DetDiffusion: Synergizing Generative and Perceptive Models for Enhanced Data Generation and Perception

Yibo Wang[1*]   Ruiyuan Gao[2*]   Kai Chen[3*]   Kaiqiang Zhou[4]   Yingjie Cai[4]

Lanqing Hong[4]   Zhenguo Li[4]   Lihui Jiang[4✉]   Dit-Yan Yeung[3]   Qiang Xu[2]   Kai Zhang[1,5✉]

[1]Tsinghua University   [2]CUHK   [3]HKUST

[4]Huawei Noah's Ark Lab   [5]Research Institute of Tsinghua, Pearl River Delta

wyb22@mails.tsinghua.edu.cn   * Equal Contribution   ✉ Corresponding authors

Abstract

Current perceptive models heavily depend on resource-intensive datasets, prompting the need for innovative solutions. Leveraging recent advances in diffusion models, synthetic data, by constructing image inputs from various annotations, proves beneficial for downstream tasks. While prior methods have separately addressed generative and perceptive models, DetDiffusion, for the first time, harmonizes both, tackling the challenges in generating effective data for perceptive models. To enhance image generation with perceptive models, we introduce perception-aware loss (P.A. loss) through segmentation, improving both quality and controllability. To boost the performance of specific perceptive models, our method customizes data augmentation by extracting and utilizing perception-aware attribute (P.A. Attr) during generation. Experimental results from the object detection task highlight DetDiffusion's superior performance, establishing a new state-of-the-art in layout-guided generation. Furthermore, image syntheses from DetDiffusion can effectively augment training data, significantly enhancing downstream detection performance.

1. Introduction

The effectiveness of current perceptive models is heavily contingent on extensive and accurately labeled datasets. However, the acquisition of such datasets is often resource-intensive. Recent advancements in generative models, especially diffusion models [35], make it possible to generate high-quality images, and thus pave the way for constructing synthetic datasets. By providing annotations such as the class labels [35], segmentation maps [41], and object bounding boxes [4], synthetic data for generative models is proved to be useful to improve the performance on downstream tasks (*e.g.*, classification [11], object detection [2, 4] and segmentation [24, 42]).
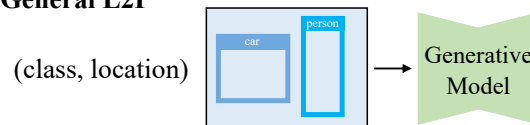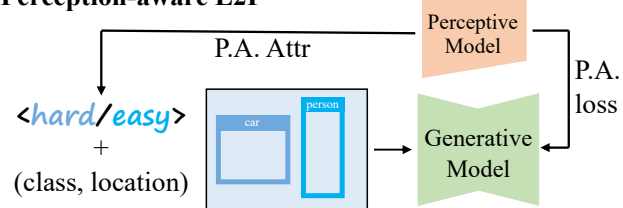

Figure 1. **Pipeline comparison** between general L2I models (*e.g.*, GeoDiffusion [4]) and our *DetDiffusion* (perception-aware L2I). Utilizing perception-aware loss (P.A. loss) and perception-aware attributes (P.A. Attr), *DetDiffusion* improves generation quality and controllability of L2I task. Perception-aware attributes further boost performance on downstream perceptive models. Moreover, perceptive model only added **1.3% of parameters**.

While most methods focus on improving generative models or perceptive models separately, the synergy between generative and perceptive models warrants a closer integration for mutual enhancement of generation and perception capabilities. In perceptive models, the challenge lies in effective data generation or augmentation, a topic previously approached mainly from a data perspective (*e.g.* OoD generalization [18, 20] and domain adaptation [10, 16, 31]). *Its potential to enhance perceptive models performance in general cases remains underexplored.* Conversely, generative model research has focused on refining models for better output quality and controllability [4, 8]. Nonetheless, it is essential to recognize that *perceptive models can also provide valuable additional insights to assist generative models in achieving better control capabilities.* This synergy between the generative and perceptive models offers a promising avenue for advancement, suggesting a need for more integrated approaches.

As the first work to investigate such synergy, we propose a novel perception-aware generation framework, namely *DetDiffusion*, as shown in Figure 1. Our framework enables generative models to harness the information from perceptive models, thereby augmenting their capacity for controlled generation. Concurrently, it facilitates the targeted generation of data based on the capabilities of perceptive models, thereby enhancing the performance of models trained on synthetic data. Specifically, for object detection tasks, we fine-tune models based on Stable Diffusion [35], employing controlled generation techniques to produce high-quality data that aids in training detection models. To elevate the quality of generation, we innovatively introduce a perception loss. By introducing a segmentation module [5] based on the intermediate feature from the UNet [36], the generated content is supervised by the object mask in conjunction with label ground truth to enhance controllability. Moreover, to further enhance the performance of detection models, we propose to extract and use object attributes from the trained detection model, and then incorporate these attributes into the training of generative models. This approach enables the generation of new data specifically tailored to produce distinctive samples, thereby significantly improving detectors' performance.

Our experiments confirm that *DetDiffusion* sets a new state-of-the-art in generation quality, achieving 31.2 mAP on COCO-Stuff. It significantly enhances detector training, increasing mAP by 0.9 mAP through the strategic use of perception-aware attribute (P.A. Attr) in training. This is largely due to *DetDiffusion*'s refined control in addressing long-tail data generation challenges. These advancements underscore *DetDiffusion* 's technical superiority and mark a pivotal advancement in controlled image generation, especially where precise detection attributes are vital.

