

# DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models

Weijia Wu<sup>13</sup> Yuzhong Zhao<sup>2</sup> Hao Chen<sup>1</sup> Yuchao Gu<sup>3</sup> Rui Zhao<sup>3</sup> Yefei He<sup>1</sup>  
 Hong Zhou<sup>1\*</sup> Mike Zheng Shou<sup>3\*</sup> Chunhua Shen<sup>1</sup>

<sup>1</sup>Zhejiang University, China

<sup>2</sup>University of Chinese Academy of Sciences, China

<sup>3</sup>National University of Singapore, Singapore

## Abstract

Current deep networks are very data-hungry and benefit from training on large-scale datasets, which are often time-consuming to collect and annotate. By contrast, synthetic data can be generated infinitely using generative models such as DALL-E and diffusion models, with minimal effort and cost. In this paper, we present DatasetDM, a generic dataset generation model that can produce diverse synthetic images and the corresponding high-quality perception annotations (*e.g.*, segmentation masks, and depth). Our method builds upon the pre-trained diffusion model and extends text-guided image synthesis to perception data generation. We show that the rich latent code of the diffusion model can be effectively decoded as accurate perception annotations using a decoder module. Training the decoder only needs less than 1% (around 100 images) manually labeled images, enabling the generation of an infinitely large annotated dataset. Then these synthetic data can be used for training various perception models for downstream tasks.

To showcase the power of the proposed approach, we generate datasets with rich dense pixel-wise labels for a wide range of downstream tasks, including semantic segmentation, instance segmentation, and depth estimation. Notably, it achieves 1) state-of-the-art results on semantic segmentation and instance segmentation; 2) significantly more robust on domain generalization than using the real data alone; and state-of-the-art results in zero-shot segmentation setting; and 3) flexibility for efficient application and novel task composition (*e.g.*, image editing). The project website and code can be found at [Website](#) and [DatasetDM](#), respectively.

## 1 Introduction

Modern deep-learning models for perception tasks often require a large amount of labeled data to achieve good performance. Unfortunately, collecting large-scale data and labeling the corresponding pixel-level annotations is a time-consuming and expensive process. For example, collecting images of urban driving scenes requires physical car infrastructure, and labeling a segmentation annotation for a single urban image in Cityscapes [13] can take up to 60 minutes. Moreover, in certain specialized domains, such as medical or human facial data, collecting relevant information can be challenging or even impossible, owing to privacy concerns or other factors. The above challenges can be a barrier to advancing artificial intelligence in computer vision.

To reduce costs, many previous researchers have primarily focused on weakly supervised [64] and unsupervised solutions [59] to address the problem. For instance, certain segmentation priors [1, 2, 39] use weak or inexpensive labels to train robust segmentation models. With the advancement of

\*H. Zhou and M. Shou are corresponding authors.

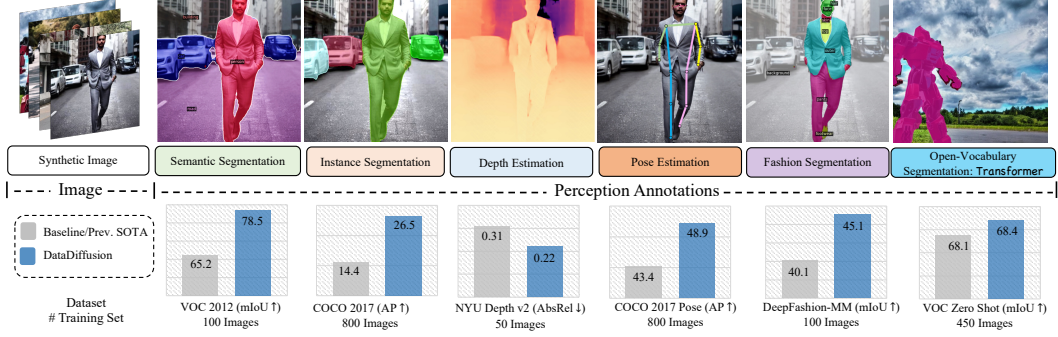


Figure 1: **Synthetic Data from DatasetDM.** The high-quality and infinitely numerous synthetic images with perception annotation can yield significant improvements for various downstream tasks.

generative models, such as DALL-E [48], Stable Diffusion [49], some researchers have begun to explore the potential of synthetic data, attempting to use it to assist in training models, or even replace real data for perception task. Most works focus on classification, face recognition [41, 57, 28], salient object detection [61] and segmentation tasks [38, 60, 65], with only a minority trying to address problems such as human pose estimation [22] or medical image analysis [27]. In the era of GANs, DatasetGAN [65] and BigDatasetGAN [36] are recognized as pioneering works that utilize the feature space of pre-trained GANs and design a shallow decoder for generating pixel-level labels in the context of segmentation tasks. Following the two works, Hands-off [62] extends this approach to multitasking scenarios, such as depth estimation. However, these methods still suffer from three major drawbacks: 1) Due to the limitations of the representation ability of early (up to 2020) GAN models, the quality of the synthesized data is often dissatisfactory, leading to an inferior performance on downstream tasks. 2) These methods primarily focus on independent downstream tasks and no one tries to explore a unified data synthesis paradigm with a generalized decoding framework. 3) The training cost is still relatively high, while these methods do not make full use of the visual knowledge contained within the latent codes of powerful text-to-image models.

By leveraging large-scale datasets of image-text pairs (*e.g.*, LAION5B [53]), recent text-to-image diffusion models (*e.g.*, Stable Diffusion [49]) present phenomenal power in generating diverse and high-fidelity images with rich texture, diverse content, and reasonable structures. The phenomenon suggests that large text-to-image diffusion models can *implicitly* learn valuable, and rich high-level and low-level visual representations from massive image-text pairs. It is natural ask: *Can we leverage the knowledge learned by these models to generate perception annotations and extend the paradigm of text-to-image generation to text-to-data generation?*

In this paper, built upon the powerful text-to-image diffusion model, we present DatasetDM, a generalized dataset generation model that can produce an unlimited number of synthetic images and the corresponding perception annotations, as shown in Fig. 1. The key to our approach is a unified perception decoder, namely P-Decoder, that decodes the latent code of the diffusion model to perception annotations. Inspired by align/instruct tuning from LLM [58], a method for inducing output following capability with minimal human-labeled, we only use less than 1% manually labeled images to train the decoder, enabling infinite annotated data generation. The generated data can subsequently be utilized to train any perception models for various downstream tasks, including but not limited to segmentation, depth estimation, and pose estimation. To maintain the robust image generation capabilities of the diffusion model, we freeze the model weights and use image inversion to extract the latent code, *i.e.*, multi-scale features, which are then fed into the unified P-Decoder. The P-Decoder accommodates various downstream tasks within a unified transformer framework, with only minor variations as depicted in Fig. 3.

To summarize, our contributions are four-fold:

- We introduce DatasetDM: a versatile dataset generation model featuring a perception decoder capable of producing an unlimited quantity of high-fidelity synthetic images, along with various perception annotations, including depth, segmentation, and human pose estimation.
- Visual align/instruct tuning, a method for inducing output following capability with minimal human-labeled data. With less than 1% of the original dataset, *i.e.*, around 100 images,

DatasetDM pushes the limits of text-to-image generation, pioneering a novel paradigm: *text-to-data generation*. This breakthrough is made possible by leveraging the rich latent code of the pre-trained diffusion model.

- Experiments demonstrate that the existing perception models trained on synthetic data generated by DatasetDM exhibit outstanding performance on **six** datasets across **five** different downstream tasks. For instance, the synthetic data delivers remarkable gains of 13.3% mIoU and 12.1% AP for semantic segmentation on VOC 2012 and instance segmentation on COCO 2017, respectively.
- Text-guided data generation allows for the generation of a diverse range of data, which has been shown to provide a more robust solution for domain generalization and long-tail data distribution. Moreover, DatasetDM offers a flexible approach for novel task composition, as exemplified by its ability to facilitate image editing (see Fig. 4).

## 2 Related Work

**Text-to-Image Generation.** Several mainstream methods exist for the task, including Generative Adversarial Networks (GANs)[24], Variational Autoencoders (VAEs)[34], flow-based models [16], and Diffusion Probabilistic Models (DPMs) [55, 49, 31, 25]. Recently, as a likelihood-based method, diffusion models have gained significant attention with promising generation abilities. They match the underlying data distribution by learning to reverse a noising process. Thanks to the high-quality image synthesis and the stability of training, diffusion models are quickly emerging as a new frontier [49, 48, 32, 51] in the field of image synthesis.

Text-guided diffusion models are used for text-conditional image generation, where a text prompt  $\mathcal{P}$  guides the generation of content-related images  $\mathcal{I}$  from a random Gaussian noise  $z$ . Visual and textual embeddings are typically fused using cross-attention. Recent large text-to-image diffusion models, such as Stable Diffusion [49] of Stability AI, DALL-E2 [48] of OpenAI, and Imagen [51] of Google, have shown powerful performance in generating diverse and high-fidelity images with rich textures and reasonable structures. Their impressive synthesis capabilities suggest that these models can implicitly learn valuable representations with the different semantic granularity from large-scale image-text pairs. In this paper, we leverage these learned representations (latent codes), to extend the paradigm of text-to-image generation to text-to-data generation.

**Synthetic Datasets for Perception Task.** Previous studies in dataset synthesis, such as Virtual KITTI [21] and Virtual KITTI 2 [8], primarily rely on 3D computer simulations to address standard 2D computer vision problems, such as object detection [46], scene understanding [52], and optical flow estimation [7]. However, these methods are limited by the domain of 3D models and cannot be generalized to arbitrary scenes and open-set categories. For instance, Virtual KITTI is exclusively focused on autonomous driving scenes and supports only 20 commonly occurring categories, which cannot be extended to open-scene domains like the COCO benchmark [40].

