InvSeg: Test-Time Prompt Inversion for Semantic Segmentation

Jiayi Lin[1], **Jiabo Huang**[2], **Jian Hu**[1], **Shaogang Gong**[1]

[1]Queen Mary University of London
[2]Sony AI
{jiayi.lin, jian.hu, s.gong}@qmul.ac.uk, raymond.huang@sony.com

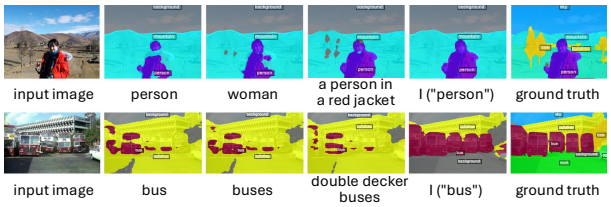
Abstract

Visual-textual correlations in the attention maps derived from text-to-image diffusion models are proven beneficial to dense visual prediction tasks, e.g., semantic segmentation. However, a significant challenge arises due to the input distributional discrepancy between the context-rich sentences used for image generation and the isolated class names typically used in semantic segmentation. This discrepancy hinders diffusion models from capturing accurate visual-textual correlations. To solve this, we propose InvSeg, a test-time prompt inversion method that tackles open-vocabulary semantic segmentation by inverting image-specific visual context into text prompt embedding space, leveraging structure information derived from the diffusion model's reconstruction process to enrich text prompts so as to associate each class with a structure-consistent mask. Specifically, we introduce Contrastive Soft Clustering (CSC) to align derived masks with the image's structure information, softly selecting anchors for each class and calculating weighted distances to push inner-class pixels closer while separating inter-class pixels, thereby ensuring mask distinction and internal consistency. By incorporating sample-specific context, InvSeg learns context-rich text prompts in embedding space and achieves accurate semantic alignment across modalities. Experiments show that InvSeg achieves state-of-the-art performance on the PASCAL VOC, PASCAL Context and COCO Object datasets.

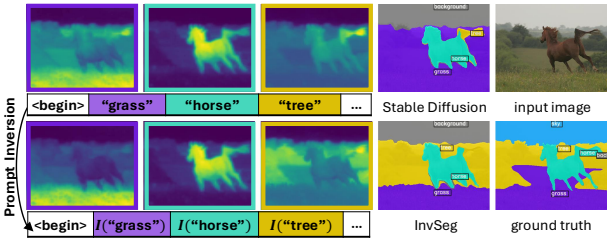
Project Page — https://jylin8100.github.io/InvSegProject

Introduction

Open-Vocabulary Semantic Segmentation (OVSS) aims to divide an image into several semantically consistent regions, corresponding to label names in a large, unrestricted vocabulary. Recently, thanks to visual-textual correlation ability learned from large scale image-text pairs, stable diffusion models (Rombach et al. 2022) have shown substantial potential in open vocabulary semantic segmentation. Some approaches employ the stable diffusion model for generating pseudo labels (Wu et al. 2023; Wang et al. 2024; Li et al. 2023; Xiao et al. 2023) or use it as a feature extractor for further training a segmenter (Xu et al. 2023b). However, among these approaches, the process of collecting per-pixel labels is costly and the performance can be limited when generalizing to new scenarios under distribution shifts. Therefore, to eliminate the need for expensive pixel-level annotation and achieve better generalization ability, several unsupervised approaches (Tian et al. 2023; Karazija et al. 2023; Wang et al. 2023) are proposed and have achieved notable results.



(a) Comparison of segmentation results for "person" or "bus" using text prompts with different levels of image-specific details.



(b) Overall process of InvSeg. We exploit the structural information from image to enrich the original isolated class names (top) into image-specific text prompt which derives more accurate segmentation results (bottom).

Figure 1: Motivation of test-time prompt inversion.

However, most current unsupervised approaches (Tian et al. 2023; Karazija et al. 2023) solely use isolated class names as the text prompt for diffusion models, which ignores the distributional discrepancy between the isolated class names for visual perception and descriptive, context-rich sentences used for image generation. This discrepancy in input text richness can hinder the performance of diffusion models in semantic segmentation, where precise image-text correlations is required. Although some approaches (Wang et al. 2023) enrich the class name by combining it with other descriptive text of current class, this process is cumbersome and may not always derive to satisfying output. As in our preliminary study (Fig. 1a), we evaluate text prompts with varying levels of sample-specific information: original class name, sample-

specific single word, sample-specific word group, where the latter two are extracted from the captions generated by a vision language model (VLM) (Li et al. 2022). We observe that containing more sample-specific information can enhance segmentation performance. However, in the second example, we observe that increasing sample-specific information does not necessarily improve segmentation results. This inconsistency shows the complexity of searching optimal text prompts for segmentation, which requires searching a carefully crafted combination of different words.

To avoid the complexity of searching multiple words for segmenting single class, we propose to optimize the original class name in text embedding space by extracting sample-specific visual context. We propose InvSeg, a test-time prompt inversion method, which inverts a single test image into the model's text prompt embedding space, where each text token captures a distinct visual concept in the image. Specifically, we introduce Contrastive Soft Clustering (CSC) to align the predicted masks with inherent image structure. CSC constructs a 4D distance matrix that measures pairwise distances between spatial points in an image, with each spatial point having its corresponding distance map illustrating its distance to other points. In each predicted score map corresponding a specific class, some high-confidence points are selected as the anchor points and their distance maps is used to aligned with current predicted mask. This alignment process involves utilizing each anchor point's distance map to modulate the probability distribution across classes. Particularly, we leverage the distance information to determine the extent to which we decrease the probability of the current class and increase the probabilities of other classes for each spatial point. By incorporating sample-specific context at test time, InvSeg learns a context-rich text prompt to achieve precise semantic alignment across modalities.

