover, training-free semantic segmentation has emerged as a promising avenue for improving model performance. This approach leverages pre-trained models and existing knowledge to achieve high-accuracy segmentation on new datasets or tasks without additional training, which is particularly advantageous in scenarios with time constraints or limited resources. For example, MaskCLIP by Zhou *et al.* [17] and FreeDA by Barsellotti *et al.* [1] employ pre-trained large vision models to perform segmentation of unseen classes without any training data, significantly reducing data collection and annotation time, thus enabling rapid deployment. Additionally, the StableSeg model introduced by Honbu *et al.* [6] utilizes Cross Attention Maps and Self Attention Maps within the U-Net architecture of the Stable Diffusion model to achieve segmentation without further supervision. This innovative approach highlights the potential of diffusion models to enhance segmentation accuracy through inherent attention mechanisms.

In conclusion, zero-shot learning and training-free semantic segmentation represent crucial methodologies for minimizing the cost and time associated with labeled data acquisition, allowing models to swiftly and efficiently adapt to new tasks and environments.

2.2 Diffusion Transformers

Diffusion models have emerged as powerful tools for generating high-quality visual content through a sequential noise removal process. These models initiate the process with a noisy version of an image, progressively denoising it to produce a clear and detailed final output. Diffusion models have demonstrated significant potential in various tasks such as image generation, editing, and semantic segmentation. Stable Diffusion, based on the Latent Diffusion Model [14], was trained on hundreds of millions of pairs of text and image data. This model employs an extended U-Net with integrated attention mechanisms as the primary noise reduction component, enabling the generation of high-quality images. The widespread availability and success of Stable Diffusion have underscored the effectiveness of this approach.

On the other hand, Peebles *et al.* [10] introduced the Diffusion Transformer (DiT), a Transformer-based noise reduction model. The DiT model has been adapted to create even higher quality image generation models, such as Stable Diffusion 3 [4], and video generation models like OpenAI Sora [2], which generate videos indistinguishable from real footage. Esser *et al.* [4] proposed the Multimodal Diffusion Transformer by integrating the Flow Matching technique [8] with the robust noise reduction capabilities of DiT. This multi-modal approach enables the generation of high-quality, coherent images aligned with free-form text prompts, pushing the boundaries of what is achievable in image generation and demonstrating the versatility and power of diffusion models and transformer-based architectures.

2.3 Attention Mechanisms in Diffusion Transformers

The evolution of diffusion models, particularly with the integration of Transformer architectures, has marked significant advancements in their ability to handle complex and multi-modal data. A crucial element driving this evolution is the cross-attention mechanism, which facilitates the interaction of diverse types of information, such as textual and visual data, during the generative process. Cross-attention mechanisms play a pivotal role in enhancing the understanding of contextual relationships within data, thereby contributing significantly to the interpretability and performance improvements of these models.

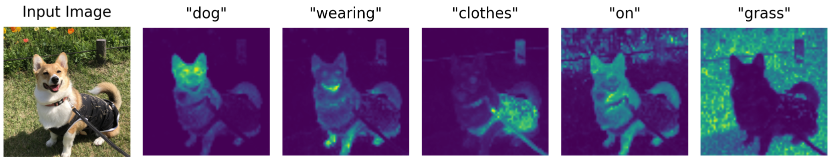


Fig. 1: Visualization of cross-attention maps between text and images. The input image is shown on the left, followed by attention maps for different text prompts. Each attention map highlights the regions of the image that correspond to the given text prompt.

Within the framework of the Multi-modal Diffusion Transformer (MM-DiT) in Stable Diffusion 3, Joint Attention layers promote the alignment and fusion of multi-modal information, enabling the generation of high-quality and coherent images that are consistent with meaningful text prompts. This alignment is achieved by mapping textual embeddings to visual features, effectively guiding the noise reduction process in accordance with the semantic content of the text prompts. Joint-Attention Maps, which capture these interactions, serve as a foundational element in our proposed methodology, providing invaluable insights for extracting meaningful segmentation masks even in zero-shot contexts. Existing research, such as that by Honbu *et al.* [6], has demonstrated the promise of utilizing attention maps within U-Net-based architectures for zero-shot training-free segmentation. However, the integration of attention mechanisms within the DiT architecture remains an underexplored area with substantial potential. Leveraging these mechanisms offers the prospect of enhancing segmentation accuracy and interpretability without the need for extensive annotated datasets.