In contrast, synthetic data generated using generation models (*i.e.*, GAN [24, 42] and Diffusion Model [55]) are more flexible and can support a wider range of tasks and open-world scenes for various downstream tasks, such as classification task [28], face recognition [28], salient object detection [61], semantic segmentation [36, 65, 4, 67], and human pose [22]. Inspired by the success of large-scale generative models, such as Stable Diffusion [49], trained on massive datasets like LAION5B [53], recent studies have begun to explore the potential of powerful pre-trained diffusion generative models. Li *et al.* [38] utilized Stable Diffusion and Mask R-CNN pre-trained on the COCO dataset [40] to design and train a grounding module for generating images and semantic segmentation masks. DiffuMask [60] produces synthetic image and semantic mask annotation by exploiting the potential of the cross-attention map between text and image from the text-supervised pre-trained Stable Diffusion model. In this paper, we take a further step by utilizing a generalized perception decoder to parse the rich latent space of the pre-trained diffusion model, enabling the generation of perception for a variety of downstream tasks, including depth, segmentation, and human pose.

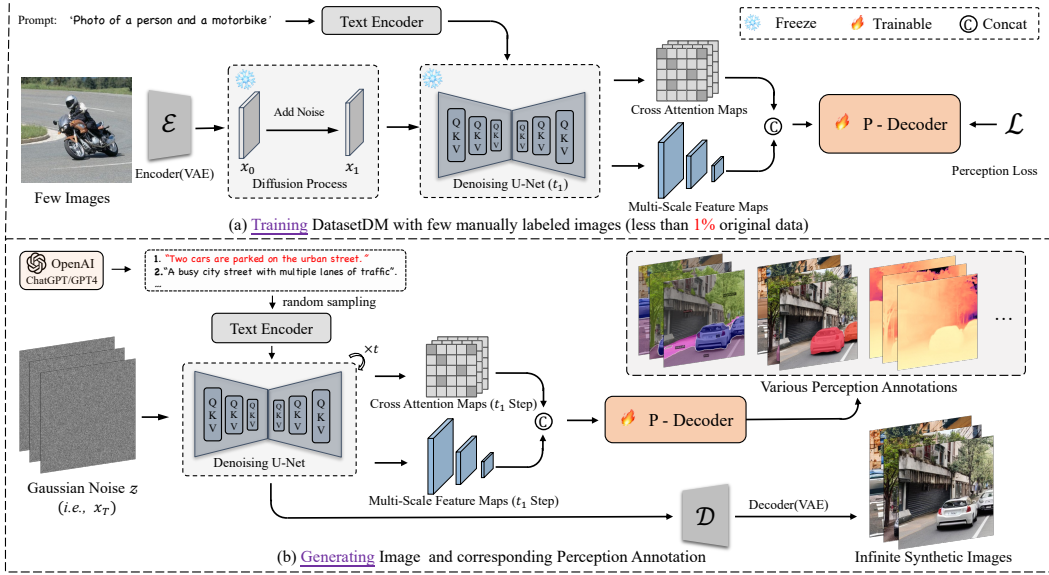


Figure 2: **The overall framework of DatasetDM.** DatasetDM consists of two main steps: 1) Training. Using diffusion inversion to extract the latent code from a small amount of data and then train the perception decoder. 2) Text-guided data generation. A large language model such as GPT-4 is utilized to prompt infinite and diverse data generation for various downstream tasks.

### 3 Methodology

#### 3.1 Formulation

Given a language prompt  $\mathcal{S}$ , text-guided diffusion models generate content-related images  $\mathcal{J} \in \mathcal{R}^{H \times W \times 3}$  from a random Gaussian noise  $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . The standard text-guided image denoising processing can be formulated as:  $\mathcal{J} = \Phi_{\text{T2I}}(z, \mathcal{S})$ , where  $\Phi_{\text{T2I}}(\cdot)$  refers to a pre-trained text-to-image diffusion model. In this paper, we adopt Stable Diffusion [49] as the base for the diffusion model, which consists of three components: a text encoder  $\tau_{\theta}(\cdot)$  for embedding prompt  $\mathcal{S}$ ; a pre-trained variational autoencoder (VAE) [18] that encodes  $\mathcal{E}(\cdot)$  and decodes  $\mathcal{D}(\cdot)$  latent code of images; and a time-conditional UNet ( $\epsilon_{\theta}(\cdot)$ ) [50] that gradually denoises the latent vectors. To fuse visual and textual embeddings, cross-attention layers are typically used in the UNet for each denoising step. The denoising process is modeled as a Markov chain:  $x_{t-1} = f(x_t, \epsilon_{\theta})$ , where  $x_t$  denote latent code at timestep  $t$ , and  $t \in [1, T]$ . The latent noise at the final timestep  $T$ , denoted as  $x_T$ , is equivalent to the random Gaussian noise  $z$ .  $f(\cdot)$  is the denoising function [31].

In this paper, we design a perception decoder that can effectively parse the latent space of the UNet  $\epsilon_{\theta}(x_t, t, \tau_{\theta}(\mathcal{S}))$ . By doing so, we extend the *text-to-image* generation approach to a *text-to-data* paradigm:

$$\{\mathcal{J}, \mathcal{P}_{1:k}\} = \Phi_{\text{T2D}}(z, \mathcal{S}), \quad (1)$$

where  $\mathcal{P}_{1:k}$  denotes the corresponding perception annotations, and  $k$  is the number of the supported downstream tasks. In fact, the paradigm can support any image-level perception task, such as semantic segmentation, instance segmentation, pose estimation, and depth estimation.

#### 3.2 Method Overview

This paper introduces a novel paradigm called *text-to-data generation*, which extends text-guided diffusion models trained on large-scale image-text pairs. Our key insight is that using a small amount of real data (using less than 1% existing labeled dataset) and a generic perception decoder to interpret the diffusion latent spaces, results in the generation of infinite and diverse annotated data. Then the synthetic data can be used to train any existing perception methods and apply them to real images.

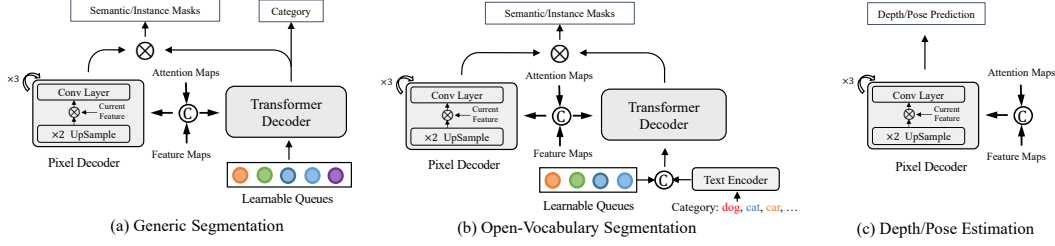


Figure 3: **Various types of tasks with our proposed P-Decoder.** The proposed decoder is a generalized architecture for the six supported tasks, with only minor variations required for different downstream applications, *i.e.*, determining whether to activate certain layers.

The proposed DatasetDM framework, presented in Fig. 2, comprises two stages. 1) The first stage—**Training**—involves using diffusion inversion (§3.3) to obtain the latent code of the real image and extract the text-image representations (§3.3). These representations and their corresponding annotations are then used to train the perception decoder (§3.4). 2) The second stage—**Inference** (§3.5)—uses GPT-4 to enhance the diversity of data and generates abundant images, while the P-Decoder produces corresponding perception annotations such as masks and depth maps.

### 3.3 Hypercolumn Representation Extraction

The first step in the training stage of DatasetDM is to extract the hypercolumn representation of real images from the latent space of the diffusion model, as shown in Fig. 2(a). To achieve this, we employ the diffusion inversion technique, which involves adding a certain level of Gaussian noise to the real image and then extracting the features from the UNet during the denoising process.

**Image Inversion for Diffusion Model.** Give a real image  $\mathcal{X} \in \mathcal{R}^{H \times W \times 3}$  from the training set, the diffusion inversion (diffusion forward process) is a process that approximates the posterior distribution  $q(\mathbf{x}_{1:T}|\mathbf{x}_0)$ , where  $\mathbf{x}_0 = \mathcal{E}(\mathcal{X})$ . This process is not trainable and is based on a fixed Markov chain that gradually adds Gaussian noise to the image, following a pre-defined noise schedule  $\beta_1, \dots, \beta_T$  [31]:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \quad (2)$$

where  $t$  represents the  $t$ -th time step, and we set it to 1 for a single forward pass in our paper. A single forward pass for visual representation extraction usually provides two advantages, *i.e.*, faster convergence and better performance [63].

**Text-Image Representation Fusion.** With the latent code  $\mathbf{x}_t$  of the real image and the corresponding language prompt  $\mathcal{S}$ , we extract the multi-scale feature maps and cross attention maps from the UNet  $\epsilon_\theta$  as follows:

$$\{\mathcal{F}, \mathcal{A}\} = \epsilon_\theta(\mathbf{x}_t, t, \tau_\theta(\mathcal{S})), \quad (3)$$

where  $\mathcal{S}$  for training set is simply defined using a template “a photo of a  $[\text{CLS}]_1, [\text{CLS}]_2, \dots$ ”. During the data generation phase, GPT-4 is used to provide diverse prompt languages. The variable  $\mathcal{F}$  denotes the multi-scale feature maps from four layers of the U-Net, corresponding to four different resolutions, *i.e.*,  $8 \times 8$ ,  $16 \times 16$ ,  $32 \times 32$ , and  $64 \times 64$ , as illustrated in Fig. 2. Additionally,  $\mathcal{A}$  represents the cross-attention maps of text-to-image from the 16 cross-attention layers in the U-Net, which implement the function  $\mathcal{A} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$ , where  $d$  is the latent projection dimension. We group the 16 cross-attention maps into 4 groups with the same resolutions, and compute their average within each group, which results in the average cross-attention maps  $\hat{\mathcal{A}}$ .

Prior works [60, 66, 30] have proved the effectiveness of class-discriminative and localization of cross-attention map between the visual embedding and the conditioning text features. Thus we concatenate the cross-attention maps  $\hat{\mathcal{A}}$  and the multi-scale feature maps  $\mathcal{F}$  to obtain the final extracted hyper-column representation, and further use a  $1 \times 1$  convolution to fuse them:  $\hat{\mathcal{F}} = \text{Conv}([\mathcal{F}, \hat{\mathcal{A}}])$ .



