Open-Vocabulary Attention Maps with Token Optimization for Semantic Segmentation in Diffusion Models

Pablo Marcos-Manchón[1, 2] Roberto Alcover-Couso[1] Juan C. SanMiguel[1] José M. Martínez[1]
[1]VPULab, Autonomous University of Madrid, Spain [2]Dynamics of Memory Formation Group, University of Barcelona, Spain
pmarcos@ub.edu, {roberto.alcover, juancarlos.sanmiguel, josem.martinez}@uam.es

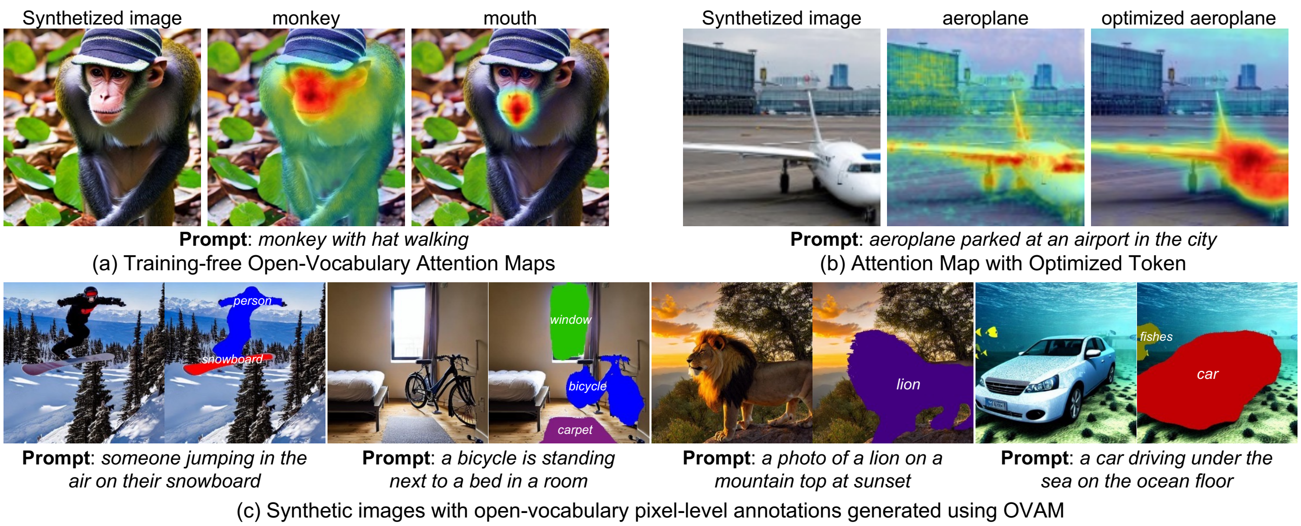

Figure 1. (a) We introduce Open-Vocabulary Attention Maps (OVAM), a training-free extension for text-to-image diffusion models to generate text-attribution maps based on open-vocabulary descriptions. Our approach overcomes the limitations of existing methods constrained by words contained within the prompt [45, 51, 55, 60]. (b) Our token optimization process enhances the creation of accurate attention maps, thereby improving the performance of existing semantic segmentation methods based on diffusion attentions [19, 45, 54, 60]. (c) Finally, we validate the utility of OVAM in producing synthetic images with precise pixel-level annotations.

Abstract

Diffusion models represent a new paradigm in text-to-image generation. Beyond generating high-quality images from text prompts, models such as Stable Diffusion have been successfully extended to the joint generation of semantic segmentation pseudo-masks. However, current extensions primarily rely on extracting attentions linked to prompt words used for image synthesis. This approach limits the generation of segmentation masks derived from word tokens not contained in the text prompt. In this work, we introduce Open-Vocabulary Attention Maps (OVAM)—a training-free method for text-to-image diffusion models that enables the generation of attention maps for any word. In addition, we propose a lightweight optimization process based on OVAM for finding tokens that generate accurate attention maps for an object class with a single annotation. We evaluate these tokens within existing state-of-the-art Stable Diffusion extensions. The best-performing model improves its mIoU from 52.1 to 86.6 for the synthetic images' pseudo-masks, demonstrating that our optimized tokens are an efficient way to improve the performance of existing methods without architectural changes or retraining. The implementation is available at github.com/vpulab/ovam.

1. Introduction

The introduction of diffusion models has led to a significant advancement in text-to-image (T2I) generation [7]. Diffusion-based models, such as Stable Diffusion [39] and other contemporary works [22, 27, 30, 32, 37, 38, 41], have been rapidly adopted across the research community and industry, owing to their ability to generate high-quality images that accurately reflect the semantics of text prompts.

The image generation in diffusion models is driven by a denoising process, which iteratively refines noise from an initial noisy vector until a coherent image is synthesized [14, 43]. To condition the image synthesis with a specific concept, usually represented by a text prompt, models utilize cross-attention mechanisms throughout the denoising steps [39, 49]. These mechanisms yield cross-attention matrices that facilitate the incorporation of semantic details into the spatial layout of images. Due to their role in fusing spatial and semantic information, these matrices have become a key part of works for interpreting the text prompt's influence on image layout [45] and for developing methods that extract pixel-level semantic annotations from the diffusion process [19, 34, 51, 54–56, 58, 60].

The use of attention matrices to relate semantic information to spatial layout draws inspiration from natural language processing, where word-specific attention has been shown to correlate with lexical attribution [5, 53]. In the context of T2I diffusion models, extracting attention matrices during image generation has become the primary method for extending the models' ability to jointly synthesize images and generate semantic segmentation pseudo-masks [19, 34, 51, 54, 60]. This method appears to be a promising approach to addressing the data scarcity challenge in semantic segmentation training, a problem arising from the high costs of pixel-level semantic annotation [21].

