

FlexDataset: Crafting Annotated Dataset Generation for Diverse Applications

Ellen Yi-Ge¹, Leo Shawn²

¹Carnegie Mellon University

²University of the Chinese Academy of Sciences

yige@andrew.cmu.edu, shanlianlei18@mails.ucas.edu.cn

Abstract

High-quality, pixel-level annotated datasets are crucial for training deep learning models, while their creation is often labor-intensive, time-consuming, and costly. Generative diffusion models have then gained prominence for producing synthetic datasets, yet existing text-to-data methods struggle with generating complex scenes involving multiple objects and intricate spatial arrangements. To address these limitations, we introduce FlexDataset, a framework that pioneers the composition-to-data (C2D) paradigm. FlexDataset generates high-fidelity synthetic datasets with versatile annotations, tailored for tasks like salient object detection, depth estimation, and segmentation. Leveraging a meticulously designed composition-to-image (C2I) framework, it offers precise positional and categorical control. Our Versatile Annotation Generation (VAG) Plan A further enhances efficiency by exploiting rich latent representations through tuned perception decoders, reducing annotation time by nearly five-fold. FlexDataset allows unlimited generation of customized, multi-instance and multi-category (MIMC) annotated data. Extensive experiments show that FlexDataset sets a new standard in synthetic dataset generation across multiple datasets and tasks, including zero-shot and long-tail scenarios.

Code — <https://github.com/EllenYiGe/FlexDataset>

Introduction

The recent surge in generative models has greatly expanded the field of computer vision, particularly in image synthesis and automated perceptual tasks. Among these advancements, text-to-image (T2I) diffusion models have emerged as a powerful technique for generating highly realistic images from textual descriptions (Ramesh et al. 2022; Rombach et al. 2022; Ge et al. 2023), offering substantial control over visual content. However, these models often struggle with generating complex scenes involving multiple objects, diverse categories, and intricate spatial arrangements. This challenge has led to the development of composition-to-image (C2I) methods, which allow users to precisely define the layout and attributes of multiple instances within a scene. Significant advancements, such as LayoutDiffusion (Zheng

et al. 2023), GLIGEN (Li et al. 2023b), and Instance Diffusion (Wang et al. 2024a), have enhanced diffusion models by incorporating composition guidance, facilitating precise extraction of instance positions within generated images. Simultaneously, the creation of high-quality, versatile annotations for perceptual tasks remains a significant challenge, as generating annotated datasets is labor-intensive, time-consuming, and costly. For example, labeling a complex scene with multiple objects can take 30 to 90 minutes (Zhang et al. 2021), emphasizing the need for innovative synthetic data generation techniques. DatasetGAN (Zhang et al. 2021) pioneered the use of GAN feature spaces for pixel-level labeling, and BigDatasetGAN (Li et al. 2022a) expanded this approach to accommodate the large class diversity in datasets like ImageNet. However, these methods are limited by their reliance on a small number of pixel-level labeled examples and often suffer from suboptimal performance due to imprecise generative masks. While powerful text-to-image diffusion models have introduced new possibilities for leveraging synthetic data to train models or even replace real data, existing methods like DiffuMask (Wu et al. 2023b) and DatasetDiffusion (Nguyen et al. 2024) are constrained by their dependence on pre-trained diffusion models and simplistic generation techniques, leading to unstable performance in more complex scenes.

In this context, synthetic annotated data has shown considerable potential. However, existing dataset generation approaches face limitations in adaptability and performance across various perceptual tasks. As highlighted in Figure 1, these methods are often constrained by their reliance on text-based generation with limited annotation control (DiffuMask, DatasetDM (Wu et al. 2023a), DatasetDiffusion), dependence heavily on pre-trained diffusion models that generate simplistic scenes, primarily focusing on single instances. Furthermore, their narrow focus on specific downstream tasks, such as semantic segmentation (e.g., DiffuMask, DetDiffusion (Wang et al. 2024b), DatasetDiffusion), restricts their broader applicability. The lack of more precise and controllable generation techniques results in unstable performance and applicability limitation in multi-instance and multi-category (MIMC) scene generation.

To address these challenges, we introduce FlexDataset, a novel framework that defines the paradigm of composition-to-data (C2D) generation. FlexDataset is meticulously de-

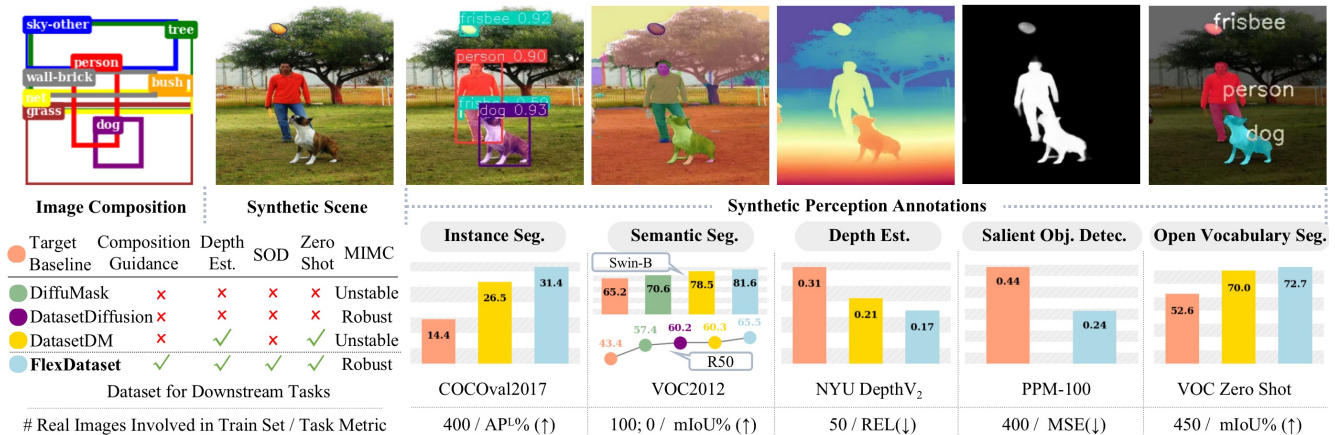


Figure 1: **Synthetic Data from FlexDataset.** FlexDataset provides high-quality, limitless images with perception annotations, leading to substantial enhancements in various downstream tasks.