The main contributions of this work contain three parts:
1. We propose *DetDiffusion*, the first framework designed to explore the synergy between perceptive models and generative models.
2. To boost generation quality, we propose a perception loss based on segmentation and object masks. To further improve the efficacy of synthetic data in perceptive models, we introduce object attributes during generation.
3. Extensive experiments on object detection task show that *DetDiffusion* not only achieves new SOTA in the layout-guided generation on COCO but also effectively prompts the performance for downstream detectors.

2. Related Work

Diffusion Models. Diffusion models, being one kind of generative model, are trained to learn the reverse denoising process after a forward transformation from the image distribution to the Gaussian noise distribution [14]. These models can employ either a Markov process [14] or a non-

Markov process [37]. Due to their adaptability and competence in managing various forms of controls [22, 35, 47] and multiple conditions [9, 15, 26, 30], diffusion models have been applied in various conditional generation tasks, such as image variation [44], text-to-image generation [35], pixel-wise controlled generation [47]. A notable variation of these models is the Latent Diffusion Model (LDM [35]). Unlike traditional diffusion models, the LDM conducts the diffusion process in a latent space, enhancing the model's efficiency. Our framework for perception data generation is based on the LDM. However, we focus on the synergy between generative models and perceptive models, proposing several designs to benefit both generation quality & controllability and performance on downstream tasks.

Layout-to-Image (L2I) Generation. Our approach focuses on converting a high-level graphical layout into a realistic image. In this context, LAMA [23] implements a locality-aware mask adaptation module for improved object mask handling during image generation. Taming [17] shows that a relatively straightforward model can surpass more complex predecessors by training in latent space. More recent developments include GLIGEN [22], which integrates additional gated self-attention layers into existing diffusion models for enhanced layout control, and LayoutDiffuse [6], which employs innovative layout attention modules tailored for bounding boxes. Our generative model shares similar architecture with GeoDiffusion [4] and Geom-Erasing [27], while *DetDiffusion* focuses on the synergy between generation and perception, and distinctively offers 1) a novel perception-aware loss (P.A. loss) that utilizes information from the segmentation head; 2) a novel object attribute mechanism (P.A. Attr) to help the training of object detectors.

Data Generation for Perceptive Models. In some L2I methods, the utility of synthetic data in enhancing object detection task performance is demonstrated, *e.g.*, GeoDiffusion [4]. Similarly, MagicDrive [8] suggests that generated images aid in 3D perception, and TrackDiffusion [21] generates data for multi-object tracking. However, they do not explore enhancing generation using perceptive models or tailoring data for specific detectors. Beyond controllable generation, several works convert generators into perceptive models by extracting annotations from generative features. DatasetDM [42] uses a Mask2Former-style P-decoder with Stable Diffusion, while Li et al. [23] develop a fusion module for open-vocabulary segmentation. These techniques, while capable of producing annotated data, are limited by their reliance on text-based generation with limited annotation control, dependency on pre-trained diffusion models restricting the cross-domain applicability, and inferior performance compared to combining diffusion models with specialized models like SAM [19].
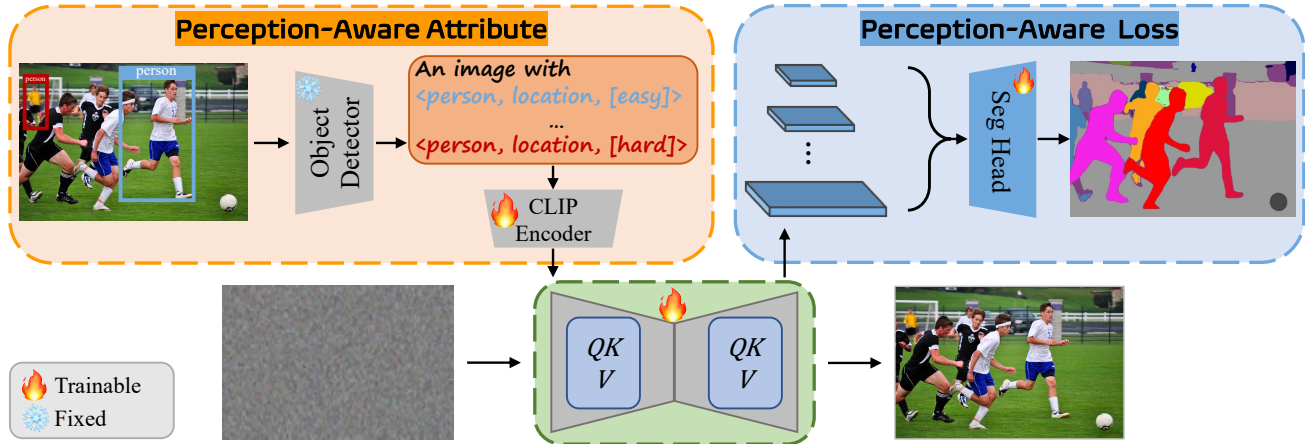
Figure 2. **Model architecture of *DetDiffusion*.** To facilitate the synergy between generative models and perceptive models, we integrate two components into L2I training pipeline. Perception-aware loss (P.A. loss) leverages the segmentation head for better generation quality and controllability. Perception-aware attribute (P.A. Attr) enables *DetDiffusion* to generate highly useable data for training augmentation.