Existing methods that directly utilize attention matrices for generating semantic segmentation are limited by text prompt tokens, requiring an association between each semantic class and a prompt word [19, 45, 51, 60]. However, as shown in Fig. 1, not all object classes are explicitly mentioned in the text prompts, which highly limits the flexibility of these methods. To address this issue, some strategies incorporate additional modules that employ these attention features for mask generation; however, this necessitates additional supervised training, thereby limiting the methods' domain and increasing computational costs [54, 55].

In response to the abovementioned challenges, we introduce Open-Vocabulary Attention Maps (OVAM), a training-free approach that generalizes the use of attention maps from image synthesis. OVAM enables the creation of semantic segmentation masks described by an open vocabulary, irrespective of the words in the text prompts used for image generation. Moreover, we propose a token optimization process based on OVAM, which allows for the learning of open-vocabulary tokens that generate accurate attention maps for segmenting an object class with just a single annotation per class (see Fig. 1). These tokens not only enhance the quality of segmentation masks produced by OVAM but also, as we demonstrate through our experiments, they can improve the performance of existing open-vocabulary segmentation methods without any modifications or additional training [19, 45, 54, 60].

2. Related work

T2I Diffusion Models. The release of pretrained T2I diffusion models [22, 27, 30, 32, 37–39, 41, 61] has advanced the application of image generation in both research and practical contexts. A key to their success has been the decomposition of the synthesis process into sub-parts: a projection of input text [35, 36], diffusion through a denoising network [14, 43], and the decoding of final images. This modularity has facilitated the adaptation of these models to new tasks. Our method is centered on adapting Stable Diffusion [39] for semantic segmentation, although it is general enough to be applied to similar models that incorporate attention mechanisms within the denoising network [49].

Use of Attention in Diffusion Models. Attention mechanisms have a key role in the control of the image generation process, making attention matrices a central component for adapting diffusion models to various tasks. These matrices are an essential element in works pertaining to image editing [2, 13, 29, 48, 63], adding layout constraints [6, 25], model interpretability [28, 45], semantic correspondence [12, 23, 44, 62], and diverse segmentation tasks [19, 26, 31, 34, 47, 51, 54–56, 58, 60]. Our work employs those attentions to generate semantic segmentation ground truth, extending diffusion models to generate synthetic images with corresponding semantic segmentation pseudo-labels.

Open-Vocabulary Segmentation. Open-vocabulary semantic segmentation aims to divide an image into regions based on textual descriptions [50]. State-of-the-art proposals mostly rely on pretrained language models [8, 18, 20, 33, 59], and recent approaches focus on T2I pretrained diffusion models. Diffusion-based systems range from training-free methods [31, 34, 45, 51, 60] to those incorporating additional trained modules that utilize diffusion features [16, 19, 26, 54, 55, 58]. Our approach introduces a training-free methodology using pretrained diffusion models, overcoming the limitations of prior training-free approaches by generating semantic segmentation masks independent of the text prompt's vocabulary constraints.

Token Optimization in Diffusion Models. Token optimization involves refining input text embeddings used for image generation while the rest of the diffusion model remains unchanged. This technique has been employed for various objectives, such as generating synthetic images imitating a target image [42, 52] or learning new concepts from a limited number of examples [1, 10, 46]. Our research applies an optimization strategy to train text-embedding tokens for class segmentation. These tokens are then used to improve the accuracy of segmentation masks created by OVAM and existing diffusion-based segmentation methods without necessitating modifications to their architecture or additional training [19, 45, 54, 60].

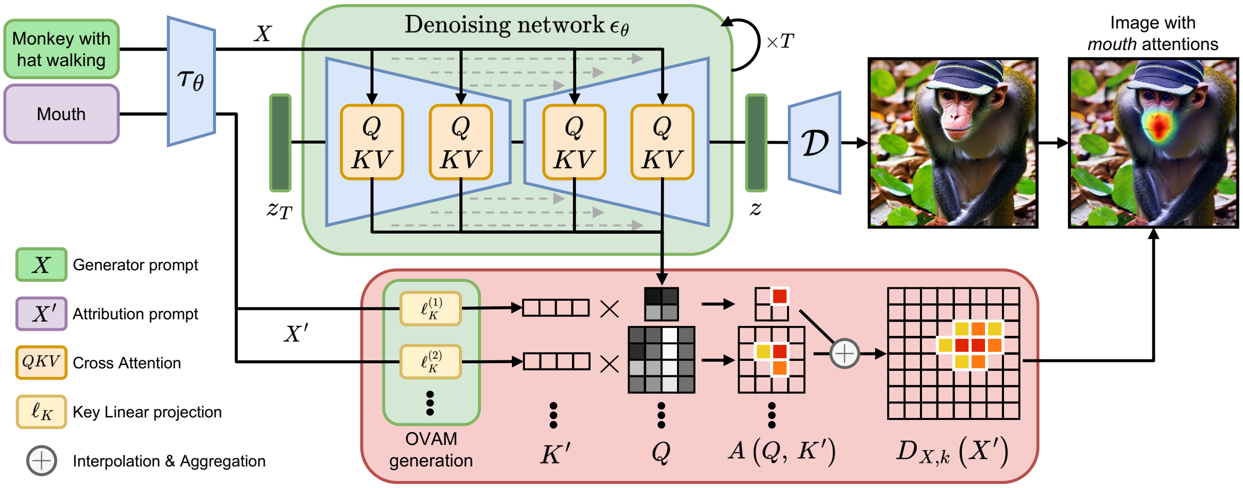

Figure 2. A schematic representation of the OVAM generation process (red module) utilizing the Stable Diffusion architecture [39] (rest of modules). The example synthesizes an image using the generator prompt *monkey with hat walking*. During the OVAM generation, pixel queries Q are extracted from the denoising network. These pixel queries are combined with the text embedding K' corresponding to the attribution prompt *mouth*, constructing the OVAM heatmap $D_{X,k}(X')$, which highlights the monkey's mouth in the synthesized image.