Figure 4: **Collaboration between GPT-4 and Diffusion Model.** Large language models, *e.g.*, GPT-4, can provide diverse and unrestricted prompts, enhancing the diversity of generated images.

### 3.4 Perception Decoder

The P-Decoder is utilized to translate the representation  $\hat{\mathcal{F}}$  into perception annotations, which are not limited to a specific type for each downstream task. To achieve this goal, inspired by previous works [10, 69], we devised a generalized architecture. This architecture is depicted in Fig. 3, with only *minor variations* (*i.e.*, whether to startup some layers) for each downstream task. For example, the pixel decoder and transformer decoder are required for generic segmentation, while only the pixel decoder is necessary for depth and pose estimation.

**Generic Image Segmentation.** In Fig. 3-(a), we present the adaptation for semantic and instance segmentation tasks, which includes two components: the pixel decoder and the transformer decoder. Similar to Mask2former [10], the transformer decoder comprises a stack of transformer layers with cross-attention and self-attention. This component refines the queries and renders the outputs. The pixel decoder is made up of multiple upsampling CNN layers, and it is used to obtain per-pixel embeddings. Given the representation  $\hat{\mathcal{F}}$  and  $N$  learnable queues  $\{Q_0, Q_1 \dots Q_T\}$  as input, the decoder outputs  $N$  binary masks  $\mathbf{O} = o_1, \dots, o_N \in \{0, 1\}^{N \times H \times W}$ , along with the corresponding category. This is achieved through simple matrix multiplication between the outputs of the transformer decoder and the pixel decoder. Following Mask2former [10], one query is responsible for a class or instance for semantic and instance segmentation, respectively. To optimize the mask prediction, we use the binary cross-entropy loss [11] and dice loss [44]. On the other hand, the cross-entropy loss is used for classification prediction.

**Open-Vocabulary Segmentation.** Based on the architecture of generic segmentation, a semantics-related query concept is proposed to support the open-vocabulary segmentation task, as shown in Fig 3-(b). During the training stage, given the category name  $\{C_0, C_1 \dots C_K\}$  (*e.g.*, dog, cat) on an image from the training set, we use the text encoder  $\tau_\theta(\cdot)$ , *e.g.*, CLIP [47] to encode them into the embedding space. Then concatenating them with the queries to equip the semantics of class vocabulary to the query as follows:

$$\hat{Q}_i = \text{MLP}([Q_i, \tau_\theta(C_j)]), \quad (4)$$

where  $Q_i$  and  $C_j$  is the  $i$ -th query embedding and  $j$ -th class. MLP refers to a learned MLP, used to fuse the class embedding and learnable query embedding. Thus DatasetDM can generate an open-vocabulary mask by incorporating a new class name, as illustrated in Fig. 5 (b).

**Depth and Pose Estimation.** For depth and pose estimation, the output format is predetermined, eliminating the need to differentiate between classes or instances. In this context, the pixel decoder is only required to predict a fixed number of maps  $\mathbf{O} \in \mathcal{R}^{M \times H \times W}$ . The value of  $M$  is set to either 1 or 17 (corresponding to 17 human key points), depending on whether the task is depth or human pose estimation. For human pose estimation, we use the mean squared error as the loss function, and the ground truth heatmaps are generated by applying a 2D Gaussian with a standard deviation of 1 pixel centered on each key point. As for depth estimation, we update the loss function from the classic scale-invariant error [37, 17].



Figure 5: **Examples of generated data for DatasetDM.** Our method can produce semantic/instance segmentation, depth, and human pose annotation across various domains.

method	VOC (Semantic Seg.)/%			COCO2017 (Instance Seg.)/%			NYU Depth V2 (Depth Est.)			COCO2017 (Pose Est.)/%		
	# real	# synth.	mIoU	# real	# synth.	AP	# real	# synth.	REL ↓	# real	# synth.	AP
Baseline	100	-	65.2	400	-	14.4	50	-	0.31	800	-	42.4
<b>DatasetDM</b>	100	40k	78.5	400	80k	26.5	50	35k	0.21	800	80k	47.5

Table 1: **Downstream Tasks.** ‘real’ and ‘synth.’ denote real and synthetic images, respectively. The backbones of baselines for four tasks are ‘Swin-B’, ‘Swin-B’, ‘Swin-L’, and ‘HR-W32’, respectively.

### 3.5 Prompting Text-Guided Data Generation

In Fig. 2 (b), we present the inference pipeline for text-guided data generation. There are two main differences compared to the training phase: firstly, the prompts come from a large language model instead of a fixed template, and secondly, the denoising process is extended to  $T$  steps to obtain the synthetic images. Large language model, *i.e.*, GPT-4, is adopted to enhance the diversity of generative data, while recent works [20, 29, 5] have proven their powerful understanding and adaptability for the real world. As shown in Fig. 4, instead of the template-based prompts from humans, we guide GPT-4 to produce diverse, and infinite prompts. For different downstream tasks and datasets, we give different guided prompts for GPT-4. For example, as for the urban scene of Cityscapes [13], the simple guided prompt is like ‘Please provide 100 language descriptions of urban driving scenes for the Cityscapes benchmark, containing a minimum of 15 words each. These descriptions will serve as a guide for Stable Diffusion in generating images.’ In this approach, we collected  $L$  text prompts, which average around 100 prompts for each dataset. For each inference, a random prompt is sampled from this set.

## 4 Experiments

### 4.1 Implementation details

**Architecture and Training Details.** Stable diffusion V1 [49] model pre-trained on the LAION5B [53] dataset is used as our text-to-image diffusion model. The decoder architecture of Mask2Former [10] was selected as the base architecture for our P-Decoder. And we use 100 queries for the segmentation task. For all tasks, we train DatasetDM for around 50k iterations with images of size  $512 \times 512$ , which only need one Tesla V100 GPU, and lasted for approximately 20 hours. Optimizer [43] with a learning rate of 0.0001 is used.

**Downstream Tasks Evaluation.** To comprehensively evaluate the generative image of DatasetDM, we conduct seven groups of experiments for the supported six downstream tasks. *Semantic Segmentation.* Pascal-VOC 2012 [19] (20 classes) and Cityscapes [14] (19 classes), as two classical benchmark are used to evaluate. We synthesized  $2k$  images for each class in both datasets, resulting in a total of 40k and 38k synthetic images for Pascal-VOC 2012 and Cityscapes, respectively. The synthetic

method	backbone	# real image	# synthetic image	AP	AP <sup>50</sup>	AP <sup>75</sup>	AP <sup>S</sup>	AP <sup>M</sup>	AP <sup>L</sup>
Baseline	R50	400	-	4.4	9.5	3.5	1.1	3.3	12.1
<b>DatasetDM</b>	R50	-	80k (R:400)	12.2	24.3	10.9	1.6	11.3	30.9
<b>DatasetDM</b>	R50	400	80k (R:400)	14.8	29.7	13.0	2.3	15.1	36.0
Baseline	Swin-B	400	-	11.3	23.0	9.6	3.2	10.1	27.1
<b>DatasetDM</b>	Swin-B	-	80k (R:400)	17.6	34.1	15.8	3.4	17.8	39.5
<b>DatasetDM</b>	Swin-B	400	80k (R:400)	23.3	43.0	22.2	7.7	26.1	48.7
Baseline	Swin-B	800	-	14.4	28.8	12.7	5.6	15.7	29.2
<b>DatasetDM</b>	Swin-B	800	80k (R:800)	26.5	46.9	25.8	7.7	29.8	53.3

Table 2: **Instance segmentation on COCO val2017.** ‘R: ’ denotes the real data used to train.

method	backbone	# real image	# synthetic image	Sampled Classes for Comparison/%					mIoU
				bird	cat	bus	car	dog	
Baseline	R50	100	-	54.8	53.3	69.3	66.8	24.2	43.4
<b>DatasetDM (ours)</b>	R50	-	40k (R:100)	84.7	74.4	86.0	79.2	63.7	60.3
<b>DatasetDM (ours)</b>	R50	100	40k (R:100)	81.7	82.3	87.7	77.9	69.3	66.1
Baseline	Swin-B	100	-	54.4	68.3	86.5	71.8	49.1	65.2
<b>DatasetDM (ours)</b>	Swin-B	-	40k (R:100)	93.4	94.5	93.8	78.8	79.6	73.7
<b>DatasetDM (ours)</b>	Swin-B	100	100 (R:100)	83.9	71.0	82.9	78.0	39.5	67.9
<b>DatasetDM (ours)</b>	Swin-B	100	400 (R:100)	86.9	92.0	90.8	82.6	86.7	76.1
<b>DatasetDM (ours)</b>	Swin-B	100	40k (R:100)	86.7	93.8	92.3	88.3	87.1	78.5
Baseline	Swin-B	11.5k (full)	-	93.7	96.5	90.6	88.6	95.7	84.3
<b>DatasetDM (ours)</b>	Swin-B	11.5k (full)	40k (R:100)	93.9	97.6	91.9	89.4	96.1	85.4

Table 3: **Semantic segmentation on VOC 2012.** ‘R: ’ refers to the number of real data used to train.

data is subsequently utilized to train Mask2former [10] and compared to its real data counterpart on a limited dataset setting (around 100 images). *Instance Segmentation.* For the COCO2017 [40] benchmark, we synthesized  $1k$  images for each class, resulting in a total of  $80k$  synthetic images. Similarly, Mask2former [10], as the baseline, is used to evaluate the synthetic data. We evaluate only the class-agnostic performance, where all the 80 classes are assigned the same class ID. *Depth Estimation.* We synthesized a total of  $80k$  synthetic images for NYU Depth V2 [54]. And using Depthformer [37]<sup>2</sup> to evaluate our synthetic data. *Pose Estimation.* We generated a set of  $30k$  synthetic images for COCO2017 Pose dataset [40] and employed HRNet [56] as the baseline model to assess the effectiveness of our approach. *Zero-Shot Semantic Segmentation.* Following Li *et al.* [38], Pascal-VOC 2012 [19] (20 classes) is used to evaluate. We train DatasetDM with only 15 seen categories, and synthesized a total of  $40k$  synthetic images for 20 categories. *Human Semantic Segmentation.* We synthesized a total of  $20k$  synthetic images for DeepFashion-MM [33] (24 classes). Mask2former [10] is used to evaluate the synthetic data. We split DeepFashion-MM into a set of 100 training images, and 12,500 testing images. Further details, such as prompts, can be found in the supplementary material.