We benchmark InvSeg on three segmentation datasets PASCAL VOC 2012 (Everingham et al. 2010), PASCAL Context (Mottaghi et al. 2014) and COCO Object (Lin et al. 2014) for OVSS task. InvSeg outperforms prior works on two datasets despite requiring less auxiliary information. In summary, our contributions are two-fold:

- To the best of our knowledge, we are the first to perform unsupervised region-level prompt inversion on diffusion models. This enables the decomposition of any image into semantic parts for visual perception tasks.
- With our proposed unsupervised constraint, we are able to obtained optimized image-specific text prompt to generate more complete and accurate attention maps and derive superior segmentation results comparing to previous unsupervised methods.

Related Work

Pre-trained Generative Models for Segmentation

Pre-trained Generative Models like Generative Adversarial Nets (Goodfellow et al. 2014; Tritrong, Rewatbowornwong, and Suwajanakorn 2021; Xu and Zheng 2021) and Stable Diffusion Models (Rombach et al. 2022) have been widely used in dense prediction like segmentation for its ability to capture fine-grained location and shape information of objects in images. Recently, there are increasing interest in exploiting diffusion models for segmentation (Wu et al. 2023; Wang et al. 2024; Li et al. 2023; Xiao et al. 2023; Wu et al. 2023). Some approaches use the model internal features to train a segmentation branch, other use diffusion models to generate new segmentation data to train a segmenter (Wu et al. 2023). However, these models require expensive per-pixel labels or labeling process, which tend to suffer from the category imbalance problem. In contrast, current unsupervised methods use attention maps from pre-trained denoising U-Nets in diffusion models. Specifically, DiffSeg (Tian et al. 2023) uses self-attention maps to get pairwise pixel similarity to cluster pixels iteratively, which however the output mask does not align to semantic category, and requires manual design of the number of clusters. DiffSegmenter (Wang et al. 2023) uses the cross-attention maps to initial the mask and self-attention maps to complete the mask for each given category. However, the text embedding of the same category in these methods remain the same for different images, while we instead propose to customize the text embedding to get more complete and accurate segmentation masks for each image.

Diffusion-based Inversion

Diffusion-based Inversion is inspired from GAN inversion (Bojanowski et al. 2017) to synthesize visual concept in the given reference images for further image editing. Specifically, DreamBooth (Ruiz et al. 2023) fine-tunes the diffusion model to learn to bind a unique identifier with that specific subject. In contrast, some other inversion approachs (Mokady et al. 2023; Gal et al. 2022) keep the model weights fixed to avoid damaging the knowledge of the pre-trained model and cumbersome tuning process. Textual Inversion (Gal et al. 2022) inverts visual concepts in the text prompt embedding space and encode the visual features as new tokens in the vocabulary. Similar in spirit, we introduce to invert new tokens for regions in an image for category segmentation. The only work that also try to learn region level prompt is SLIME (Khani et al. 2023), which uses mask annotation of few sample to learn region-level prompt and transfer to similar images. However, this method needs annotation data and requires the test images to be similar to the training image, which greatly inhibits the model's transferability to broader diverse images. Therefore, the method is only tested on a few part segmentation datasets like horse, face and car part segmentation. In contrast, InvSeg can be applied for more diverse images in an unsupervised way.

Test Time Prompt Learning

Test Time Prompt Learning (Shu et al. 2022; Hu et al. 2024; Abdul Samadh et al. 2024) is a subtopic of Test Time Adaptation (Sun et al. 2020), which narrows the distribution gap between the training and test data during test time. Test-time Prompt Tuning (TPT) (Shu et al. 2022) learns adaptive prompts of a large VLM CLIP (Radford et al. 2021) for one test sample, by forcing consistent predictions across different augmented views of each test sample with an entropy minimization objective. PromptAlign (Abdul Samadh et al. 2024) further introduce a distribution alignment objective for CLIP and explicitly aligns the train and test sample distributions

