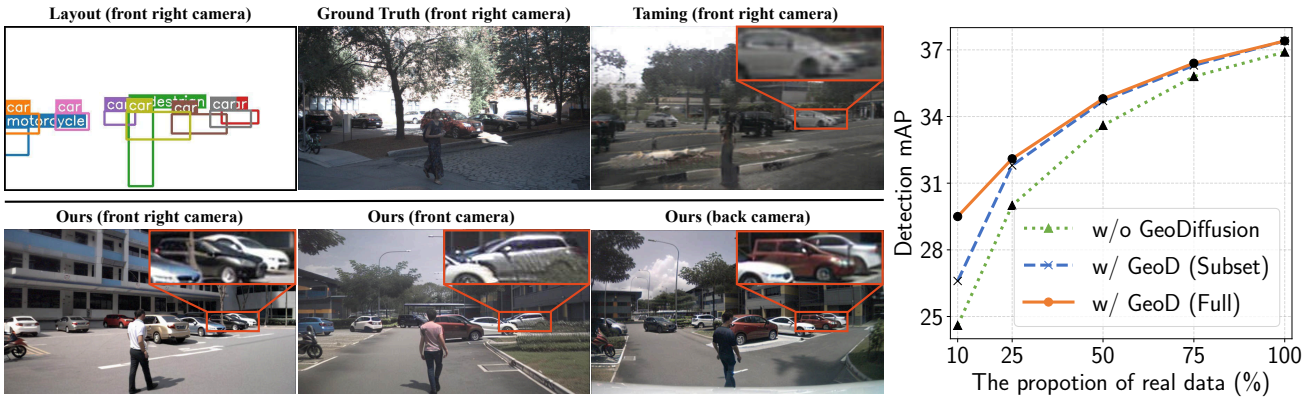
arXiv:2306.04607v1 [cs.CV] 7 Jun 2023

Integrating Geometric Control into Text-to-Image Diffusion Models for High-Quality Detection Data Generation via Text Prompt

Kai Chen[1*] Enze Xie[2*] Zhe Chen[3] Lanqing Hong[2†] Zhenguo Li[2] Dit-Yan Yeung[1]
[1]Hong Kong University of Science and Technology [2]Huawei Noah's Ark Lab [2]Nanjing University
kai.chen@connect.ust.hk {xie.enze,honglanqing,li.zhenguo}@huawei.com
chenzhe98@smail.nju.edu.cn dyyeung@cse.ust.hk
https://kaichen1998.github.io/geodiffusion/



(a) **Qualitative comparison** between *GeoDiffusion* and the state-of-the-art layout-to-image (L2I) generation method [28]. See more qualitative comparisons in Fig. 8.

(b) **The curve of mAP** with respect to the portion (%) of real data usage (*c.f.* Tab. 4).

Figure 1: (a) *GeoDiffusion* can generate high-quality object detection data by encoding various geometric conditions (*e.g.*, **bounding boxes** (bottom left) and **camera views** (bottom middle & right)) with a unified architecture. (b) **For the first time**, we demonstrate that L2I-generated images can benefit the training of object detectors [44], especially under annotation-scarce circumstances. See more detailed discussion about the experimental settings and results in Sec. 4.2.2.

Abstract

Diffusion models have attracted significant attention due to their remarkable ability to create content and generate data for tasks such as image classification. However, the usage of diffusion models to generate high-quality object detection data remains an underexplored area, where not only the image-level perceptual quality but also geometric conditions such as bounding boxes and camera views are essential. Previous studies have utilized either copy-paste synthesis or layout-to-image (L2I) generation with specifically designed modules to encode semantic layouts. In this paper, we propose GeoDiffusion, a simple framework that can flexibly translate various geometric conditions into text prompts and empower the pre-trained text-to-image (T2I) diffusion models for high-quality detection data generation. Unlike previous L2I methods, our GeoDiffusion is able to encode not only bounding boxes but also extra geometric conditions such as camera views in self-driving scenes. Extensive experiments demonstrate GeoDiffusion outperforms previous L2I methods while maintaining 4× training time faster. To the best of our knowledge, this is the first work to adopt diffusion models for layout-to-image generation with geometric conditions and demonstrate that L2I-generated images can be beneficial for improving the performance of object detectors.