3. Method

Our objective is to enhance the generation quality from a perceptive perspective and facilitate downstream perceptive tasks. Designing proper and strong supervision is of great importance in tackling this challenging problem, we propose to integrate easily accessible but previously neglected perceptive information i.e., perception-aware attribute (P.A. Attr) and loss (P.A. loss), into the generation framework to promote the information interaction between perceptive models and generative models. We first introduce the preliminaries in Section 3.1 and expand the perception-aware attribute (P.A. Attr) in detail (Section 3.2), which is generated via an object detector and designed as special tokens to assist diffusion models. In Section 3.3, a tailored perception-aware loss (P.A. loss) is introduced. The overall architecture is depicted in Figure 2.

3.1. Preliminaries

Diffusion Models (DMs) have emerged as prominent text-to-image generation models, characterized by their effectiveness in generating realistic images. A notable variation, the Latent Diffusion Model (LDM) [35], innovatively transfers the diffusion process of standard DMs into a latent space. This transition is significant, as LDMs demonstrate the ability to maintain the original model's quality and flexibility, but with a substantially reduced computational resource requirement. This efficiency gain is primarily attributed to the reduced dimensionality of the latent space, which facilitates faster training times without compromising the generative capabilities of the model.

Stable Diffusion, an exemplary implementation of the Latent Diffusion Model (LDM), utilizes a distinctive pipeline for text-to-image (T2I) generation. The process commences with the encoding of the original image x into latent space using a pre-trained Vector Quantized Varia-

tional AutoEncoder (VQ-VAE) [40], resulting in a latent representation $z = \mathcal{E}(x) \in \mathcal{R}^{H' \times W' \times D'}$, where is much smaller than original dimension. Concurrently, the text condition y undergoes encoding via a pre-trained CLIP [33] text encoder $\tau_\theta(\cdot)$. At a given timestep t, random noise is integrated into the latent variable z to form z_t. The noise prediction is executed by a UNet $\epsilon_\theta(\cdot)$, which incorporates both resnet and transformer networks of varying dimensions for enhanced generative capability. The integration of the condition variables $\tau_\theta(y)$ with the UNet is achieved through cross-attention mechanisms. The formulation of the objective function can be expressed as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t} \|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2. \quad (1)$$

This equation represents the mean-squared error between the original noise ϵ and the noise predicted by the model, encapsulating the core learning mechanism of the Stable Diffusion model.

3.2. Perception-Aware Attribute as Condition Input

To enhance the performance of detection models, this study introduces a novel approach centered around the generation of perception-aware realistic images. The methodology involves a two-step process: initially, object attributes are extracted from a pre-trained detector. These attributes encapsulate critical visual characteristics essential for accurate object detection. Subsequently, the extracted attributes are integrated into the training regime of a generative model. This integration aims to ensure that the generated images not only exhibit high realism but also align closely with the perceptive criteria crucial for effective detection. By doing so, the generative model is tailored to produce images that are more conducive to training robust detectors, potentially leading to significant improvement in detection accuracy and reliability.

Perception-Aware Attribute. We define the perception-aware attribute (P.A. Attr) of an object as d. For each image x, a pre-trained detector $\mathcal{D}(\cdot)$, such as Faster R-CNN [34] or YOLO series [1] detectors, is employed to yield n predicted bounding boxes, represented as $b = [b_1, ..., b_n] = \mathcal{D}(x)$. To refine the selection of bounding boxes, a filtering criterion based on a confidence score threshold γ is applied. This process effectively retains a subset of bounding boxes that meet the threshold, resulting in a reduced set $b' = [b_1, ..., b_{n'}]$, where n' is significantly smaller than n. This selective approach ensures that only bounding boxes with a high likelihood of accurate object detection are considered, thereby enhancing the following reliability of the perception-aware attribute (P.A. Attr) extracted.

Furthermore, for each image x, there are m ground truth objects bounding boxes, represented as $o = [o_1, ..., o_m]$. The detection difficulty of each ground truth box o_i is assessed based on its intersection with the n' predicted boxes. Specifically, for each ground truth box, if any predicted bounding box b_j, where $i \in [1, ..., n']$, has an intersection over union (IoU) with the ground truth box exceeding a threshold β, it is classified as $[easy]$ to detect. Conversely, the ground truth boxes without such overlapping predicted boxes are labeled as $[hard]$. This classification mechanism is encapsulated in the following expression:

$$d_i = \begin{cases} [easy], & \text{if exists } i, \text{ IoU}(b_j, o_i) > \beta, \\ [hard], & \text{else.} \end{cases} \quad (2)$$

Perception-Aware Attribute as Prompt Token. In this approach, each ground truth box within an image is characterized by three attributes. These include the pre-existing attributes of category (c_i) and location (l_i), along with the newly introduced attributes perception difficulty attribute (d_i). The category attribute c_i is the category text itself. For the location attribute l_i, the original representation is continuous coordinates. Here we discretize it via partitioning the pixel image space into a grid of location bins, and each location bin corresponds to a unique token (check more details in [4]). So that we can feed the specific location token into a text encoder of L2I diffusion model [4, 46] to obtain the final location attribute l_i. In this way, attributes representing location, category, and ease of detection are organized into a unified representation.