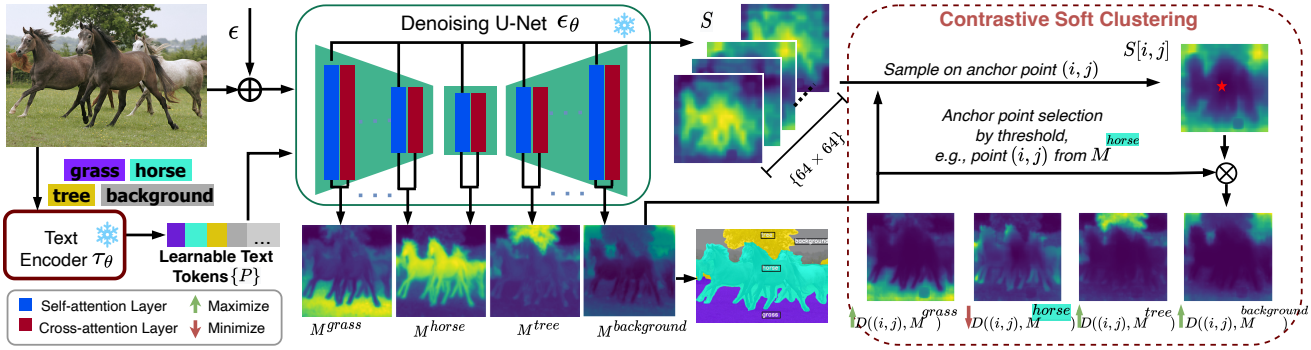


Figure 2: Overview of InvSeg framework. Our proposed Contrastive Soft Clustering method can achieve region-level prompt inversion. The text tokens are first initialized with the pretrained text encoder from the diffusion model (dashed box on left) and then are used as the only learnable parameters during the test time training. After the adaption process, the learned text tokens can be used to derive a more accurate and complete refined attention maps $\{M\}$ for segmentation.

to mitigate the domain shift. We exploit the entropy minimization constraint during the optimization of text prompt, to stabilize the adaptation process.

Methodology

InvSeg aims at inverting the image-specific fine-grained text prompts from a given test image using pretrained diffusion model, where the inverted prompts are further used to produce high quality segmentation masks. We first exploit the original text prompts (category names) to generate segmentation masks with diffusion models, which provide initial segmentation masks for further adaptation. Then we introduce use Contrastive Soft Clustering to achieve region-level prompt inversion in an unsupervised way, which helps segment complete and disjoint masks. Finally, we stabilize the adaption process with Entropy Minimization.

Given a test image $I \in \mathbb{R}^{h \times w \times 3}$ and a list of categories from the test dataset. We first use Vision Language Model (Li et al. 2022) to filter out the C categories (including one category as "background") in current image following previous works (Tian et al. 2023; Wang et al. 2023). Then the filtered category are directly combined as the text prompt input.

Diffusion Models

The text-conditioned diffusion model models a data distribution conditioned a natural language text prompt. The text prompt y is first tokenized and then encoded by a text encoder τ_θ to obtain text token embeddings $\tau_\theta(y) = \{P_0^k\}$, where k is the index of each token. On the other hand, the image $I \in \mathbb{R}^{h \times w \times 3}$ is first encoded to auto-encoder latent space, and then added with a standard Gaussian noise ϵ for t time step to obtain $\mathcal{I}_t$. Finally, the objective is to predict the added noise with diffusion model noted as ϵ_θ, where the text encoder τ_θ and diffusion model ϵ_θ are optimized simultaneously:

$$L_{LDM} = \mathbb{E}_{\mathcal{I},y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(\mathcal{I}_t, t, \tau_\theta(y)\right)\|_2^2\right] \quad (1)$$

In InvSeg, we keep the parameters of the models τ_θ, ϵ_θ fixed, but optimize $\{P_0^k\}$ directly to obtain an image-specific prompt $\{P_I^k\}$ for image I.

Attention Map Generation in Diffusion Models. Diffusion model employs a UNet structure of four resolutions $\{8, 16, 32, 64\}$, where the attention modules (Vaswani et al. 2017) are applied on each resolution for multiple times. These attention modules include self-attention and cross-attention. Self-attention captures pixel-level affinities within the image, while cross-attention captures the relationship between the text tokens and image pixels in embedding space. For each attention module, there are three components: query Q, key K, and value V of dimension d. The output of attention is $O = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \cdot V$.

Here we use a normalized attention map $S = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)$ for further computation following previous work (Wang et al. 2023), which shows the similarity between query Q and key K. Therefore, in a self-attention module, we obtain S^{self} that capture pixel-level similarities in the image embedding space. Similarly, in a cross-attention module, S^{corss} is obtained to measure the similarities between each text token and image pixels. We further aggregate

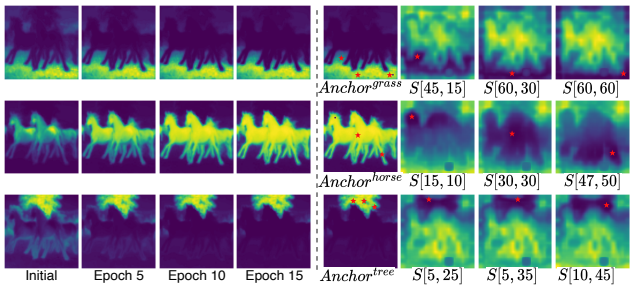


Figure 3: Illustration of the soft selection (with probability) of anchor points for each category c: $Anchor^c$ during different optimization steps (left) and the distance matrix S on certain anchor points (right). On the right sub-figure, we sample 3 anchor points for each category, showing the distance from each anchor point to other pixels in the image. Darker areas represent smaller distances (higher similarity) to the anchor.

the all the $\{S_l^{self}\}$ from different self-attention modules l respectively for further computation, which is also performed on $\{S_l^{cross}\}$ following (Vaswani et al. 2017). Taking $\{S_l^{self}\}$ as an example, we first average those with the same resolution and interpolate them to the resolution of 64×64. Then we use a weighted sum among all the resolutions to obtain a final self-attention map $\mathcal{A}^{self} \in \mathbb{R}^{HW \times HW}$, where $H = W = 64$. In a similar way, we obtain $\mathcal{A}^{cross} \in \mathbb{R}^{HW \times K}$, where K is the number of tokens. Finally, we use a mask refinement mechanism in DiffSegmenter (Vaswani et al. 2017) which uses $\{S_l^{self}\}$ to incomplete $\mathcal{A}^{cross}$ to get a refined cross-attention map:

$$M = \text{norm}\left(\mathcal{A}^{self} \cdot \mathcal{A}^{cross}\right), \tag{2}$$

where $\text{norm}(\cdot)$ is min-max normalization. Therefore, for each token k, we have a corresponding refined cross-attention map M^k. Furthermore, for each category, we use notation c as the index. Therefore, we can further obtain a set of refined cross-attention map M^c from M^k.