## 4.2 Main Results

Tab. 1 provides a basic comparison of the selected four downstream tasks. More additional experiments can be found in Tab. 2 and Tab. 3, as well as in the Supplementary Material (*i.e.*, Pose Estimation, Depth Estimation, Zero-Shot Segmentation, Fashion Segmentation, and others).

**Semantic Segmentation.** Tab. 3 displays the performance on VOC 2012. Using only 100 real images (5 images per class), training with purely synthetic data from DatasetDM achieves a 73.7% mIoU, an 8.5% improvement compared to using real data alone. Moreover, when jointly training with the 100 real images, further improvement is observed, resulting in a mIoU of 78.5%.

**Instance Segmentation.** Tab. 2 presents three distinct training settings, encompassing variations in the backbone and the number of training images. Regardless of the chosen setting, DatasetDM consistently achieves an improvement of approximately 10%. Employing 800 training images (10 images per class) and the Swin-B backbone, DatasetDM yields a 12.1% increase in Average Precision (AP), resulting in a final AP of 26.5%.

**Depth Estimation.** Tab. 1 presents a concise comparison between synthetic and real data on the NYU Depth V2 dataset [54]. Detailed information (*e.g.*, backbone, other metrics) can be found in the

<sup>2</sup><https://github.com/zhyever/Monocular-Depth-Estimation-Toolbox>

step	car	dog	mIoU	step	car	dog	mIoU	# train im.	syn.	joint	prompt (# num.)	car	dog	mIoU
1	88.3	87.1	78.5	-	88.1	87.0	77.6	60	71.4	77.1	Human (100)	84.9	84.5	76.6
100	88.1	87.2	78.5	1	88.3	87.1	78.5	100	73.7	78.5	GPT-4 (100)	85.2	86.1	77.1
200	88.0	86.6	78.3	200	87.7	87.0	78.0	200	74.4	79.4	GPT-4 (200)	88.0	86.2	77.3
500	87.2	84.9	76.8	500	87.2	86.3	77.5	400	76.4	80.4	GPT-4 (500)	88.1	87.1	78.5
800	86.3	83.4	76.1	800	87.1	86.1	77.1	1,000	78.4	81.1	GPT-4 (1k)	88.3	87.1	78.5

(a) **Visual Features  $\mathcal{F}$ .** (b) **Cross Attention  $\hat{\mathcal{A}}$ .** (c) **Size of Train Set.** (d) **Prompt Candidates.**

Table 4: **DatasetDM Ablation on Pascal-VOC 2012 for semantic segmentation.** Swin-B is used as the backbone. 100 real images are used for (a), (b), and (d). ‘Syn.’ and ‘Joint’ denote training with only synthetic data and joint training with real data, respectively.

supplementary material. When trained with 50 images, DatasetDM can achieve a 10% improvement compared to training solely with real images.

**Human Pose Estimation.** For the human pose estimation task on the COCO2017 dataset, DatasetDM demonstrates significant improvements compared to the baseline trained on 800 real images, achieving a 5.1% increase, as illustrated in Tab. 1. Similar to depth estimation, additional information can be found in the supplementary material.

**Zero Shot and Long-tail Segmentation.** Tab. 5 displays the results of experiments related to zero-shot and long-tail segmentation. Our model, DatasetDM, notably alleviates the effects of long-tail distribution by synthesizing a substantial amount of data for rare classes, leading to an improvement of up to 20% in mIoU. Details for both tasks can be found in the supplementary material.

### 4.3 Ablation studies

Method	Zero-Shot Setting			Long-tail Setting		
	seen	unseen	harm.	head	tail	mIoU
Baseline(no Syn.)	61.3	10.7	18.3	61.2	44.1	52.6
DiffuMask [60]	71.4	65.0	68.1	-	-	-
DatasetDM	78.8	60.5	68.4	73.1	66.4	70.0

Table 5: **Zero Shot and Long-tail Segmentation on VOC 2012.** Zero Shot: following priors [38, 60], we train DatasetDM with only 15 seen categories, and tested for 20 categories. Long-tail Setting: the 20 categories are divided into head (10 classes, 20 images each class) and tail classes (10 classes, 2 images each class).

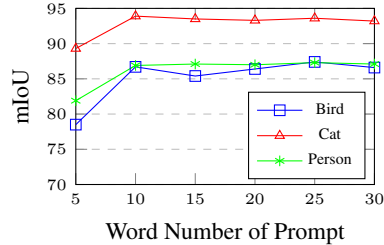


Figure 6: **Ablation of Prompt Length.**

**Diffusion Time Steps.** Tab. 4a depicts the influence of visual features derived from various diffusion time steps, and the maximum sampling step is 1000. We observe that a large step results in adverse outcomes, whereas the performance with a smaller step tends to remain relatively stable.

**Cross-Attention Fusion.** Fig. 4b demonstrates that the cross attention maps,  $\mathcal{F}$  in step 1 can yield a modest improvement, roughly 1%. Interestingly, it appears that as the step size increases, the benefit becomes less pronounced. Indeed, when the step size surpasses 500, it may even result in detrimental effects. From this perspective, the utility of the cross-attention map is limited.

**Training Set Size.** Additional training data for DatasetDM can further improve synthetic data, as shown in Tab. 4c. The increase of the training data from 60 to 400 images precipitates the most conspicuous improvement, subsequently reaching a saturation point. With 1k training images, the performance escalates to an impressive 81%, demonstrating competitive prowess for the application. Notably, 1k training images representing roughly 10% of the original data, is still relatively diminutive.

**Prompt from Language Model.** *Candidate Number.* We also investigated the impact of the number of prompt candidates, as depicted in Tab. 4d. With the current configurations, an increase in the number of prompts can potentially lead to a performance improvement of 2%. *Word Number of Each Prompt.* We simply study the effect of the length of prompt in Fig. 6. An increase in text length from 5 to 10 yields approximately 4% enhancement. When the text length surpasses 10, the performance appears to plateau. We argue that the upper limit is due to the current capacity of generative models.

Task	Dataset	Full Real Data	Used for DatasetDM	# synthetic image
Instance Segmentation	COCO 2017 [40]	118.3k	400 (0.3%)	80k
Semantic Segmentation	VOC 2012 [19]	11.5k	100 (0.87%)	40k
Semantic Segmentation	Cityscapes [13]	2.9k	9 (0.3%)	38k
Semantic Segmentation	DeepFashion-MM [33]	12.7k	120 (0.9%)	38k
Zero-Shot Segmentation	VOC 2012 [19]	11.5k	450 (3.9%)	40k
Depth	NYU Depth V2 [54]	24.2k	50 (0.2%)	35k
Human Pose	COCO 2017-Pose [40]	118.3k	800 (0.6%)	80k

Table 6: **Comparison of Data size.** With less than 1% manually labeled images, DatasetDM can enable the generation of an infinitely large annotated dataset.

method	backbone	input size	# real im.	# synthetic im.	AP	AP <sup>50</sup>	AP <sup>75</sup>	AP <sup>M</sup>	AP <sup>L</sup>	AR
Baseline	R50	256 × 192	800	-	31.3	62.0	27.7	30.7	32.0	36.2
<b>DatasetDM</b>	R50	256 × 192	-	80k (R:800)	11.4	28.2	7.8	6.9	17.7	14.3
<b>DatasetDM</b>	R50	256 × 192	800	80k (R:800)	36.4	66.7	35.0	33.0	40.8	40.1
Baseline	HR-W32	256 × 192	800	-	42.4	73.3	42.1	39.5	47.0	46.7
<b>DatasetDM</b>	HR-W32	256 × 192	-	80k (R:800)	13.4	30.9	9.9	8.0	21.7	17.7
<b>DatasetDM</b>	HR-W32	256 × 192	800	80k (R:800)	47.5	75.6	49.3	44.2	52.6	51.2
Baseline	HR-W32	384 × 288	800	-	43.4	72.2	44.7	40.5	47.9	47.5
<b>DatasetDM</b>	HR-W32	384 × 288	800	80k (R:800)	48.9	76.7	51.4	44.6	55.0	52.4

Table 7: **Human Pose Estimation on COCO va12017.** ‘R:’ refers to the number of real data used for training DatasetDM.

## 5 Conclusion

In this study, we investigate using a perception decoder to parse the latent space of an advanced diffusion model, extending the text-to-image task to a new paradigm: text-guided data generation. Training the decoder requires less than 1% of existing labeled images, enabling infinite annotated data generation. Experimental results show that the existing perception models trained on synthetic data generated by DatasetDM exhibit exceptional performance across six datasets and five distinct downstream tasks. Specifically, the synthetic data yields significant improvements of 13.3% mIoU for semantic segmentation on VOC 2012 and 12.1% AP for instance segmentation on COCO 2017. Furthermore, text-guided data generation offers additional advantages, such as a more robust solution for domain generalization and enhanced image editing capabilities. We hope that this research contributes new insights and fosters the development of synthetic perception data.

## A Implementation details

### A.1 Dataset Details

- Pascal-VOC 2012 [19] (20 classes) is a popular dataset for semantic segmentation in computer vision. It contains thousands of annotated images featuring 20 different object classes, such as animals, vehicles, and furniture.
- Cityscapes [13] (19 classes) is a benchmark dataset for semantic urban scene, primarily focusing on semantic segmentation tasks in computer vision. It contains high-quality pixel-level annotations of 5,000 images from 50 different cities, captured at various times of the day and under diverse weather conditions. There are 2,975 images for training and 500 images for validation.
- COCO 2017 (Common Objects in Context) [40] is a widely-used benchmark dataset for computer vision tasks, such as object detection, segmentation, and human pose estimation. It contains over 200,000 labeled images with 1.5 million object instances belonging to 80 object categories
- NYU Depth V2 [54] is designed for indoor scene understanding tasks in computer vision, specifically for depth estimation task. The NYU Depth V2 dataset contains 1,449 labeled images and 407,024 unlabeled frames, captured from 464 diverse indoor scenes.
- DeepFashion-MM [33] (24 classes) is a benchmark dataset designed for the task of clothing synthesis in the field of computer vision. It consists of 24 different clothing classes.