In the following section, we describe our proposed method, which utilizes cross-attention maps extracted during the inference stages of MM-DiT to generate segmentation masks. This technique aligns with the principles established by prior research while pioneering new possibilities achievable with state-of-the-art diffusion transformers.

3 Methodology

3.1 Overview

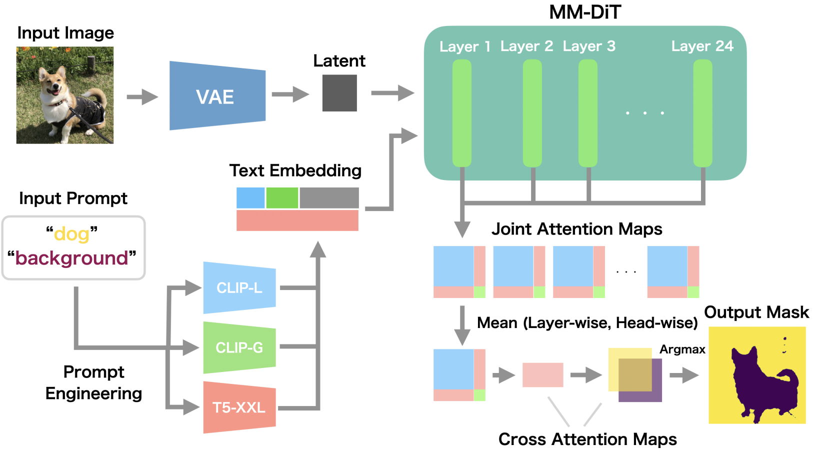


Fig. 2: Overview of the Proposed Method

Inspired by the concept of utilizing Cross-Attention Maps as demonstrated by Honbu *et al.* [6], in this paper, we propose a novel approach for training-free semantic segmentation by leveraging MM-DiT within the Stable Diffusion 3 framework. Unlike its predecessors in the v1 and v2 series, Stable Diffusion 3 employs three separate text encoders to precisely capture the features of input prompts and generate text-aligned images. The architecture of Stable Diffusion 3 repeatedly applies MM-DiT through Joint-Attention mechanisms that synergize image and text embeddings.

Our proposed method encompasses two primary steps: (1) generation of text embeddings using three distinct text encoders, and (2) usage of Joint-Attention for segmenting regions within the image. A single inference step within MM-DiT facilitates training-free semantic segmentation.

Figure 2 provides an overview of the proposed method's workflow. This diagram illustrates the sequential steps involved in the methodology, highlighting key processes such as data acquisition, preprocessing, feature extraction, model training, and evaluation. Each step is critical for the successful implementation and validation of the proposed approach.

3.2 Generation of Text Embeddings

In Stable Diffusion 3, three models—CLIP/L-14, CLIP/G-14, and T5-XXL—are utilized for text encoding. Here, we describe the generation of text embeddings needed for DiT inference. Given a set of k class labels $\{c_0, c_1, \ldots, c_{k-1}\}$ representing target segmentation classes, initial preprocessing involves appending a prefix prompt like "a photo of" to each class label, resulting in modified prompts $\{p_0 + c_0, p_1 + c_1, \ldots, p_{k-1} + c_{k-1}\}$. Each modified prompt is fed into the three respective text models (CLIP/L-14, CLIP/G-14, T5-XXL) to generate corresponding text embeddings.

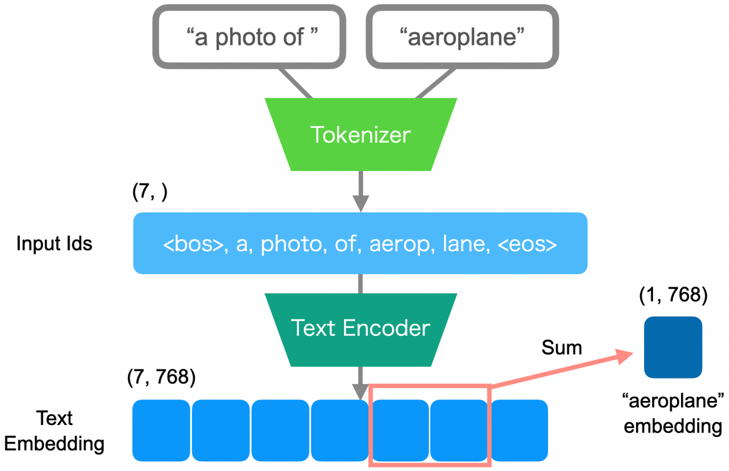


Fig. 3: Overview of the method for generating text embeddings for classes. A prefix prompt like 'a photo of' is added to the class category, and the sum of the portions corresponding to the class category from the generated embeddings is utilized as the embedding for that category.