Contrastive Soft Clustering

As the refined cross-attention map M^c derived directly from initial text token embeddings P_0^c can only locate part of the object (Fig. 1b) and tend to be distracted by salient category when locating background category (Fig. 1a), which leads to unsatisfying segmentation results. We propose Contrastive Soft Clustering to achieve region-level inversion for each text token (category) to generate complete and disjoint attention map for each category by exploiting the structure information in the pretrained diffusion models.

Specifically, after the refined cross-attention maps M^c are obtained, we further optimize P^c by ensuring that its corresponding M^c is aligned with the objectness information indicating in self-attention maps. As illustrated in previous works (Tian et al. 2023; Wang et al. 2023), the self-attention maps contain specific structural information of object shape and location. Based on this observation, we build a 4D distance matrix S to measure the distance between each pair of pixels in the refined cross-attention map M^c. We measure the distance between two attention maps using KL divergence following (Tian et al. 2023). Therefore, the distance between pixel (i, j) and pixel (k, l) is:

$$\begin{aligned}\mathcal{S}[i,j,k,l] = &\text{KL}\left(\mathcal{A}^{\text{self}}[i,j] \| \mathcal{A}^{\text{self}}[k,l]\right) \\ &+ \text{KL}\left(\mathcal{A}^{\text{self}}[k,l] \| \mathcal{A}^{\text{self}}[i,j]\right)\end{aligned} \tag{3}$$

Based on the distance matrix S and M^c, we can measure the weighted distance between pixel (i, j) and a group of weighted pixels in M^c as follows:

$$D((i,j), M^c) = \frac{\sum_{(k,l)\in Q^c} \left(\mathcal{S}[i,j,k,l] \cdot M^c[k,l]\right)}{\sum_{(k,l)\in Q^c} M^c[k,l]} \tag{4}$$

where Q^c is the set of all the pixels in M^c.

To measure the distance of all the pixels in M^c, we use sigmoid to binarize the value in M^c, which gets a soft selection of high confident pixels corresponding to the current category c. We denote this soft selected set as $Anchor^c \in \mathbb{R}^{H \times W}$, where the pixels with value close to 1 is used as an anchor of category c. Therefore, the distance within M^c:

$$\begin{aligned}&D(Anchor^c, M^c) = \\ &\qquad \frac{\sum_{(i,j)\in Q^c} \left(Anchor^c[i,j] \cdot D((i,j), M^c)\right)}{\sum_{(i,j)\in Q^c} Anchor^c[i,j]}.\end{aligned} \tag{5}$$

Similarly, we can easily calculate the distance between different categories c and c', that is $D(Anchor^c, M^{c'})$. Therefore, for all C classes, the total distance within each class is:

$$D_{intra} = \sum_{c=1}^{C} D(Anchor^c, M^c) \tag{6}$$

and the total distance of all pairs of two different classes is:

$$D_{inter} = \sum_{c'=1}^{C-1} \sum_{c=c'+1}^{C} D(Anchor^c, M^{c'}) \tag{7}$$

In this way, we are able to perform a contrastive soft clustering by minimizing the distance within each category and maximizing the distance among different categories. The loss function of contrastive soft clustering is denoted as:

$$\mathcal{L}_{Cluster} = \frac{D_{intra}}{C} - \frac{2 * D_{inter}}{C * (C-1)} \tag{8}$$

Stablizing Adaptation Process

Entropy Minimization. Inspired by TPT (Shu et al. 2022), we augment the input image into different views, and try to encourage consistent predictions across views. To do this, we use different augmentation functions Aug_i on the input image, and get the corresponding attention maps $M^c_{Aug_i}$. Then we reverse the augmentation of align pixels between different maps $M^c_{Aug_i}$ to get $M\prime^c_{Aug_i}$. Further, we average the $M\prime^c_{Aug_i}$ over different augmented view to get $\overline{M}^c$. Finally, we minimize the entropy of $\overline{M}^c$ for all categories: $\mathcal{L}_{Etrp} = -\sum_{c=1}^{C} \overline{M}^c \cdot \log \overline{M}^c$.

Overall Optimization Process. For each input image, we optimize the text embeddings with 15 steps, using a loss function $\mathcal{L} = \mathcal{L}_{Cluster} + \alpha * \mathcal{L}_{Etrp}$, where α is the coefficient of the loss functions. Finally, we generate a segmentation mask by performing an argmax M^c across differnet categories c and interpolate it to match the original size of input image.