signed to produce synthetic datasets with versatile annotations tailored for a wide range of downstream tasks. Unlike previous text-to-data approaches like DatasetDM, FlexDataset leverages a groundbreaking MIMC C2I framework, enabling the generation of complex, realistic scenes where multiple objects coexist and interact within a coherent global context. This is achieved by encoding pixels, context-aware categorical embeddings, and the entire image with composition guidance. Additionally, to reduce computational time, we introduce the Versatile Annotation Generation (VAG) *Plan A*. This approach directly utilizes the rich latent representations from the C2I model for VAG using perception decoders, rather than processing generated image features through entire perception models. This innovation reduces annotation synthesis time by nearly fivefold while maintaining high-quality pixel-level annotation synthesis.

In summary, our contributions are four-fold:

- We introduce *FlexDataset*, a comprehensive framework that redefines high-fidelity annotated dataset generation using a composition-guided generative approach. FlexDataset produces unlimited pixel-level synthetic images with versatile annotations for tasks like salient object detection (SOD), depth estimation, and generic segmentation, including zero-shot and long-tail settings.
- We propose the Versatile Annotation Generation (VAG) *Plan A*, which enhances annotation synthesis speed and quality by leveraging latent representations from the MIMC C2I model with optimized perception decoders. *VAG Plan A* accelerates synthesis nearly fivefold while maintaining high quality.
- FlexDataset provides precise control over semantic and spatial attributes, seamlessly integrating multiple subjects into customized images. It supports adjustments such as bounding box resizing, repositioning, and category alteration, enabling countless scene variations. Using less than 1% labeled data, it generates extensive synthetic datasets closely resembling real-world MIMC conditions, significantly reducing annotation efforts.

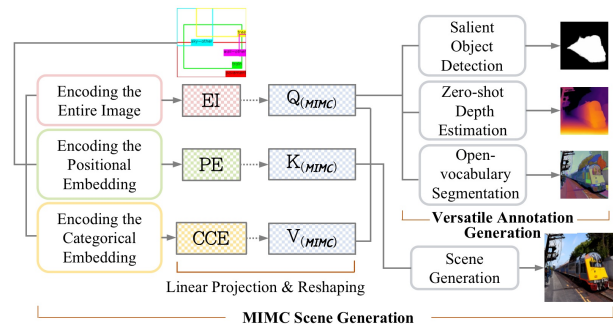


Figure 2: **The overall framework of FlexDataset.** Users input scene compositions with categories and bounding boxes. These features generate the Query, Key, and Value for the MIMC process. $Q_{(MIMC)}$ is then used for versatile annotation generation via tuned perception decoders for downstream tasks.

- Experiments show that perception models trained on FlexDataset’s synthetic data achieve outstanding results across five datasets and six tasks. For example, FlexDataset reduces SOD mean squared error by 20.1% on the PPM-100 dataset.

Related Work

Composition-guided Image Generation

Composition-guided methods generate images based on layouts that specify the arrangement and attributes of multiple instances. Unlike traditional text-to-image (T2I) methods, which struggle with controlling complex scenes with multiple objects, some diffusion models [(Li et al. 2023b), (Wang et al. 2023), (Zheng et al. 2023)] allow for composition guidance. For example, LayoutDiffusion (Zheng et al. 2023) and GLIGEN (Li et al. 2023b) input bounding box positions and labels into the diffusion model to learn layout information. DenseDiffusion (Kim et al. 2023) modulates attention maps

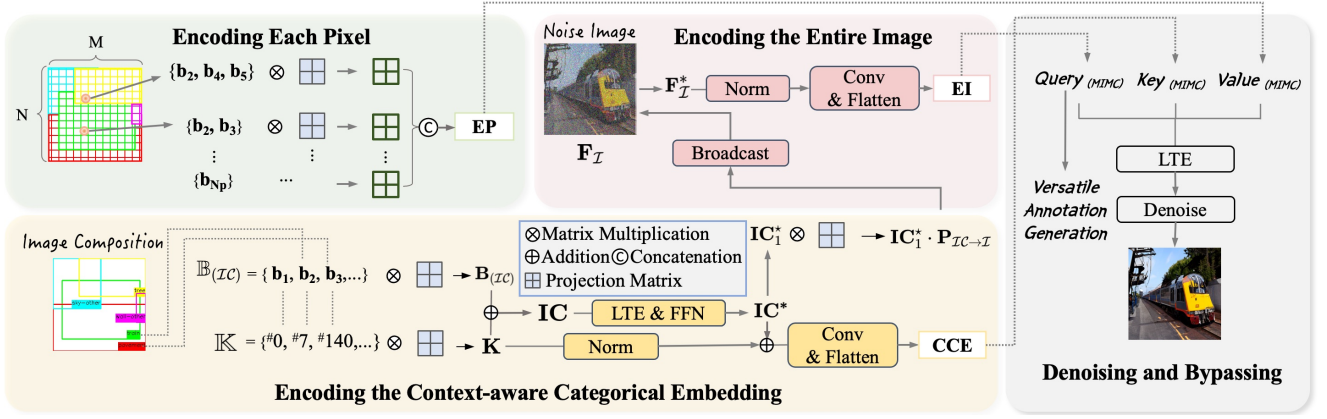


Figure 3: **Model architecture of MIMCSC.** (a) Encoding the Entire Image: The image is encoded into feature maps, normalized, convolved, and flattened to generate $\text{Query}_{(MIMC)}$. (b) Encoding Each Pixel: Each pixel is projected into a matrix based on its corresponding instance bounding boxes and contribute to the entire matrix of the image as $\text{Key}_{(MIMC)}$. (c) Encoding the Context-aware Categorical Embedding: The image composition and category information are projected, combined, normalized, and convolved to form $\text{Value}_{(MIMC)}$. (d) These encoded features undergo Linear Transformer Encoding (LTE) and denoising to synthetic scenes while $\text{Query}_{(MIMC)}$ is utilized for versatile annotation generation. The dashed lines with arrow represent linear projection and reshaping.