From these generated embeddings, as illustrated in Figure 3, we extract text embeddings that correspond to each class label c_i. Taking into account the presence of special tokens such as <bos> and <eos> in the CLIP tokenizer, and defining n_i as the number of tokens in each prefix prompt p'_i, the following operations are performed to obtain the class-specific text embeddings:

$$\mathcal{E}_{\text{CLIP-L},i} = \text{Sum}\left(\text{CLIP-L}(p'_i)[n_i + 1 : -1]\right) \tag{1}$$

$$\mathcal{E}_{\text{CLIP-G},i} = \text{Sum}\left(\text{CLIP-G}(p'_i)[n_i + 1 : -1]\right) \tag{2}$$

$$\mathcal{E}_{\text{T5},i} = \text{Sum}\left(\text{T5}(p'_i)[n_i : -1]\right) \tag{3}$$

For all three models, the embeddings for class tokens 0 to $k-1$ are summed according to the above operations, while the pad token embeddings are inserted for positions from k to 76. This results in the final embeddings $\mathcal{E}_{\text{CLIP-L}}$, $\mathcal{E}_{\text{CLIP-G}}$, $\mathcal{E}_{\text{T5}}$.

Subsequently, the two CLIP embeddings are concatenated along the embedding dimension and padded to match the dimensions of the T5 embeddings.

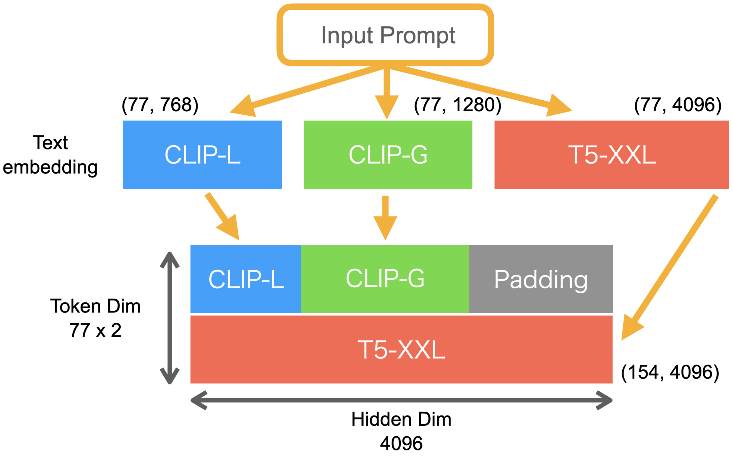


Fig. 4: Procedure for creating a text embedding from the three text encoders

These are then concatenated along the token dimension, resulting in the final text embedding $\mathcal{E}'$ for DiT inference.

3.3 Segmentation using Joint-Attention

For an input image dimensioned at 1024×1024, the VAE encoder in Stable Diffusion 3 generates a 16-channel latent variable z. The latent variable z is patched into a 2×2 grid and positional embeddings are added to form image features x. Assuming the image contains noise corresponding to predefined timestep t, the MM-DiT model predicts the next step of noise.

Stable Diffusion 3 does not separately compute Self-Attention for image and text features or Cross-Attention between them. Instead, it combines them into a unified Query, Key, and Value for computing Joint-Attention. If the linear transformation layers for image and text features are l_{IQ}, l_{IK}, l_{IV}, and l_{TQ}, l_{TK}, l_{TV} respectively, Joint-Attention Map (JAMap) is computed as follows:

$$Q = \mathrm{Concat}(l_{IQ}(x),\ l_{TQ}(\mathcal{E}')) \tag{4}$$

$$K = \mathrm{Concat}(l_{IK}(x),\ l_{TK}(\mathcal{E}')) \tag{5}$$

$$V = \mathrm{Concat}(l_{IV}(x),\ l_{TV}(\mathcal{E}')) \tag{6}$$

$$\mathrm{JAMap} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{7}$$

The JAMap, as shown in Figure 5, can be divided into four quadrants representing Self-Attention for image features, Self-Attention for text features, and

