Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/86/20241585

Attribute binding enhancement improvement for diffusion model based zero-shot image segmentation

Nanli Wu

School of Software Engineering, Beihang University, Beijing, China

20241032@buaa.edu.cn

Abstract. As diffusion model's potential ability to accomplish perception tasks being discovered, many researchers have tried to apply diffusion model in segmentation tasks and achieved good results. However, there are not many methods to optimize the diffusion model for segmentation tasks by improving the text side. Even if some research have pointed out that diffusion model sometimes 'misunderstand' prompt and bind attributes to wrong objects. With the development of artificial intelligence and the gradual entry of large models into people's daily lives, in addition to the performance of the model, the model's interactivity and ability to understand natural language are also more important. Based on existing zero-shot diffusion model based segmentation method, this work introduces a new method to enhance attribute binding in the embedding of prompt improve the performance of the model. Through this method, more descriptive text will get better segmentation results, which to some extent, improves the segmentation performance of the model for inputs that are more in line with natural language description habits.

Keywords: diffusion model, image segmentation, zero-shot, attribute binding.

1. Introduction

Nowadays, in the field of computer vision, people must attach labels and annotations to thousands of images in dataset. Therefore, dataset labeling has become a time and effort consuming problem for many researchers. That's why researchers are more and more interested and dig deeper and deeper into zero-shot ways to deal with perception tasks in computer vision. This work introduces an attribute enhancing method to improve the performance of zero-shot diffusion model based image segmentation method.

Diffusion model, one of the most state-of-art generative model. However, researchers have recently found the potential of diffusion model in perception tasks. As a multi-modal large model, diffusion model accept text as input and generate images, indicating its ability to bridge the gap between the modal differences of text and image.

In image segmentation task, we can make full use of that ability to achieve zero-shot segmentation goals. Since diffusion model can generate images of specific objects according to given description, diffusion model must have learnt feature of a great number of objects during its training process. Through using pre-trained diffusion model, target objects can be separated from images.

In this paper, the author would like to dig deeper into cross-attention based zero-shot image segmentation using diffusion model and use attribute binding to improve the performance and understanding of natural language of current segmentation method. With the development of artificial

DOI: 10.54254/2755-2721/86/20241585

intelligence and the gradual entry of large models into people's daily lives, in addition to the performance of the model, the model's interactivity and ability to understand natural language are also more important.

As Feng pointed out in his paper introducing the method called Structured Diffusion Guidance, one of the flaws of diffusion model is that it often mis-bind attributes to wrong objects. Thus, they applied special process to prompt text to enhance attribute binding and improved the generation performance [1, 2]. The author transfers attribute binding method from generation task to segmentation task. Through pre-processing prompt text by noun chunk replacement and unconditioned embedding averaging and enhance attribute binding, the author manages to get better segmentation result. Apart from better performance, the attribute binging enhanced image segmentation may also come into use in specific tasks which require better understanding of prompt.

2. Related works

Recently, researchers have discovered that diffusion model can be applied in many kinds of perceptual challenges, for example, image classification, object detection and semantic segmentation and can have better performance than traditional methods.

As for what part of diffusion model is the key factor that enables diffusion to bridge the gap between text and image, the Prompt-to-Prompt image editing method pointed out that cross-attention maps record the prompt information and determine the spatial layout of the generated images.

There are many works exploring the potential of diffusion model. The DiffuMask use diffusion model to generate dataset and get the annotation for each picture by dealing with cross-attention maps during diffusion step [3]. Through the generated high accuracy annotated dataset, DiffuMask can train segmentation model on that dataset. Though generated dataset is much more effective than mannually labelled one, generating a whole dataset and train another segmentation model is still complex. OVDiff is an open-vocabulary segmentation model which generates lots of images with diffusion model and extract prototypes of target images from these images [4]. This method also requires a generation process and needs to extract prototypes for each class. ODISE also uses the cross-attention map of diffusion model to generate mask embedding, but it still needs supervision [5]. SegDiff combines input image and the current segmentation map and get segmentation result through multiple iterations, but it also needs the supervision of the real segmentation result [6]. Different from those mothods using cross-attention map to generate segmentation result, the DiffSeg method find out that self-attention map contains the information of objects, because two regions associated with the same object have higher self-attention values [7]. However, DiffSeg cannot distinguish which class each segmentaion result belongs to. DiffSegmenter is a zero-shot open-vocabulary segmentation model that generates the initial score map by aggregating cross-attention map during diffusion steps, completes the score map with self-attention map and achieves the segmentation result according to the score map [8]. It is completely zero-shot without need for partly annotated data or extra segmentation network. However, there still is improvement that can be made upon this method. That is fixing the attribute binding flaw of diffusion model.

