FreeGen: Bridging Visual-Linguistic Discrepancies Towards Diffusion-based Pixel-level Data Synthesis

Wenzhuang Wang^{1,2}, Mingcan Ma^{2}, Yong Chen^{2}, Changqun Xia^{3*}, Zhenbao Liang^{2}, Jia Li^{1*}

[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University
[2] Geely Automobile Research Institute
[3] Pengcheng Laboratory
{wz_wang, jiali}@buaa.edu.cn, xiachq@pcl.ac.cn

Abstract

Text-to-image diffusion model has inspired research into text-to-data synthesis without human intervention, where spatial attentions correlated with semantic entities in text prompts are primarily interpreted as pseudo-masks. However, these vannila attentions often deliver visual-linguistic discrepancies, in which the associations between image features and entity-level tokens are unstable and divergent, yielding inferior masks for realistic applications, especially in more practical open-vocabulary settings. To tackle this issue, we propose a novel text-guided self-driven generative paradigm, termed FreeGen, which addresses the discrepancies by recalibrating intrinsic visual-linguistic correlations and serves as a real-data-free method to automatically synthesize open-vocabulary pixel-level data for arbitrary entities. Specifically, we first learn an Attention Self-Rectification mechanism to reproject the inherent attention matrices to achieve robust semantic alignment, thereby obtaining class-discriminative masks. A Temporal Fluctuation Factor is present to assess mask quality based on its variation over uniform sampling timesteps, enabling the selection of reliable masks. These masks are then employed as self-supervised signals to support the learning of an Entity-level Grounding Decoder in a self-training manner, thus producing open-vocabulary segmentation results. Extensive experiments show that the existing segmenters trained on FreeGen narrow the performance gap with real data counterparts and remarkably outperform the state-of-the-art methods.

Introduction

Modern semantic segmenters typically demand vast volumes of visual images paired with dense annotations to achieve satisfactory fine-grained recognition. However, the procedure for collecting large-scale images and labelling them with dense labels is extremely cost-prohibitive for primary manual efforts. Moreover, due to concerns over data privacy in real-world applications, acquiring a considerable number of training data poses a formidable barrier. The above challenges significantly hinder further performance improvements and urge scholars to rethink *model-centric* semantic segmentation from a *data-centric* perspective.

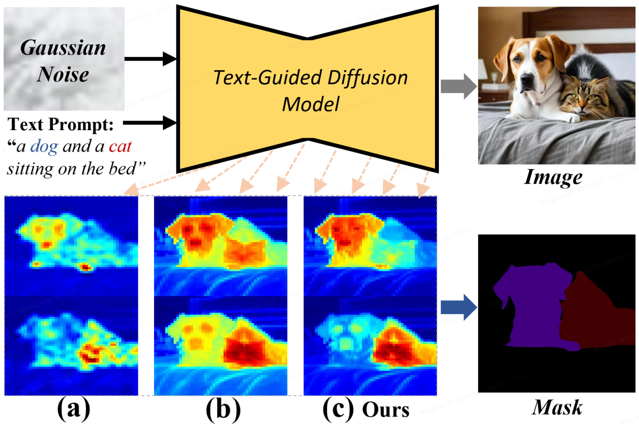

Gaussian Noise
Text-Guided Diffusion Model
Text Prompt:
"a dog and a cat sitting on the bed"
Image
Mask
(a)
(b)
(c) Ours

Figure 1: The visualization of visual-linguistic correspondences between our FreeGen and other methods. Compared to (a) DiffuMask (Wu et al. 2023b) and (b) Dataset Diffusion (Nguyen et al. 2024), Our FreeGen (c) can obtain better visual-linguistic correlations.

In light of generative models, an alternative strategy aims to harness them to synthesize images with ground truth, thereby augmenting or even replacing the role for real data in downstream tasks. Predominant methods (Zhang et al. 2021; Li et al. 2022b) involves utilizing GANs (Goodfellow et al. 2014; Brock, Donahue, and Simonyan 2018) to generate pixel-level data. However, due to the unattainable Nash equilibrium, these works often produce monotonous images with imprecise masks. Inspired by the remarkable success of latent diffusion model (LDM) (Rombach et al. 2022), researchers are unlocking their possibilities in generating image-mask pairs based on free-form text prompts without manual intervention, i.e., *text-to-data synthesis*.

In line with a common thought, the concurrent works Diffusionseg (Ma et al. 2023), DiffuMask (Wu et al. 2023b) and Attn2Mask (Yoshihashi et al. 2023) directly mine cross-attention associations, which bind visual features and entity tokens in the text prompt, to roughly identify object positions and generate semantic labels. Nevertheless, as

[*]Correspondence should be addressed to Changqun Xia (xiachq@pcl.ac.cn) and Jia Li (jiali@buaa.edu.cn).

shown in Fig.1 (a), cross-attention matrices often exhibit sparse correlations, leading to subpar masks with ambiguous edges. Meanwhile, Dataset Diffusion (Nguyen et al. 2024) further introduces self-attention exponentiation to enhance mask quality to a certain extent. While the exponential dot-product sometimes exacerbates the discrepancy when the cross-attention matrices are coarse, especially when encountering multiple objects in the synthetic image. As depicted in Fig.1 (b), visual regions associated with the entity “dog” unexpectedly incorporate elements of “cat”. Indeed, the visual-linguistic attentions often suffer from weak alignment and ambiguous edges throughout the denoising steps, referred to as “visual-linguistic discrepancies” in our paper, which imposes a low-quality bottleneck on synthetic masks. Particularly, this discrepancy leads to more pronounced limitations in practical open-vocabulary mask generation. Although there are several initial attempts (Wu et al. 2023a; Li et al. 2023), they either rely on real data for supervision or lack flexibility due to pre-training on specific domains. Naturally, an urgent question emerges: *Could we tackle the visual-linguistic discrepancies to generate entity-free image-mask pairs without any manual effort?*

To resolve this problem, we propose a novel self-driven generative paradigm, dubbed ***FreeGen***, which exploits the intrinsic knowledge of LDM (Rombach et al. 2022) to tackle visual-linguistic discrepancies and facilitate the automatic pipeline of open-vocabulary pixel-level data synthesis in segmentation. Specifically, we adopt a two-step learning strategy. ***First***, we aim to learn an *Attention Self-Rectification (ASR)* mechanism to reproject the internal visual-linguistic correlations, which involves applying a learnable cross-head interaction matrix to multi-head self-attention within the latent space to capture informative relations between visual regions. These relations are then utilized to re-weight spatial cross-attentions, thereby aggregating class-discriminative masks. We introduce a *Temporal Fluctuation Factor* to evaluate mask quality and filter out those low-quality synthetic data based on its variations over uniform sampling timesteps. ***Second***, these masks serve as supervisory sources to learn an *Entity-level Grounding Decoder (EGD)* in a hardness-aware self-training fashion, further enhancing external visual-linguistic correlations and resembling open-vocabulary generative abilities. Once the two-step training is complete, the EGD is capable of generating pseudo-masks for arbitrary entities in text prompts.