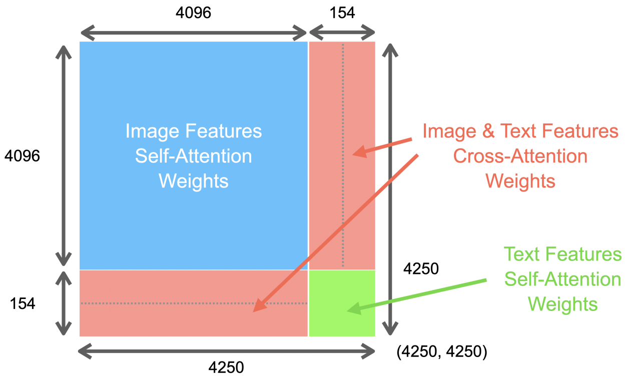


Fig. 5: Visualization of the Joint-Attention Map. The JAMap consists of four regions corresponding to the Self Attention and Cross Attention between image and text features.

two Cross-Attention components between them. The regions in JAMap corresponding to Cross-Attention are extracted, transposed, and averaged to yield the Cross-Attention Map (CAMap) for the i-th DiT layer.

$$\mathrm{CAMap}_i = \frac{\mathrm{JAMap}[:, d_i :, : d_i] + \mathrm{JAMap}[:, : d_i, d_i :]^T}{2}, \tag{8}$$

where d_i denotes the dimension of image features. Each CAMap is calculated using 24 Multi-Head Attention layers. Given the token dimensions and image feature dimensions d_t and d_i, respectively, the Head-dimension-averaged CAMap reshapes into $(2, d_t, d_i, d_i)$, representing CAMaps from CLIP and T5.

If there are k class labels, the final Cross-Attention-based segmentation map M is computed as follows, considering token indices impacted by special tokens.

$$\mathrm{CAMap}_{\mathrm{CLIP},i} = \mathrm{CAMap}[0, 1 : k+1] \tag{9}$$

$$\mathrm{CAMap}_{\mathrm{T5},i} = \mathrm{CAMap}[1, : k] \tag{10}$$

After extracting CAMaps for CLIP and T5, and averaging them over 24 heads, the segmentation mask is generated by applying Argmax across the class dimension, C. The mean map from all CAMaps sourced from the three text encoders is referred to as the Cross Attention Probability Map (CAPM).

$$\mathrm{CAPM}_i = \frac{\mathrm{CAMap}_{\mathrm{CLIP}} + \mathrm{CAMap}_{\mathrm{T5}}}{2} \tag{11}$$

To obtain the final mask image M, the cross-attention probability maps (CAPM) from selected layers k are summed and then the Argmax function is applied across the class dimension. Here, the index set k can be any subset of the layers from 1 to 24.

$$M = \mathrm{Argmax}_C \left(\sum_{j \in k} \mathrm{CAPM}_j \right) \tag{12}$$

Here, k_j represents the index of each selected layer from which the cross-attention probability maps are obtained.

4 Experiments

4.1 Experimental Setup

In this study, we utilize Stable Diffusion 3, which employs a DiT-based architecture, as our model. We used the 'stabilityai/stable-diffusion-3-medium-diffusers' checkpoint from the diffusers library by Hugging Face [11]. Unless otherwise specified, all input images are resized to a uniform size of 1024 × 1024. For the class labels, which are considered to be known, we prepend the prefix prompt "a photo of" when inputting into the text encoders. Quantitative evaluation is performed using the standard segmentation datasets PASCAL VOC [5] and Cityscapes [3].

In this paper, we perform segmentation using the MM-DiT model of Stable Diffusion 3 by specifying a single timestep out of a total of 1000 image generation steps. The appropriate timestep value is experimentally determined and will be discussed in detail in Section 4.3. Additionally, the choice of layers from which to extract attention maps significantly impacts segmentation accuracy. This choice will be explored in detail in Section 4.2.

4.2 Layer-wise Segmentation Differences in MM-DiT

Visualization of the cross-attention maps between text and image features across the 24 layers of MM-DiT is presented in Figure 6.