during inference without additional training. Instance Diffusion (Wang et al. 2024a) and MIGC (Zhou et al. 2024) extend layout-conditioned diffusion to generate multiple objects with precise quantities.

Synthetic Data for Perceptual Annotation Generation

Synthetic data generated by GANs (Goodfellow et al. 2020; Ling et al. 2021) and diffusion models (Sohl-Dickstein et al. 2015) offer flexibility for a range of tasks and open-world scenes. DiffuMask (Wu et al. 2023b), for instance, uses cross-attention maps from Stable Diffusion to produce synthetic images and semantic masks. Composition-based methods like GeoDiffusion (Chen et al. 2023), MagicDrive (Gao et al. 2023), and TrackDiffusion (Li et al. 2023a) enhance object detection by generating 3D-aided masks. However, these methods do not optimize generation for specific detectors. Other works convert generators into perceptual models by extracting annotations from generative features, such as DatasetDM (Wu et al. 2023a), DetDiffusion (Wang et al. 2024b), and Dataset Diffusion (Nguyen et al. 2024). These techniques, while capable of producing annotated data, are limited by reliance on text-based generation, dependency on pre-trained diffusion models that generate simplistic scenes, and a narrow focus on specific tasks. In contrast, FlexDataset employs complex image compositions rather than text prompts, enabling the coexistence and interaction of multiple objects.

Methodology

We pioneer an innovative paradigm, composition-to-data (C2D) generation, enhancing composition-guided diffusion models through training on image-composition pairs. Our FlexDataset hinges on two key insights: **[Sustainable]** By

leveraging less than 1% of an existing labeled dataset and utilizing enhanced yet lightweight perception decoders tailored to various downstream applications to interpret the diffusion latent space, we can generate infinite and diverse annotated data. This allows state-of-the-art methods to train on our synthetic datasets, significantly reducing labor costs; **[MIMC Crafting]** FlexDataset enables the creation of complex and realistic scenes. It efficiently generates customized images through a sophisticated C2I process, providing precise semantic and positional control over multi-category instances. Figure 2 demonstrates the overall framework.

MIMC Composition-guided Scene Generation (MIMCSG)

In multi-instance and multi-category (MIMC) scene generation, users specify the composition of N instances within the image through their layout bounding boxes $\mathbb{B}_{(\mathcal{IC})} = \{\mathbf{b}_1, \dots, \mathbf{b}_N\} \in \mathbb{R}^{N \times 4}$, where $\mathbf{b}_i = (x_{i1}, y_{i1}, x_{i2}, y_{i2})^T$, \mathcal{IC} representing the context of 'Image Composition', and the corresponding categories $\mathbb{K} = \{\mathbf{k}_1, \dots, \mathbf{k}_N\}^T$, where $\mathbf{k}_i \in$ distinct category set $\{dk_1, dk_2, \dots, dk_{N_D}\}$. Note that \mathbf{b}_1 is designed to be $(0, 0, 1, 1)^T$ representing the entire image coverage for further calculation. The pipeline then generates an image where each instance adheres to the user-defined category and ensures a coherent global scene alignment.

Unified Content-aware Embedding for Image Composition. To achieve a unified representation that is both content-aware and position-aware for the composition, we utilize projection matrices $\mathbf{P}_{\mathbb{B}} \in \mathbb{R}^{4 \times d_{\mathcal{IC}}}$, $\mathbf{P}_{\mathbb{K}} \in \mathbb{R}^{1 \times d_{\mathcal{IC}}}$ to map $\mathbb{B}_{(\mathcal{IC})}$ and \mathbb{K} into a unified space, where $d_{\mathcal{IC}}$ is the dimension of the unified embedding. The corresponding positional encoded embedding is denoted as $\mathbf{B}_{(\mathcal{IC})} = \mathbb{B}_{(\mathcal{IC})} \cdot \mathbf{P}_{\mathbb{B}}$ while the content-aware encoded embedding is denoted as $\mathbf{K} = \mathbb{K} \cdot \mathbf{P}_{\mathbb{K}}$, where $\mathbf{B}_{(\mathcal{IC})}, \mathbf{K} \in \mathbb{R}^{N \times d_{\mathcal{IC}}}$. Define the uni-

fied embedding \mathbf{IC} as follows:

$$\mathbf{IC} = \mathbf{B}_{(\mathcal{IC})} + \mathbf{K}, \quad (1)$$

where \mathbf{IC} , $\mathbf{B}_{(\mathcal{IC})}$, $\mathbf{K} \in \mathbb{R}^{N \times d_{\mathcal{IC}}}$. \mathbf{IC} reflects the alignment and integration of spatial and categorical information within the image composition.