Extensive experiments on 3 benchmarks show that the existing segmenters trained on our *FreeGen* significantly outperforms Dataset Diffusion (Nguyen et al. 2024) by a substantial margin of 9.3% in mIoU on VOC 2012 (Everingham et al. 2010). The contributions of our paper lies in four-folds:

- We propose a self-driven generative *FreeGen*, which exploits the inherent diffusion knowledge to achieve open-vocabulary pixel-level data synthesis without manual efforts via a two-stage training strategy.
- We design an *Attention Self-Rectification* mechanism, which strives to recalibrate robust correspondences between visual features and semantic entities.
- To further enhance open-vocabulary segmentation abilities, an *Entity-level Grounding Decoder* is developed to generate dense masks for arbitrary objects in a hard-aware self training loop.
- Experiments show that the segmenters trained on FreeGen achieve competitive performance against real counterparts. And in open-vocabulary settings, FreeGen delivers SOTA results on unseen classes of VOC 2012.

Related Work

Semantic Segmentation. Semantic segmentation aims to simultaneously predict class labels and corresponding masks for objects within the images. Most traditional semantic segmenters (Long, Shelhamer, and Darrell 2015; Chen et al. 2017; Zhang et al. 2020a,b; Cheng et al. 2022) are optimised on numerous image-mask pairs and restricted to a predefined set of categories present in the training set. To further elevate the generalised segmentation capabilities of segmenters, recent literatures (Ding et al. 2022; Li et al. 2022a; Liang et al. 2023) attempt to identify arbitrary object categories appearing in the image, known as open-vocabulary semantic segmentation, which often rely on a large-scale multimodal model, i.e., CLIP (Radford et al. 2021), to learn the alignment maps between visual features and text embeddings of class entities. In this paper, we propose to extend the ability of “model-centric” semantic segmenters by training them on synthetic data to realize “data-centric” open-vocabulary pixel-level data generation.

Diffusion Models for Semantic Segmentation. Diffusion models (Ho, Jain, and Abbeel 2020) represent a likelihood-based generative strategy that fundamentally revolutionises image synthesis and spawns a series of text-guided generative models (Nichol et al. 2021; Rombach et al. 2022; Saharia et al. 2022; Chang et al. 2023) and their potential in computer vision (Chen et al. 2023; Bandara, Nair, and Patel 2022; Chen, Sun, and Lin 2024), especially in semantic segmentation (Ji et al. 2023; Zhao et al. 2023; Kondapaneni et al. 2024). Currently, there are two methodologies for harnessing generative diffusion models for discriminative segmentation. One approach involves using it as a feature extractor or directly tuning the denoising U-Net conditioned on visual images to predict dense labels. The other approach uses diffusion models to generate photo-realistic images with dense annotations for downstream auxiliary training. For instance, DatasetDM (Wu et al. 2023a) leverages a few real data to train a P-decoder following Stable Diffusion, thereby producing images along with perceptual labels. While other efforts (Wu et al. 2023b; Nguyen et al. 2024; Yoshihashi et al. 2024; Li et al. 2023) seek to construct synthetic segmentation datasets without any manual involvement, achieving promising results comparable to real-data-required counterparts.

However, these works often struggle with unstable intermediate attentions in the latent space to roughly generate suboptimal masks. In response, our paper highlights a novel self-driven paradigm that leverages inherent diffusion knowledge to rectify visual-linguistic correspondences, and additionally incorporates a temporal fluctuation factor to select feasible synthetic masks.

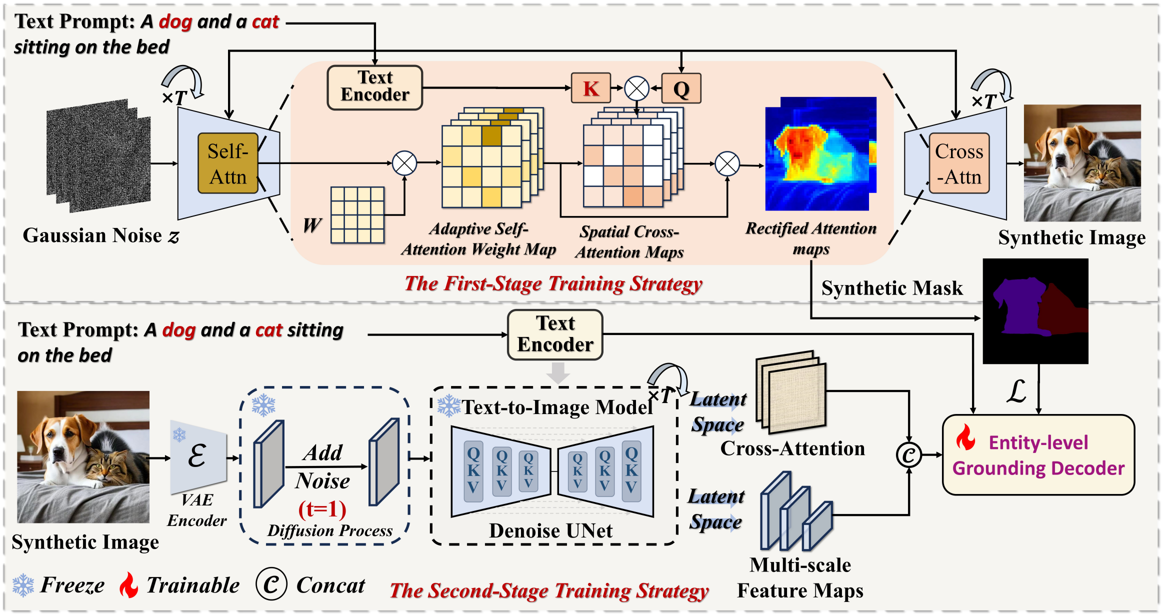


Figure 2: Overall pipeline of our FreeGen: (1) The first-stage involves learning an *Attention Self-Rectification* mechanism for recalibrating self- and cross-attention maps, achieving strong correspondences. (2) While second-stage involves learning an *Entity-level Grounding Decoder* for open-vocabulary mask generation.