Though being one of the most state-of-art generative model, diffusion model still has its flaw. According to Structured Diffusion Guidance method, when facing multiple objects generating task, diffusion model will make some mistake, for example, mis-bind the attribute of one object to another or a few objects in the prompt text are missing in the generated image [1]. This work replaces the embedding of each noun chunk in the prompt with the independent embedding of that very noun chunk. In this way, the key and value tensor are changed and the attribute binding of the cross-attention map has become tighter. Another method named SynGen also noticed this kind of flaw of diffusion model. SynGen introduces a new loss function to encourage the cross-attention maps of semantically related words to be similar [9].

This paper refers to the some of the methods in the above papers. Specifically, this paper refers to the DiffSegmenter method for zero-shot open semantic segmentation and strengthens the attribute binding of the diffusion model by transferring and improving the Structured Diffusion Guidance method to achieve better segmentation accuracy.

DOI: 10.54254/2755-2721/86/20241585

3. Methodology

3.1. Zero-shot open-vocabulary segmentation

The work in this paper mainly bases on DiffSegmenter (Figure 1). On the image side, the input image is simple encoded with a VAE encoder and passed to diffusion model. On the text side, the target object and the image are given to a BILP model, which outputs a prompt describing the target object in the image with detail. Through the procession on the text side, diffusion model can have more information about the target object and segment better. After the encoded image and text embedding are passed to the diffusion model, all the attention maps during each diffusion step are collected. Finally, DiffSegmenter generates the initial score map by aggregating cross-attention map during diffusion steps, completes the score map with self-attention map and achieves the segmentation result according to the score map. That is the basic construction of zero-shot open-vocabulary segmentation based on DiffSegmenter.

3.2. Noun chunk replacement

Noun chunk replacement was first applied in the Structured Diffusion Guidance method. Initially, the noun chunk replacement was used to solve the flaw of the built-in CLIP text encoder of diffusion model. For instance, given the noun chunks "red orange and green strawberry", diffusion model always prefer to generate red strawberry than the target green one. Researchers believe that rare combination of attribute and object are more likely to be mis-bind and covered by more common combination like the "red strawberry". Structured Diffusion Guidance pointed out that this kind of mistake are probably caused by the CLIP text encoder [1]. To solve that problem and enhance attribute binding, the noun chunk replacement was applied. Though noun chunk replacement was meant to be use in generation task, the author transferred noun chunk replacement to segmentation task.

Given the detailed description of the target object generated by BILP, the first step is to put that prompt P into the built-in CLIP text encoder of diffusion model and get a sequence of embeddings. Here, E is the complete encoding of the entire prompt, and E_i is the embedding of the token at position i in the prompt. The original embedding of P looks like this:

$$E = [E_1, E_2, \dots, E_i, \dots, E_n],\ i = 1,2, \dots, n = len(P) \tag{1}$$

Next for segmentation, the noun chunk containing the target object should be extracted from the prompt. Through calling the NLP model spaCy, the prompt can be parsed into a syntax tree. From the syntax tree, the noun chunk containing the target object can be extracted. Suppose that the position of the tokens of the noun chunk containing the target object is a list called POS, the noun chunk will be encoded independently into embedding R:

$$R = \left[R_i, R_j, R_k, \dots\right],\ i, j, k, \dots\ in\ POS \tag{2}$$

Then the position of the noun chunk in E will be replaced by the independent embedding of the noun chunk R. The result of the replacement will be the embedding $\overline{E}$:

$$\overline{E} = \left[E_1, E_2, \dots, E_m, \dots, R_i, R_j, R_k, \dots, E_n\right],\ m = 1,2, \dots, n = len(P),\ i, j, k, \dots\ in\ POS \tag{3}$$