Furthermore, in contrast to existing methods [22, 46] utilizing captions as the text prompt. We design an effective text prompt equipped with multiple pairs of perception-aware attributes. Specifically, the prompt is "`An image with {objects}`", where `objects` are $[(c_1, l_1, d_1), ..., (c_m, l_m, d_m)]$ and m is number of ground truth bounding boxes. This comprehensive attribute set and effective prompt aim to encapsulate a more holistic understanding of each object's characteristics, potentially providing a much richer description for perception.
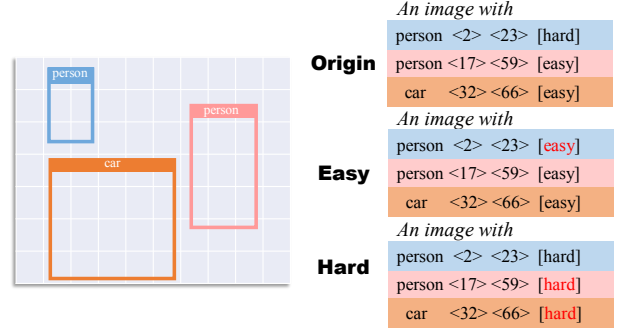

Figure 3. **Three strategies for attribute application.** Check the detailed definition in Sec. 4.1.

3.3. Perception-Aware Loss as Supervision

During training diffusion generation model, the objective is to minimize the reconstruction distance between the predicted image (or noise) and its ground truth counterpart. Traditional generation methodologies predominantly utilize L1 or L2 losses for this purpose. However, these standard loss functions often fall short of producing images with high-resolution details and precise control over image attributes. To address this limitation, a novel *perception-aware loss* (P.A. loss) is proposed. This loss function is constructed to leverage rich visual features, thereby facilitating more nuanced image reconstruction.

Visual Features. Recent studies have indicated that the feature maps in the UNet model exhibit effectiveness in class-discriminative and localization tasks [12, 42, 43, 48]. Inspired by these papers, we try to extract multi-scale feature maps $f = [f_1, f_2, f_3, f_4]$ from four layers of the U-Net $\epsilon_\theta(\cdot)$ of the encoding and decoding path during the training process, corresponding to resolutions of ($\frac{H}{8} \times \frac{W}{8}, \frac{H}{4} \times \frac{W}{4}, \frac{H}{2} \times \frac{W}{2}, H \times W$). And we upsample $\text{Upsample}(\cdot)$ and mix convolution layers $\text{Conv}(\cdot)$ to concatenate the final multi-scale feature maps, which can be expressed as:

$$F_n = \begin{cases} \text{Conv}(f_n, \text{Upsample}(F_{n-1})), & n = 2, 3, 4 \\ f_n, & n = 1 \end{cases} \quad (3)$$

Thus, we have obtained the visual features $\mathcal{F} = F_4$.

Perception-Aware Loss. To achieve a higher level of precision in image generation, this study introduces a custom-designed loss function, utilizing the rich information from perception learning. Central to this formulation is the use of a segmentation head, which processes these features to produce instance masks, represented as $M = [m_1, .., m_k]$. In optimizing the model's high-dimensional feature space, the loss function incorporates two key components: the mask loss $\mathcal{L}_m$ and a dice loss $\mathcal{L}_d$ following [42]. The integration of these losses, specifically tailored for perceptive enhancement, serves to finely tune the U-Net's capabilities. This

approach allows for more granular control over the generated images, leveraging the perceptive information embedded within the features for more accurate image synthesis.

Considering that noise has been added to the visual features $\mathcal{F}$, we assign different scales to the loss term, easing the burden of model training. Specifically, we adopt $\bar{\alpha}_t$ from Denoising Diffusion Probabilistic Models (DDPM), as delineated in Ho et al. [14], to mitigate this effect. This scheme is designed to counterbalance the noise component in the features, ensuring the integrity and utility of $\mathcal{F}$. Given the DDPM scheduler adds Gaussian noise to the data according to a variance schedule $\beta_1, \dots, \beta_t$, and then using the notation $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. The perception-aware loss (P.A. loss) can be expressed as:

$$\mathcal{L}_p = \sqrt{\bar{\alpha}_t}(\mathcal{L}_m + \mathcal{L}_d), \qquad (4)$$

where the $\sqrt{\bar{\alpha}_t}$ is designed to reduce the impact of feature maps with higher noise levels, thereby emphasizing feature maps with lower noise (*i.e.*, smaller time steps).

Objective Function. Ultimately, our objective function combines the perception-aware loss with the foundational loss function of the Latent Diffusion Model (LDM). This integration is mathematically represented as follows:

$$\mathcal{L} = \mathcal{L}_{LDM} + \lambda \mathcal{L}_p \qquad (5)$$

For the purposes of this model λ is set to 0.01, ensuring a balanced incorporation of the perception-aware components while maintaining the primary structure and goals of the LDM loss function. This calibrated approach allows for a nuanced optimization that leverages the strengths of both losses, thereby enhancing the model's performance in generating high-quality, perception-aligned images.

4. Experiments

4.1. Experiment Settings

Dataset. We employ the widely recognized COCO-Thing-Stuff benchmark [3, 23, 25] for the L2I task, which includes 118,287 training images and 5,000 validation images. Each image is annotated with bounding boxes and pixel-level segmentation masks for 80 categories of objects and 91 categories of stuff. In line with previous works [4, 6, 46], we ignore objects belonging to crowds or occupying less than 2% of the image area.