Experiments

Datasets and Metrics. We evaluate InvSeg on three commonly used benchmarks, namely, PASCAL VOC 2012 (Everingham et al. 2010), PASCAL Context (Mottaghi et al. 2014) and COCO Object (Lin et al. 2014), containing 20,59,80 foreground classes, respectively. The experiments are performed on the validation sets, including 1449, 5105, and 5000 images. Following prior works (Tian et al. 2023; Karazija et al. 2023; Wang et al. 2023), we use mean intersection over union (mIoU) to measure segmentation performance. We also use the metric mean accuracy (mAcc) in our ablation study.

Methods	Training dataset	mIOU		
		PASCAL VOC	PASCAL Context	COCO Object
DeiT (Touvron et al. 2021)	IN-1K	53.0	35.9	-
MoCo (He et al. 2020)	IN-1K	34.3	21.3	-
DINO (Caron et al. 2021)	IN-1K	39.1	20.4	-
ViL-Seg (Liu et al. 2022)	CC12M	33.6	15.9	-
MaskCLIP (Zhou, Loy, and Dai 2022)	LAION	38.8	23.6	20.6
GroupViT (Xu et al. 2022)	CC12M	52.3	22.4	-
ZeroSeg (Chen et al. 2023)	IN-1K	40.8	20.4	20.2
TCL (Cha, Mun, and Roh 2023)	CC3M+CC12M	51.2	24.3	30.4
ViewCo (Ren et al. 2023)	CC12M+YFCC	52.4	23.0	23.5
CLIPpy (Ranasinghe et al. 2022)	HQITP-134M	52.2	-	32.0
SegCLIP (Luo et al. 2023)	CC3M+COCO	52.6	24.7	26.5
OVSegmentor (Xu et al. 2023a)	CC4M	53.8	20.4	25.1
SimSeg (Yi et al. 2023)	CC3M+CC12M	57.4	26.2	29.7
Generative models based				
DiffSegmenter (Wang et al. 2023)	-	<u>60.1</u>	<u>27.5</u>	**37.9**
Diffusion Baseline⋆ (Rombach et al. 2022)	-	59.6	25.0	34.5
InvSeg (Ours)	-	**63.4**	**27.8**	<u>36.0</u>

Table 1: **Comparison with existing methods.** Models in the first three rows are finetuned on target datasets while the rest approaches do not require mask annotations. **Bold fonts** refer to the best results and <u>underline fonts</u> refer to the second best. ⋆ notes our implementation.

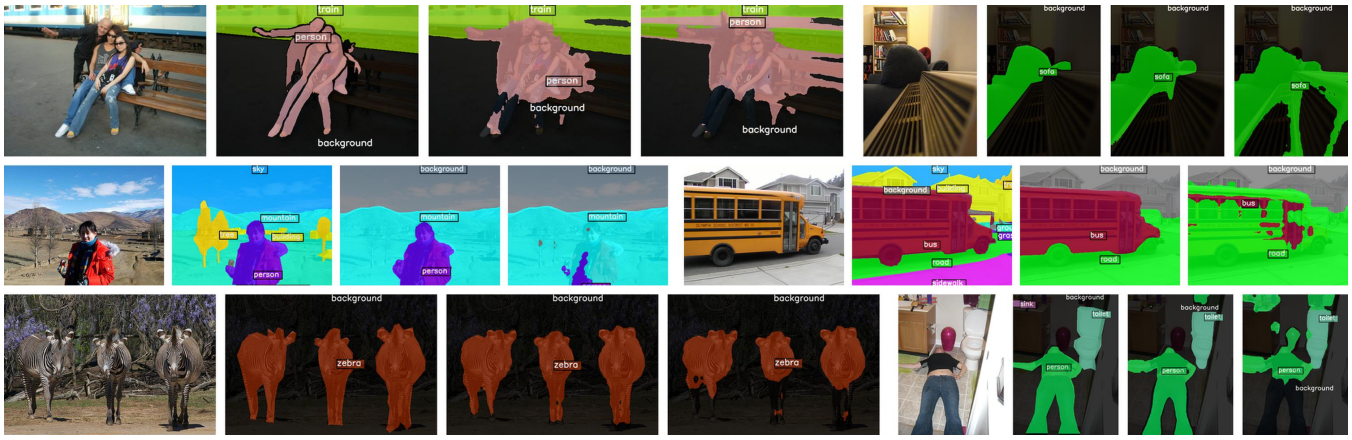


Figure 4: Examples of Segmentation on VOC (top), Context (middle) and COCO (bottom). For each sample (image group of four), from left to right is input, GT, InvSeg, Diffusion baseline.

Implementation Details. To obtain category names, we follow DiffSegmenter (Wang et al. 2023) that use BLIP (Li et al. 2022) and CLIP (Radford et al. 2021) to generate the category names out of all the candidate categories in the dataset. We validate our approach using Stable Diffusion v2.1 (Rombach et al. 2022) with frozen pre-trained parameters. Each test image is augmented for 2 times using random resized crops with minimum crop rate 0.6, which construct a batch of size 2 in each optimization step. We employ the Adam optimizer with a learning rate of 0.01. The weights for $\mathcal{L}_{Etrp}$ is $\alpha = 1$. We optimize each image for 15 steps on single H100 GPU, with the inference time of around 7.9 seconds per image (not including the category name extraction), which comparative to existing test-time prompt tuning methods like our adapted TPT (Shu et al. 2022) for segmentation with 7.2 seconds for inference. The running memory was 32.4G on GPU with full precision (32-bit floating point). As for time step in diffusion model, during adaption, the time step for each iteration is sampled from a range [5, 300] where the model can learn a more robust prompt from different time steps. While the time step for inference is 50, falling in the range during adaption.