3.3. Unconditioned embedding averaging

Though noun chunk replacement is probably enough in generation tasks to enhance attribute binding, noun chunk replacement only does not make obvious improvement in the enhancement of attribute binding and segmentation result. This is most likely because the model pays too much attention to the target object, resulting in the modification of the target object's attributes having little effect on the segmentation result. Therefore, it is probably necessary to appropriately reduce the impact of the target object embedding in the final embedding. Through averaging the embedding of the token of the target object with unconditioned text embedding at a certain ratio, we can increase the weight of the modification to the object in a proper extent, so that the attribute binding will be further strengthened.

DOI: 10.54254/2755-2721/86/20241585

Here, the unconditioned embedding is supposed as U. The ratio of the mixed average of the unconditional text encoding and the target word text embedding is assumed to be α. The position of the token related with the target object is q. The updated text embedding of the target object is E_q and looks as follow:

$$\overline{E_q} = \alpha \times R_q + (1 - \alpha) \times \mathrm{U} \tag{4}$$

And if we replace the text embedding of the target object in $\overline{E}$ with $\overline{E_q}$, the final embedding of the prompt would be:

$$\overline{E} = [E_1, E_2, \ldots, E_m, \ldots, R_i, R_j, \overline{E_q}, R_k, \ldots, E_n],\ m = 1,2, \ldots, n = len(P),\ i, j, q, k, \ldots\ in\ POS \tag{5}$$

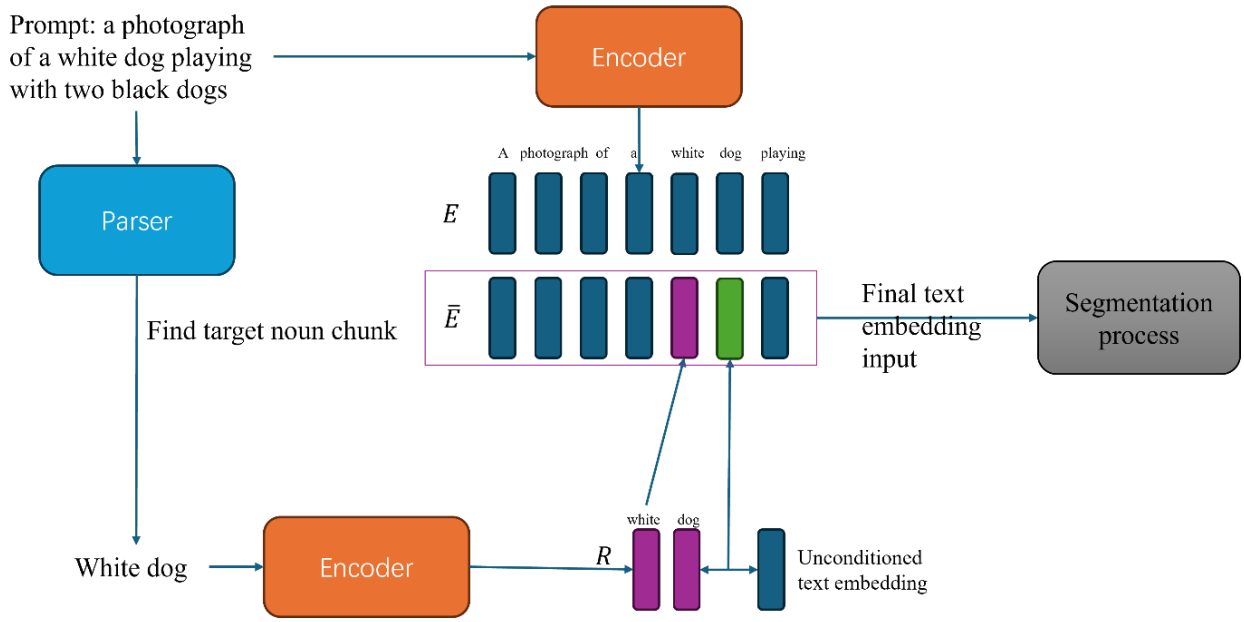


Figure 1. Overall architecture diagram (Photo/Picture credit : Original)

