Spatially-Guided Self-Attention Refinement for Zero-Shot Hair Segmentation (Student Abstract)

Suin Kim[1], Jihoon Lee[2], Moonsung Kang[1], Doheun Cha[2], Sangtae Ahn[2*]

[1]School of Electronics Engineering, Kyungpook National University
[2]School of Electronic and Electrical Engineering, Kyungpook National University
{tndls142, leejh98123, vb123, chadoheun, stahn}@knu.ac.kr

Abstract

Recent advances in diffusion-based models have significantly broadened their scope, extending well beyond image generation to encompass zero-shot segmentation tasks. In this work, we introduce a novel, training-free approach that harnesses both self- and cross-attention maps to achieve highly detailed hair segmentation. Our method demonstrates remarkable efficacy in producing fine-grained results without the need for additional training.

Introduction

Diffusion-based models generate attention maps that enable semantic segmentation without extra training. Focusing on hair segmentation, which requires fine detail, we note that cross-attention maps capture high-level semantics but lack spatial precision, while self-attention maps preserve spatial detail with limited semantics. Building on a previous zero-shot method such as DiffSegmenter (Wang et al. 2025), we introduce a training-free, unsupervised framework for zero-shot hair segmentation. Our method identifies a target cross-attention map representing hair and iteratively selects self-attention maps by minimizing their Kullback-Leibler Divergence (KLD) with the target, resulting in hair masks that are sharper, spatially consistent, and semantically meaningful.

Methodology

Our approach consists of four steps, illustrated in Figure 1. Each step is described in detail as follows.

Prompt and Initialization

Given an input image, we provide a fixed prompt that naturally describes a photo while including the target token (e.g., "a photo of a man with detailed hair") and perform a single denoising step at timestep $T = 0$. This step initializes the attention maps used for segmentation.

Cross-Attention Map Aggregation

We extract cross-attention maps from the U-Net at multiple upsampling layers with different resolutions $\{8 \times 8, 16 \times 16, 32 \times 32, 64 \times 64\}$. We extract and merge the cross-attention maps corresponding to the "Hair" token. Each map is resized to 64×64 and combined with resolution-specific weights to form a representative cross-attention map $\mathcal{A}_{\text{cross}} \in \mathbb{R}^{64\times64}$.

Upsample & Duplicate Self-Attention Maps

Self-attention maps are extracted from multiple upsampling layers of the U-Net, at resolutions corresponding to those of the cross-attention maps. Following the procedure proposed in DiffSeg, we iteratively upsample and merge lower-resolution maps with the highest-resolution maps, so that all self-attention maps are aligned as $\mathcal{A}_{self} \in \mathbb{R}^{64\times64\times64\times64}$.

Self-Attention Map Extraction and Manipulation

The main contribution of our method lies in how we utilize the representative cross-attention map to guide self-attention extraction. Each candidate self-attention map is compared against the cross-attention map utilizing a modified version of the KLD, since the standard KLD suffers from asymmetry issues. Prior to computing the divergence, both $\mathcal{A}_{cross}$ and $\mathcal{A}_{self}$ are L1-normalized to ensure valid probability distributions, as shown in Eq.(1), Eq.(2):

$$\tilde{\mathcal{A}}_{\text{cross}} = \frac{\mathcal{A}_{\text{cross}}}{||\mathcal{A}_{cross}||}, \tag{1}$$

$$\tilde{\mathcal{A}}_{\text{self}}[i,j,:,:] = \frac{\mathcal{A}_{\text{self}}[i,j,:,:]}{||\mathcal{A}_{self}[i,j,:,:]||} \tag{2}$$

The divergence between the normalized self and cross-attention maps is then computed as in Eq. (3):

$$\begin{aligned} D[i,j] = \frac{1}{2}\Big(&\mathcal{KL}\big(\tilde{\mathcal{A}}_{\text{cross}} \parallel \tilde{\mathcal{A}}_{\text{self}}[i,j,:,:]\big) \\ &+ \mathcal{KL}\big(\tilde{\mathcal{A}}_{\text{self}}[i,j,:,:] \parallel \tilde{\mathcal{A}}_{\text{cross}}\big)\Big) \end{aligned} \tag{3}$$

, where i and j index the first and second dimensions of the self-attention tensor $\mathcal{A}_{\text{self}}$, with $i, j \in \{0, 1, \ldots, 63\}$. Only maps with divergence below a predefined threshold τ are retained, where we set $\tau = 0.9$ in our experiments. The

*Corresponding author: Sangtae Ahn (stahn@knu.ac.kr)