Method

In this section, we will elaborate our FreeGen based on LDM (Rombach et al. 2022), as shown in Fig. 3. Specifically, we develop a two-stage learning strategy, where the first-stage training involves learning an *Attention Self-Rectification* mechanism with a temporal fluctuation factor to deal with the visual-linguistic discrepancies, constructing and selecting a synthetic segmentation dataset on base classes with class-discriminative annotations, and the second-stage training involves learning an *Entity-level Grounding Decoder* to generate open-vocabulary masks for novel classes in a hardness-aware self-training manner.

Problem Formulation

Given a set of textual captions $\mathcal{P}$, a text-guided diffusion model ϕ, which consists of a text encoder, a variational autoencoder and the temporal U-Net, refers to the generation of realistic images $\mathcal{I}$ from Gaussian noise $z \sim \mathcal{N}(0,1)$ via a denoising Markov chain of length T, i.e., $\mathcal{I} = \phi(z, \mathcal{P}, T)$. To achieve high-fidelity image synthesis, intra-visual and inter-visual-linguistic interplays occur continuously at multiple levels within the temporal U-Net. For each denoising timestep $t \in [1, T]$, this procedure delivers available intermediate representations within the latent space, including self-attention maps $\mathcal{A}_s$, cross-attention maps $\mathcal{A}_c$ and multi-scale features $\mathcal{F}$. Our key insight is that the intermediate representations accumulate the intrinsic knowledge in the context of the diffusion model, allowing data synthesis without any manual intervention. Our FreeGen, driven by the intrinsic knowledge, extends text-to-image synthesis to more challenging open-vocabulary text-to-data synthesis, i.e. $\{\mathcal{I}, \mathcal{M}_*\} = \phi(z, \mathcal{P}, T)$, via a two-stage training strategy. $\mathcal{M}_*$ denotes open-vocabulary dense annotations for arbitrary semantic entities. Finally, we can get obtain a pixel-level open-vocabulary dataset, i.e., $\mathcal{D}_{train} = \{(\mathcal{I}, \mathcal{P}, \mathcal{M}_*)\}$.

First-stage: Attention Self-Rectification

Recalibrate Correspondences from Disentangle Attention Maps. Most existing works (Wu et al. 2023b) apply coarse cross-attentions that may deviate from entity tokens to finish pseudo-labels, which often deliver unreliable masks. In this regard, we aim to learn an Attention Self-Rectification mechanism to recalibrate visual-linguistic correspondences from disentangled two parts: self- and cross-attention maps. The former encodes pairwise similarities within intra-visual regions and the latter expresses inter-visual-linguistic correlations. To obtain class-discriminative masks with object locations and categories, we adopt the former to reweight the latter in an adaptive learning manner.

Adaptive Self-attention Weight Map. Multi-head self-attention matrices capture global dependencies among visual regions, in which each head highlights specific object positions and maintaining their contextual information. For the timestep t, we introduce a learnable cross-head interaction matrix $W \in R^{H \times N}$ to dynamically aggregate them and generate a new adaptive self-attention weight map $\mathcal{A}_{s,t}$,

$$\mathcal{A}_{s,t,h} = W_h \odot Softmax(\frac{Q^v_{h,t}(K^v_{h,t})^\top}{\sqrt{d_h}}), h \in [1, H], \quad (1)$$

where $Q^v_{h,t}$ and $K^v_{h,t} \in R^{H \times N \times C}$ are flattened vectors of sequence length N from latent visual features. C is the channel dim, H denotes the number of attention heads and d_h is the h-th head dim.

Spatial Cross-attention Re-weighting. Cross-attention matrices focus on the associations between latent image features and semantic entities in text prompts. However, as mentioned above, the vanilla cross-attention maps are coarse and sparse, which means that the visual-linguistic correlations suffer from weak alignment, leading to voids for the synthetic mask only by simple thresholding. In this context, we introduce a learnable linear layer $\Upsilon \in R^{C\times C}$ to reproject the key vectors $K_t^y \in R^{H\times N^{'}\times C}$ ($N^{'}$ is the number of class entities) corresponding to the class-entities y, allowing for a new cross-attention map $\mathcal{A}_{c,t} \in R^{H\times N\times N^{'}}$ for any given entity at timestep t:

$$\mathcal{A}_{c,t,h} = Softmax(\frac{Q^v_{h,t}(\Upsilon_h(K^y_{h,t}))^\top}{\sqrt{d_h}}), h \in [1, H], \quad (2)$$

Afterwards, we exploit the adaptive $\mathcal{A}_{s,t}$ to re-weight the $\mathcal{A}_{c,t,h}$ to obtain spatially informative rectified attention maps $\mathcal{A}_{*,t} \in R^{H\times N\times N^{'}}$, which can be formalized as:

$$\mathcal{A}_{*,t} = \mathcal{A}_{s,t} \cdot \mathcal{A}_{c,t}. \quad (3)$$

Temporal Fluctuation Factor. To further bridge visual-linguistic discrepancies, we design a Temporal Fluctuation Factor (TFF) for data evaluation and filtering. As shown in Fig. 3, the synthetic masks across timesteps often show large variance fluctuations in the ambiguous regions. Observing this, TFF evaluates the quality of synthetic image-mask pairs based on their variations over uniform sampling timesteps.