3. Proposed methodology

3.1. Cross-Attention Formulation

The denoising process, applied during image synthesis in diffusion models, takes place in a lower-dimensional latent space with shape $W \times H \times C$. Within this space, a latent vector—z_T—is transformed by the denoising network until they obtain a latent representation of the final image [14].

In Stable Diffusion [39], the UNet used as the denoising network contains convolutional downsampling and upsampling blocks [40]. These blocks process the latent vector z_t at each time step, iteratively generating the image.

At each time step t, the i-th convolutional block outputs a vector, denoted as $h_{i,t} \in \mathbb{R}^{\lceil \frac{W}{r^{(i)}} \rceil \times \lceil \frac{H}{r^{(i)}} \rceil \times C^{(i)}}$, where $r^{(i)}$ is a reduction factor tied to the block's resolution. These vectors contain the spatial information of the process. A text encoder, τ_θ, is used to convert the input text prompt into a text embedding $X \in \mathbb{R}^{l_E \times l_X}$ that consists of l_X tokens, each of dimension l_E. This embedding, which contains the semantic information, is subsequently combined using multi-headed cross-attention layers at each block's output.

Each cross-attention takes three inputs: a query Q, a key K, and a value V [49]. While K and V are computed from linear projections of the text embedding X, the input Q is sourced from a projection of the convolutional blocks' outputs. These linear projections, denoted as $\ell^{(i)}$, project X and $h_{i,t}$ into $l_H^{(i)}$ attention heads to build Q, K and V:

$$Q_{i,t} = \ell_Q^{(i)}\left(h_{i,t}\right),\, K_i = \ell_K^{(i)}\left(X\right),\, V_i = \ell_V^{(i)}\left(X\right). \quad (1)$$

The components Q, K, V are combined in the cross-attention blocks. As a result, during the process a cross-attention matrix $A\left(Q_{i,t},\, K_i\right) \in \mathbb{R}^{\lceil \frac{W}{r^{(i)}} \rceil \times \lceil \frac{H}{r^{(i)}} \rceil \times l_H^{(i)} \times l_X}$, is computed for each block i and time step t:

$$\text{CrossAttention}(Q_{i,t},\, K_i,\, V_i) = A\left(Q_{i,t},\, K_i\right) \cdot V_i\,, \quad (2)$$

$$A\left(Q_{i,t},\, K_i\right) = \text{softmax}\left(\frac{Q_{i,t} K_i^T}{\sqrt{d}}\right). \quad (3)$$

These cross-attention matrices weigh the influence of each token from X on the image's pixels, establishing a direct correlation between the spatial layout and the semantic content of the text. Studies like DAAM [45] utilize these token-aggregated matrices to discern the influence of each prompt word on the resulting image. Moreover, these matrices are employed as input features for open-vocabulary semantic segmentation systems based on diffusion models [19, 51, 54, 55, 58, 60]. However, direct extraction restricts open-vocabulary mask generation to the tokens within X. To overcome this limitation, our approach introduces Open-Vocabulary Attention Maps (OVAM), which generalize these matrices and eliminate this constraint.

3.2. Open-Vocabulary Attention Maps

For the construction of Open-Vocabulary Attention Maps (OVAM), we introduce a second text prompt, which we refer to as the *attribution prompt*. This text allows us to control the attention heatmaps using open vocabulary (see Fig. 2), eliminating the constraint of using the text prompt employed for image generation. The diffusion model's text encoder τ_θ is employed to transform the attribution prompt to produce an associated text embedding $X' \in \mathbb{R}^{l_E \times l_{X'}}$. This embedding consists of $l_{X'}$ tokens, which may not match the length of X, with dimension l_E. For generating the attention matrices, we project X' using the key projections $\ell_K^{(i)}$ to generate attribution keys K':

$$K'_i = \ell_K^{(i)}\left(X'\right) . \tag{4}$$

These attribution keys K' are combined with the pixel queries $Q_{i,t}$ computed during the denoising process (refer to Eqn. 1), creating the open-vocabulary attention matrices as $A\left(Q_{i,t}, K'_i\right)$ (see Eqn. 2). These matrices, which capture the influence of tokens on the image, have dimensions $\lceil \frac{W}{r^{(i)}} \rceil \times \lceil \frac{H}{r^{(i)}} \rceil \times l_H^{(i)} \times l_{X'}$, including two spatial dimensions, the attention heads, and the tokens of X'. To generate the OVAM corresponding to the k-th token of X', we aggregate the matrices across blocks, timestamps, and attention heads, following the same procedure used in [45]:

$$D_{X,k}\left(X'\right) = \sum_{i,t,h} \text{resize}\left(A_{h,k}\left(Q_{i,t},\, K'_i\right)\right) \in \mathbb{R}^{W \times H}, \tag{5}$$

where the notation $A_{h,k}$ refers to the slice associated with the h-th attention head and the k-th token. For matrices of varying resolutions, bilinear interpolation is used for resizing to a common resolution. Figure 2 shows the construction process of these maps for text attribution.

This approach can be viewed as a generalization of the DAAM method [45] and others that directly extract cross-attentions to generate semantic segmentation masks [51, 55, 60], offering a more versatile framework. When both X and X' are identical, the heatmaps $D_{X,k}\left(X'\right)$ are equivalent to directly extracting and aggregating the cross-attention matrices computed during image synthesis.

3.3. Token Optimization via OVAM

The proposed OVAM (Eqn. 5) relies on spatial information from cross-attention queries $Q_{i,t}$, derived from X and the initial state of the diffusion process, alongside semantic information from keys K'_i, computed from X'. When $Q_{i,t}$ is fixed, OVAM acts as a mapping from X' to attribution maps for each token it contains. One challenge in open-vocabulary segmentation lies in selecting the most effective descriptors. For example, is "mouth" the best descriptor for the monkey's mouth (see Figure 2)?. While some open-vocabulary segmentation studies address this by averaging synonyms and using prompts in varied contexts [20], we formulate descriptor selection as an optimization over X'.