4. Experiments

4.1. Datasets

To evaluate the performance of the improved model and verify the feasibility of the methodology, the author tested the model on two datasets, COCO and RefCOCO. Specifically, the author used Microsoft Common Objects in Context 2014 version [10]. This dataset contains 80 categories of objects. There exist two versions of this dataset based on different split. One version has train, val and test split containing 83k, 41k, 41k images separately. The other version is the version the author chooses, which is only split into a validation set containing 50k images and the rest images for training. Since the method of this work is zero-shot, the author only used the 50k validation set to evaluate the model's ability with common image segmentation task. However, common dataset dataset like COCO only has one word to refer to the target object. Under that circumstance, the improved model in this work only performs the same as the base line, since there are no attributes to describe the object and assist segmentation. The main improvement of this work requires a dataset with a detailed description of the attributes of the target object to be reflected. Therefore, the author also evaluated the model with the referring expression generation RefCOCO. RefCOCO is a dataset extracted from COCO and has referring expression labeled for all the targets [11]. To be specific, there are 50k objects and 19,994 images contained in RefCOCO, labeled with 142,209 refer expressions. With RefCOCO, the feasibility of the methodology and whether there really is any improvement made can finally be validated.

DOI: 10.54254/2755-2721/86/20241585

4.2. Platform

kernel: Linux vcg42 5.4.0-182-generic
OS: Ubuntu 20.04.3 LTS
CPU: Intel(R) Xeon(R) Silver 4116 CPU @ 2.10GHz x 2
Mem: 251Gi
storage: 110T
GPU: NVIDIA GeForce RTX 4090 x 7

4.3. Non referring expression segmentation

When segmenting objects without referring expression, the method in this work reverts to its baseline, which basically operates and performs the same as the method the author based on, the DiffSegmenter. We can tell from table 1 that the baseline already outperforms the other methods on COCO dataset, whether they are completely training free or not.

Table 1. Segmentation results on COCO dataset[8]

Method	mIOU
ReCo(Shin, Xie, and Albanie 2022)	15.7
MaskCLIP(Zhou, Loy, and Dai 2022)	20.6
TCL(Cha, Mun, and Roh 2023)	30.4
CLIPpy(Ranasinghe et al. 2022)	32.0
ViewCo(Ren et al. 2023)	23.5
SegCLIP(Luo et al. 2023)	26.5
OVSegmentor(Xu et al. 2023a)	25.1
OVDiff(Karazija et al. 2023a)	34.8
Baseline(DiffSegmenter)	37.9
Ours	37.9

4.4. Referring expression segmentation

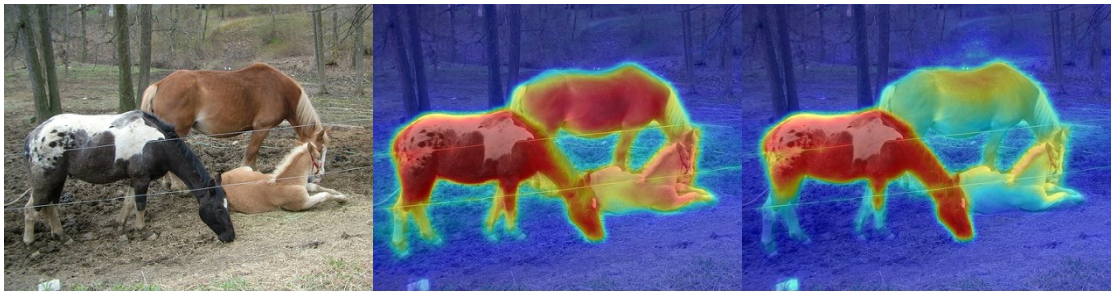
With RefCOCO dataset and its referring expression for target objects, we can finally evaluate the improvement upon the baseline. From table 2, we can tell that the model preprocessed with noun chunk replacement and unconditioned embedding averaging has better performance across all thresholds. The mIOU has increased 0.6% compared to the baseline. It seems that the overall improvement is not obvious. It appears that he attribute binding method increases the accuracy obviously for part of the pictures, while does not make much difference for part of the others and make things worse for a few pictures.