Specifically, we use the text-guided diffusion model ϕ to output the initial latent state $z \in \mathbb{R}^{\mathcal{H}\times\mathcal{W}\times C}$, where $\mathcal{H}$, $\mathcal{W}$ are height and width. Then we denoise z in the uniform sampling timesteps $S = \{T_1, T_2, \ldots, T_{\mathcal{K}}\}$ within the U-Net to obtain different masks of $\mathcal{A}_*$ over $\mathcal{K}$ timesteps. Finally, the TFF can be calculated from the variance fluctuations between different masks, which can be formulated as:

$$TFF = \frac{1}{\mathcal{K}\cdot\mathcal{H}\cdot\mathcal{W}} \sum_{i=1}^{\mathcal{K}}\sum_{h=1}^{\mathcal{H}}\sum_{w=1}^{\mathcal{W}} |\mathcal{M}_{i,h,w} - \overline{\mathcal{M}}_{i,h,w}|, \quad (4)$$

where $\mathcal{M}$ and $\overline{\mathcal{M}}$ denote the synthetic masks and the mean of them, respectively. After this, we can select high-quality synthetic data, as shown **in the Supp**, and ultimately harvest better synthetic masks to train our ASR mechanism.

Training Strategy. We use only the $\mathcal{A}_{*,t}$ from the last timestep, averaging along the head dimension to, extract class-discriminative heat maps, which are then upsampled and reshaped to image width $\mathcal{H}$ and height $\mathcal{W}$, with each heat map corresponding to a class entity. To learn the ASR mechanism, we also require pseudo-labeled segmentation maps $\mathcal{G}_* \in R^{\mathcal{H}\times\mathcal{W}\times N^{'}}$. Unlike the existing work (Li et al. 2023), which exploits the segmenter pre-trained on specific domain datasets, we adopt a class-agnostic segmenter, i.e., SAM (Kirillov et al. 2023), to generate accurate semantic masks via point prompts. In particular, we choose three points with the strongest response and the farthest distance in each channel of $\mathcal{A}_{*,t}$ without ASR to obtain $\mathcal{G}_*$. Our training objective is to learn rectified attention maps that are consistent with pseudo-labeled segmentation maps:

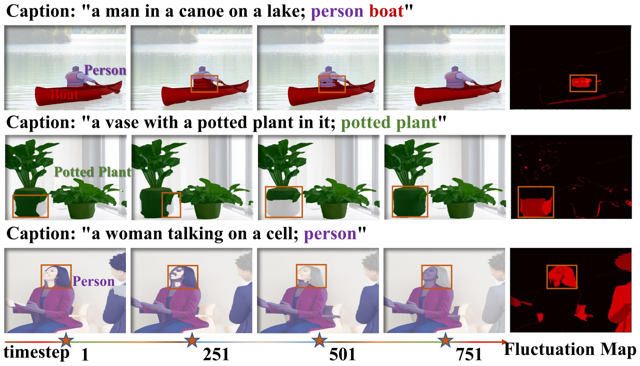

Caption: "a man in a canoe on a lake; person boat"
Person
Boat
Caption: "a vase with a potted plant in it; potted plant"
Potted Plant
Caption: "a woman talking on a cell; person"
Person
timestep
1
251
501
751
Fluctuation Map

Figure 3: Visualization of the temporal fluctuation factor computation principle, which shows the masks at different timesteps and the fluctuation heatmaps among them.

$$\mathcal{L}_1(\mathcal{A}_*, \mathcal{G}_*) = \sum_{k=1}^{N^{'}} BCE(\mathcal{A}_*(...,k), \mathcal{G}_*(...,k)), \quad (5)$$

where BCE is the Binary Cross-Entropy (Jadon 2020) loss. The training strategy allows us to generate pseudo-masks for multi-class objects in the synthetic image.

Second-stage: Entity-level Grounding Decoder

After the first-stage training, we can transform heat-maps $\mathcal{A}_*$ into a set of class-discriminative masks $\mathcal{M}$. Assume we have a synthetic training set for base class entities (i.e., bus, cat) consisting of B triplets, i.e., $D_{train} = \{(\mathcal{I}_1, p_1, \mathcal{M}_1), .., (\mathcal{I}_B, p_B, \mathcal{M}_B)\}$, where p_i indicates the text prompt. The synthetic training set is utilized to support the learning of our EGD, which comprises three modules: a text adapter, an entity-guided feature alignment module, and an open-vocabulary grounding decoder. The second-stage training further reinforces the connections between visual features and open-vocabulary semantic entities, providing our FreeGen with the ability to fine-generate novel classes.

Entity-guided Feature Alignment. Given a synthetic image $\mathcal{I} \in R^{\mathcal{H}\times\mathcal{W}\times 3}$ from the training set, we can extract the multi-scale feature maps $\mathcal{F}$ within the temporal U-Net, corresponding to four resolutions, i.e., 64×64, 32×32, 16×16 and 8×8, in a single forward diffusion process ($t = 1$). We concatenate visual features with the same resolution and compress their channel dimensions with 3×3 convolution blocks. Then, $\mathcal{F}$ and linguistic embeddings $\mathcal{Y}$ can be interacted bidirectionally layer by layer as follows:

$$\mathcal{F}_{l+1} = BiAttn(\varphi(\mathcal{F}_l), \mathcal{Y}_l) + \mathcal{F}_l, \quad (6)$$
$$\mathcal{Y}_{l+1} = BiAttn(\xi(\mathcal{Y}_l), \mathcal{F}_l) + \mathcal{Y}_l, \quad (7)$$

where $\varphi(\cdot)$ denotes multi-scale deformable self-attention (Zhu et al. 2020), linguistic embeddings are further refined by vanilla transformer $\xi(\cdot)$ layers (Vaswani 2017), and $BiAttn$ denotes the bidirectional attention blocks.