To perform the optimization, we define X' with one token for each class we aim to optimize. Initially, X' is initialized with tokens corresponding to the class names for each segmentation class we target, including a background class initialized with $\langle SoT \rangle$, known to effectively capture background information in Stable Diffusion [45]. Additionally, we require ground truth $G_X \in \mathbb{R}^{W \times H \times l_{X'}}$ for an image

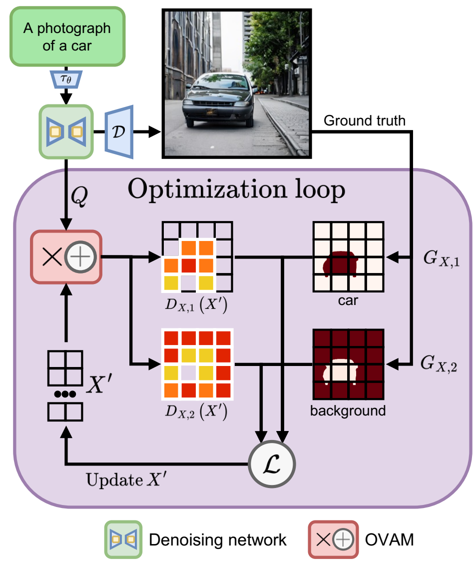

Figure 3. Diagram illustrating the optimization process for an X' composed of two tokens: a *background* and a *car* token. OVAM generates one heatmap for each token, and the optimization updates X' to align the attentions generated with the target mask.

synthesized from prompt X. This ground truth is a semantic segmentation map comprising $l_{X'}$ classes, $G_{X,k}$, with each class corresponding to a token of X'.

For the general case with multiple images, consider a set $\left\{\left(X_j, G_{X_j}\right)\right\}_j$, where each X_j is a prompt embedding used for image synthesis, and G_{X_j} is the corresponding ground truth. Our goal is to jointly optimize the tokens of X' to segment images consistent with the ground truth:

$$X^* = \underset{X'}{\text{argmin}} \sum_j \mathcal{L}\left(D_{X_j}\left(X'\right), G_{X_j}\right) , \tag{6}$$

where $\mathcal{L}$ is a loss function measuring the discrepancy between the OVAM heatmap and the ground truth mask. We employ the binary cross-entropy [15] as training loss:

$$\mathcal{L}\left(D_{X_j}\left(X'\right), G_{X_j}\right) = \sum_{k=1}^{l_{X'}} \text{BCE}\left(D_{X_j,k}\left(X'\right), G_{X_j,k}\right) . \tag{7}$$

Given that the objective is differentiable with respect to X', gradient descent is used for optimization. This method is similar to the conventional training of semantic segmentation models but computationally efficient since it involves learning a reduced number of parameters and can be completed in less than a minute on a single GPU. Figure 3 illustrates the optimization process, and Figure 4 shows examples of attention maps from optimized tokens. The experimental section further evaluates the efficacy of these tokens in open-vocabulary segmentation and their adaptability to different systems without requiring architectural changes.

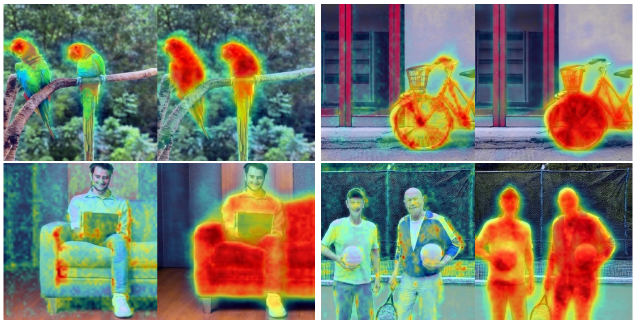

Figure 4. Comparative visualization of attention maps. Left images show attention using class name tokens (*bird*, *bicycle*, *sofa* and *person*) while the images on the right use optimized tokens with a training set that does not contain these images.

3.4. Mask Binarization

To generate binary segmentation masks from OVAM, we apply a refinement based on self-attentions and binarize the continuous heatmaps using a fixed threshold.

The denoising network of Stable Diffusion [39], besides integrating cross-attention mechanisms for merging spatial and semantic information, also includes self-attention blocks [49]. Unlike cross-attention, self-attention mechanisms do not directly relate semantics with spatial layout but capture object groupings information. For this reason, they are utilized to generate and refine segmentation masks in diffusion-based studies [34, 47, 51, 56]. Since we seek the most granular information to refine the masks, we exclusively extract self-attention matrices from the highest-resolution blocks, $W \times H$, and aggregate them across blocks, heads, and time steps to produce a fused map:

$$\mathcal{A}_\alpha = \underset{[\alpha,1]}{norm}\left(\sum_{i,h,t} A^{self}_{t,i,h}\right) \in \mathbb{R}^{W\times H}. \tag{8}$$

To combine these self-attention maps with OVAM heatmaps, we first normalize their values using min-max scaling, setting the range between α and 1. This normalization allows us to fuse the self-attention map with the OVAM heatmap multipliying both maps. The hyperparameter α allow the control of self-attention's impact on the final mask. Subsequently, we apply a threshold binarization relative to the peak value of the combined heatmap:

$$D^{\mathbb{I}_{\alpha,\tau}}_{X,k}(X') = \mathbb{I}\left(\mathcal{A}_\alpha \cdot D_{X,k}(X') \geq \tau M\right), \tag{9}$$

where $M = \max \mathcal{A}_\alpha \cdot D_{X,k}(X')$.

Finally, these binary masks are refined using dense Conditional Random Fields [17]. This process uses the image's geometry to improve the masks, adjusting their alignment with objects and improving the accuracy of the semantic segmentation.

4. Experimental Results

4.1. Evaluating Generated Pseudo-Masks

First, we perform an experiment to assess the quality of OVAM-generated pseudo-masks and compare them with those extracted by other Stable Diffusion-based methods.