## A.2 Baseline for Downstream Tasks

- **Semantic/Instance Segmentation.** We use Mask2former [10] as the baseline for comparing synthetic data to real data. We employ the official code<sup>3</sup>, maintaining all network settings, loss functions, and configurations as presented in the original code. To evaluate the effectiveness of synthetic data, we establish three settings: 1) training with purely real data, 2) training with purely synthetic data, and 3) joint training with both synthetic and real data.
- **Open-Vocabulary Semantic Segmentation.** Similar to the generic semantic segmentation, we use Mask2former [10] as the baseline. We train DatasetDM on 15 seen categories and generate a total of 40k synthetic images for 20 categories. Subsequently, we utilize these data to train the Mask2former model and evaluate its performance on the 20 categories of VOC 2012.
- **Depth Estimation.** DepthFormer [37]<sup>4</sup>, serving as the baseline, is employed to assess our approach. We adhere to all network settings, loss functions, configurations, and training strategies outlined in the original implementation.
- **Pose Estimation.** We adopt HRNet [56] and its official code<sup>5</sup> for evaluating the pose estimation task on synthetic data generated by DatasetDM. Currently, we focus on single-person scenarios in each synthetic image and guide GPT-4 to generate corresponding images accordingly.

## A.3 Training Setup for DatasetDM

All experiments for training DatasetDM were carried out on a single V100 GPU, while downstream task baselines (*i.e.* Mask2former, Depthformer) were trained using 8 V100 GPUs. Training our DatasetDM for 50k iterations with just one V100 GPU takes merely a day, showcasing its efficacy and efficiency. For all tasks, we employ the Adam optimizer [43] with a learning rate of 0.001 and a batch size of 1. The loss function and data augmentations vary for different tasks.

- **Semantic/Instance Segmentation.** During the training phase of DatasetDM, we primarily utilize two data augmentation techniques: random cropping to a size of  $512 \times 512$  pixels, and random scaling.
- **Depth Estimation.** For depth estimation, we employ four data augmentation methods: random flipping, cropping, brightness-contrast adjustment, and hue-saturation value manipulation.
- **Pose Estimation.** For pose estimation, we use four data augmentation techniques: random scaling, cropping, flipping, and rotation.

## A.4 Details for Training Data of DatasetDM

**Quantities of training read data.** Tab. 6 provides a comprehensive comparison of the quantities of training read data and synthetic data used for each downstream task in this study. Notably, with the exception of the seen class in the zero-shot segmentation setting, training with our DatasetDM requires less than 1% of the available real data. This efficiency potentially reduces the implementation costs of perception algorithms and significantly improves data utilization.

**Class Split for Zero-Shot and Long-Tail Segmentation.** Tab. 9 provides a comprehensive overview of the class distribution for both zero-shot and long-tail scenarios. The division for zero-shot classes is consistent with previous studies [6, 60, 38]. The configuration for long-tail data distribution is firstly established in this paper.

# B Experiments

## B.1 Comparison with Other Data Augmentation Methods.

From a certain perspective, the proposed DatasetDM is more akin to an efficient data augmentation technique, and thus we compare it with some previous data augmentation schemes, as shown

<sup>3</sup><https://github.com/facebookresearch/MaskFormer>

<sup>4</sup><https://github.com/zyeever/Monocular-Depth-Estimation-Toolbox>

<sup>5</sup><https://github.com/HRNet/HRNet-Human-Pose-Estimation>

method	backbone	# real image	# synthetic image	8 Classes/%		19 Classes/%			
				vehicle	mIoU	car	bus	bicycle	mIoU
Baseline	R50	9	100k+ (R:16)	84.3	71.5	82.8	22.3	42.4	36.8
HandsOff [62]	R101	16	100k+ (R:16)	-	55.1	-	-	-	-
HandsOff [62]	R101	50	100k+ (R:16)	-	60.4	-	-	-	-
<b>DatasetDM (ours)</b>	R50	-	38k (R:9)	86.9	69.5	83.3	8.3	53.5	34.2
<b>DatasetDM (ours)</b>	R50	9	38k (R:9)	88.6	76.7	85.6	28.9	56.5	42.1
<b>DatasetDM (ours)</b>	R101	9	38k (R:9)	88.9	77.5	85.9	27.9	60.4	43.7
Baseline	Swin-B	9	-	84.1	74.5	83.3	27.7	42.0	41.1
<b>DatasetDM (ours)</b>	Swin-B	-	38k (R:9)	85.7	73.3	84.3	20.3	29.1	37.3
<b>DatasetDM (ours)</b>	Swin-B	9	38k (R:9)	89.4	80.0	87.2	30.2	66.5	47.4

Table 8: **Semantic segmentation on Cityscapes for two different split settings: 8 and 19 categories.** ‘vehicle’, ‘car’, ‘bus’, and ‘bicycle’ are sampled classes for presentation. ‘R: ’ refers to the number of real data used for training DatasetDM.

Zero-Shot Segmentation		Long-Tail Segmentation	
Seen Class	Unseen Class	Head Class	Tail Class
aeroplane (0), bicycle (1), bird (2), boat (3), bottle (4), bus (5), car (6), cat (7), chair (8), cow (9), diningtable (10), dog (11), horse (12), motorbike (13), person (14)	potted plant (15), sheep (16), sofa (17), train (18), tvmonitor (19)	aeroplane (0), bicycle (1), bird (2), boat (3), bottle (4), bus (5), car (6), cat (7), chair (8), cow (9)	diningtable (10), dog (11), horse (12), motorbike (13), person (14), potted plant (15), sheep (16), sofa (17), train (18), tvmonitor (19)

Table 9: **Details for Zero-Shot and Long-tail Segmentation on VOC 2012 [19].**

in Tab. 11. Compared with flip and color augmentation, DatasetDM demonstrates a substantial advantage, bringing significant improvements, around 10% increase, which is significant for the computer vision community.

## B.2 Ablation Study for Baseline of Downstream Tasks.

In addition, the synthetic data generated by DatasetDM can seamlessly integrate with any existing downstream task model. To substantiate this claim, we tested our model with several other benchmark models, such as DeepLabV3, the results of which are detailed in Tab. 11. Notably, our synthetic data was able to enhance the performance of DeepLabV3 by approximately 10%, underscoring the robustness of our approach.

## B.3 Human Pose Estimation on COCO val2017.

Tab. 7 presents the results of human pose estimation on the COCO 2017 dataset. Following the approach of HRNet [56], we established three distinct experimental settings, including variations in the backbone and input size, to evaluate the synthetic data from our model. Irrespective of the specific setting, our method consistently achieved an improvement of approximately 5% in Average Precision (AP), which is a significant increase. Finally, it is noteworthy that our model attained competitive performance, with an AP of 48.9%, using merely 800 training images.

## B.4 Semantic segmentation on Cityscapes.

To compare with previous methods [62], we conducted experiments with a 9-classes division. Furthermore, we also provided experiments with a 19-class division, which presents a more challenging setting. DatasetDM demonstrates consistent advancements over the baseline or prior state-of-the-art (SOTA), achieving up to a 10% improvement in mean Intersection over Union (mIoU) under each experimental setup.

## B.5 Semantic segmentation on DeepFashion-MM.

Tab. 12 showcases the performance of semantic segmentation on the DeepFashion-MM dataset [33]. Like our other experiments, we have conducted two sets of experiments using different backbones. Regardless of the setup, the joint training with synthetic data consistently outperforms the baseline that uses purely synthetic data, with an approximate improvement of 5% mIoU.

Collapsed label (8)	Cityscapes (Fine annotations) original labels
Void	Unlabeled (0), ego vehicle (1), rectification border (2), out of ROI (3), static (4), dynamic (5), ground (6), sidewalk (8), parking (9), rail track (10)
Road	Road (7)
Construction	Building (11), wall (12), fence (13), guard rail (14), bridge (15), tunnel (16)
Object	pole (17), polegroup (18), traffic light (19), traffic sign (20)
Nature	Vegetation (21), terrain (22)
Sky	Sky (23)
Human	Person (24), rider (25)
Vehicle	UCar (26), truck (27), bus (28), caravan (29), trailer (30), train (31), motorcycle (32), bicycle (33), license plate (-1)

Table 10: **Details for 8 and 19 categories on Cityscapes [13].**

data aug.	baseline	backbone	# real image	# synthetic image	mIoU	Improv.
crop	Mask2former [10]	R50	100	-	41.5	-
flip, crop, color	Mask2former [10]	R50	100	-	43.4	+1.9
crop, DatasetDM	Mask2former [10]	R50	100	40k (R:100)	65.2	+23.7
flip, crop, color, DatasetDM	Mask2former [10]	R50	100	40k (R:100)	66.1	+24.6
crop	Mask2former [10]	Swin-B	100	-	64.1	-
flip, crop, color	Mask2former [10]	Swin-B	100	-	65.2	+1.1
crop, DatasetDM	Mask2former [10]	Swin-B	100	40k (R:100)	77.8	+13.7
flip, crop, color, DatasetDM	Mask2former [10]	Swin-B	100	40k (R:100)	78.5	+14.4
crop	DeepLabV3+ [9]	Mobilenet	100	-	39.1	-
crop, DatasetDM	DeepLabV3+ [9]	Mobilenet	100	40k (R:100)	45.3	+6.2
flip, crop, color	DeepLabV3+ [9]	Mobilenet	100	-	40.5	+1.4
flip, crop, color, DatasetDM	DeepLabV3+ [9]	Mobilenet	100	40k (R:100)	46.1	+7.0
crop	DeepLabV3+ [9]	R50	100	-	45.1	-
crop, DatasetDM	DeepLabV3+ [9]	R50	100	40k (R:100)	55.3	+10.2
flip, crop, color	DeepLabV3+ [9]	R50	100	-	46.3	+1.2
flip, crop, color, DatasetDM	DeepLabV3+ [9]	R50	100	40k (R:100)	56.9	+11.8

Table 11: **Ablation Study for Baseline and Data Augmentation.** ‘R: ’ refers to the number of real data used to train. ‘crop’, ‘flip’, and ‘color’ refer to the ‘random crop’, ‘random horizontal flip’, and ‘color augmentation’, respectively.