Open-vocabulary Grounding Decoder. Based on the Mask2former head (Cheng et al. 2022), we concatenate

Segmenter	Backbone	Training Set		Semantic Segmentation (IoU) for Sampled Classes /%										mIoU
		# Real	***# Synthetic***	aeroplane	bird	boat	bus	car	cat	chair	cow	dog	sofa	
Train with Full Real Data														
DeepLabV3	R50	VOC (R:10.6*k*)	-	89.2	89.9	74.1	94.5	87.7	92.4	34.7	89.1	88.1	43.5	77.4
DeepLabV3	R101	VOC (R:10.6*k*)	-	93.6	93.4	67.0	95.9	90.7	94.9	36.8	89.4	89.7	58.7	79.8
Mask2former	R50	VOC (R:10.6*k*)	-	87.5	94.4	70.6	95.5	87.7	92.2	44.0	85.4	89.1	53.6	77.2
Mask2former	Swin-B	VOC (R:10.6*k*)	-	**97.0**	93.7	71.5	91.7	89.6	96.5	**57.5**	**95.9**	**96.8**	65.6	84.3
Train with Pure Synthetic Data														
Mask2former	R50	-	DiffuMask (S: 60.0*k*)	80.7	86.7	56.9	81.2	74.2	79.3	14.7	63.4	65.1	27.8	57.4
Mask2former	Swin-B	-	DiffuMask (S: 60.0*k*)	90.8	92.9	67.4	88.3	82.9	92.5	27.2	92.2	86.0	49.8	70.6
Mask2former	R50	-	DatasetDM (S: 40.0*k*)	-	84.7	-	86.0	79.2	74.4	-	-	63.7	-	60.3
Mask2former	Swin-B	-	DatasetDM (S: 40.0*k*)	-	93.4	-	93.8	78.8	94.5	-	-	79.6	-	73.7
DeepLabV3	R50	-	Attn2mask (S: 80.0*k*)	65.7	82.5	64.7	87.0	76.0	83.2	25.0	65.3	73.0	13.6	62.2
DeepLabV3	R101	-	Dataset Diffusion (S: 40.0*k*)	81.6	73.3	62.2	85.5	64.8	78.2	21.6	69.2	71.8	41.8	64.6
DeepLabV3	R50	-	FreeGen (S: 40.0*k*)	86.7	84.2	68.2	92.8	78.3	79.6	28.8	78.5	66.5	42.1	67.5
DeepLabV3	R101	-	FreeGen (S: 40.0*k*)	88.0	89.7	75.6	93.9	81.6	85.3	32.6	82.6	73.0	45.9	70.3
Mask2former	R50	-	FreeGen (S: 40.0*k*)	86.6	84.3	67.6	93.2	79.2	86.8	34.6	82.3	75.0	46.9	69.6
Mask2former	Swin-B	-	FreeGen (S: 40.0*k*)	90.8	89.5	**81.2**	94.6	76.5	93.7	39.4	92.7	88.3	52.5	76.3
Train with Synthetic Data and Finetune on Real Data														
Mask2former	R50	VOC (R:5.0*k*)	DiffuMask (S: 60.0*k*)	85.4	92.8	74.1	92.9	83.7	91.7	38.4	86.5	86.2	39.8	77.6
Mask2former	Swin-B	VOC (R:5.0*k*)	DiffuMask (S: 60.0*k*)	95.6	94.4	72.3	**96.9**	**92.9**	**96.6**	51.5	96.7	95.5	70.2	84.9
Mask2former	R50	VOC (R:5.0*k*)	FreeGen (S: 40.0*k*)	89.6	88.1	71.5	89.1	82.4	88.3	42.2	90.0	84.6	60.2	78.3
Mask2former	Swin-B	VOC (R:5.0*k*)	FreeGen (S: 40.0*k*)	96.8	94.8	76.2	96.0	91.8	96.5	52.4	95.0	94.4	**69.2**	**85.1**
Train with Real & Synthetic Data														
DeepLabV3	R50	VOC (R:10.6*k*)	FreeGen (S: 20.0*k*)	91.7	91.6	75.3	93.5	86.7	93.0	38.1	91.5	91.1	52.3	78.5
DeepLabV3	R101	VOC (R:10.6*k*)	FreeGen (S: 20.0*k*)	92.7	91.2	79.1	95.9	88.9	94.4	39.9	91.7	90.9	55.9	80.5
Mask2former	R50	VOC (R:10.6*k*)	FreeGen (S: 20.0*k*)	86.5	90.4	76.3	93.0	91.8	93.7	39.9	78.6	87.3	54.4	78.3
Mask2former	Swin-B	VOC (R:10.6*k*)	FreeGen (S: 20.0*k*)	95.3	**96.7**	78.7	96.8	92.2	**96.6**	44.7	92.4	95.1	68.1	84.8

Table 1: Performance results of segmenters DeepLabV3 and Mask2former on VOC 2012 *val*. 'R' and 'S' denote the "Real" and "Synthetic" dataset. The best results are in **bold** and the second best results are underlined.

the linguistic embeddings from the aforementioned alignment module with the learnable queries to learn the semantic concepts for arbitrary entities, thereby supporting the open-vocabulary mask generation. The classification score for each object category is obtained by the dot product between query and linguistic features, followed by a Sigmoid activation. More details are **in the Supp.**

Hardness-aware Self-training. To guarantee the learning effectiveness of EGD, we adopt a hardness-aware self-training strategy inspired by curriculum learning (Bengio et al. 2009), which advocates that the decoder starts learning from simple synthetic image-mask pairs and gradually advancing to complex ones. To implement this idea, we leverage the EGD, pre-trained once on synthetic data, to compute the loss map for each image-mask pair. To be specific, we calculate the hardness value $\mathcal{S}$ based on pixel-level cross-entropy losses for N object categories in synthetic images:

$$\mathcal{S}=\sum_{j=1}^{N}\sum_{i=1}^{h_j\times w_j}\frac{[(\mathcal{M}_*^i==j)\times L_i]}{h_j\times w_j}, \tag{8}$$

where L_i is the loss map, $h_j \times w_j$ indicate the area of j-th object category in the synthetic mask $\mathcal{M}_*$. According to the hardness values $\mathcal{S}$. We can sort the synthesized data and select the top 50% for training the decoder, and utilize it to regenerate the segmentation masks for the remaining 50% of synthesized images. This has two advantages: first, it helps to train the decoder effectively, and second, the synthesised mask can be further refined. To achieve optimisation, we use dice (Milletari, Navab, and Ahmadi 2016) and cross-entropy losses for classification, and BCE losses for mask prediction.