Dataset Generation. Two datasets were produced using varying prompting strategies to compare the quality of generated pseudo-masks. Initially, a reduced dataset named *VOC-sim* was created to measure mask quality in a simplified context, partially replicating experiments from [19]. This dataset was generated using Stable Diffusion [39] with prompt templates of the form *A photograph of a ⟨classname⟩*, utilizing the 20 VOC classes [9] to create a total of 600 images, with an equal number of images per class. To generate different images using the same prompt, the diffusion model's seed was varied. Subsequently, a more complex dataset, *COCO-cap*, was generated. It was based on 1,100 text captions sampled from COCO [3] containing one of the VOC classes objects. These richer text prompts were used to generate images with complex scenes.

Token Optimization. For each class, we optimized a single token using only one image using the prompt *A photograph of a ⟨classname⟩*, employing distinct seeds from those for the evaluation set. We conducted separate optimizations for each class with X' consisting of two tokens, representing *background* and the *class* (refer to Fig. 3). The optimization was performed over 500 epochs on an A40 GPU, with the process for each token taking less than a minute. These tokens, corresponding to the 20 VOC classes, were used in all further experiments.

Masks Generation with OVAM. When generating synthetic images using Stable Diffusion 1.5, we produced OVAM binary masks, with both optimized and non-optimized variants, for comparison. For the non-optimized evaluation, we utilized the attribution prompt X' = *A photograph of ⟨classname⟩*, extracting the heatmap associated with the classname. This approach was applied in the *VOC-sim* evaluation, where it matches the text prompt, as well as in *COCO-cap*, where the prompts differ. For binarization, we empirically determined the hyperparameters $\tau = 0.4$ (also used in [45]) and $\alpha = 0.85$ (Eqn. 9), which were suitable choices in initial experiments with preliminary datasets. In evaluations with optimized tokens, we used the attribution prompts X' obtained from the optimization process (see Fig. 3), setting parameters to $\tau = 0.8$ and $\alpha = 0.95$. We use the dCRF [17] implementation *SimpleCRF* [11] with the default parameters.

Mask Generation with Other Methods. To benchmark OVAM's performance, we replicate the experimental protocol with the same prompts and seeds using various Stable Diffusion-based pseudo-mask generation methods. We compared OVAM with the training-free methods

DAAM [45] and Attn2Mask [60], which generate segmentation masks using cross-attentions. Due to the absence of a public implementation, we recreated Attn2Mask based on details from the original work. In the *COCO-cap* dataset, the class name does not always appear in the prompt; sometimes, a synonym is used or inferred from the text context. Therefore, these methods, which generate masks using cross-attention linked to prompt words, cannot create masks for all images in *COCO-cap*. To address this, we modified them to employ Open-Vocabulary Attention Maps, yielding identical results when the word is in the text prompt and enabling generation across all images. Furthermore, we assessed the masks created by Grounded Diffusion [19] and DatasetDM [54]. As these methods incorporate additional trained modules for mask generation, we employ the pre-trained weights provided by the authors for creating masks for VOC classes.

Evaluation. The pseudo-masks were manually annotated for the evaluation. Due to different parameters of Stable Diffusion required for some methods (e.g. the number of timesteps), synthesized images contain small variations. Nevertheless, we annotated each unique image, including those with subtle variations, to ensure a comparable evaluation. Three annotators were tasked with labeling the primary object in each image, excluding any without a clear object. We evaluated the mask of this main object against the corresponding automatically generated pseudo-masks using $\text{mIoU} = \frac{1}{20}\sum_c \text{IoU}_c$, where $\text{IoU}_c = \frac{TP_c}{TP_c+FP_c+FN_c}$.

Results. Table 1 summarizes the results. We observe that training-free methods achieve comparable outcomes, with variations likely due to distinct processing applied to attention maps. OVAM with token optimization outperforms models that require additional training, even though it relies on only a single annotation per class. The performance of Grounded Diffusion is noteworthy; however, it is skewed due to lower performance in several classes, which is addressed in the subsequent experiment.

Method	Vocab.	Dataset (mIoU %)	
		VOC-sim	COCO-cap
Training-free methods			
DAAM [45]	Prompt	66.2	48.4
Attn2Mask [60]	Prompt	68.7	55.0
OVAM (w/o token opt.) (ours)	Open	**70.4**	**58.2**
Methods with additional training			
Grounded Diffusion [19]	Open	62.1	50.2
DatasetDM [54]	Open	80.3	59.3
OVAM + token opt. (ours)	Open	**82.5**	**69.2**

Table 1. Comparing pseudo-mask generation performance on *VOC-sim* and *COCO-cap*. For *COCO-cap*, methods bound by prompt vocabulary have been adapted to employ attention from attribution prompts for evaluating instances where the class name is not explicitly present in the prompts.

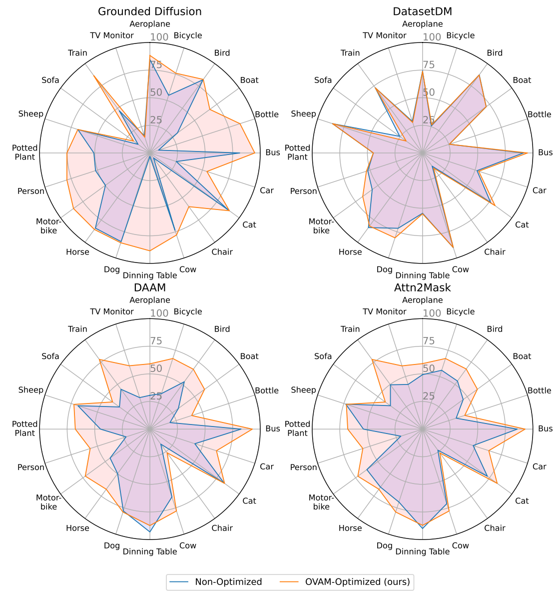

Figure 5. Class-performance comparison (IoU) of pseudo-masks generated by existing methods with/without the OVAM's token optimization, for the classes of the synthetic dataset *COCO-cap*.