## B.6 Depth Estimation on NYU Depth V2 val dataset.

Tab. 13 presents the depth estimation experiment conducted on the NYU Depth V2 validation dataset [54]. Two training strategies have been devised based on variations in the training data. Independent of the data volume utilized, our approach consistently yields substantial improvements, specifically 0.1 and 0.02 respectively.

## B.7 Zero-Shot Semantic Segmentation on VOC 2012

Consistent with preceding studies [38, 60], we conduct an experiment on zero-shot (open-vocabulary) semantic segmentation tasks using the VOC 2012 dataset [19]. Tab. 14 offers a comparative analysis with existing approaches to zero-shot semantic segmentation. In this experiment, our model is trained on a mere 450 images, with 30 images allocated for each of the 15 seen classes, and testing is conducted across all 20 categories. Despite the limited dataset in comparison to the complete set of 10.6k images, our model continues to exhibit competitive performance. In relation to methods employing synthetic data, our model achieves state-of-the-art (SOTA) performance, reaching 68.4% mIoU.

## B.8 Domain Generalization across Different Domains

Following DiffuMask [60], we further assess the domain generalization capabilities of synthetic data produced by DatasetDM, as depicted in Fig. 15. When compared with the previous state-of-the-art (SOTA) method, DatasetDM demonstrates superior effectiveness in domain generalization. For instance, DatasetDM achieves a score of 73.6%, as opposed to DiffuMask’s score of 69.5% on the VOC 2012 val set. Compared to real data, DatasetDM exhibits enhanced robustness in terms of generalization. It is reasonable that synthetic data exhibits greater diversity, especially when integrated with language models, as shown in Fig. 10. In terms of diversity and robustness, it far surpasses real datasets.

method	backbone	# real image	# synthetic image	Sampled Classes for Comparison/%					mIoU
				outer	dress	headwear	belt	socks	
Baseline	R50	100	-	58.2	65.2	19.2	24.3	0	31.2
<b>DatasetDM (ours)</b>	R50	-	40k (R:100)	53.1	57.2	0.4	0.4	0	28.9
<b>DatasetDM (ours)</b>	R50	100	40k (R:100)	53.1	59.3	34.3	59.1	3.2	36.7
Baseline	Swin-B	100	-	58.1	56.1	64.3	33.4	7.2	40.1
<b>DatasetDM (ours)</b>	Swin-B	100	40k (R:100)	70.0	70.8	72.0	32.8	5.9	45.1

Table 12: **Semantic segmentation on DeepFashion-MM [33].** ‘R: ’ refers to the number of real data used to train.

method	backbone	# real image	# synthetic image	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL $\downarrow$	Sq REL $\downarrow$	RMS $\downarrow$	RMS log $\downarrow$
Baseline	Swin-L	50	-	0.59	0.84	0.93	0.31	0.37	0.81	0.30
<b>DatasetDM</b>	Swin-L	-	35k (R:50)	0.68	0.90	0.97	0.22	0.19	0.60	0.23
<b>DatasetDM</b>	Swin-L	50	35k (R:50)	0.68	0.91	0.98	0.21	0.18	0.63	0.23
Baseline	Swin-L	250	-	0.79	0.96	0.99	0.16	0.11	0.51	0.19
<b>DatasetDM</b>	Swin-L	-	35k (R:250)	0.78	0.96	0.99	0.17	0.11	0.52	0.19
<b>DatasetDM</b>	Swin-L	250	35k (R:250)	0.80	0.97	0.99	0.14	0.09	0.47	0.18

Table 13: **Depth Estimation on NYU Depth V2 val dataset.** Measurements are made for the depth range from  $0m$  to  $10m$ .

## B.9 More Qualitative Results

To demonstrate the high-quality synthetic data, we visualized synthetic data from two domains: human-centric and urban city, as shown in Fig. 7 (human-centric) and Fig. 8 (urban city scenario). The human-centric domain predominantly encompasses datasets related to human activity, such as COCO 2017, Cityscapes, and DeepFashion-MM. On the other hand, the urban city scenario pertains specifically to datasets like Cityscapes and COCO 2017. To the best of our knowledge, our work is the first to support multi-task synthesis of data. We believe that unified annotation synthesis is meaningful and can support interactions between different modalities. Recent works, *e.g.* ImageBind [23] have already demonstrated its feasibility and necessity. Our method also has many advantages, such as the ability to custom design datasets for a specific domain or to address bad case scenarios, and it is particularly effective in solving problems related to long-tail data distribution. This is straightforward; we can achieve it simply by adjusting our prompts.

## C Details on the Architecture of Perception Decoder

We show the detailed architecture of our P-Decoder in Fig. 9, which consists of pixel decoder, text encoder, transformer decoder.

### C.1 Text Encoder for Open-Vocabulary Segmentation.

In the open-vocabulary setting, for each class, we encode the corresponding class name (*i.e.* *cat*, *dog*) into a  $d$ -dimensional vector using the CLIP encoder. For a word corresponding to two text tokens, we average them into one token. Subsequently, this token is replicated  $n$  times, resulting in an  $n \times d$  matrix. The matrix is then concatenated with a learnable query embedding of dimensions  $n \times 768$ . Ultimately, the concatenation is processed through a Multilayer Perceptron (MLP) layer to fuse the elements.

### C.2 Semantic and Instance Segmentation.

With the representation  $\hat{\mathcal{F}}$ , which is fused from multi-scale features and cross-attention maps, we employ a pixel decoder and a transformer decoder to derive the per-pixel embedding  $C \times H \times W$  and mask embedding  $C \times N$ . As per the method outlined by Li *et al* [38], the pixel decoder consists of several straightforward up-sampling layers. Each layer comprises four types of computations: 1)  $1 \times 1$  Conv for adjusting feature dimensionality, 2) Upsample using simple linear interpolation to upscale the feature to a higher spatial resolution, 3) Concat for merging features from different layers, and 4) Mix-conv for blending features from varying spatial resolutions, which includes two  $3 \times 3$  Conv. Similar to Mask2former [10], the transformer decoder comprises a stack of transformer layers

methods	backbone	Train Set/%			mIoU/%		
		# real image	# synthetic image	categories	seen	unseen	harmonic
ZS3 [6]	-	10.6k	-	15	78.0	21.2	33.3
CaGNet [26]	-	10.6k	-	15	78.6	30.3	43.7
Joint [3]	-	10.6k	-	15	77.7	32.5	45.9
STRICT [45]	-	10.6k	-	15	82.7	35.6	49.8
SIGN [12]	-	10.6k	-	15	83.5	41.3	55.3
ZegFormer [15]	-	10.6k	-	15	86.4	63.6	73.3
Li <i>et al.</i> [38]	ResNet101	-	10.0k (R:110k, COCO)	15+5	62.8	50.0	55.7
DiffuMask [60]	ResNet101	-	200.0 k (R:0)	15+5	62.1	50.5	55.7
DiffuMask [60]	Swin-B	-	200.0k (R:0)	15+5	71.4	65.0	68.1
DatasetDM	ResNet101	-	40k (R:450, VOC)	15+5	65.1	51.1	57.1
DatasetDM	Swin-B	-	40k (R:450, VOC)	15+5	78.8	60.5	68.4

Table 14: **Zero-Shot Semantic Segmentation on PASCAL VOC 2012.** ‘Seen’, ‘Unseen’, and ‘Harmonic’ denote the mIoU of seen, and unseen categories, and their harmonic mean. ‘R:’ refers to the number of real data from VOC 2012 or COCO 2017 used to train the generation model.

Train Set	Test Set	mIoU/%			
		Car	Person	Motorbike	mIoU
Cityscapes [13]	VOC 2012 [19] val	26.4	32.9	28.3	29.2
ADE20K [68]	VOC 2012 [19] val	73.2	66.6	64.1	68.0
DiffuMask [60]	VOC 2012 [19] val	74.2	71.0	63.2	69.5
DatasetDM	VOC 2012 [19] val	77.9	72.9	70.1	73.6
VOC 2012 [19]	Cityscapes [13] val	85.6	53.2	11.9	50.2
ADE20K [68]	Cityscapes [13] val	83.3	63.4	33.7	60.1
DiffuMask [60]	Cityscapes [13] val	84.0	70.7	23.6	59.4
DatasetDM	Cityscapes [13] val	85.6	58.9	12.7	52.4

Table 15: **Performance for Domain Generalization between different datasets.** Mask2former [10] with ResNet50 is used as the baseline. Person and Rider classes of Cityscapes [13] are consider as the same class, *i.e.*, Person in the experiment.

with cross-attention, self-attention, and masked attention. The final mask predictions of dimensions  $N \times H \times W$  can be obtained by performing a simple matrix multiplication of the per-pixel embedding of dimensions  $C \times H \times W$  and the mask embedding of dimensions  $C \times N$ .

### C.3 Human Pose and Depth Estimation.

By expanding the segmentation architecture with the addition of two convolutional layers to the pixel decoder, we are able to efficiently handle the associated tasks of pose and depth estimation. Consequently, we derive two predictive outputs, denoted by  $\mathbf{O} \in M \times H \times W$  and  $\mathbf{O} \in M \times H \times W$ , corresponding to the human pose and depth estimation tasks, respectively.

## D Synthetic Dataset

### D.1 Prompts from GPT-4

Here, we also demonstrate the detailed process of prompt generation, guided by GPT-4, as shown in Fig. 10. Throughout the process, human only need to provide a small number of prompts to guide GPT-4. With a cost of no more than 50 words prompt clue, we can accomplish the generation of a massive number of prompts for a downstream task dataset. It is worth mention that text-guided data is extremely flexible. We can customize the generation of certain attributes of data domain. For instance, if we need to enhance the variation in the number of objects, we can provide a prompt like More variation in number. This is extremely flexible and convenient.