Implementation Details. We fine-tune *DetDiffusion* from the Stable Diffusion v1.5 [32] checkpoint. We introduce location tokens into the text encoder and initialize the embedding matrix of the location tokens with 2D sine-cosine embedding. With the VQ-VAE [40] fixed, we fine-tune all parameters of the text encoder and use AdamW [28] optimizer with a cosine learning rate schedule of $1e^{-4}$. And the linear warm-up is adopted in the first 3000 steps. The

Method	Epoch	FID↓	mAP↑	AP$_{50}$↑	AP$_{75}$↑
LostGAN	200	42.55	9.1	15.3	9.8
LAMA	200	31.12	13.4	19.7	14.9
TwFA	300	22.15	-	28.2	20.1
Frido	200	37.14	17.2	-	-
L.Diffuse†	60	22.20	11.4	23.1	10.1
L.Diffusion†	180	22.65	14.9	27.5	14.9
ReCo†	100	29.69	18.8	33.5	19.7
GLIGEN	86	21.04	22.4	36.5	24.1
ControlNet†	60	20.37	24.8	36.6	27.7
GeoDiffusion	60	20.16	29.1	*38.9*	33.6
DetDiffusion $_{origin}$	60	**19.28**	*29.8*	38.6	*34.1*
DetDiffusion $_{hard}$	60	19.72	25.3	33.7	29.1
DetDiffusion $_{easy}$	60	*19.66*	**31.2**	**40.2**	**35.5**

Table 1. Evaluation of image quality and correspondence to layout on COCO val-set. The best results are in **bold** and the second best results are *underlined italic*. † Implemented by ourselves.

text prompt is replaced with a null text for unconditional generation with a probability of 10%. The model is trained on 8×32GB GPUs with a batch size of 32, requiring about 20 hours for 60 epochs. We sample images using DPM-Solver [29] scheduler for 50 steps with CFG at 3.5.

Strategy for Attribute Application. Upon completion of the training process, it is flexible to apply perception-aware attribute (P.A. Attr) during the generation. For the purpose of simple yet effective validation, we adopt three attribute strategies in Fig 3: 1) *DetDiffusion* $_{origin}$: the original perception-aware attribute (P.A. Attr). We obtain the attributes of each object in the image using a detector and use them directly for generation. 2) *DetDiffusion* $_{hard}$: all objects are assigned the $[hard]$ attribute. All objects are treated as difficult samples for perception. 3) *DetDiffusion* $_{easy}$: all objects are assigned the $easy$ attribute. All objects are treated as easy samples for perception.

4.2. Main Results

The L2I generation requires the generated objects to be as consistent as possible with the original image while ensuring high-quality image generation. Therefore, we will first comprehensively analyze the fidelity experiment in Section 4.2.1. Additionally, an important purpose of generating target detection data is its applicability to downstream target detection. We present the trainability experiment in Section 4.2.2.

4.2.1 Fidelity

Set up. To evaluate fidelity, we utilize two primary metrics on the COCO-Thing-Stuff validation set. The Fréchet Inception Distance (FID) [13] assesses the overall visual quality of the generated image. It measures the distinc-
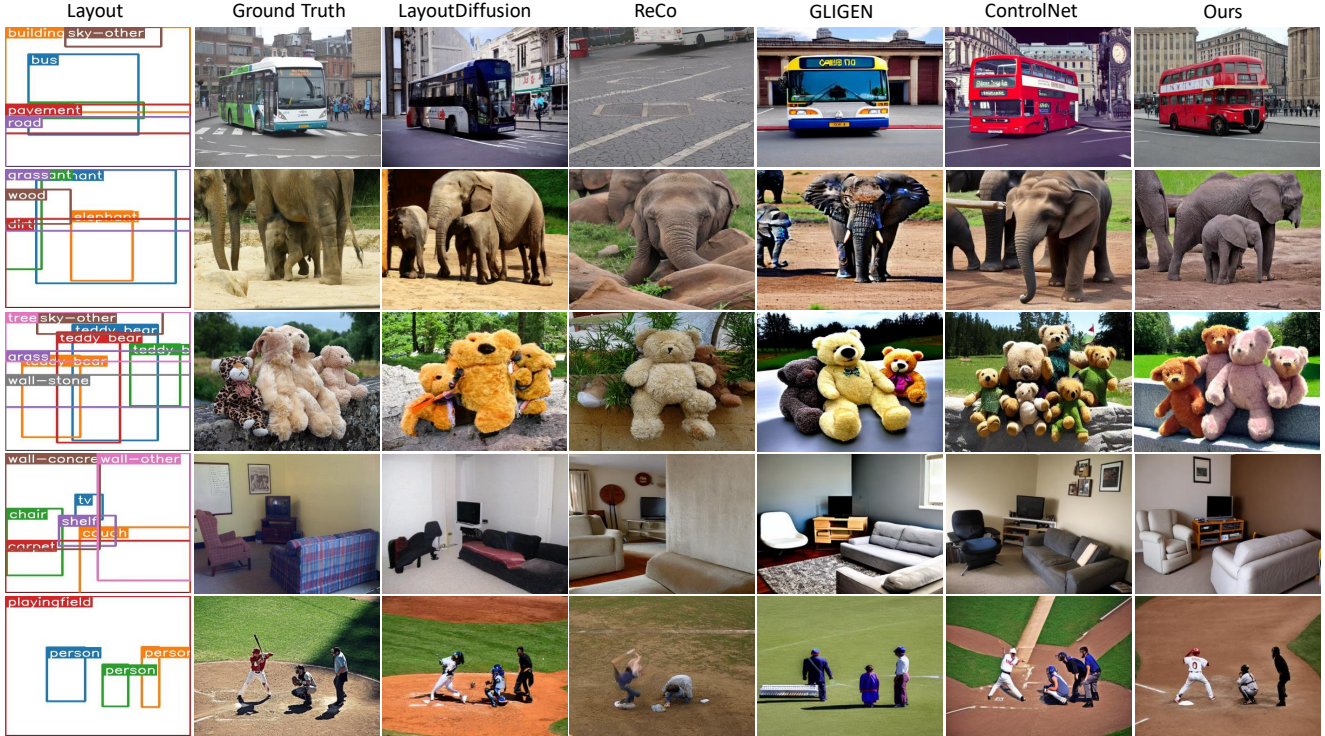
Figure 4. **Qualitative comparison on the Microsoft COCO dataset.** Our *DetDiffusion* can generate highly realistic images consistent with the provided semantic layouts.