4.2. Token Optimization with Different Methods

Use of Token Optimization in Other Methods. We replicated the evaluation of the prior experiment, applying token optimization across all techniques. Methods with additional training [19, 54], contains modules capable of evaluate open-vocabulary tokens for mask description. Instead of using the name of the class, as used in original papers, we utilized OVAM-optimized tokens. This integration is effective because all the methods are based on the cross-attention mechanisms of Stable Diffusion. For the training-free methods [45, 60], we employed the adaptations compatible with open-vocabulary attention maps, which allow for the utilization of any arbitrary tokens to describe the masks.

Evaluation. Like changing the token used as descriptor of the pseudo-mask does not change the image generation, we repeat the evaluation performed in the previous experiment using *VOC-sim* and *COCO-cap* synthetic datasets.

Results. Figure 5 displays the results for all classes, and Table 2 offers detailed outcomes for selected classes. Token optimization has improved the performance across all methods tested, improving the mIoU on all classes in methods based exclusively on cross-attentions [19, 45, 60]. DatasetDM [54] shows a minor improvement, as its mask generation is partly based on additional features from the diffusion process. Grounded Diffusion [19] showed notable improvement, likely due to low performance caused by misaligment between class names and the target concept on some classes, being corrected by the more precise class descriptions provided by the optimized tokens.

Method	Selected classes (COCO-cap IoU %)										Dataset (mIoU %)	
	aeroplane	bicycle	boat	bus	car	cat	cow	dog	motorbike	person	VOC-sim	COCO-cap
Training-free methods												
DAAM [45]	30.6	33.8	31.9	82.6	42.8	83.0	64.6	77.9	44.6	22.7	66.2	48.4
DAAM + *token optimization*	**59.1**	**67.2**	**61.2**	**92.8**	**63.4**	**83.6**	**77.9**	**79.0**	**72.4**	**56.9**	**79.7**	**66.1**
Attn2Mask [60]	49.3	56.0	45.5	85.8	48.2	72.6	71.0	69.4	62.4	20.7	68.7	55.0
Attn2Mask + *token opt.*	**59.3**	**67.2**	**61.4**	**92.9**	**63.1**	**83.6**	**77.9**	**78.9**	**72.5**	**56.9**	**81.9**	**66.1**
OVAM (ours)	65.1	64.3	51.9	84.9	47.5	67.9	76.5	65.8	69.4	19.7	70.4	58.2
OVAM + *token optimization*	**67.8**	**68.4**	**64.6**	**94.5**	**63.2**	**87.6**	**82.4**	**81.9**	**74.2**	**60.9**	**82.5**	**69.2**
Methods with additional training												
DatasetDM [54]	**74.1**	25.7	**71.3**	91.4	51.9	76.2	**90.1**	71.7	56.4	52.2	80.3	59.3
DatasetDM + *token opt.*	73.7	**26.7**	71.2	**95.1**	**53.5**	**81.2**	89.9	**80.8**	**67.0**	**53.5**	**80.6**	**60.5**
Grounded Diffusion [19]	84.6	54.9	30.8	81.4	25.1	87.3	73.7	84.4	49.8	51.8	62.1	50.2
Grounded Diffusion + *token opt.*	**88.3**	**75.9**	**67.1**	**95.0**	**54.3**	**89.0**	**78.1**	**85.5**	**85.6**	**79.1**	**86.6**	**73.3**

Table 2. Class-performance comparison of selected state-of-the-art methods employing the proposed token optimization process in OVAM.

4.3. Training a Semantic Segmentation Model

Dataset Generation. For training the semantic segmentation model, we created a large synthetic dataset by sampling 20,000 COCO captions of images containing a main object from a VOC class. Following the generation protocol from Experiment 4.1, we generated images, and the corresponding pseudo-mask for the main class, using OVAM with optimized tokens. To automatically filter out low-quality images that could negatively impact training, we used a filter based on CLIP [35] similarity. For this filtering, we generated a CLIP embedding for each image and calculated the cosine similarity with the CLIP embedding of the text *A photograph of a ⟨classname⟩*. Images with lower similarity scores were more likely to contain scenes unrelated to the class and were therefore filtered out. We removed the bottom 30% of images for each class based on these similarity scores. Additionally, we discarded images where the pseudo-mask covered more than 95% or less than 5% of the image area. After filtering, obtaining a final dataset with 13,484 images.

Training setup. Our training scheme involves Mask2former [4] (transformer-based) and Upernet [57] (convolutional-based) models trained on the VOC12 dataset [9], augmented with synthetic data from OVAM, exploring two distinct scenarios. In the first, we simulate a scarcity of real data by selecting 250 images per category, totaling 5000 real images for training. In the second scenario, we utilize the entire VOC12 dataset. For experiments combining real and synthetic images, a fine-tuning protocol is applied. Specifically, starting with models trained on synthetic images, we conduct further training with an initial learning rate ten times smaller. Training settings, including a batch size of two, are adopted from MMSegmentation [24].

Evaluation. Once the semantic segmentation model is trained, we follow the official VOC evaluation protocol on the VOC validation set [9].

Results. Table 3 compiles the results of training different models with real data, synthetic data from OVAM, and their combinations. This experiment leads to two key observations: Firstly, when real data is scarce, adding synthetic data generated through OVAM can match the results obtained using double the real data. Secondly, when more real data is available, incorporating synthetic data from OVAM can improve models performance by up to a 6.9% in mIoU.

4.4. Ablation Study

To understand the impact of different components in OVAM, we conducted several ablation studies, each replicating the evaluation from Experiment 4.1 to analyze the effects of post-processing, layer, and time step selection.

Post-processing effect. Table 4 summarizes the findings from our study to quantify the impact of post-processing stages — specifically, the application of dCRF and self-attention refinement — on the creation of OVAM masks.