Upon examining each layer, we observe that attention maps closer to the initial or final layers show weaker responses and appear noisier concerning the regions corresponding to the text. In contrast, the maps from the middle layers exhibit strong responses to image keypoints that correspond to the text tokens. This leads us to hypothesize that the choice of attention maps significantly affects segmentation accuracy. Accordingly, we evaluate the performance on the PASCAL VOC dataset using different layers' attention maps, as shown in Figure 7.

Starting with layers 11 and 12, we incrementally included more layers around the center to determine the optimal range. We expanded the range to include layers 10 and 13, and ultimately found that the highest scores were obtained when using layers 9 through 14. This confirms that the choice of layers significantly impacts segmentation performance.

As observed in Figure 7, the evaluation scores are highest when selecting multiple layers from the central part of MM-DiT. This can be attributed to the

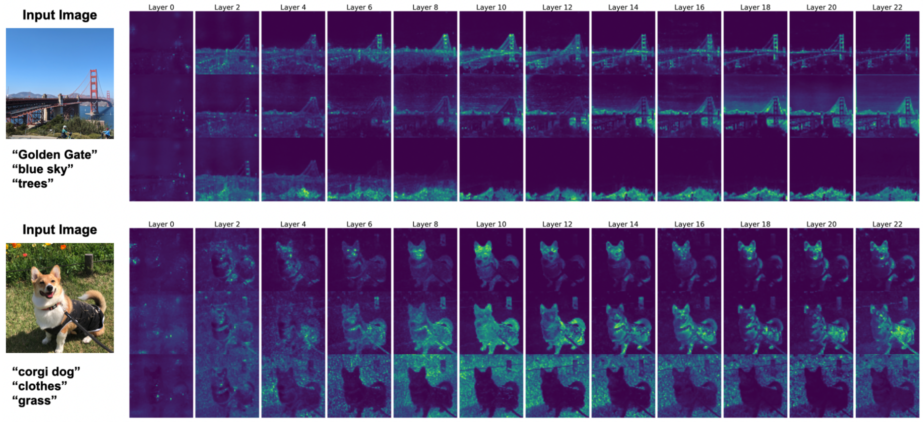


Fig. 6: Qualitative segmentation results from different layers of MM-DiT.

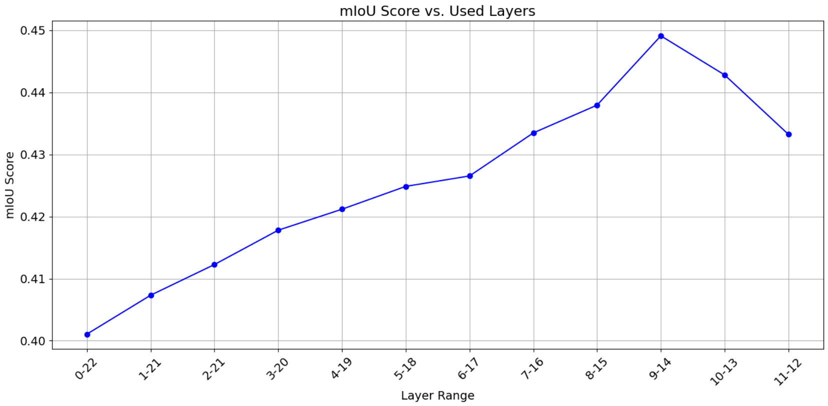
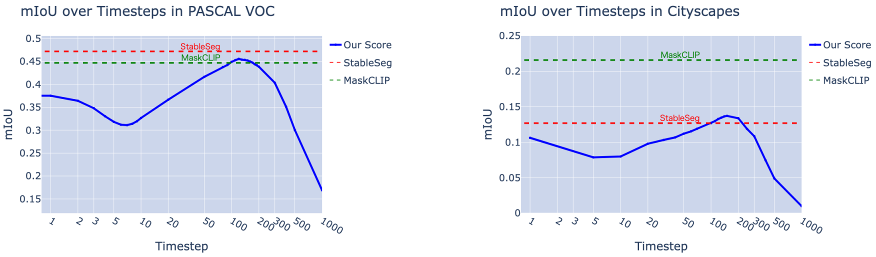


Fig. 7: Quantitative segmentation results on PASCAL VOC by utilizing different layers of MM-DiT, showing how the choice of layers from 0 to 23 affects the mIoU score.

fact that, similar to U-Net-based Stable Diffusion, MM-DiT also encapsulates more semantic information of the image features in layers closer to the model's center. The incremental inclusion of layers around the center—starting with layers 11 and 12, and expanding to include layers 9 to 14—showed a noticeable improvement in segmentation accuracy, supporting our hypothesis. As the experiments progressed, we decided to use this optimal range of layers (9 to 14) for subsequent evaluations.