tion in feature distribution between the real images and the generated images using an ImageNet-pretrained Inception-V3 [39] network. The YOLO Score [23] in LAMA [23] uses the mean average precision (mAP) of 80 object categories' bounding boxes on generated images. It is achieved using a pre-trained YOLOv4 [1] model, demonstrating the precision of object detection in a generated model. Our model is trained on the image size of 256×256. Following previous work, we utilize images containing between 3 to 8 objects, resulting in 3,097 images during validation.

Results. we evaluated our models with three attribute strategies on the COCO-Thing-Stuff validation set and compared them with state-of-the-art models for L2I task such as LostGAN [38], LAMA [23], TwFA [45], Frido [7], LayoutDIffuse [6], LayoutDiffusion [49], Reco [46], GLI-GEN [22], GeoDiffusion [4], and ControlNet [47]. As shown in Table 1, our *DetDiffusion* $_{origin}$ and *DetDiffusion* $_{easy}$ strategies outperformed all competitors across all metrics. The *DetDiffusion* $_{origin}$ strategy achieved the best FID (19.28) and outperformed other models in YOLO Score. This indicates that using perception-aware loss (P.A. loss) and treating the perception-aware attribute (P.A. Attr) as an additional condition obtained from perception can generate more realistic images. The *DetDiffusion* $_{easy}$ strategy achieved a YOLO Score exceeding the best model by 2.1mAP, and showed significant improvement compared to the *DetDiffusion* $_{origin}$ strategy that demonstrates generated

examples are easy for the detector to perceive. The *DetDiffusion* $_{hard}$ strategy is designed to generate examples that are more challenging for the detector, and the results are in line with our expectations. The YOLO Score decreased compared to the *DetDiffusion* $_{origin}$ strategy, indicating that generated examples are more difficult. The significance of the *DetDiffusion* $_{hard}$ strategy lies in its impact on trainability, which is demonstrated in Section 4.2.2.

The enhanced FID and YOLO Score achieved with the *DetDiffusion* $_{origin}$ approach illustrates the effectiveness of incorporating P.A. Attr and P.A. loss in regulating the performance of the images. Furthermore, the *DetDiffusion* $_{easy}$ and *DetDiffusion* $_{hard}$ strategies demonstrate our model's ability to comprehend and manipulate the attributes of [*easy*] and [*hard*] from the perceptual models, thus enabling control over the difficulty level.

4.2.2 Trainability

Set up. This section explores the potential advantages of using generated images from *DetDiffusion* for training object detectors. The evaluation of trainability includes using a pre-trained L2I model to create a new synthetic training set from the original annotations. Both the original and synthetic training sets are then employed to train a detector.

COCO Trainability. To establish a reliable baseline, we utilize the COCO2017 dataset, selectively choosing images
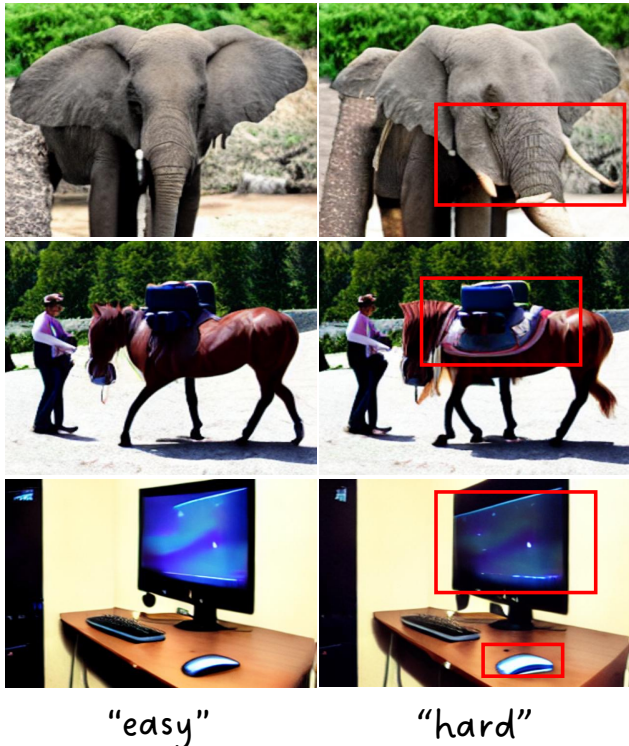
"easy"　　　　　"hard"

Figure 5. **Qualitative comparison on the perception-aware attribute.** Although provided with the exact same semantic layouts, simply changing the perception-aware attribute (P.A. Attr) among [*easy*] (left) and [*hard*] (right) can effectively alter the low-level image pattern of generated images. The former achieves better detector recognizability, while the latter performs as better augmentation samples, as demonstrated in Table 1 and 2 respectively.

containing 3 to 8 objects to enhance synthetic image quality and maintain fidelity. This process yielded a training set comprising 47,429 images with 210,893 objects. The goal is to showcase the improvements *DetDiffusion* can contribute to downstream tasks, maintaining fixed annotations for various model comparisons. For training efficiency and focused evaluation of data quality's impact on training, we adopt a modified 1× schedule, reducing the training period to 6 epochs. *DetDiffusion* is trained on images resized to 800×456, its maximum supported resolution, to reconcile resolution differences with COCO.