### D.2 Prompts for Each Datasets

As shown in Tab. 16, we also provide some prompt cases of our method for each dataset, and we will open-source these prompts along with the corresponding code. For tasks that distinguish between classes, *i.e.* semantic and instance segmentation, we will guide GPT-4 to generate around 100

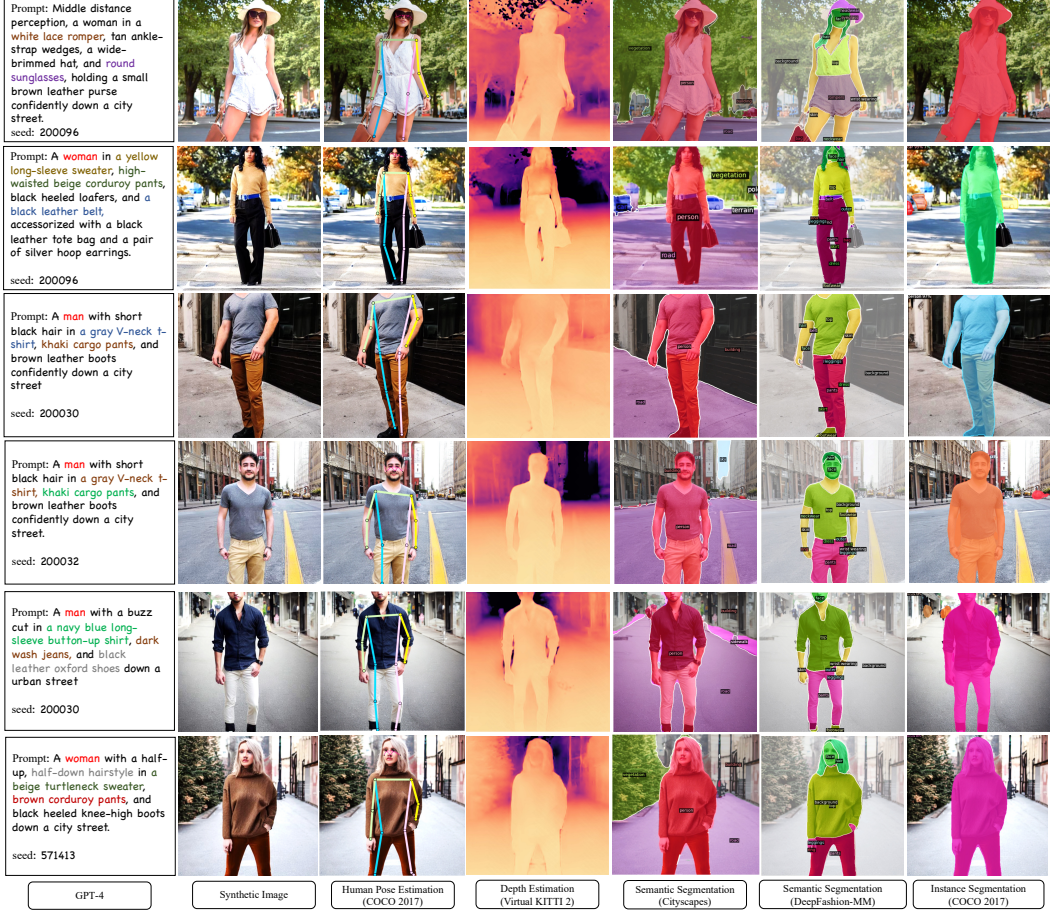


Figure 7: **Examples of Human-Centric Generated Data for DatasetDM.** Our method is capable of generating high-accuracy, high-diversity, and unified perceptual annotations.

descriptions specifically for each class. For tasks and datasets that are not class-sensitive, *e.g.* pose and depth estimation, we guide GPT-4 to generate a large number of descriptions all at once.

## E Limitation & Future Work

### E.1 Potential Negative Societal Impacts

As with other projects involving synthetic image generation, the potential adverse societal implications of our work largely revolve around ethical considerations. Utilizing the Stable Diffusion model, trained on the 5-billion image LION dataset, raises notable private copyright concerns due to the nature of the dataset. Nonetheless, we maintain that such potential repercussions should not overshadow the applicability and value of our research. Numerous text-guided text-to-image diffusion works exist, of which ours is but one. We also emphasize that it’s feasible to mitigate these ethical issues through the careful crafting of specific prompts, serving as an effective countermeasure. Additionally, we can mitigate this impact from an algorithmic perspective [35], such as eliminating certain concepts (which may infringe on personal privacy) from the pretrained model.

### E.2 Limitation & Future Work

**Limitation.** The main limitation of this study is that the quality and complexity of the synthesized data still cannot compare with real data. If certain companies and organizations could invest substantial resources to collect and manually annotate massive amounts of training data, better results could be achieved. However, this is actually the main limitation of the Stable diffusion model. We could also

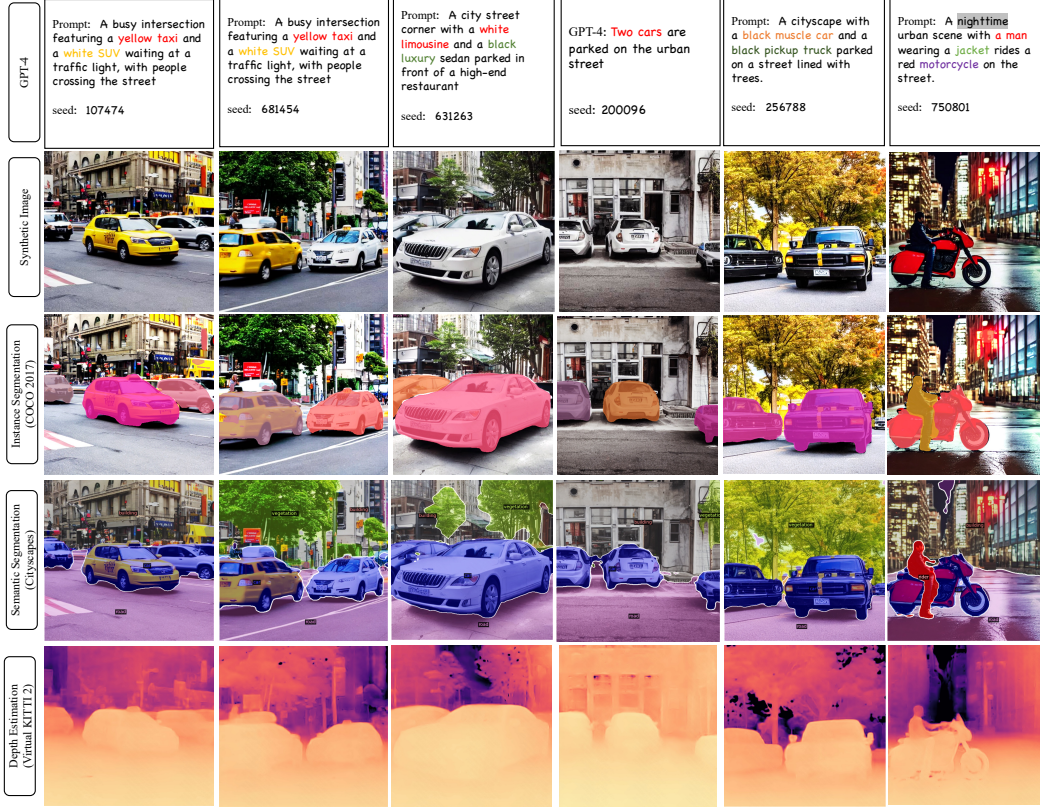


Figure 8: Examples of Generated Data for Urban City Scenario from DatasetDM.

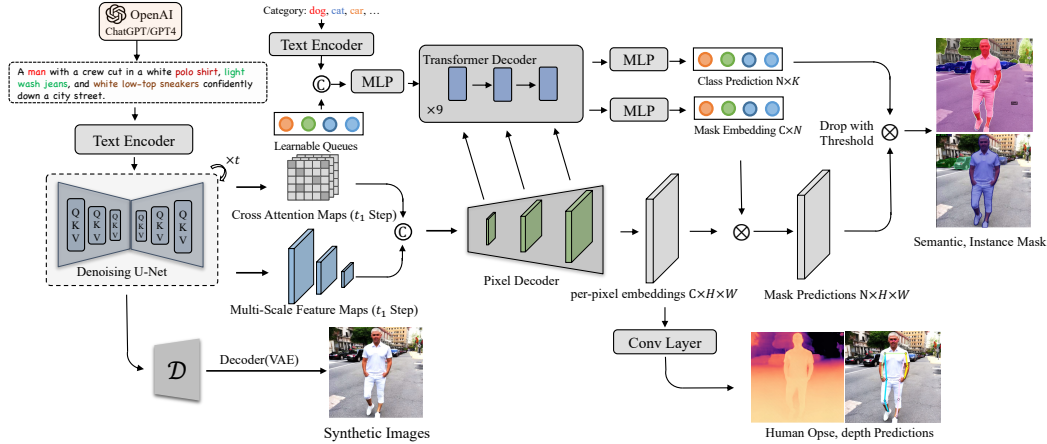


Figure 9: **Details for P-Decoder.** The whole framework of decoder consists of text encoder, pixel decoder, and transformer decoder. For different downstream task, we only need to adjust *minor variations* i.e. whether to startup some layers.

consider using more powerful diffusion models to alleviate this issue, as shown in Fig. 11. Deepfloyd IF<sup>6</sup> is a more powerful text-guided image generation model, which significantly outperforms Stable diffusion in two main aspects. First, it excels in semantic alignment - given a lengthy text description, the IF model can generate related images more accurately. Second, the IF model can synthesize images at a higher resolution, specifically 1024, while the resolution of Stable diffusion is only 512. We believe that our method, in combination with the DeepFloyd IF model, can lead to further improvements and make a greater contribution. Due to time constraints (DeepFloyd IF released on May 2023), we are unable to provide related experiments, but this does not affect the validation of