Context-aware Embedding with Intra-attention. Although \mathbf{IC} incorporates content-aware and position-aware embedding, it lacks inter-instance dependencies and relationships, limiting the understanding of the scene, especially when objects intersect or obscure each other. To address this, we integrate intra-attention into \mathbf{IC} to form a context-aware representation. To effectively fuse the \mathbf{IC} embedding, we employ a Linear Transformer Encoder (LTE) (Katharopoulos et al. 2020) utilizing multiple layers of linearized self-attention, with output of each layer then undergoing a position-wise feed-forward neural network $FFN(\cdot) : FFN(\cdot) = (ReLU(\cdot)\mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2$, where \mathbf{W}_1 , \mathbf{W}_2 are weight matrices and b_1 , b_2 are biases. The final content-aware, position-aware and context-aware embedding is the output of LTE(\mathbf{IC}), denoted as $\mathbf{IC}^* \in \mathbb{R}^{N \times d_{\mathcal{IC}}}$. LTE captures the intra-interaction within the spatial and categorical information for all instances in the composition, with computational complexity decreased to a linear scale.

MIMC Query, Key & Value Construction. Evidently, the initial formulations highlight processing the semantic and spatial MIMC information both independently and through cross-reference. As illustrated in Figure 3, we then construct the final \mathbf{Q} , \mathbf{K} , \mathbf{V} through three encoding ways utilizing our formulations:

Encoding the entire image: Denote $\mathbf{F}_{\mathcal{I}} \in \mathbb{R}^{H \times W \times d_{\mathcal{I}}}$ as the feature map of the entire image. Let $\mathbf{IC}_1^* \in \mathbb{R}^{1 \times d_{\mathcal{IC}}}$ denote the first row of \mathbf{IC}^* , semantically representing the background coverage, the projected embedding is given by $\mathbf{IC}_1^* \cdot \mathbf{P}_{\mathcal{IC} \rightarrow \mathcal{I}} \in \mathbb{R}^{1 \times d_{\mathcal{I}}}$, where $\mathbf{P}_{\mathcal{IC} \rightarrow \mathcal{I}}$ presents the projection matrix. Then we broadcast the projected embedding to the feature map to get $\mathbf{F}_{\mathcal{I}}^*$. The output of the process of encoding the entire image, denoted as \mathbf{EI} , is calculated as $\mathbf{EI} = Conv(\text{Norm}(\mathbf{F}_{\mathcal{I}}^*))$. We further investigate the impact of incorporating text prompts in dataset synthesis. The corresponding embeddings are concatenated with $\mathbf{F}_{\mathcal{I}}^*$. A detailed analysis is provided in the ablation study.

Encoding each pixel: Each pixel (x, y) belongs to a set of instances including background, thus associated with a matrix of size $4 \times d(x, y)$, where 4 represents the dimensions of the bounding box for instances, and $d(x, y)$ is the number of instances the pixel belongs to. Each pixel can be transformed into a matrix of size $4 \times d_{\text{pixel}}$ with projection matrix $\mathbf{P}_{\text{pixel}} \in \mathbb{R}^{d(x, y) \times d_{\text{pixel}}}$. Given an image \mathcal{I} of size $M \times N$, we obtain a final matrix of size $(N_P, 4 \times d_{\text{pixel}})$, denoted as \mathbf{EP} , where $N_P = M \times N$. Unlike the structural patch encoded in LayoutDiffusion (Zheng et al. 2023), our encoding offers more precise pixel-wise representation.

Encoding the context-aware categorical embedding: We obtain CCE as following: $\mathbf{CCE} = Conv(\mathbf{IC}^* + \text{Norm}(\mathbb{K}))$. The key and value embeddings are derived from the content-aware embedding \mathbf{K} , emphasizing categorical information, and the fused image

composition embedding \mathbf{IC}^* , capturing intra-interactions among instances. Averaging \mathbb{K} and \mathbf{IC}^* yields a representation that integrates both general layout and specific characteristics of instances.

Inspire by (Zhou et al. 2024), to confine the context of each instance to a designated spatial domain, we propose an enhancement to the conventional attention mask, denoted as \mathbf{M} . The adjustment involves the bilateral neglect of tokens within both the query and key matrices, applied specifically for the i_{th} instance as follows:

$$M_i(x, y) = \begin{cases} 1, & \text{if } x_{i1} \leq x \leq x_{i2} \text{ and } y_{i1} \leq y \leq y_{i2}, \\ -inf, & \text{otherwise,} \end{cases} \quad (2)$$

where the background mask M_1 is defined as the area of the entire image excluding the union of all instance masks:

$$M_1(x, y) = 1 - \min\left(1, \sum_{i=2}^N M_i(x, y)\right). \quad \text{The final linear attention is derived through:}$$

$$\mathbf{A}_{final} = (\phi(\mathbf{Query}_{(\mathcal{MIMC})})\phi(\mathbf{Key}_{(\mathcal{MIMC})})^T \odot \mathbf{M}) \mathbf{Value}_{(\mathcal{MIMC})} \quad (3)$$

Herein, the combined mask tensor \mathbf{M} is formulated by stacking the individual masks along the third dimension, represents the amalgamation of all subject-specific masks: $\mathbf{M} = [M_1, M_2, \dots, M_N]$; \odot denotes the Hadamard product; $\phi(\cdot) = \text{elu}(\cdot) + 1$, and $\text{elu}(\cdot)$ denotes the exponential linear unit (Clevert, Unterthiner, and Hochreiter 2016) activation function; $\mathbf{Query}_{(\mathcal{MIMC})}$, $\mathbf{Key}_{(\mathcal{MIMC})}$, $\mathbf{Value}_{(\mathcal{MIMC})}$ are obtained from \mathbf{EI} , \mathbf{CCE} , and \mathbf{EP} , respectively, through linear projection and reshaping. The attention mechanism ensures that each pixel only attends to others within the same instance region. This maintains instance-specific features and avoids attribute leakage between instances. By integrating these masks into the linear attention framework, we ensure that the attention mechanism respects the instance boundaries, thereby preserving the instance-specific features and improving the overall quality of the generated images. This comprehensive approach is crucial for generating coherent and realistic multi-instance scenes. $\mathbf{Query}_{(\mathcal{MIMC})}$, $\mathbf{Key}_{(\mathcal{MIMC})}$, and $\mathbf{Value}_{(\mathcal{MIMC})}$ then undergo LTE followed by denoising and image rendering in align with LayoutDiffusion (Zheng et al. 2023). For the other bypassing, $\mathbf{Query}_{(\mathcal{MIMC})}$ serves as a perception task query for versatile annotation generation.