4.3 Segmentation Variability with Different Timesteps

We evaluated the impact of different timesteps on segmentation accuracy. Specifically, we conducted experiments using comparable training-free segmentation methods, such as MaskCLIP and StableSeg [6,17]. Notably, for a fair comparison, we assessed StableSeg—which employs U-Net-based Stable Diffusion—under identical conditions by utilizing only the Cross-Attention Maps as we do. The evaluation was carried out on benchmarks including PASCAL VOC and Cityscapes, with results illustrated in Figures 8a and 8b.



(a) PASCAL VOC

(b) Cityscapes

Fig. 8: Comparison of mIoU values for different methods with varying timesteps.

These results indicate that the optimal segmentation accuracy is attained when assuming noise is applied around $t = 150$. We hypothesize that at timesteps approximating this value, the model accentuates depicting objects according to the textual description rather than merely denoising images, reflecting a different interpretation phase of the diffusion model's learning process. As the experiments progressed, we decided to use this optimal timestep (around $t = 150$) for subsequent evaluations. Additionally, the results of region segmentation at different time steps are shown in Figure 9.

4.4 Comparison with Existing Training-free Methods

Using the optimal timestep and hyperparameters for the layer of the MM-DiT determined from the above experiments, we conducted a quantitative comparison with the CLIP-based method MaskCLIP [17] and the Stable Diffusion v1 based

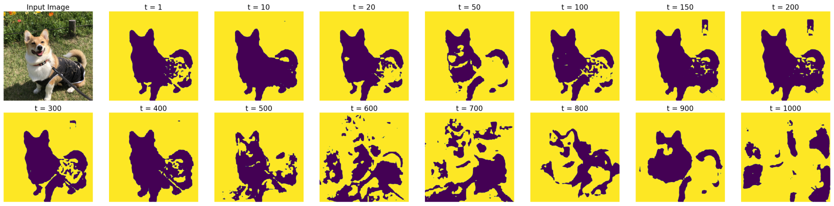


Fig. 9: Results of region segmentation for 'clothed corgi' and 'background' at different time steps.

method StableSeg [6]. The results are presented in Table 1. The evaluation settings are consistent with those described in Section 4.3. These results indicate that our method achieves comparable segmentation accuracy to the two existing methods.

Table 1: Comparison of mIoU scores for different methods across two datasets

Method	**PascalVOC**	**Cityscapes**
Ours	0.452	0.137
MaskCLIP	0.447	0.216
StableSeg	0.472	0.127

4.5 Qualitative Comparison of Different Text Encoders

Stable Diffusion 3 leverages text embeddings generated by three text encoders: two instances of CLIP and one T5. To assess the impact of these distinct text encoders on the resultant segmentation, we qualitatively analyzed their cross-attention maps as demonstrated in Figure 10. While foreground extraction of images containing a single foreground object displayed high accuracy, segmentation performance deteriorated in scenarios where multiple foreground objects were present, occasionally leading to segmentation failures.

4.6 Open Vocabulary Segmentation

Leveraging the Stable Diffusion 3 model trained on extensive web-based data, our method enables open vocabulary segmentation, transcending the limitations of conventional class labels typically found in standard segmentation datasets. The segmentation results are illustrated in Figure 11. This approach allows for flexible segmentation of classes, including sentences containing adjectives and proper nouns.