Results. As shown in Table 2, ReCO [46], GeoDiffusion [4], and all our three strategies can be beneficial for the training of downstream detectors, with the synthetic images generated by the strategies showing a more significant gain for the detector (exceeding 35.0 mAP). Furthermore, compared to the "origin" strategy, the "hard" strategy exhibits the most improvement across all detector metrics. This is attributed to the generation of more challenging instances through the "hard" strategy, which often represents the long-tail data in real datasets or serves as a form of stronger data augmentation. Overall, our model's generated data significantly enhances the training of downstream

Method	mAP	AP_{50}	AP_{75}	AP^m	AP^l
Real only	34.5	55.5	37.1	37.9	44.3
L.Diffusion	34.0	54.5	36.5	37.2	43.6
GLIGEN	34.3	54.8	36.7	37.4	44.3
ControlNet	34.4	54.5	36.9	37.8	45.0
ReCo	33.6	53.2	36.2	36.7	44.0
GeoDiffusion	34.8	55.3	37.4	38.2	45.4
DetDiffusion $_{origin}$	*35.3*	*55.7*	*38.2*	*38.4*	*46.5*
DetDiffusion $_{hard}$	**35.4**	**55.8**	**38.3**	**38.5**	**46.6**
DetDiffusion $_{easy}$	35.2	55.5	37.9	38.3	46.3

Table 2. Comparison of trainability on COCO. *DetDiffusion* leads to better improvements by emphasizing hard objects in augmentation. The best results are in **bold** and the second best results are *underlined italic*.
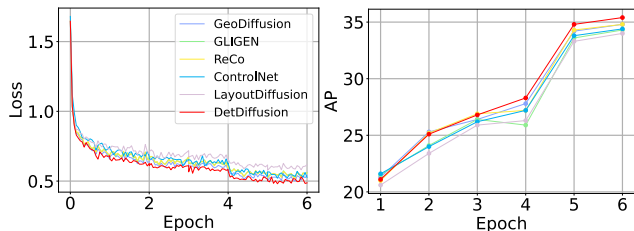
(a) Training loss curve.　　(b) Validation mAP curve.

Figure 6. **Training loss and validation mAP curves.**

detectors, surpassing all other L2I models, and reveals that information obtained through perception can further benefit downstream training.

To verify the *training effectiveness with equal training costs*, we plot the training loss curve and val mAP curve in Figure 6a and 6b respectively. Our *DetDiffusion* achieves the best performance throughout the training procedure.

We present more results on trainability in Table 3, focusing on **less frequent categories** in the COCO dataset such as parking meter, scissor and microwave, each accounting for less than 0.2% of the dataset. It can be observed that our hard strategy yields gains across all categories, with particularly significant improvements for long-tail categories.

4.3. Qualitative Results

Fidelity. Figure 4 displays examples that validate our model's fidelity and accuracy in image generation. LayoutDiffusion's chaotic results stem from its extra control modules clashing with the diffusion process. ReCo, dependent on high-quality captions, often suffers quality reduction and misses details. GLlGEN and ControlNet, despite high-quality outputs, lack precise object supervision, leading to insufficient detail and variable object quantities. Our implementation of P.A. loss and P.A. Attr enhances object quality, ensuring consistent quantities and controlled generation, as reflected in the alignment of generated object numbers with P.A. Attr.

Method	Average Precision↑										
	mAP	parking.	scissor	micro.	mouse	keyboard	hot dog	baseball.	sandwich	train	skateboard	tv
real only	34.5	41.6	17.8	49.2	55.6	43.8	23.6	31.3	29.8	51.9	41.7	49.8
easy	35.2	42.5	17.8	47.3	54.5	45.0	24.6	32.0	30.3	53.8	41.9	50.4
orgin	35.3	40.4	17.7	51.5	55.5	44.3	25.4	30.9	30.2	53.8	43.4	50.2
hard	**35.4**	**43.3**	**19.4**	**53.2**	**56.3**	**45.5**	**25.4**	**32.7**	**31.4**	**54.1**	**44.1**	**50.6**

Table 3. **Rare categories results of trainability on COCO2017.** The best results are in **bold**. The "parking.", "micro." and "baseball." suggest parking meter, microwave and baseball gloves.

P.A. Attr	P.A. loss	FID↓	mAP↑
		20.16	29.1
✓		19.92	30.4
✓	✓	**19.66**	**31.2**

Table 4. **Ablations on essential components of *DetDiffusion*** for perception-awareness. Best results are achieved when both components are adopted.

Detector	Method	mAP↑	AP$_{50}$↑	AP$_{75}$↑
FCOS	Real only	33.7	52.9	35.7
	DetDiffusion $_{origin}$	34.9	53.8	**37.0**
	DetDiffusion $_{easy}$	34.8	53.8	36.7
	DetDiffusion $_{hard}$	**35.0**	**54.0**	36.9
ATSS	Real only	36.3	53.9	39.1
	DetDiffusion $_{origin}$	37.2	54.8	40.2
	DetDiffusion $_{easy}$	37.1	54.6	40.0
	DetDiffusion $_{hard}$	**37.4**	**55.0**	**40.5**

Table 5. **Trainability for more detectors on COCO.**

	FID↓	mAP↑	AP$_{50}$↑	AP$_{75}$↑
Faster R-CNN	19.99	29.5	39.2	33.8
YOLOv4	**19.92**	**30.4**	**40.8**	**35.1**

Table 6. **Perception aware attribute from different detectors.** Note the results are both evaluated with YOLOv4. Better performance is achieved if the perception-aware attribute is provided by the evaluated detector specifically.