Method	Segmentor	Backbone	COCO 2017		VOC 2012	
			# Synthetic	mIoU	***# Synthetic***	mIoU
DiffuMask	Mask2former	R50	-	-	S: 60*k*	57.4
	Mask2former	Swin-B	-	-	S: 60*k*	70.6
Attn2mask	DeepLabV3	R50	-	-	S: 80*k*	62.2
	Mask2former	Swin-B	-	-	S: 80*k*	71.0
Dataset Diffusion	DeepLabV3	R50	S: 80*k*	32.4	S: 40*k*	61.6
	DeepLabV3	R101	S: 80*k*	34.2	S: 40*k*	64.8
	Mask2former	R50	S: 80*k*	31.0	S: 40*k*	60.5
DatasetDM	Mask2former	R50	-	-	S: 40*k*	60.3
	Mask2former	Swin-B	-	-	S: 40*k*	73.7
FreeGen	DeepLabV3	R50	S: 80*k*	36.2	S: 40*k*	67.5
	DeepLabV3	R101	S: 80*k*	37.4	S: 40*k*	70.3
	Mask2former	R50	S: 80*k*	34.5	S: 40*k*	69.6
	Mask2former	Swin-B	S: 80*k*	**44.3**	S: 40*k*	**76.3**

Table 2: Comparisons in mIoU between training on synthetic data and testing on VOC 2012 and COCO 2017 *val*.

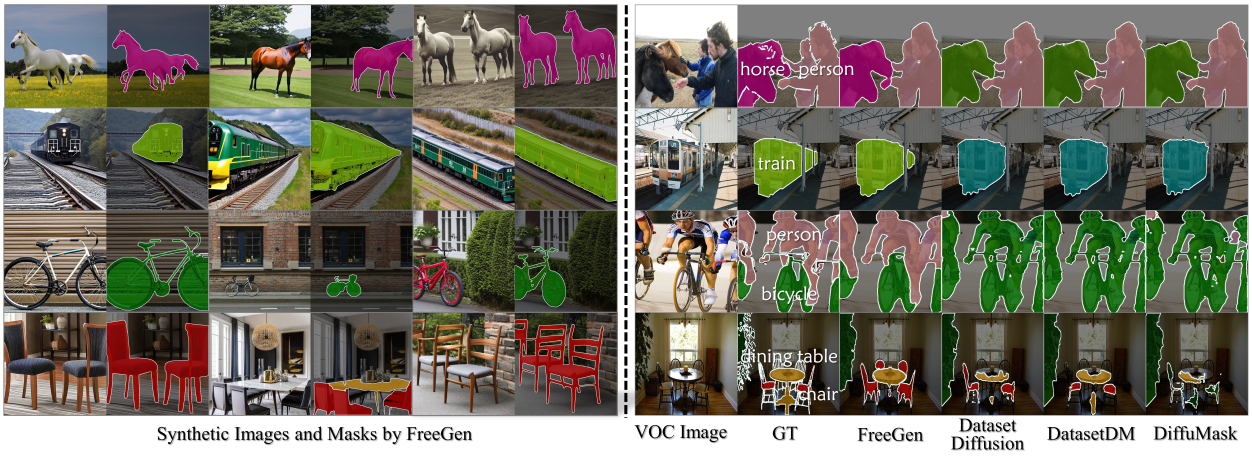
Experiments

Datasets and Experimental Settings

Datasets. We conduct experiments on 3 benchmarks: VOC 2012 (Everingham et al. 2010), augmented with SBD (Hariharan et al. 2011) for 10.6k training and 1,449 validation images in 20 classes; COCO 2017 (Lin et al. 2014) with 118,287 training and 5k validation images in 80 classes; and Cityscapes (Cordts et al. 2016), an urban scene dataset with 2,975 training and 500 validation images in 19 classes.

Evaluation Metrics. Following previous efforts (Nguyen et al. 2024; Wu et al. 2023b), we train existing segmenters on synthetic data and test them on real *val* data, assessing mean Intersection over Union (mIoU). For open-vocabulary setting, the mIoU is measured on seen and unseen classes.

Experimental Setting. To thoroughly assess the generative capabilities of our FreeGen, we follow three experimen-



(a) Generation Results from our FreeGen

(b) Segmentation Results on VOC 2012's validation set

Figure 4: Showcase of FreeGen generation and qualitative analysis with SOTA methods. The visualisation results show that our FreeGen can produce more accurate semantic masks for semantic entities, especially in more challenging multi-class scenarios.

tal protocols: semantic segmentation, open-vocabulary segmentation, and domain generalization, as previously utilized in (Wu et al. 2023b; Li et al. 2023).

Implementation Details

Our FreeGen, based on the Stable Diffusion V2.1 pre-trained on LAION5B (Schuhmann et al. 2022), generates 512×512 image-mask pairs via 50 denoising steps. For fair comparison, we generate 40*k*, 80*k*, and 80*k* images for VOC 2012, COCO 2017, and Cityscapes. We choose the vanilla SAM (Kirillov et al. 2023) to obtain primary masks and set the uniform sampling timesteps to $\mathcal{K} = 4$ to get temporal fluctuation factor. We train our ASR and EGD using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 20 epochs, where the batch size is 4 and the learning rate is 1e-4. We train DeepLabV3 and Mask2former on synthetic data following MMSegmentation's default settings (Contributors 2020) and compare them with models trained on real data. Following (Nguyen et al. 2024), to enhance the variety of textual guidance, we adopt the image captioner BLIP (Li et al. 2022c) and ChatGPT (Achiam et al. 2023) to derive more text prompts for object categories.

Main Quantitative & Qualitative Results

Protocol-I: Semantic Segmentation

VOC 2012. Tab. 1 systematically compares the mIoU results of the DeepLabV3 and Mask2former segmenters. From Tab. 1, FreeGen boosts the existing semantic segmenters by a large margin under four training settings. Especially, FreeGen achieves a competitive result of 69.6 mIoU against the full real data of 77.2 mIoU. Without any manual annotation, FreeGen shows amazing results (within 4% gap) close to the performance trained on real data for many classes, i.e., *bird, cat, cow, horse*. Furthermore, FreeGen significantly outperforms DiffuMask, Attn2Mask and Dataset Diffusion by **7.4%** on VOC 2012 *val*. A point worth highlighting is that Mask2former trained on FreeGen delivers a 2.6 mIoU improvement over DatasetDM (Wu et al. 2023a), which needs a few real data for supervision. When fine-tuned on a few real data (5*k*) or trained jointly with real & synthetic data, FreeGen consistently delivers performance improvements for segmenters, enabling them to surpass their real data counterparts, i.e., 78.3 *v.s* 77.2 for Mask2former (R50). Fig. 4 illustrates the synthetic image-mask pairs produced by FreeGen and the segmentation results, highlighting our method's superior capability to resolve visual-linguistic discrepancies in complex multi-class data synthesis.