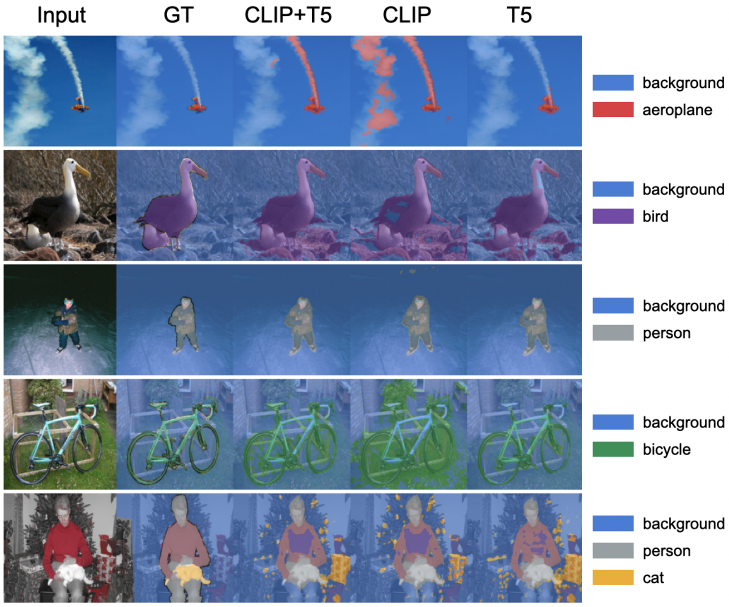


Fig. 10: Qualitative comparison showing how different text encoders affect segmentation outcomes on PASCAL VOC. The upper four images exhibit accurate segmentation, while the bottom image shows a failure case where the cat's position is misaligned.

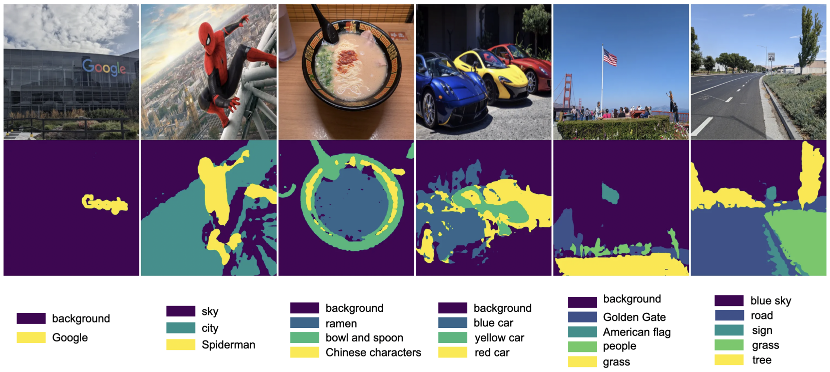


Fig. 11: Open vocabulary segmentation results, demonstrating capability beyond traditional class labels, including segmentation of classes such as proper nouns.

5 Conclusion

In this paper, we proposed a novel training-free semantic segmentation method leveraging the Multi-modal Diffusion Transformer (MM-DiT) architecture within the Stable Diffusion 3 framework. Through extensive experiments, we demonstrated that our approach achieved the comparable performance to existing CLIP-based and U-Net-based methods within the Stable Diffusion framework, albeit with relatively modest direct segmentation scores. Our contribution lies in the exploration and utilization of cross-attention maps in image generation diffusion models, representing a significant step toward enhancing the interpretability and accuracy of zero-shot segmentation methods.

Our findings indicate that selecting attention maps from the middle layers of the MM-DiT model significantly improves segmentation results, emphasizing the importance of these layers in capturing semantic information. Additionally, our experiments revealed the impact of varying timesteps on segmentation accuracy, with optimal results achieved around the timesteps where the model transitions from merely denoising to aligning with textual descriptions.

Qualitative comparisons further validated the influence of different text encoders on the segmentation process. While the model exhibited high performance in single-object scenarios, its performance deteriorated in complex scenes with multiple foreground objects, highlighting areas for future improvement.

In conclusion, this research provides foundational insights into the utilization of cross-attention mechanisms within diffusion-based models for training-free segmentation. These insights pave the way for future research to refine and enhance the performance and applicability of training-free techniques across diverse and practical domains.

References

1. Barsellotti, L., Amoroso, R., Cornia, M., Baraldi, L., Cucchiara, R.: Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In: CVPR (2024)
2. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), `https://openai.com/research/video-generation-models-as-world-simulators`
3. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
4. 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)
5. 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)
6. Honbu, Y., Yanai, K.: Training-free region prediction with stable diffusion. In: ACM MM (2024)

7. 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)
8. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2022)
9. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: CVPR. pp. 7086–7096 (2022)
10. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023)
11. 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)
12. 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)
13. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: CVPR (2022)
14. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
15. Tian, J., Aggarwal, L., Colaco, A., Kira, Z., Gonzalez-Franco, M.: Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. CVPR (2024)
16. 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)
17. 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)