Our approach grants users granular control over individual objects within the generated image, facilitating precise manipulation of each object. By defining the composition, the user can ensure that each object is positioned and sized according to customization, thus enhancing the accuracy and relevance of the generated image.

Versatile Annotation Generation (VAG)

It is crucial to explore how the latent representation $\mathbf{Query}_{(\mathcal{MIMC})}$ can be translated into perception annotations across various downstream tasks. The primary distinction in our VAG approach lies in directly passing $\mathbf{Query}_{(\mathcal{MIMC})}$ through perception decoders (denoted as *VAG Plan A*), rather than using synthetic image features

that must go through a complete pre-trained perception model—comprising both encoder and decoder—as in previous approaches (denoted as *VAG Plan B*) like DiffuMask (Wu et al. 2023b). Drawing inspiration from previous works on perception models (Pang et al. 2020; Yang et al. 2024; Zou et al. 2024), we develop a pipeline for multi-task annotation generation that relies solely on the perception decoders from these methods. Specifically, in (Pang et al. 2020), the image features must undergo five-layer VGG-16 blocks (Simonyan and Zisserman 2015) as the encoder, followed by the aggregation interaction, self-interaction, and fusion unit modules to generate the SOD annotation. Our optimized method simplifies this process by directly feeding $\text{Query}_{(MIMC)}$ into a two-layer aggregation interaction and subsequent modules, bypassing the need for a full encoder stack. Similarly, for depth estimation, we eliminate the necessity of processing generated image features through the Depth-Anything shared encoder (Yang et al. 2024). Instead, $\text{Query}_{(MIMC)}$ is fed directly into the depth decoder. Likewise, for segmentation, $\text{Query}_{(MIMC)}$ is used directly in the self- and cross-attention modules of SEEM (Zou et al. 2024) to generate segmentation annotations, omitting the step of processing generated image features through the image encoder. Ultimately, we prioritize using perception decoders to translate latent information from the C2I process over processing entire models with generated images since *VAG Plan A* reduces annotation synthesis time by nearly 5-fold while maintaining high quality, as proved in our ablation studies. Detailed comparisons between entire perception models and perception decoders are in the appendix.

Optimization Objectives

Composition-Conditional Image Generation Loss for MIMCSG. To support composition-conditioned image generation, we adopt a technique called classifier-free guidance (Zheng et al. 2023). This method interpolates between the predictions of a diffusion model with and without condition input. We first construct a padding composition $Com_\phi = \{ins_{Com}, ins_1, \dots, ins_N\}$. During training, the composition condition Com of the diffusion model is replaced with Com_ϕ with a fixed probability. Define a pixel as $px_0 \sim q(px_0)$, where $q(\cdot)$ denotes the Markovian noising process, we can obtain the noised samples from px_1 to px_T , where T denotes the maximum steps of $q(\cdot)$. The training loss is calculated by:

$$\mathcal{L} = \mathbb{E}_{t \sim [1, T], px_0 \sim q(px_0), \epsilon \sim \mathcal{N}(0, \mathbf{I})} [\|\epsilon - \epsilon_\theta(px_t, t)\|^2] \quad (4)$$

When sampling, the composition-conditional image is generated using:

$$\hat{\epsilon}_\theta(px_t, t | Com.) = (1-s) \cdot \epsilon_\theta(px_t, t | Com_\phi) + s \cdot \epsilon_\theta(px_t, t | Com.), \quad (5)$$

where s scales the gap between $\epsilon_\theta(px_t, t | Com_\phi)$ and $\epsilon_\theta(px_t, t | Com.)$ to enhance conditional guidance.

Perception Decoder Tuning for VAG. To utilize the perception decoders, we need to train the decoders in *VAG Plan A* using annotations generated from *VAG Plan B*. The loss between the two sets of generated annotations fine-tunes the decoders, enhancing their adaptability and precision for downstream tasks.

Target Baseline Training Loss on Synthetic Dataset.

We train various perception models as baselines including Mask2Former (Cheng et al. 2022), MODNet (Ke et al. 2022), and DepthFormer (Li et al. 2022) for different downstream tasks using synthetic datasets generated by FlexDataset. The loss functions of these models are based on the specific perception tasks.

By incorporating these optimization objectives, we enhance FlexDataset for effective high-fidelity MIMC image generation and accurate annotation synthesis, improving the performance of baselines across various downstream tasks by training on our synthetic dataset.

Experiments

Dataset

For training C2I model and perception decoders, following the methodology of LayoutDiffusion (Zheng et al. 2023), we employ the COCO 2017 Stuff Segmentation Challenge subset. Each image contains bounding boxes and pixel-level segmentation masks for 80 categories of things and 91 categories of stuff. From these, we select images that feature between 3 to 8 objects, each covering more than 2% of the image area and not belonging to a crowd.

Implementation details

Reverse Tuning for MIMCSG. In MIMCSG process, we begin by synthesizing images with reverse tuning technique, that leverages the real bounding boxes and category labels of a tiny sub-dataset (e.g.: 100,400,800 images) to train generative models. As mentioned, we tune the C2I model with composition-conditional image generation loss. For all tasks, we train FlexDataset for approximately 200 iteration using images of size 512×512 on a single Tesla V100 GPU. We use the optimizer from (Loshchilov and Hutter 2017) with a learning rate of 0.0002.