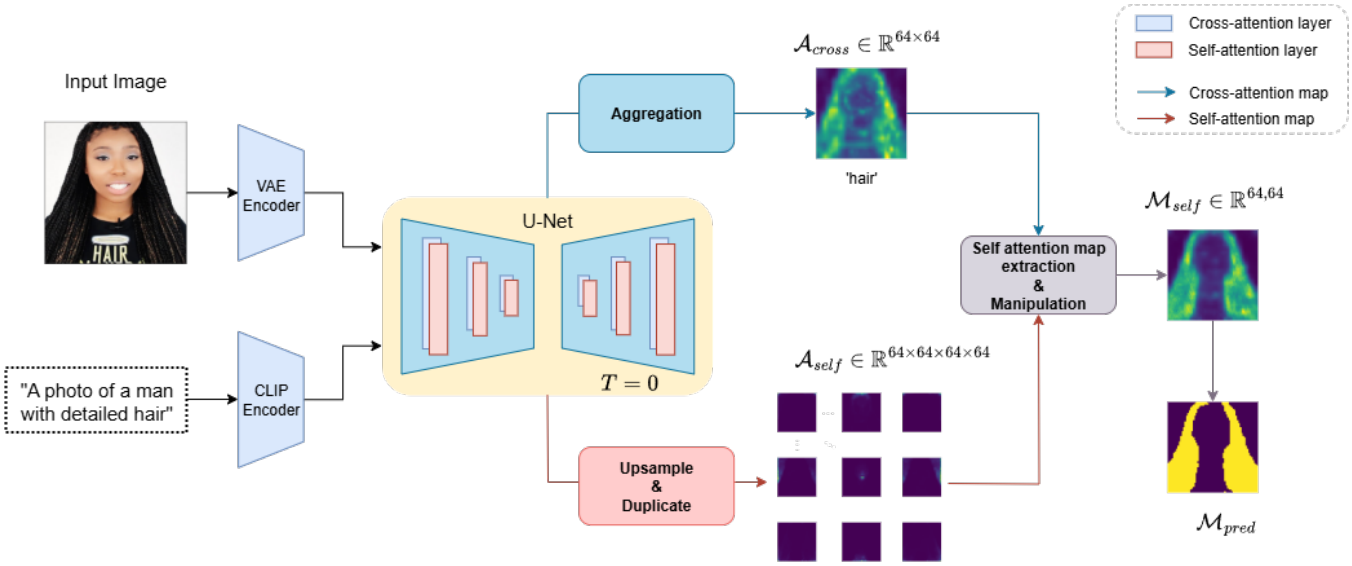

Figure 1: Pipeline for zero-shot hair segmentation: Cross-attention maps guide the selection of self-attention maps, which are then averaged and binarized to obtain the final mask.

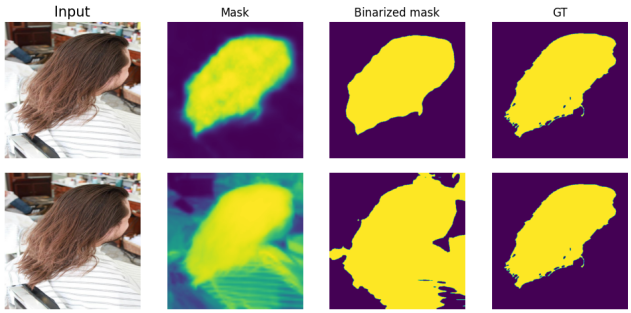

Figure 2: Visual results produced by different models on the Figaro-1k dataset. First row: Proposed method; Second row: DiffSegmenter. From left to right: input image, mask, binarized mask, and ground truth.

threshold τ was empirically determined based on preliminary observations. The selected maps from $\tilde{\mathcal{A}}_{\text{self}}$ are then averaged to obtain $\mathcal{M}_{self} \in \mathbb{R}^{64\times64}$, which is subsequently binarized utilizing Otsu's method and resized to produce the final mask $\mathcal{M}_{pred}$.

Experimental Setup and Results

Dataset and Evaluation Metrics

We evaluate our method on the Figaro-1k dataset. Performance is measured utilizing the standard evaluation metrics: mIoU, Accuracy, Precision, and F1-score.

Results

Figure 2 presents the visualization results of our proposed method. Also, we evaluate the performance of using only the binarized cross-attention map, which shows significantly lower performance than that of our method. This is likely due to the fact that cross-attention captures semantic information rather than spatial locality. Next, we compare our method against DiffSegmenter, which shows consistently better performance across all metrics as shown in Table 1.

Methods	Dataset	mIoU	Accuracy	Precision	F1-Score
DiffSegmenter	Train	0.8066	0.8963	0.8845	0.8836
	Test	0.8070	0.8956	0.8872	0.8847
Ours (w/o SA)	Train	0.7730	0.8852	0.8736	0.8638
	Test	0.7764	0.8865	0.8768	0.8668
Ours	Train	**0.8642**	**0.9360**	**0.9190**	**0.9232**
	Test	**0.8768**	**0.9418**	**0.9269**	**0.9323**

Table 1: Quantitative results on the Figaro-1k dataset. Since the model is training-free, results are presented on both the training and test sets.

Conclusion

We present a zero-shot segmentation approach that employs cross-attention to capture semantic information and self-attention to preserve spatial details. Th method effectively demonstrates the potential of attention maps for segmentation without requiring additional training.

Acknowledgements

This research, undertaken at Kyungpook National University, was supported by the Regional Innovation System & Education (RISE) program through the Daegu RISE Center, funded by the Ministry of Education (MOE) and the Daegu Metropolitan City, Republic of Korea (2025-RISE-03-001). This research was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-02214941)

References

Wang, J.; Li, X.; Zhang, J.; Xu, Q.; Zhou, Q.; Yu, Q.; Sheng, L.; and Xu, D. 2025. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *IEEE Transactions on Image Processing*.