To visually express the effect of the improved method, some examples of improvement is given (Figure 2, Figure 3, Figure 4, Figure 5 and Figure 6). As shown in figure 2, after applying the noun chunk replacement and unconditioned embedding averaging, the segmentation result is much better. Without noun chunk replacement and unconditioned embedding averaging, though the prompt is "a photograph of a black and white horse", the segmentation result of DiffSegmenter, the baseline, still includes the brown horses. However, after the improvement of attribute binding enhancement, the model basically did not classify the brown dog into the segmentation results. This shows that with attribute binding enhancement, the segmentation result of the model would be more accurate and more precise.

Table 2. Segmentation results on RefCOCO dataset

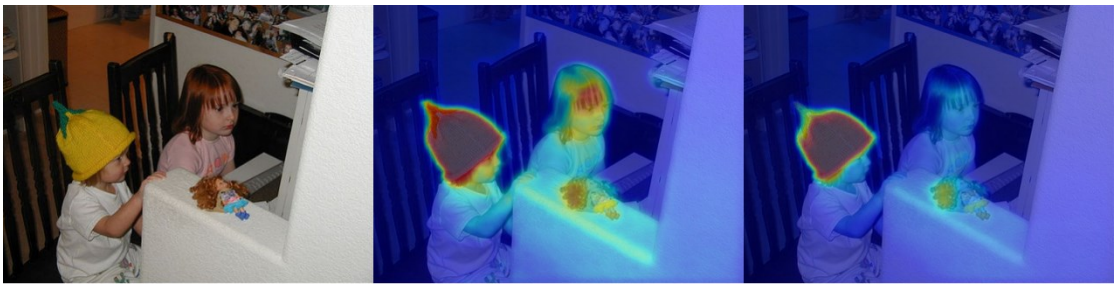
	threshold	mIOU
Baseline	0.5	22.4
Ours	0.5	23.0

DOI: 10.54254/2755-2721/86/20241585



Origin image DiffSegmenter Ours

Figure 2. heat map of segmentation result with prompt "a photograph of a black and white horse"(Photo/Picture credit : Original)

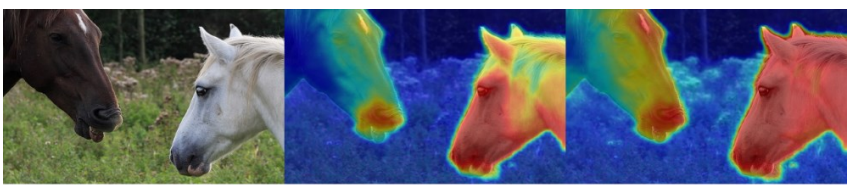


Origin image DiffSegmenter Ours

Figure 3. heat map of segmentation result with prompt "a photograph of yellow hat on a baby's head" (Photo/Picture credit : Original)

Origin image DiffSegmenter Ours

Figure 4. heat map of segmentation result with prompt "a photograph of children in red" (Photo/Picture credit : Original)



Origin image DiffSegmenter Ours

Figure 5. heat map of segmentation result with prompt "a photograph of white horse"(Photo/Picture credit : Original)

DOI: 10.54254/2755-2721/86/20241585

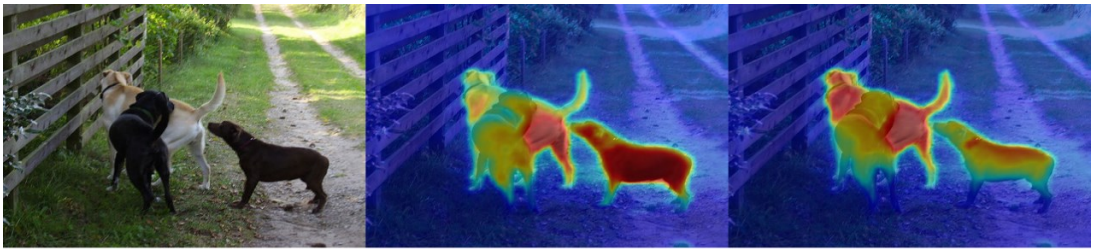


Figure 6. heat map of segmentation result with prompt "a photograph of white horse" (Photo/Picture credit : Original)