Segmenter	Backbone	Training Set		Sampled Classes /%		mIoU
		#Real	***#Synthetic***	Human	Vehicle	
		Train with Pure Real Data				
Mask2former	R50	R:3.0*k*	-	83.4	94.5	89.0
Mask2former	Swin-B	R:3.0*k*	-	85.5	96.0	90.8
Mask2former	Swin-B	R:1.5*k*	-	84.6	95.3	90.0
		Train with Pure Synthetic Data				
Mask2former	R50	-	DiffuMask(S:100.0*k*)	70.7	85.3	78.0
Mask2former	SwinB	-	DiffuMask(S:100.0*k*)	72.1	87.0	79.6
Mask2former	R50	-	FreeGen(S:80.0*k*)	71.2	86.1	78.7
Mask2former	SwinB	-	FreeGen(S:80.0*k*)	72.8	87.9	80.4
		Finetune on Real Data				
Mask2former	R50	R:1.5*k*	DiffuMask(S:100.0*k*)	84.6	95.5	90.1
Mask2former	Swin-B	R:1.5*k*	DiffuMask(S:100.0*k*)	86.4	96.4	91.4
Mask2former	R50	R:1.5*k*	FreeGen(S:80.0*k*)	85.3	96.0	90.7
Mask2former	Swin-B	R:1.5*k*	FreeGen(S:80.0*k*)	**86.6**	**97.1**	**91.9**

Table 3: Comparisons in mIoU between training on synthetic data and testing on Cityscapes *val*.

COCO 2017. Tab. 2 reports the comparison results between FreeGen and SOTA methods on more challenging COCO dataset. For DeepLabV3, our FreeGen demonstrates a +3.2%∼+3.8% absolute improvement in mIoU over Dataset Diffusion. For Mask2former, FreeGen achieves the best performance with a 3.5% improvement. These results further demonstrate the applicability of our FreeGen for multi-class data synthesis. More qualitative results and detailed IoUs for 80 classes are provided **in the Supp**.

Cityscapes. Tab. 3 presents the comparison results on Cityscapes *val*, which involves complex segmentation of urban street scenes. Following (Wu et al. 2023b), we evaluate

Methods	Train Set / #Categories		mIoU %		
	Real / 15	Synthetic / 15+5	Seen	Unseen	Harmonic
Manual **Annotation** Supervision					
ZS3	✔	✗	78.0	21.2	33.3
CaGNet	✔	✗	78.6	30.3	43.7
Joint	✔	✗	77.7	32.5	45.9
STRICT	✔	✗	82.7	35.6	49.8
SIGN	✔	✗	83.5	41.3	55.3
ZegFormer	✔	✗	86.4	63.6	73.3
Text Prompt Supervision					
Li et al. (R101)	✗	✔	62.8	50.0	55.7
DiffuMask (R50)	✗	✔	60.8	50.4	55.1
DiffuMask (R101)	✗	✔	62.1	50.5	55.7
DiffuMask (Swin-B)	✗	✔	71.4	65.0	68.1
DatasetDM (R101)	✗	✔	65.1	51.1	57.1
DatasetDM (Swin-B)	✗	✔	78.8	60.5	68.4
FreeGen (R50)	✗	✔	70.1	51.9	59.6
FreeGen (R101)	✗	✔	69.0	53.3	60.1
FreeGen (Swin-B)	✗	✔	77.9	64.9	70.8
Manual Annotation & Text Prompt Supervision					
Li et.al (R101)	✔	✔	83.0	71.3	76.7
FreeGen (R50)	✔	✔	77.7	63.4	69.8
FreeGen (R101)	✔	✔	80.3	71.5	75.6
FreeGen (Swin-B)	✔	✔	**84.5**	**76.5**	**80.3**

Table 4: Performance Comparisons for Zero-Shot Semantic Segmentation on VOC 2012 *val*.

Train Set	Test Set	Sampled Classes mIoU %			mIoU
		Car	Person	Motorbike	
Cityscapes	VOC 2012 *val*	26.4	32.9	28.3	29.2
ADE20K	VOC 2012 *val*	73.2	66.6	64.1	68.0
DiffuMask	VOC 2012 *val*	74.2	71.0	63.2	69.5
DatasetDM	VOC 2012 *val*	77.9	72.9	70.1	73.6
FreeGen	VOC 2012 *val*	**79.2**	**79.9**	**80.8**	**80.0**

Table 5: Comparisons for domain Generalization. *Person* and *Rider* of Cityscapes are regarded as the same class.

FreeGen on two general classes, Human and Vehicle, where "human" contains two subclasses: person and rider, while "vehicle" includes four subclasses: car, bus, truck and train. From Tab. 3, we can observe that, when trained on our FreeGen and fine-tuned on real data, Mask2former achieves effective performance improvements, i.e., 91.9 *v.s* 90.0 mIoU. Besides, with less synthetic data (80*k* *v.s* 100*k*), FreeGen outperforms DiffuMask by 0.7% on the vehicle class.

Protocol-II: Open-vocabulary Segmentation. Tab. 4 delves into the open vocabulary segmentation capabilities of generative FreeGen. Compared to zero-shot semantic segmenters (Bucher et al. 2019; Gu et al. 2020; Baek, Oh, and Ham 2021; Pastore et al. 2021; Cheng et al. 2021; Ding et al. 2022) trained on real *manual annotation*, our *text-prompts-supervised* FreeGen achieves the SOTA results on *Unseen* classes, i.e., 64.9 *v.s* 63.6. Moreover, when fine-tuned on the real data of *Seen classes*, the harmonic mean increases to a promising **80.3**. The performance improvement under this setting validates the open-set generalization capability of our entity-level grounding decoder. More qualitative results on open-vocaublary segmentation are available **in the Supp.**

Protocol-III: Domain Generalization. Tab. 5 shows the cross-dataset generalization of Mask2former (R50) trained on different datasets. Compared to the real data, FreeGen exhibits superior domain generalization ability, i.e., **80.0** with FreeGen *v.s* 68.0 with ADE20K (Zhou et al. 2017). In particular, FreeGen outperforms DatasetDM with a remarkable improvement of 6.4% mIoU. For *car* class, mask2former trained with Cityscapes achieves only 26.4, but with FreeGen is 79.2. We attribute this gap to differences in object size and domain shifts between the two datasets.