Results on Open-vocabulary Semantic Segmentation

Tab. 1 shows a comparison between InvSeg and previous works in open-vocabulary semantic segmentation. These prior approaches can be broadly categorized into two main

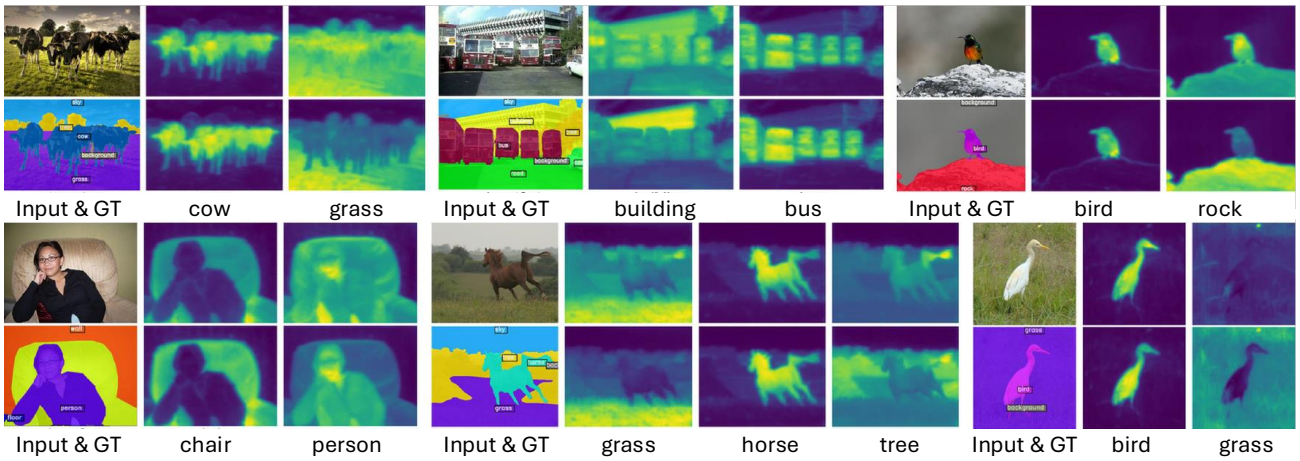


Figure 5: Visualization of refined cross-attention maps derived from text prompts before (top) and after (bottom) prompt inversion. Before prompt inversion, the segmentation of background elements such as "grass" or "trees" is influenced by foreground objects like "cow" or "horse", resulting in mistakenly ignoring background classes or segmenting foreground (and background) classes. After applying prompt inversion, this phenomenon is suppressed by improving the distinction between foreground and background through proposed Contrastive Soft Clustering.

groups: those based on discriminative models (such as CLIP (Radford et al. 2021)) and those based on generative models (stable diffusion models (Rombach et al. 2022)). Note that a direct comparison with most diffusion model-based methods would not be equitable. This is because many of these methods use extra mask annotations or synthesis mask which usually requires a pretrained segmenter, such as generating pseudo labels (Wu et al. 2023; Wang et al. 2024; Li et al. 2023; Xiao et al. 2023; Marcos-Manchón et al. 2024) or using diffusion models as a backbone to train a model in a supervised manner (Xu et al. 2023b; Zhao et al. 2023). Therefore, we have focused our comparison on a select few diffusion model-based methods that, do not require mask annotations: DiffSegmenter (Wang et al. 2023), the original Stable Diffusion model (referred to as Diffusion Baseline). We observe that InvSeg achieves competitive performance compared to previous CLIP based models. InvSeg demonstrates state-of-the-art performance on both VOC and Context datasets. When compared to the diffusion baseline that utilizes the original text prompt, InvSeg shows significant improvements in mIOU, with gains of up to 3.8% on VOC and 2.8% on Context. However, InvSeg still lags behind DiffSegmenter on the COCO dataset. This discrepancy may be caused by the fundamental difference in the segmentation strateg. DiffSegmenter employs a sequential strategy, predicting masks for each class individually through binary segmentation. In contrast, InvSeg efficiently segments all classes simultaneously, which is a more complex challenge.

Ablation Studies and Analyses

We conduct a comprehensive analysis of various parameters and strategies on InvSeg. Our experiments are carried out on two widely-used datasets: PASCAL VOC 2012 and COCO Object, including: loss weights, anchor points selection, text initialization strategies, and diffusion model parameters.

Method's Variants		PASCAL VOC		COCO Object	
CSC	α	mIOU	mAcc	mIOU	mAcc
1	0	61.9	79.4	35.0	58.2
1	0.1	62.2	79.4	35.3	57.9
1	**1**	**63.4**	79.2	**36.0**	**58.8**
1	10	62.4	77.0	35.8	56.2
0	1	61.9	**79.5**	35.7	55.8

Table 2: Ablating loss weights for $\mathcal{L}_{Cluster}$ and $\mathcal{L}_{Etrp}$.

Method's Variants	PASCAL VOC		COCO Object	
scale	mIOU	mAcc	mIOU	mAcc
2	63.1	78.4	35.9	57.3
4	**63.4**	**79.2**	**36.0**	58.8
8	63.0	**79.2**	35.8	59.8
16	62.7	79.0	35.5	**59.9**

Table 3: Ablating softness of anchor selection in InvSeg.