Effect of layer selection. The denoising UNet of Stable Diffusion 1.5, use for the experiments, contain attention block of different resolution: 64x64, 32x32 and 16x16. In our proposed OVAM implementation, we aggregate attention from all resolutions. To quantify the impact of different blocks, we repeat the evaluation using different combinations of blocks. Results are compiled in Table 5. We observe that 16x16 blocks have better performance, which can be explained by the fact that they belong to a deeper part of the UNET [40], in charge of processing mainly semantic information.

Effect of denoising time step selection. In our method, we aggregate attention across all time steps of the denoising process. We explored the impact of three time step selection strategies: choosing a single time step $t = T$, selecting the initial time steps $t \leq T$, and opting for the latter time steps $t \geq T$. Figure 6 illustrates the results of this ablation study, varying the number of selected time steps throughout the process for these three strategies. The findings indicate that aggregating across all time steps yields the best per-

Model (Backbone)	Training set	Selected classes (VOC validation set IoU %)											mIoU %
		aeroplane	bicycle	boat	bus	car	cat	chair	cow	dog	person	tv	
Upernet (ResNet50)	OVAM (S: 13.5K)	45.2	29.5	19.7	68.6	46.5	60.5	12.2	55.2	45.0	50.2	26.2	42.1
	VOC (R: 11.5K)	75.9	31.3	33.5	**83.6**	**76.9**	**77.6**	23.3	49.6	**66.8**	**73.1**	25.9	54.3
	OVAM (S: 13.5K) + VOC (R: 5K)	72.2	22.9	29.1	79.9	75.4	76.8	23.3	**58.8**	66.2	72.4	25.3	54.4
	OVAM (S: 13.5K) + VOC (R: 11.5K)	**80.8**	**47.1**	**51.8**	52.3	72.5	71.1	**74.5**	21.3	57.3	73.0	**39.8**	**58.1**
Mask2Former (ResNet50)	OVAM (S: 13.5K)	54.9	30.1	40.3	70.8	47.2	65.0	3.1	51.1	46.8	39.9	0.0	41.5
	VOC (R: 11.5K)	88.9	53.0	62.9	83.0	83.0	90.9	27.2	**71.8**	73.2	86.2	59.6	69.7
	OVAM (S: 13.5K) + VOC (R: 5K)	85.4	53.5	**65.6**	**87.7**	82.6	82.6	26.9	66.6	70.1	**86.9**	**67.0**	69.8
	OVAM (S: 13.5K) + VOC (R: 11.5K)	**89.7**	**63.0**	60.6	84.9	**83.4**	**91.0**	**28.5**	69.9	**74.2**	86.5	62.8	**70.5**
Mask2Former (Swin-B)	OVAM (S: 13.5K)	66.1	27.9	57.3	70.8	51.7	81.1	15.0	78.9	68.9	46.4	0.0	53.5
	VOC (R: 11.5K)	**94.9**	**82.0**	81.2	91.5	90.3	**95.4**	42.1	85.4	**89.4**	90.3	**86.7**	82.8
	OVAM (S: 13.5K) + VOC (R: 5K)	91.7	79.7	79.1	**95.3**	89.6	93.5	**44.9**	91.8	89.0	90.6	82.6	82.9
	OVAM (S: 13.5K) + VOC (R: 11.5K)	94.7	78.2	**81.4**	93.9	**91.3**	94.2	41.7	**94.2**	88.2	**90.6**	79.7	**83.8**

Table 3. Model Performance Metrics on the VOC Validation Set [9]. In the table, *S* represents the number of OVAM-generated synthetic training samples, while *R* denotes the count of real training samples from the VOC dataset. Best results are indicated in bold.

Post-processing		VOC-sim (% mIoU)		COCO-cap (% mIoU)	
Self-Att.	dCRF	w/o opt.	w/ opt.	w/o opt.	w/ opt.
		66.2	79.7	48.4	66.1
✓		67.7	80.0	55.0	66.3
	✓	68.4	81.9	50.7	69.0
✓	✓	**70.4**	**82.5**	**58.2**	**69.2**

Table 4. Post-processing ablation study for the OVAM proposal.

Block resolution			VOC-sim (% mIoU)		COCO-cap (% mIoU)	
64	32	16	w/o opt.	w/ opt.	w/o opt.	w/ opt.
✓			40.2	43.5	26.2	24.4
	✓		30.5	71.7	16.1	56.0
		✓	**71.5**	78.9	57.0	68.2
✓	✓		50.5	74.9	37.2	53.6
✓		✓	70.5	82.2	**60.3**	**70.3**
	✓	✓	67.0	80.7	53.5	68.2
✓	✓	✓	70.4	**82.5**	58.2	69.2

Table 5. Cross-Attention blocks ablation study.

formance. Moreover, when token optimization is used, a similar level of performance can be achieved by only extracting attentions from $t = 12$, located at the midpoint of the diffusion process.

5. Conclusions

In conclusion, our work introduces Open-Vocabulary Attention Maps (OVAM), extending text-to-image diffusion models like Stable Diffusion for generating synthetic images with semantic segmentation pseudo-masks through open vocabulary descriptors. Our approach adapts existing Stable Diffusion-based segmentation methods, which were previously limited to text prompt-linked masks, to recognize any arbitrary word. Moreover, our token optimization technique notably enhances the precision of attention maps for class segmentation. Experimental results show that

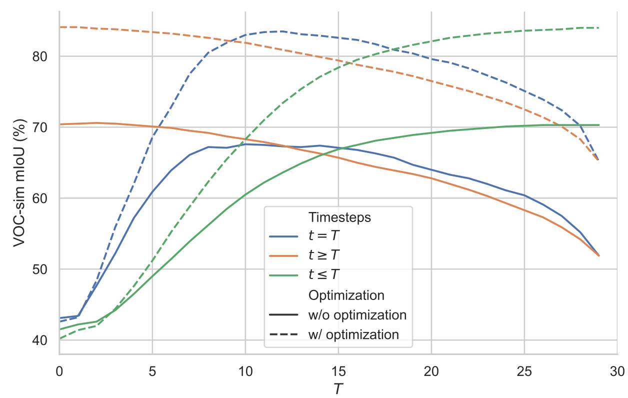

Figure 6. Time-step selection study for the OVAM proposal.

this optimization leads to significant performance gains, increasing OVAM pseudo-masks by +12.2 mIoU and improving other diffusion-based pseudo-mask generation methods by as much as +24.5 mIoU.