Downstream Task Evaluation To comprehensively evaluate the generative image of FlexDataset, we conduct experiments across six supported downstream tasks. The corresponding annotations are generated with tuned perception decoders. We primarily benchmark our work against the state-of-the-art text-to-data method, DatasetDM (Wu et al. 2023a). *Salient Object Detection.* We evaluate FlexDataset on the PPM-100 benchmark (Ke et al. 2022) with MODNet (Ke et al. 2022) serving as the SOD baseline to assess the effectiveness of our generated data. FlexDataset uses 80k synthetic images based on 400 real images. The evaluation metrics are Mean Squared Error (MSE) and Mean Absolute Deviation (MAD). In alignment with DatasetDM (Wu et al. 2023a), we retained the same settings for other downstream tasks including *Semantic Segmentation*, *Instance Segmentation*, *Depth Estimation*, *Zero-Shot Semantic Segmentation*, and *Long-tail Semantic Segmentation* to ensure a fair comparison. Further details can be found in the appendix.

Main Results

Table 1 compares the four downstream tasks. Additional results are provided in Tables 3, 4, 5, and the appendix.

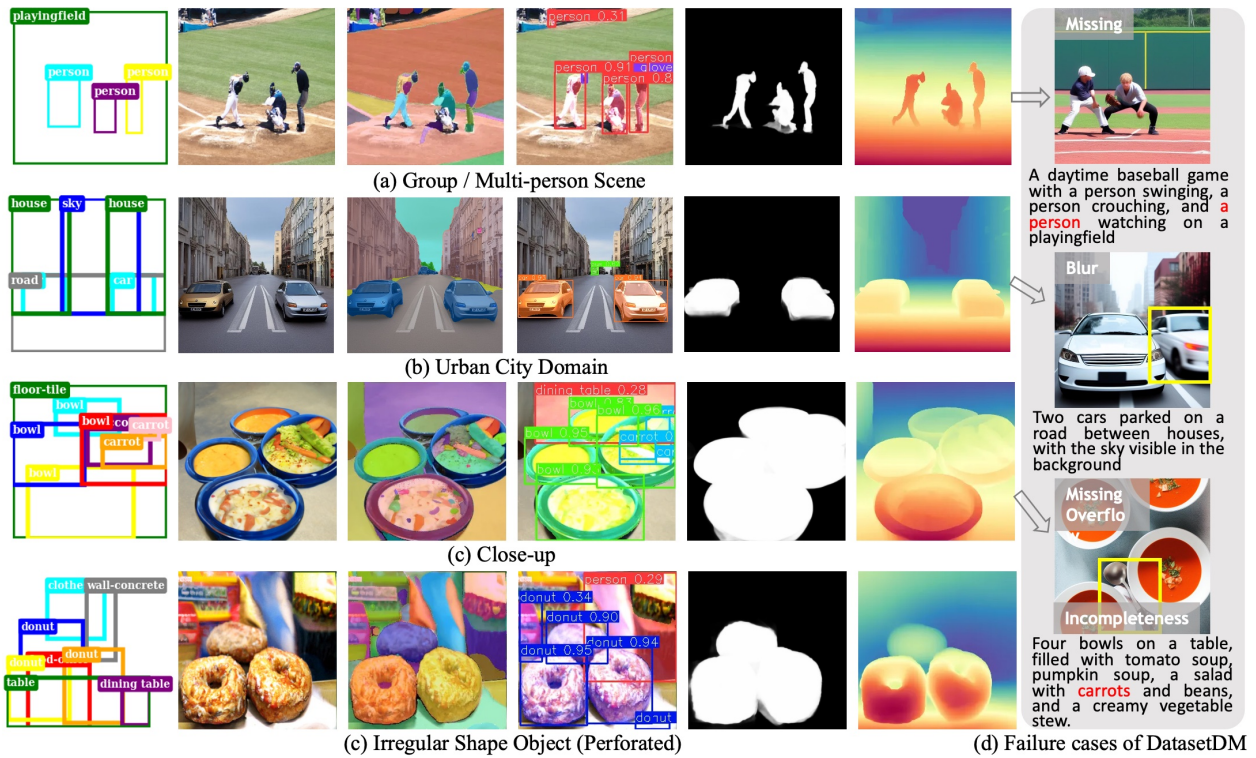


Figure 4: **Examples of annotated data generated from FlexDataset** with various MIMC configurations. Text prompts are created from generated scene using ControlCap (Zhao et al. 2024) for a fair comparison with DatasetDM. **Red**: missed classes; **Yellow boxes**: blurred or overflowed classes.

method	VOC (Semantic Seg.)/%		COCO2017 (Instance Seg.)/%			NYU Depth V2 (Depth Est.)			PPM-100 (Salient Obj. Det.)			
	# real	# synth.	mIoU	# real	# synth.	AP	# real	# synth.	REL ↓	# real	# synth.	MSE ↓
Baseline	100	-	65.2	400	-	14.4	50	-	0.31	full	-	0.44
DiffuMask	-	60k	70.6	-	-	-	-	-	-	-	-	-
DatasetDiffusion	-	40k	60.2	-	-	-	-	-	-	-	-	-
DatasetDM	100	40k	78.5	400	80k	26.5	50	35k	0.21	-	-	-
FlexDataset	100	40k	81.6	400	80k	31.4	50	35k	0.17	-	40k	0.24

Table 1: **Downstream Tasks.** 'real' and 'synth' represent real and synthetic images, respectively. The baseline backbones for the four tasks are 'Swin-B', 'Swin-B', and 'Swin-L'. DatasetDiffusion employs ResNet50 as its backbone.



Figure 5: **Annotation Prediction Results on NYU Depth V2.** FlexDataset can enhance the performance of the targeting perception model (e.g. DepthFormer (Li et al. 2022)).