Loss weights. To verify the effectiveness of the objective function in prompt inversion, we conducted experiments with different loss weights. In the first row, we observe that solely performing Contrastive Soft Clustering already surpasses the current state-of-the-art DiffSegmenter by 1.8% in mIOU on VOC. Furthermore, applying the entropy minimization constraint brings more than 1% improvement in mIOU on both datasets, which illustrates its effectiveness in stabilizing the optimization process. These results demonstrate the effectiveness of our objective function, combining Contrastive Soft Clustering and entropy minimization to achieve superior segmentation performance across different datasets.

Anchor points selection. When selecting anchors from score maps, we rescale the score maps before applying sigmoid

Method's Variants	PASCAL VOC		COCO Object	
Strategies	mIOU	mAcc	mIOU	mAcc
DiffSegmenter (Wang et al. 2023)	63.4	79.2	36.0	58.8
BLIP (Li et al. 2022)	46.3	53.1	32.3	41.6
LLaVA (Liu et al. 2023)	56.4	70.4	33.6	50.3
COCO Caption (Lin et al. 2014)	-	-	38.3	52.8
Ground Truth	**69.4**	**83.5**	**40.3**	**63.5**

Table 4: Ablating different text initialization strategies.

Method's Variants				PASCAL VOC		COCO Object	
Res8	Res16	Res32	Res64	mIOU	mAcc	mIOU	mAcc
✓				18.4	54.3	10.5	45.4
	✓			**63.4**	**79.2**	**36.0**	58.8
		✓		46.5	69.6	23.1	53.6
			✓	22.9	57.4	12.3	45.7
✓	✓			54.7	70.4	28.7	54.3
	✓	✓		60.7	77.4	33.0	**59.2**
✓	✓	✓		57.0	74.5	30.6	53.3
	✓	✓	✓	58.4	76.6	31.1	55.9
✓	✓	✓	✓	54.8	72.5	29.3	56.9

Table 5: Ablating layers to use for training in diffusion model.

to control the steepness of the curve, which determines how much we allow the anchor selection to be soft (having non-binary values), which would result in binary selection of anchors if the scale is infinitely large. We show that the model reaches optimal performance when scale is 4. When the scale is 8 or larger, the performance drops, indicating that an overly rigid or near-binary selection of anchors is suboptimal for our segmentation task. Conversely, when the scale is less than 4, the relatively soft selection (allowing lower probabilities) can also hinder performance. This demonstrates the importance of finding the right balance in anchor selection softness.

Text initialization strategies. We evaluated the effect of using different text initialization strategies, containing the method used in DiffSegmenter, extracting class names from captions (generated by BLIP, LLaVA or GT caption), and using ground truth class names. We observe that the performance of different strategies varies significantly. The method in DiffSegmenter achieves the best overall performance, only falling behind the one using COCO captions on the COCO Object dataset. Among the remaining methods, LLaVA shows superior results compared to BLIP, mainly due to its more detailed captions. All these methods lag behind the one using ground truth class names, indicating substantial room for improvement in obtaining candidate category names.

Diffusion model parameters. We conducted ablation studies on the following key parameters of our diffusion model: 1) layers to use: using different layers to optmize the text prompt can cause different result obviuosly. We aggregate the layers with same resolution together and only evaluate different resolutions. We show that when resolution is 16, we get the best performance. Combining mutiltple resolution do not bring better results. 2)time steps for training: InvSeg only runs a single pass through the diffusion process during each iteration. different time step reveals different information in the attention map. sampling in the range of 5 to 300 get the best performance. more or less will drop. 3) time steps for inference: The model achieves its best performance when t=50, with similar results at t=25. We observe that fewer time steps retain more details of the original image, with less noise added, resulting in better segmentation results.

Method's Variants	PASCAL VOC		COCO Object	
step	mIOU	mAcc	mIOU	mAcc
100	61.8	78.1	35.5	58.5
200	62.0	78.2	35.7	58.5
300	**63.4**	**79.2**	**36.0**	**58.8**
500	62.3	78.4	35.8	58.2

Table 6: Ablating training step in diffusion model.

Method's Variants	PASCAL VOC		COCO Object	
step	mIOU	mAcc	mIOU	mAcc
25	63.3	78.5	**36.0**	58.4
50	**63.4**	79.2	**36.0**	58.8
75	63.3	79.6	**36.0**	59.1
100	63.2	**79.9**	35.9	**59.5**
200	62.5	78.1	35.5	58.5
300	60.7	78.5	34.3	58.1

Table 7: Ablating time step for inference in diffusion model.

Visualization. We present a comparison of the final segmentation results in Fig. 4. Compared to the diffusion baseline, InvSeg produces more accurate and complete segmentation masks. To further illustrate the improvements, we provide a visualization of the refined attention maps $\{M\}$ for both the Diffusion Baseline and InvSeg in Fig. 5. This comparison reveals a notable difference in how each method handles various image elements. In the Diffusion Baseline, before prompt inversion is applied, we observe that background categories (such as "grass") tend to be overshadowed or suppressed by more salient foreground categories (like "cow" or "horse"). This imbalance in attention can lead to less accurate segmentation, particularly for less prominent image elements. InvSeg, on the other hand, demonstrates a more balanced and comprehensive attention distribution across both foreground and background elements. This improved attention mechanism contributes significantly to the enhanced accuracy and completeness of the segmentation masks produced.