4.5. Hyper parameter tuning

The new hyper parameter the author introduced in this work is the α during the unconditioned embedding averaging, which adjusts how much unconditioned text embedding should be combined to the embedding of the word corresponding to the target object. The larger α is, the less unconditioned text embedding will be added. As we can tell from table 3 and Figure 7, when $\alpha = 0.9$, the best result is achieved and the performance drops rapidly as α decreases. It's not hard to understand why this would happen. As explained in the methodology, the enhancement of attribute binding is achieved by unconditioned embedding averaging, which indirectly reduces the weight of the target word in the prompt. So as to say, when α becomes excessively low, the meaning of the target word would partly disappear from the prompt. No wonder the performance of the model drops as α decreases. The unconditioned embedding averaging only takes effect when the weight of the target word in the prompt is reduced to a proper extent that the modifiers become more influential in the prompt and the attention to the target word is still high enough.

Table 3. Segmentation result with different α

α	threshold	mIOU
0.1	0.5	17.6
0.3	0.5	19.1
0.5	0.55	20.2
0.7	0.5	21.9
0.8	0.5	22.5
0.85	0.5	22.6
0.9	0.5	23.0
0.95	0.5	22.8
1	0.55	22.6

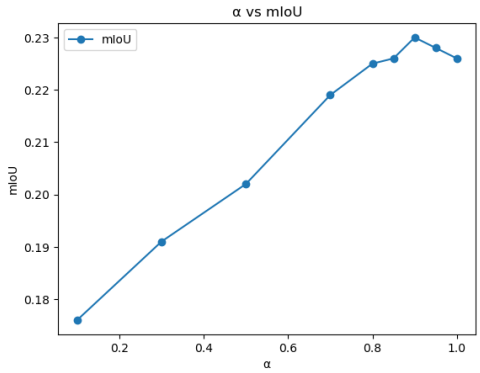


Figure 7. Line chart of segmentation result with different α (Photo/Picture credit : Original)

DOI: 10.54254/2755-2721/86/20241585

4.6. Ablation Studies and Analyses

As shown in table 4 the noun chunk replacement improved the mIOU by 0.2% percent from 22.4% to 22.6%. And the unconditioned embedding averaging made another 0.4% improvement upon the noun chunk replacement and increased the mIOU to 23.0%.

From Figure 8 we can visually feel the effects of different improvements. With the prompt as “a photograph of white dog playing with two black dogs”, the baseline only segments a part of the white dog and includes one of the black dogs into the segmentation result. With the same prompt and the noun chunk replacement improvement, the segmentation result of the white dog becomes better, counting the head of the white dog into the segmentation result. However, with noun chunk replacement improvement, another black dog is still included in the segmentation result. With unconditioned embedding averaging added, the segmentation result of the white dog is much more precise and the black dog is barely included in the segmentation result.

From table 4, we can tell that the main improvement is made by unconditioned embedding averaging. This leads to the conclusion that compared to the mis-binding or missing of attribute, the main problem is that the target word is too influential in the prompt which covers the meaning of the attributes modifying it. That's why unconditioned embedding averaging takes more effect in the attribute binding enhancement and performance improvement.

However, this also leads to the limitation of this work. Since unconditioned embedding averaging indirectly reduces the weight of the target word in the prompt, the segmentation result of the baseline determines the effect of attribute binding enhancement. If the segmentation result of the baseline is bad in the first place, for instance recognizing unrelated objects as the target, the improvement method based on attribute binding enhancement will only make things worse by reducing the weight of the target word in the prompt. Unconditioned embedding averaging is mainly suitable for circumstances that the model has basically recognized the target object, but the segmentation result has additional parts that do not exist in the prompt or is missing parts required by the prompt. Under this circumstance, unconditioned embedding averaging can help the model "understand" the prompt more and give a more precise segmentation.