Salient Object Detection. Table 2 shows that FlexDataset achieves the lowest MSE (0.24) and MAD (0.79) on PPM-100, outperforming other methods that use real images. *Instance Segmentation.* As shown in Table 3, FlexDataset with the R50 backbone and 400 real images achieves 17.7% AP, outperforming DatasetDM (12.2%). With 80,000 synthetic images, it reaches 19.4%, and with the Swin-B back-

method	backbone	# real image	# synthetic image	MSE↓	MAD↓
LFM	R50	400	-	0.94	1.58
HAtt	R50	400	-	0.67	1.37
BSHM	R50	400	-	0.63	1.14
MODNet	R50	400	-	0.44	0.86
FlexDataset	R50	-	80k (R:400)	0.24	0.79

Table 2: **Salient Object Decton on PPM-100** 'R:' represents the training data sourced from real datasets.

bone and 800 real images, it achieves 31.4%, surpassing DatasetDM's 26.5%. *Semantic Segmentation.* From Table 4, with 100 real images, FlexDataset improves mIoU by 5.2% over DatasetDM. Using the Swin-B backbone, it achieves 88.1% in full training. *Depth Estimation.* Table 1 compares synthetic vs. real data on the NYU Depth V2 dataset, where FlexDataset outperforms DatasetDM by 0.04 in depth estimation. *Zero Shot and Long-tail Segmentation.* As seen in

method	backbone	# real / # synth	AP	AP ^S	AP ^M	AP ^L
Mask2Former	R50	400 / -	4.4	1.1	3.3	12.1
DatasetDM	R50	- / 80k (R:400)	12.2	1.6	11.3	30.9
FlexDataset	R50	- / 80k (R:400)	17.7	8.5	17.3	36.7
DatasetDM	R50	400 / 80k (R:400)	14.8	2.3	15.1	36.0
FlexDataset	R50	400 / 80k (R:400)	19.4	6.3	17.2	39.4
Mask2Former	Swin-B	400 / -	11.3	3.2	10.1	27.1
DatasetDM	Swin-B	- / 80k (R:400)	17.6	3.4	17.8	39.5
FlexDataset	Swin-B	- / 80k (R:400)	27.3	10.5	21.6	45.3
DatasetDM	Swin-B	400 / 80k (R:400)	23.3	7.7	26.1	48.7
FlexDataset	Swin-B	400 / 80k (R:400)	30.4	12.9	32.0	53.2
Mask2Former	Swin-B	800 / -	14.4	5.6	15.7	29.2
DatasetDM	Swin-B	800 / 80k (R:800)	26.5	7.7	29.8	53.3
FlexDataset	Swin-B	800 / 80k (R:800)	31.4	13.4	33.7	57.3

Table 3: Instance segmentation results on COCO **val2017**. ‘R:’ indicates the real data utilized for training.

method	backbone	# real / # synth	Sampled Classes			mIoU
			Bird	Cat	Car	
Mask2Former	R50	100 / -	54.8	53.3	66.8	43.4
DiffuMask	R50	- / 60k	86.7	79.3	74.2	57.4
DatasetDiffusion	R50	- / 40k (R:100)	-	-	-	60.2
DatasetDM	R50	- / 40k (R:100)	84.7	74.4	79.2	60.3
FlexDataset	R50	- / 40k (R:100)	89.2	79.9	85.3	65.5
DatasetDM	R50	100 / 40k (R:100)	81.7	82.3	77.9	66.1
FlexDataset	R50	100 / 40k (R:100)	84.9	85.8	81.2	70.9
Mask2Former	Swin-B	100 / -	54.4	68.3	71.8	65.2
DiffuMask	Swin-B	- / 60k	92.9	92.5	82.9	70.6
DatasetDM	Swin-B	- / 40k (R:100)	93.4	94.5	78.8	73.7
FlexDataset	Swin-B	- / 40k (R:100)	94.7	96.2	85.8	76.4
DatasetDM	Swin-B	100 / 40k (R:100)	86.7	93.8	88.3	78.5
FlexDataset	Swin-B	100 / 40k (R:100)	87.4	95.0	89.6	81.6
Mask2Former	Swin-B	full / -	93.7	96.5	88.6	84.3
DiffuMask	Swin-B	5k / 60k	94.4	96.6	92.9	84.9
DatasetDM	Swin-B	full / 40k (R:100)	93.9	97.6	89.4	85.4
FlexDataset	Swin-B	full / 40k (R:100)	94.2	97.8	93.5	88.1

Table 4: Semantic segmentation results on VOC 2012. ‘R:’ indicates the real data utilized for training.

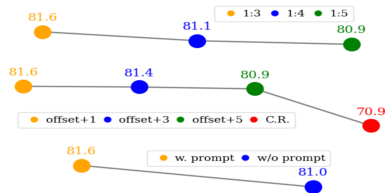


Figure 6: VAG performance (mIoU%) on semantic segmentation under different conditions. The top, middle, and bottom line charts correspond to ablation studies 2-4, respectively.

Table 5, FlexDataset improves mIoU by up to 2.7% over DatasetDM by addressing long-tail distribution challenges.

Ablation Study

Annotation Synthesis Technique: Using tuned perception decoders for annotation generation (*VAG Plan A*) is preferable due to their substantial computational efficiency. Table 6 shows no notable difference in annotation generation quality between the two methods. However, the lat-

method	Zero-Shot Configuration			Long-tail Configuration		
	seen	unseen	harm.	head	tail	mIoU/%
Baseline	61.3	10.7	18.3	61.2	44.1	52.6
Li <i>et al.</i>	62.8	50.0	55.7	-	-	-
DiffuMask	71.4	65.0	68.1	-	-	-
DatasetDM	78.8	60.5	68.4	73.1	66.4	70.0
FlexDataset	83.6	77.5	80.4	75.2	70.3	72.7

Table 5: Zero Shot and Long-tail Segmentation on VOC 2012. For Zero Shot, consistent with priors (Li *et al.* 2023c; Wu *et al.* 2023b,a), FlexDataset is trained using only 15 seen categories and evaluated across all 20 categories. In the Long-tail configuration, the 20 categories are divided into head classes (10 classes, 20 images per class) and tail classes (10 classes, 2 images per class).