Conclusion

We highlight the importance of customizing image-specific text prompts to boost the potential of generative text-to-image diffusion models to segment more diverse visual concepts. To this end, we introduce region-level prompt inversion using contrastive soft clustering, leveraging the structural information embedded in pretrained diffusion models. To the best of our knowledge, this is the first unsupervised region-level prompt inversion approach. The proposed InvSeg is able learn more detailed spatial information for each visual concept within an image, leading to competitive segmentation results compared to previous unsupervised methods.

Acknowledgments

This work was supported by Veritone, the China Scholarship Council and Queen Mary University of London's Apocrita HPC facility from QMUL RESEARCH-IT.

References

Abdul Samadh, J.; Gani, M. H.; Hussein, N.; Khattak, M. U.; Naseer, M. M.; Shahbaz Khan, F.; and Khan, S. H. 2024. Align Your Prompts: Test-Time Prompting with Distribution Alignment for Zero-Shot Generalization. *Advances in Neural Information Processing Systems*, 36.

Bojanowski, P.; Joulin, A.; Lopez-Paz, D.; and Szlam, A. 2017. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Cha, J.; Mun, J.; and Roh, B. 2023. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11165–11174.

Chen, J.; Zhu, D.; Qian, G.; Ghanem, B.; Yan, Z.; Zhu, C.; Xiao, F.; Culatana, S. C.; and Elhoseiny, M. 2023. Exploring Open-Vocabulary Semantic Segmentation from CLIP Vision Encoder Distillation Only. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 699–710.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

Hu, J.; Lin, J.; Gong, S.; and Cai, W. 2024. Relax Image-Specific Prompt Requirement in SAM: A Single Generic Prompt for Segmenting Camouflaged Objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12511–12518.

Karazija, L.; Laina, I.; Vedaldi, A.; and Rupprecht, C. 2023. Diffusion Models for Zero-Shot Open-Vocabulary Segmentation. *arXiv preprint arXiv:2306.09316*.

Khani, A.; Asgari, S.; Sanghi, A.; Amiri, A. M.; and Hamarneh, G. 2023. SLiMe: Segment Like Me. In *The Twelfth International Conference on Learning Representations*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Z.; Zhou, Q.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7667–7676.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Liu, Q.; Wen, Y.; Han, J.; Xu, C.; Xu, H.; and Liang, X. 2022. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *European Conference on Computer Vision*, 275–292. Springer.

Luo, H.; Bao, J.; Wu, Y.; He, X.; and Li, T. 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, 23033–23044. PMLR.

Marcos-Manchón, P.; Alcover-Couso, R.; SanMiguel, J. C.; and Martínez, J. M. 2024. Open-Vocabulary Attention Maps with Token Optimization for Semantic Segmentation in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9242–9252.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.-G.; Lee, S.-W.; Fidler, S.; Urtasun, R.; and Yuille, A. 2014. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 891–898.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ranasinghe, K.; McKinzie, B.; Ravi, S.; Yang, Y.; Toshev, A.; and Shlens, J. 2022. Perceptual grouping in vision-language models. *arXiv preprint arXiv:2210.09996*.

Ren, P.; Li, C.; Xu, H.; Zhu, Y.; Wang, G.; Liu, J.; Chang, X.; and Liang, X. 2023. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. *arXiv preprint arXiv:2302.10307*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35: 14274–14289.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, 9229–9248. PMLR.

Tian, J.; Aggarwal, L.; Colaco, A.; Kira, Z.; and Gonzalez-Franco, M. 2023. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint arXiv:2308.12469*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Tritrong, N.; Rewatbowornwong, P.; and Suwajanakorn, S. 2021. Repurposing gans for one-shot semantic part segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4475–4485.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Li, X.; Zhang, J.; Xu, Q.; Zhou, Q.; Yu, Q.; Sheng, L.; and Xu, D. 2023. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*.

Wang, Y.; Sun, R.; Luo, N.; Pan, Y.; and Zhang, T. 2024. Image-to-Image Matching via Foundation Models: A New Perspective for Open-Vocabulary Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3952–3963.

Wu, W.; Zhao, Y.; Shou, M. Z.; Zhou, H.; and Shen, C. 2023. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*.

Xiao, C.; Yang, Q.; Zhou, F.; and Zhang, C. 2023. From text to mask: Localizing entities using the attention of text-to-image diffusion models. *arXiv preprint arXiv:2309.04109*.

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18134–18144.

Xu, J.; Hou, J.; Zhang, Y.; Feng, R.; Wang, Y.; Qiao, Y.; and Xie, W. 2023a. Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2935–2944.

Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; and De Mello, S. 2023b. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2955–2966.

Xu, J.; and Zheng, C. 2021. Linear semantics in generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9351–9360.

Yi, M.; Cui, Q.; Wu, H.; Yang, C.; Yoshie, O.; and Lu, H. 2023. A Simple Framework for Text-Supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7071–7080.

Zhao, W.; Rao, Y.; Liu, Z.; Liu, B.; Zhou, J.; and Lu, J. 2023. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5729–5739.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *European Conference on Computer Vision*, 696–712. Springer.