1. Introduction

The cost of real data collection and annotation has been a longstanding problem in the field of deep learning. As a more cost-effective alternative, data generation techniques have been investigated for potential performance improvement [1, 42]. However, the effectiveness of such techniques has not met the expectation, mainly limited by the quality of

[*] Equal contribution.
[†] Corresponding author.

Paradigm	Realistic	Layout Control	Extra Geo. Control
Copy-paste [57]	✗	✓	✗
Layout-to-image [56]	✓	✓	✗
GeoDiffusion	✓	✓	✓

Table 1: **Key differences between our *GeoDiffusion* and existing works.** *GeoDiffusion* can generate highly realistic detection data with flexible fine-grained geometric control.

generated data. Recently, diffusion models (DMs) [26, 40] have emerged as one of the most popular generative models, owing to the remarkable ability to create content. Moreover, as demonstrated by He *et al*. [22], DMs can generate high-quality images to improve the performance of classification models. However, the usage of DMs to generate data for complex perception tasks (*e.g*., object detection [3, 20, 31]) has been rarely explored, which requires to consider about not only image-level perceptual quality but also geometric controls (*e.g*., bounding boxes and camera views). Thus, there is a need to investigate how to effectively utilize DMs for generating high-quality data for such perception tasks.

Existing works have primarily utilized two techniques to employ generative models for detection data generation: 1) copy-paste synthesis [13, 57] and 2) layout-to-image (L2I) generation [51, 56]. Copy-paste synthesis [13, 57] involves generating foreground objects and placing them on a pre-existing background image. Although proven beneficial for detectors, it only combines different parts of images instead of generating a complete scene, leading to less realistic images. The L2I generation [51, 56], on the other hand, adopts classical generative models (*e.g*., GAN [19] and VAE [30]), to directly generate realistic detection data conditional on the semantic layouts. However, it still remains a question of how to apply diffusion models to L2I generation. Moreover, L2I generation relies on specifically designed modules (*e.g*., RoI Align [23, 56] and layout-mask-image module [33, 51]) to encode layouts, limiting its flexibility to incorporate extra geometric conditions such as camera views. Therefore, the question arises: *Can we utilize a pre-trained powerful text-to-image (T2I) diffusion model to encode various geometric conditions for high-quality detection data generation?*

Inspired by the recent advancements of language models (LM) [2, 7, 41], this work proposes *GeoDiffusion*, a simple framework to translate different geometric conditions as a "foreign language" via text prompts to empower pre-trained text-to-image diffusion models [45] for high-quality object detection data generation. Different from the previous L2I methods which can only encode bounding boxes, our work can encode various additional geometric conditions flexibly benefiting from translating conditions into the text prompts. For example, *GeoDiffusion* is able to control image generation conditioning on camera views in self-driving scenarios. Considering the extreme imbalance among foreground and background regions, we further propose the foreground re-weighting mechanism which adaptively assigns higher loss weights to foreground regions while considering the area difference among foreground objects at the same time. Despite its simplicity, our *GeoDiffusion* can generate highly realistic images consistent with the geometric layouts, significantly surpassing previous L2I methods (**+21.85 FID** and **+27.1 mAP**). **For the first time**, we demonstrate generated images of L2I models can be beneficial for the training of object detectors, particularly in annotation-scarce scenarios. Furthermore, our *GeoDiffusion* can be utilized to generate novel, unseen images for the simulation purposes.

The main contributions of this work contain three parts:

1. We propose *GeoDiffusion*, an embarrassing simple framework to integrate geometric conditions into pre-trained text-to-image diffusion models for high-quality detection data generation. To our best knowledge, this is the first work to utilize diffusion models for layout-to-image generation with geometric conditions.