<sup>6</sup><https://github.com/deep-floyd/IF>

Dataset	Category	Prompts
VOC 2012 [19]	Car	A classic red convertible parked near a sandy beach, its vibrant color contrasting with the clear blue sky. ...
	Person	A young woman jogging in a park, wearing athletic clothing and listening to music through her earphones. ...
	Dog	A playful Golden Retriever, its fur gleaming in the sunlight, splashing in the water at a dog-friendly beach. ...
Cityscapes [13]	Car	A sleek, black car cruises down a busy urban street lined with towering skyscrapers. ...
	Person	A woman in a red dress is crossing the street at the crosswalk while cars wait for her. ...
	Bus	A red double-decker bus drives through the heart of the city on a busy urban street, the passengers admiring the sights from the upper level. ...
COCO 2017 [40]	Car	Two red cars parked on a busy city street in the afternoon. ...
	Person	A group of people playing volleyball on a beach, with the ocean in the background. ...
	Bus	A big red bus parked on the side of the road with a tree behind it. ...
NYU Depth V2 [54]	-	A kitchen with white cabinets, stainless steel appliances, and a wooden table. A bedroom with a queen-size bed, dresser, and nightstand. ...
COCO 2017 Pose [40]	-	a person with a backpack, wearing a green jacket and khaki pants. a middle-aged woman wearing a red blazer, black slacks, and pumps. ...
DeepFashion-MM [33]	-	A woman wearing a loose-fitting white blouse with ruffled sleeves, paired with high-waisted, wide-leg navy blue pants and black ankle-strap stiletto heels. A man dressed in a classic white button-up shirt, khaki chinos with a slim fit, and brown suede desert boots. ...

Table 16: **Prompts for Different Datasets.** For different data domains, we will guide GPT-4 to generate corresponding prompts. We will release the code and corresponding prompts files.

the effectiveness of our method. Our primary contribution lies in using a unified decoder to parse the latent space of the pre-trained diffusion model, not in enhancing the quality of image synthesis.

**Future Work** This study is intriguing and innovative, possessing profound exploratory significance. We identify several avenues for future enhancement: firstly, employing a more robust text-guided image generation model may yield substantial improvements. Secondly, augmenting the efficiency of prompt generation, or designing prompts that better align with the target domain could prove beneficial. For example, synthesizing specific prompts corresponding to the COCO 2017 dataset could be viable.

## References

- [1] J. Ahn and S. Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018.
- [2] P. Akiva and K. Dana. Towards single stage weakly supervised semantic segmentation. *arXiv preprint arXiv:2106.10309*, 2021.
- [3] D. Baek, Y. Oh, and B. Ham. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In *Proc. ICCV*, 2021.
- [4] D. Baranchuk, I. Rubachev, A. Voynov, V. Khruklov, and A. Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.
- [5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [6] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez. Zero-shot semantic segmentation. *NeurIPS*, 2019.

Can GPT-4 produce good prompts for text-to-image generation model?

**Human:** Hi, do you know COCO2017 benchmark?

**GPT-4:** Yes, I am familiar with the COCO2017 benchmark. The Common Objects in Context (COCO) dataset is a popular large-scale dataset used for object detection, segmentation, and captioning tasks. The 2017 version of the COCO dataset is an update of previous iterations that provides a more extensive set of annotations, images, and challenges. COCO2017 contains over 200,000 labeled images, covering 80 object categories, and is divided into training, validation, and testing subsets. The dataset is widely used by researchers and developers to train and evaluate computer vision models, particularly in tasks such as object detection, semantic segmentation, and instance segmentation. It also includes a captioning task, where models generate textual descriptions of images. The COCO dataset and its benchmarks have been instrumental in driving the development of new algorithms and improvements in object detection, segmentation, and captioning tasks. Many state-of-the-art models are evaluated using COCO2017 as a benchmark to assess their performance.

**Human:** Could you provide multiple descriptions for the car category in the COCO benchmark, with at least 10 variations for each description? These descriptions will be used as a guide for Stable Diffusion to generate corresponding images.

**GPT-4:** Sure, here are several distinct descriptions for the car category in the COCO benchmark, each with **three variations**:

- 1a. A red sports car speeding down a city street.
- 1b. A blazing red sports car racing along an urban road.
- 1c. A bright red sports car zooming through a bustling cityscape.
- 2a. A blue family sedan parked in a suburban driveway.
- 2b. A sky-blue family car parked outside a suburban home.
- 2c. A navy blue family vehicle resting in a suburban neighborhood driveway.
- ...



**Human:** More intricate and diverse descriptions.

**GPT-4:** Here are several more complex and diverse descriptions for the car category in the COCO benchmark:

- 1a. A sleek, dark gray electric sports car silently gliding down a rain-soaked city street, reflecting the neon lights from the surrounding buildings.
- 1b. A futuristic electric sports car with a deep gray hue speeding on wet urban roads, casting reflections of the vibrant neon city lights.
- 1c. A cutting-edge, graphite-colored electric sports car driving effortlessly on damp city streets, mirroring the luminous neon signs nearby.
- 2a. A vintage mint-green station wagon, packed with surfboards and beach gear, parked near a sandy shore with the ocean waves crashing in the background.
- ...



**Human:** More variation in numbers and scenarios.

**GPT-4:** Here are several more complex and diverse descriptions for the car category in the COCO benchmark:

- 1a. A trio of vintage race cars, their vibrant colors and classic designs, speeding around a historic racetrack during a nostalgic event.
- 1b. Three eye-catching vintage race cars, showcasing their timeless beauty and speed, as they compete on a legendary circuit in a retro-themed race.
- ...



Figure 10: **Prompts of diffusion model from GPT-4.** By providing some simple cues, GPT-4 can generate a vast and diverse array of prompts.

- [7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012.
- [8] Y. Cabon, N. Murray, and M. Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.
- [9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
- [10] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.
- [11] B. Cheng, A. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.
- [12] J. Cheng, S. Nandi, P. Natarajan, and W. Abd-Almageed. Sign: Spatial-information incorporated generative network for generalized zero-shot semantic segmentation. In *Proc. ICCV*, 2021.
- [13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [15] J. Ding, N. Xue, G.-S. Xia, and D. Dai. Decoupling zero-shot semantic segmentation. In *Proc. CVPR*, 2022.
- [16] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.



Figure 11: **Stronger Diffusion Model, Greater Potential.** With the advancements in generative models, synthetic data will have greater potential and possibilities for perception tasks. A simple solution is to replace Stable Diffusion with DeepFloyd IF directly.

- [17] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.
- [18] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.
- [19] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [20] L. Floridi and M. Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.
- [21] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.
- [22] Y. Ge, J. Zhao, and L. Itti. Pose augmentation: Class-agnostic object pose transformation for object recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 138–155. Springer, 2020.
- [23] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.
- [24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [25] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023.
- [26] Z. Gu, S. Zhou, L. Niu, Z. Zhao, and L. Zhang. Context-aware feature generation for zero-shot semantic segmentation. In *ACM MM*, 2020.
- [27] K. Han, Y. Xiong, C. You, P. Khosravi, S. Sun, X. Yan, J. Duncan, and X. Xie. Medgen3d: A deep generative framework for paired 3d image and mask generation. *arXiv preprint arXiv:2304.04106*, 2023.
- [28] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.
- [29] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*, 2023.
- [30] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.
- [31] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [32] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022.
- [33] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

- [34] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [35] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu. Ablating concepts in text-to-image diffusion models. *arXiv preprint arXiv:2303.13516*, 2023.
- [36] D. Li, H. Ling, S. W. Kim, K. Kreis, S. Fidler, and A. Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340, 2022.
- [37] Z. Li, Z. Chen, X. Liu, and J. Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022.
- [38] Z. Li, Q. Zhou, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. Guiding text-to-image diffusion model towards grounded generation. *arXiv preprint arXiv:2301.05221*, 2023.
- [39] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016.
- [40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [41] Z. Lin, H. Liang, G. Fanti, and V. Sekar. Raregan: Generating samples for rare classes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7506–7515, 2022.
- [42] H. Ling, K. Kreis, D. Li, S. W. Kim, A. Torralba, and S. Fidler. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34:16331–16345, 2021.
- [43] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [44] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.
- [45] G. Pastore, F. Cermelli, Y. Xian, M. Mancini, Z. Akata, and B. Caputo. A closer look at self-training for zero-label semantic segmentation. In *Proc. CVPRW*, 2021.
- [46] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3362–3369. IEEE, 2012.
- [47] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [48] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [50] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [51] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
- [52] S. Satkin, J. Lin, and M. Hebert. Data-driven scene understanding from 3d models. 2012.
- [53] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.
- [54] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgb-d images. *ECCV (5)*, 7576:746–760, 2012.
- [55] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.
- [56] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

- [57] X. Wang, L. Jing, Y. Lyu, M. Guo, J. Wang, H. Liu, J. Yu, and T. Zeng. Deep generative mixture model for robust imbalance classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [58] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.
- [59] G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.
- [60] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023.
- [61] Z. Wu, L. Wang, W. Wang, T. Shi, C. Chen, A. Hao, and S. Li. Synthetic data supervised salient object detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5557–5565, 2022.
- [62] A. Xu, M. I. Vasileva, A. Dave, and A. Seshadri. Handsoff: Labeled dataset generation with no additional human annotations. *arXiv preprint arXiv:2212.12645*, 2022.
- [63] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. *arXiv preprint arXiv:2303.04803*, 2023.
- [64] M. Zhang, Y. Zhou, J. Zhao, Y. Man, B. Liu, and R. Yao. A survey of semi-and weakly supervised semantic segmentation of images. *Artificial Intelligence Review*, 53:4259–4288, 2020.
- [65] Y. Zhang, H. Ling, J. Gao, K. Yin, J.-F. Lafleche, A. Barriuso, A. Torralba, and S. Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021.
- [66] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023.
- [67] Y. Zhao, Q. Ye, W. Wu, C. Shen, and F. Wan. Generative prompt model for weakly supervised object localization. *arXiv preprint arXiv:2307.09756*, 2023.
- [68] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.
- [69] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan, et al. Generalized decoding for pixel, image, and language. *arXiv preprint arXiv:2212.11270*, 2022.