Easy and Hard. In Figure 5, we present perception-aware attribute (P.A. Attr) selections, comparing "easy" and "hard" instances. The "easy" images, exemplified by elephants, horses, monitors, and keyboards, are generated with an emphasis on intrinsic object features, ensuring clarity and lack of noise. Conversely, the "hard" examples, such as elephants with tusks, saddled horses, dim monitors, and reflective mice, incorporate additional elements that introduce noise through occlusions, lighting, and other complexities. These attributes make object recognition more challenging. Notably, there are both clearly distinguishable and subtly different "easy" and "hard" cases, highlighting the nuanced impact on the detection process. This indicates the identification of challenging examples without prior knowledge. For further illustrations, see Appendix D.

4.4. Ablation Study

Model components. We sequentially integrate two modules into the baseline model to evaluate our model's key elements. For a clear demonstration of P.A. loss effects, all attributes are set as $[easy]$. As Table 4 indicates, adding P.A. Attr notably enhances image fidelity and YOLO Score. This implies that perceptual information inclusion aids in producing more realistic and recognizable images. Furthermore, implementing P.A. loss, which oversees potential features in intermediately generated images, significantly improves the model's precision in image generation, especially in positional accuracy.

Trainability. We further conduct experiments on FCOS and ATSS. As shown in Table 5, images generated by *DetDiffusion* achieve significant improvement regardless of the detector models, consistently with results in Table 2.

Detector. We explore two widely recognized detectors [1, 34] for acquiring P.A. Attr in experiments that omit the use of P.A. loss. Table 6 demonstrates the detector choice significantly affects P.A. Attr quality, with YOLOv4 outperforming in this aspect. Therefore, YOLOv4 serves as a primary detector for Fidelity, while Faster R-CNN is used for trainability due to its role as a trained downstream detector.

5. Conclusion

This paper proposes *DetDiffusion*, a simple yet effective architecture to utilize the intrinsic synergy between generative and perceptive models. By incorporating detector-awareness into geometric-aware diffusion models via P.A. Attr as conditional inputs and P.A. loss as supervision, *DetDiffusion* can generate detector-customized images for better recognizability and trainability.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 4, 6, 8

[2] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018. 1

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. 5

[4] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. *arxiv preprint arXiv:2306.04607*, 2023. 1, 2, 4, 5, 6, 7

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2

[6] Jiaxin Cheng, Xiao Liang, Xingjian Shi, Tong He, Tianjun Xiao, and Mu Li. Layoutdiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023. 2, 5, 6

[7] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. In *AAAI*, 2023. 6

[8] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 1, 2

[9] Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, and Qiang Xu. DiffGuard: Semantic mismatch-guided out-of-distribution detection using pre-trained diffusion models. In *ICCV*, 2023. 2

[10] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu. Soda10m: Towards large-scale object detection benchmark for autonomous driving. *arXiv preprint arXiv:2106.11118*, 2021. 1

[11] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 1

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 4

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 5

[15] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[16] Benedikt T Imbusch, Max Schwarz, and Sven Behnke. Synthetic-to-real domain adaptation using contrastive unpaired translation. In *CASE*. IEEE, 2022. 1

[17] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. *arXiv preprint arXiv:2105.06458*, 2021. 2

[18] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N. Murali. Deflating dataset bias using synthetic data augmentation. In *CVPRW*, 2020. 1

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[20] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022. 1

[21] Pengxiang Li, Zhili Liu, Kai Chen, Lanqing Hong, Yunzhi Zhuge, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Trackdiffusion: Multi-object tracking data generation via diffusion models. *arXiv preprint arXiv:2312.00651*, 2023. 2

[22] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 2, 4, 6

[23] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *ICCV*, 2021. 2, 5, 6

[24] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, 2023. 1

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[26] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 2

[27] Zhili Liu, Kai Chen, Yifan Zhang, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, and James Kwok. Geom-erasing: Geometry-driven removal of implicit concept in diffusion models. *arXiv preprint arXiv:2310.05873*, 2023. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[29] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 5

[30] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 2

[31] Koustav Mullick, Harshil Jain, Sanchit Gupta, and Amit Arvind Kale. Domain adaptation of synthetic driving datasets for real-world autonomous driving. *arXiv preprint arXiv:2302.04149*, 2023. 1

[32] Suraj Patil, Pedro Cuenca, Nathan Lambert, and Patrick von Platen. Stable diffusion with diffusers. *Hugging Face–The AI community building the future.*, 2022. 5

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 4, 8

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 2

[38] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *ICCV*, 2019. 6

[39] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 6

[40] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 3, 5

[41] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 1

[42] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. In *NeurIPS*, 2023. 1, 2, 4

[43] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv preprint arXiv:2303.11681*, 2023. 4

[44] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *ICCV*, 2023. 2

[45] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *CVPR*, 2022. 6

[46] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *CVPR*, 2023. 4, 5, 6, 7

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 6

[48] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023. 4

[49] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *CVPR*, 2023. 6