Moreover, we demonstrate the practical value of OVAM-generated synthetic data for training semantic segmentation models. When this data is used for training, models with half the amount of real data can achieve competitive results on the VOC Challenge, comparable to those trained with the full set of real data. Furthermore, when combined with the entirety of real data, certain models exhibit performance improvements of up to 6.9% in mIoU.

Our findings affirm the viability of OVAM not only in enhancing existing diffusion-based segmentation methods but also as a valuable approach for synthetic data generation to train robust semantic segmentation models.

Acknowledgement This work has been partially supported by the SEGA-CV (TED2021-131643A-I00) and the HVD (PID2021-125051OB-I00) projects of the Ministerio de Ciencia e Innovación of the Spanish Government.

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 2

[2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*, 2023. 2

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289, 2022. 7

[5] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, 2019. 2

[6] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2174–2183, 2023. 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Conference on Neural Information Processing Systems (NIPS)*, volume 34, pages 8780–8794, 2021. 1

[8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11583–11592, 2022. 2

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 5, 7, 8

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[11] Healthcare Intelligence Laboratory. SimpleCRF. https://github.com/HiLab-git/SimpleCRF, 2017. 5

[12] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. In *Conference on Neural Information Processing Systems (NIPS)*, 2023. 2

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NIPS)*, volume 33, pages 6840–6851, 2020. 2, 3

[15] Shruti Jadon. A survey of loss functions for semantic segmentation. In *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, 2020. 4

[16] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*, 2023. 2

[17] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Conference on Neural Information Processing Systems (NIPS)*, volume 24, pages 109–117, 2011. 5

[18] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[19] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7667–7676, 2023. 1, 2, 3, 5, 6, 7

[20] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7061–7070, 2023. 2, 4

[21] Hubert Lin, Paul Upchurch, and Kavita Bala. Block annotation: Better image annotation with sub-image decomposition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[22] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2309.06380*, 2023. 1, 2

[23] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Conference on Neural Information Processing Systems (NIPS)*, 2023. 2

[24] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark, 2020. https://github.com/open-mmlab/mmsegmentation. 7

[25] Chong Mou, Xintao Wang, Liangbin Xie, Jing Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[26] Minheng Ni, Yabo Zhang, Kailai Feng, Xiaoming Li, Yiwen Guo, and Wangmeng Zuo. Ref-diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*, 2023. 2

[27] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya

Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, volume 162, pages 16784–16804, 2022. 1, 2

[28] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. In *Conference on Neural Information Processing Systems (NIPS)*, 2023. 2

[29] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2023. 2

[30] Pablo Pernias, Dominic Rampas, and Marc Aubreville. Wuerstchen: Efficient pretraining of text-to-image models. *arXiv preprint arXiv:2306.00637*, 2023. 1, 2

[31] Koutilya PNVR, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, and David Jacobs. Ld-znet: A latent diffusion approach for text-based image segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4157–4168, 2023. 2

[32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2

[33] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19446–19455, 2023. 2

[34] Nguyen Quang Ho, Vu Truong Tuan, Tran Anh Tuan, and Nguyen Khoi. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In *Conference on Neural Information Processing Systems (NIPS)*, 2023. 2, 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2, 7

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67, 2020. 2

[37] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 2

[38] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023. 1, 2

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 3, 5

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and ComputerAssisted Intervention (MICCAI)*, pages 234–241, 2015. 3, 7

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Conference on Neural Information Processing Systems (NIPS)*, volume 35, pages 36479–36494, 2022. 1, 2

[42] Idan Schwartz, Vésteinn Snæbjarnarson, Hila Chefer, Serge Belongie, Lior Wolf, and Sagie Benaim. Discriminative class tokens for text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22725–22735, 2023. 2

[43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, volume 37, pages 2256–2265, 2015. 2

[44] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 2

[45] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1, pages 5644–5659, 2023. 1, 2, 3, 4, 5, 6, 7

[46] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. *Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2023. 2

[47] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint arXiv:2308.12469*, 2023. 2, 5

[48] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 2

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NIPS)*, volume 30, pages 6000–6010, 2017. 2, 3, 5

[50] Chunlei Wang, Wenquan Feng, Xiangtai Li, Guangliang Cheng, Shuchang Lyu, Binghao Liu, Lijiang Chen, and Qi

Zhao. Ov-vg: A benchmark for open-vocabulary visual grounding. *arXiv preprint arXiv:2310.14374*, 2023. 2

[51] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 1, 2, 3, 4, 5

[52] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In *Conference on Neural Information Processing Systems (NIPS)*, 2023. 2

[53] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019. 2

[54] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Conference on Neural Information Processing Systems (NIPS)*, 2023. 1, 2, 3, 6, 7

[55] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1206–1217, 2023. 1, 2, 3, 4

[56] Changming Xiao, Qi Yang, Feng Zhou, and Changshui Zhang. From text to mask: Localizing entities using the attention of text-to-image diffusion models. *arXiv preprint arXiv:2309.01369*, 2023. 2, 5

[57] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 7

[58] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 2, 3

[59] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision (ECCV)*, 2021. 2

[60] Ryota Yoshihashi, Yuya Otsuka, Kenji Doi, and Tomohiro Tanaka. Attention as annotation: Generating images and pseudo-masks for weakly supervised semantic segmentation with diffusion. *arXiv preprint arXiv:2309.01369*, 2023. 1, 2, 3, 4, 6, 7

[61] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative AI: A survey. *arXiv preprint arXiv:2303.07909*, 2023. 2

[62] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023. 2

[63] Yuechen ZHANG, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. In *Conference on Neural Information Processing Systems (NIPS)*, 2023. 2