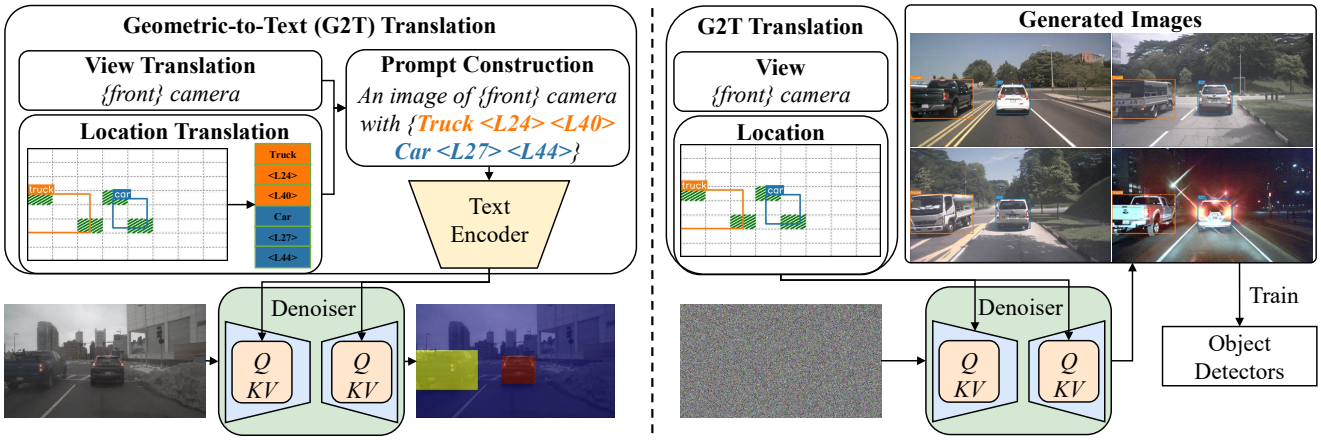
2. With extensive experiments, we demonstrate that our *GeoDiffusion* outperforms previous L2I methods by a significant margin while maintaining highly efficient (approximately 4$\times$ training acceleration).

3. For the first time, we demonstrate that the generated images of layout-to-image models can be beneficial to training object detectors, especially for the annotation-scarce circumstances in object detection datasets.

2. Related Work

Diffusion models. Recent progress in deep generative models has witnessed the success of diffusion models [26, 48, 49, 50], which generates image data through a progressive denoising diffusion procedure starting from a normal distributed random sample. These models, as evidenced by various studies [10, 27, 45], have demonstrated exceptional capabilities in image generation and potential applications, including text-to-image synthesis [39, 43], image-to-image translation [46, 47], inpainting [54], and text-guided image editing [24, 39]. Given the impressive success, employing the diffusion models to generate perception-centric training data holds significant promise for exploiting the boundaries of perceptual model capabilities. Thus, this paper explores the possibility of exploiting state-of-the-art text-to-image diffusion models to generate high-quality object detection data for training better object detectors.

Copy-paste synthesis. Object detection models require a large amount of data frequently. Considering this situation, the replication of image samples, also known as copy-paste [13, 14, 16, 17, 18], has emerged as a straightforward way



(a) Fine-tuning T2I diffusion models for L2I generation

(b) Generated images can be utilize to train object detector

Figure 2: **Model architecture of *GeoDiffusion*.** (a) *GeoDiffusion* encodes various geometric conditions via text prompts to empower text-to-image (T2I) diffusion models for generalized layout-to-image generation with various geometric conditions. (b) *GeoDiffusion* can generate highly realistic and diverse detection data to benefit the training of object detectors.

to improve the data efficiency of object detection models. Nikita *et al*. [13] first introduce Copy-Paste as an effective data augmentation for object detection, and observe that an appropriate context model is crucial for finding realistic object locations in new scenes automatically. Instaboost [16] differs from prior work by jittering existing instances rather than pasting the instances from other images. Simple Copy-Paste [18] presents a simple random placement strategy and yields solid improvements on strong baseline models. Cut-Paste-and-Learn [14] proposes to extract object instances, blend and paste them on diverse backgrounds, and train on augmented images in addition to the original dataset. Recently, [17, 57] perform the copy-paste synthesis by first generating foreground objects which are copied and pasted on a background image. Although beneficial for detectors, synthesized images are: a) *not realistic*; b) *no controllable* on fine-grained geometric conditions (*e.g*., camera views). Thus, we focus on integrating various geometric conditions into pre-trained text-to-image diffusion models.