Two Stage		One Stage			mIoU
HST	EGD	TFF	ASR	SAM	
✔	✔	✔	✔	✔	69.6
✔	✔	✔	✔	✗	66.2
✗	✔	✔	✔	✔	68.4
✗	✗	✔	✔	✔	67.1
✗	✗	✗	✔	✔	66.4
✗	✗	✗	✗	✔	63.8
✗	✗	✗	✗	✗	60.5

Table 6: Ablation experiments of FreeGen on VOC 2012.

Ablation Study

Effect of the Components of FreeGen. Tab.6 reports the effectiveness of the components of FreeGen using Mask2former (R50) on VOC 2012 *val*, where HST denotes hardness-aware self-training. We can see that the raw result is 60.5% similar to Dataset Diffusion. After one-stage training for visual-linguistic discrepancy and evaluation for synthetic masks, the mIoU increases to 67.1%, demonstrating the necessity of our ASR. Two-stage self-training gives a further improvement of 3.4% on the VOC 2012 **val**. It should be noted that without SAM, our FreeGen still achieves a competitive result of 66.2% mIoU, confirming the effectiveness of our generative strategy.

Impact of Different Annotations on Synthetic Masks. We also investigate the impact of three annotations on synthetic masks, including manual label, pseudo-label and FreeGen-label. 'Pseudo-label' refers to the annotations from Mask2former (R50) pre-trained on VOC 2012. Experimental results indicate that FreeGen-label is comparable to manual-label and significantly exceeds the pseudo-label. More ablation results are **in the Supp**.

Conclusion

This paper introduces a novel text-guided self-driven data synthesis method, FreeGen, which employs a two-stage training strategy. The first stage focuses on leveraging intrinsic diffusion knowledge to resolve visual-linguistic inconsistencies, enhanced by a temporal fluctuation factor that selects more reliable synthetic masks. The second stage refines visual-linguistic alignment and equips FreeGen with the ability to perform open-vocabulary segmentation in a self-training loop. Our extensive experiments on three benchmarks demonstrate that FreeGen significantly narrows the performance gap between real and synthetic data across various training configurations. We expect that FreeGen will accelerate the transition from reliance on real data to synthetic data in dense vision tasks.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant Nos. 62132002, 62102206) and Major Key Project of PCL (Grant No. PCL2023A10-1).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Baek, D.; Oh, Y.; and Ham, B. 2021. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9536–9545.

Bandara, W. G. C.; Nair, N. G.; and Patel, V. M. 2022. Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models. *arXiv preprint arXiv:2206.11892*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Bucher, M.; Vu, T.-H.; Cord, M.; and Pérez, P. 2019. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 19830–19843.

Chen, Z.; Sun, K.; and Lin, X. 2024. CamoDiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1272–1280.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cheng, J.; Nandi, S.; Natarajan, P.; and Abd-Almageed, W. 2021. Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9556–9566.

Contributors, M. 2020. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark. https://github.com/open-mmlab/mmsegmentation.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11583–11592.

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.

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.

Gu, Z.; Zhou, S.; Niu, L.; Zhao, Z.; and Zhang, L. 2020. Context-aware feature generation for zero-shot semantic segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1921–1929.

Hariharan, B.; Arbeláez, P.; Bourdev, L.; Maji, S.; and Malik, J. 2011. Semantic contours from inverse detectors. In *2011 international conference on computer vision*, 991–998. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jadon, S. 2020. A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*, 1–7. IEEE.

Ji, Y.; Chen, Z.; Xie, E.; Hong, L.; Liu, X.; Liu, Z.; Lu, T.; Li, Z.; and Luo, P. 2023. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21741–21752.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kondapaneni, N.; Marks, M.; Knott, M.; Guimaraes, R.; and Perona, P. 2024. Text-image alignment for diffusion-based perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13883–13893.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022a. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

Li, D.; Ling, H.; Kim, S. W.; Kreis, K.; Fidler, S.; and Torralba, A. 2022b. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21330–21340.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022c. 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.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7061–7070.

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.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Ma, C.; Yang, Y.; Ju, C.; Zhang, F.; Liu, J.; Wang, Y.; Zhang, Y.; and Wang, Y. 2023. Diffusionseg: Adapting diffusion towards unsupervised object discovery. *arXiv preprint arXiv:2303.09813*.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. Ieee.

Nguyen, Q.; Vu, T.; Tran, A.; and Nguyen, K. 2024. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Pastore, G.; Cermelli, F.; Xian, Y.; Mancini, M.; Akata, Z.; and Caputo, B. 2021. A closer look at self-training for zero-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2693–2702.

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.

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.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Vaswani, A. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wu, W.; Zhao, Y.; Chen, H.; Gu, Y.; Zhao, R.; He, Y.; Zhou, H.; Shou, M. Z.; and Shen, C. 2023a. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36: 54683–54695.

Wu, W.; Zhao, Y.; Shou, M. Z.; Zhou, H.; and Shen, C. 2023b. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1206–1217.

Yoshihashi, R.; Otsuka, Y.; Doi, K.; Tanaka, T.; and Kataoka, H. 2024. Exploring Limits of Diffusion-Synthetic Training with Weakly Supervised Semantic Segmentation. arXiv:2309.01369.

Yoshihashi, R.; Otsuka, Y.; Tanaka, T.; et al. 2023. Attention as annotation: Generating images and pseudo-masks for weakly supervised semantic segmentation with diffusion. *arXiv preprint arXiv:2309.01369*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020a. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; and Sun, Q. 2020b. Feature pyramid transformer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 323–339. Springer.

Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.-F.; Barriuso, A.; Torralba, A.; and Fidler, S. 2021. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10145–10155.

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, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 633–641.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*.