VAG Method	Sem. Seg. %	ST (h)	Depth Est. %	ST (h)
Entire PM	81.1	257.3	0.20	92.3
Tuned PD	81.6	138.4	0.17	18.6

Table 6: VAG performance of ablation study 1. Entire Perception Models (PM) vs. Tuned Perception Decoders (PD). ‘Sem. Seg.’ and ‘Est.’ denote Semantic Segmentation and Estimation, respectively. ST (h) represents synthesis time in hours.

ter greatly enhances computational efficiency five times; **Proportion of Single-Category and Multi-Category Instances in Synthetic Images:** We examine how different proportions of single-category and multi-category objects in synthetic images affect training performance. Figure 6 shows that MIMC configurations maintain robust performance. Balancing these proportions ensures that FlexDataset closely mirrors real-world scenarios; **Bounding Box Offset:** We introduced various levels of bounding box offsets in synthetic images. Figure 6 shows that slight offsets enhance generation performance, indicating improved generalization and robustness to real-world variations in image compositions. **Impact of Text Prompt Supervision:** We investigated whether incorporating text prompts enhances mask generation. A CLIP text encoder (Radford *et al.* 2021) projects category prompts (e.g., car, tree) into sequence embeddings, which are concatenated with F_T^* . Location tokens are added to CLIP, initialized with 2D sine-cosine embeddings. Figure 6 shows a 0.6% improvement in generation.

Conclusion

We have presented FlexDataset, a framework for generating high-fidelity synthetic datasets tailored to diverse perceptual tasks such as salient object detection, depth estimation, and generic segmentation. FlexDataset pioneers a composition-to-data (C2D) generation paradigm, enabling the creation of complex, multi-instance and multi-category (MIMC) scenes that closely resemble real-world environments. Our Versatile Annotation Generation (*VAG Plan A*) enhances annotation synthesis efficiency by nearly five-fold. Comprehensive experiments demonstrate that FlexDataset surpasses existing text-to-data methods, underscoring its potential to transform dataset creation.

References

- Chen, K.; Xie, E.; Chen, Z.; Hong, L.; Li, Z.; and Yeung, D.-Y. 2023. Integrating Geometric Control into Text-to-Image Diffusion Models for High-Quality Detection Data Generation via Text Prompt. *arXiv preprint arXiv: 2306.04607*.
- 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.
- Clevert, D.-A.; Unterthiner, T.; and Hochreiter, S. 2016. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv:1511.07289*.
- Gao, R.; Chen, K.; Xie, E.; Hong, L.; Li, Z.; Yeung, D.-Y.; and Xu, Q. 2023. MagicDrive: Street View Generation with Diverse 3D Geometry Control. *arXiv preprint arXiv:2310.02601*.
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.
- Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020. Transformers Are Rnns: Fast Autoregressive Transformers with Linear Attention. In *International Conference on Machine Learning*, 5156–5165. PMLR.
- Ke, Z.; Sun, J.; Li, K.; Yan, Q.; and Lau, R. W. 2022. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1140–1147.
- Kim, Y.; Lee, J.; Kim, J.-H.; Ha, J.-W.; and Zhu, J.-Y. 2023. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7701–7711.
- Li, P.; Liu, Z.; Chen, K.; Hong, L.; Zhuge, Y.; Yeung, D.-Y.; Lu, H.; and Jia, X. 2023a. TrackDiffusion: Multi-object Tracking Data Generation via Diffusion Models. *arXiv preprint arXiv:2312.00651*.
- Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.
- Li, Z.; Chen, Z.; Liu, X.; and Jiang, J. 2022. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*.
- Li, Z.; Zhou, Q.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023c. Guiding Text-to-Image Diffusion Model Towards Grounded Generation. *arXiv preprint arXiv:2301.05221*.
- Ling, H.; Kreis, K.; Li, D.; Kim, S. W.; Torralba, A.; and Fidler, S. 2021. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 34: 16331–16345.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- 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.
- Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9413–9422.
- 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.
- Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*.
- Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2256–2265. PMLR.
- Wang, X.; Darrell, T.; Rambhatla, S. S.; Girdhar, R.; and Misra, I. 2024a. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6232–6242.
- Wang, Y.; Gao, R.; Chen, K.; Zhou, K.; Cai, Y.; Hong, L.; Li, Z.; Jiang, L.; Yeung, D.-Y.; Xu, Q.; et al. 2024b. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7246–7255.
- Wang, Z.; Sha, Z.; Ding, Z.; Wang, Y.; and Tu, Z. 2023. Tokencompose: Grounding diffusion with token-level supervision. *arXiv preprint arXiv:2312.03626*.
- 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. *arXiv preprint arXiv:2303.11681*.
- Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.
- Zhao, Y.; Liu, Y.; Guo, Z.; Wu, W.; Gong, C.; Wan, F.; and Ye, Q. 2024. ControlCap: Controllable Region-Level Captioning. *arXiv:2401.17910*.
- Zheng, G.; Zhou, X.; Li, X.; Qi, Z.; Shan, Y.; and Li, X. 2023. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22490–22499.

Zhou, D.; Li, Y.; Ma, F.; Zhang, X.; and Yang, Y. 2024. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6818–6828.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2024. Segment Everything Everywhere All at Once. *Advances in Neural Information Processing Systems*, 36.