Table 4. Ablations study

	threshold	mIOU
Baseline	0.5	22.4
Noun chunk replacement	0.55	22.6
Unconditioned embedding averaging	0.5	23.0

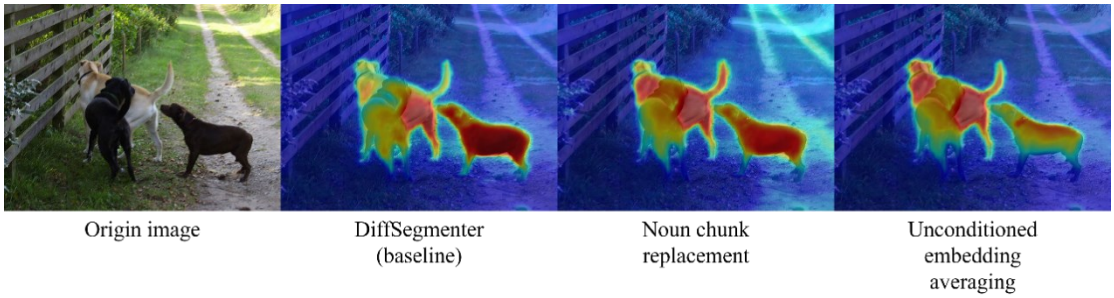


Figure 8. heat map of segmentation result with prompt “a photograph of white dog playing with two black dogs”(Photo/Picture credit : Original)

5. Conclusion

This paper presents an attribute enhancing method to improve the performance of diffusion model based image segmentation model when dealing with referring expression. The author transferred the attribute binding enhancing method applied in diffusion model generation tasks called noun chunk replacement and came up with a new attribute binding enhancing method called unconditioned embedding averaging.

DOI: 10.54254/2755-2721/86/20241585

Through experiments on COCO dataset and RefCOCO dataset, it is proved that with noun chunk replacement and unconditioned embedding averaging, the attribute binding in prompt embedding can be enhanced if the prompt contains descriptive attributes about the target object. As the development of artificial intelligence and big models, artificial intelligence is integrating into people's lives. A image segmentation model that can accept referring expression and natural language as input and better 'understand' the prompt would probably be a demand of the future. The limitation of this work is that the effect of attribute bind enhancing method is restricted by the performance of the basic image segmentation method. If the segmentation method of the baseline is bad, noun chunk replacement and unconditioned embedding averaging can barely make any improvement. In the future, as the performance of image segmentation models generally rise, the attribute bind enhancing method is likely to become useful.

References

[1] Feng W, He X, Fu T J, et al. 2022 Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv preprint arXiv:2212.05032.

[2] Hertz A, Mokady R, Tenenbaum J, et al. 2022, Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626.

[3] Wu W, Zhao Y, Shou M Z, et al. 2023, Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models, Proceedings of the IEEE/CVF International Conference on Computer Vision: 1206-1217.

[4] Karazija L, Laina I, Vedaldi A, et al. 2023, Diffusion models for zero-shot open-vocabulary segmentation, arXiv preprint arXiv:2306.09316.

[5] Xu J, Liu S, Vahdat A, et al. 2023, Open-vocabulary panoptic segmentation with text-to-image diffusion models, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.: 2955-2966.

[6] Amit T, Shaharbany T, Nachmani E, et al. 2021, Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390.

[7] Tian J, Aggarwal L, Colaco A, et al. 2024, Diffuse Attend and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 3554-3563.

[8] Wang J, Li X, Zhang J, et al. 2023, Diffusion model is secretly a training-free open vocabulary semantic segmenter. arXiv preprint arXiv:2309.02773.

[9] Rassin R, Hirsch E, Glickman D, et al. 2024, Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. Advances in Neural Information Processing Systems, 36.

[10] Lin T Y, Maire M, Belongie S, et al. 2014, Microsoft coco: Common objects in context, Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing: 740-755.

[11] Kazemzadeh S, Ordonez V, Matten M, et al.2014, Referitgame: Referring to objects in photographs of natural scenes, Proceedings of the 2014 conference on empirical methods in natural language processing. 787-798.