Layout-to-image generation takes a textual or graphical input of a high-level layout and generates a corresponding photorealistic image. Despite significant recent progress on generative models [26, 49], controllable generation depicting complex object layouts is still difficult. To address that, Zhao *et al*. [56] proposes Layout2Im, a layout-based generation approach, which can generate a set of realistic images which have the correct objects in the desired locations by giving the coarse spatial layout. LostGAN [51] presents an end-to-end trainable architecture to generate from a reconfigurable layout and style. Taming [28] demonstrates that a conceptually simple model can outperform previous highly specialized systems when trained in the latent space. CAL2IM [23] adopts a novel context feature transformation module and a location-sensitive appearance representation to improve existing L2I methods, while LAMA [33] designs a locality-aware mask adaption module to adapt overlapped or nearby object masks during generation. Aforementioned methods all achieve promising generation performance, but still lack the capability to utilize extra geometric conditions. Moreover, the L2I-generated images have not been proven beneficial for object detectors, which is a huge advantage of our proposed *GeoDiffusion*.

3. Method

In this section, we first introduce the basic formulation of our *generalized layout-to-image* (GL2I) generation problem with geometric conditions and diffusion models (DMs) [26] in Sec. 3.1.1 and 3.1.2 separately. Then, we discuss how to flexibly encode the geometric conditions via text prompts to utilize pre-trained text-to-image (T2I) diffusion models [45] and build our *GeoDiffusion* in Sec. 3.2 and 3.3.

3.1. Preliminary

3.1.1 Generalized Layout-to-image Generation

Let $L = (v, \{(c_i, b_i)\}_{i=1}^{N})$ be a *geometric layout* with N bounding boxes, where $c_i \in \mathcal{C}$ denotes the semantic class, and $b_i = [x_{i,1}, y_{i,1}, x_{i,2}, y_{i,2}]$ represents the locations of the bounding box (*i.e*., top-left and bottm-right corners). $v \in \mathcal{V}$ can be any extra *geometric conditions* associated with the layout. Without loss of generality, we take camera views as an example in this paper. Therefore, the goal of *generalized layout-to-image* generation is to learn a $\mathcal{G}(\cdot, \cdot)$ to generate high-quality images $I \in \mathcal{R}^{H \times W \times 3}$ conditional on the given geometric layout L as,

$$I = \mathcal{G}(L, z), \tag{1}$$

where $z \sim \mathcal{N}(0, 1)$ is a random Gaussian noise.

3.1.2 Conditional Diffusion Models

Different from typical generative models like GAN [19] and VAE [30], diffusion models [26] learn data underlying distribution by conducting a T-step denoising process from normally distributed random variables, which can also be considered as learning an inverse process of a fixed Markov Chain of length T. Given a noisy input x_t at the time step $t \in \{1, ..., T\}$, the model $\epsilon_\theta(x_t, t)$ is trained to recover its clean version x by predicting the random noise added at time step t, and the objective function can be formulated as,

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

Latent diffusion models (LDM) [45] instead perform the diffusion process in the latent space of a pre-trained Vector Quantized Variational AutoEncoder (VQ-VAE) [52]. The input image x is first encoded into the latent space of the VQ-VAE encoder as $z = \mathcal{E}(x) \in \mathcal{R}^{H'\times W'\times D'}$, and then taken as clean samples in Eqn. 2. To facilitate conditional generation, LDM further introduces a conditional encoder $\tau_\theta(\cdot)$, and the objective function can be formulated as,

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2, \quad (3)$$

where y in the introduced condition (*e.g.*, text in LDM [45]).

3.2. Geometric Conditions as a Foreign Language

In this section, we explore encoding various geometric conditions via text prompts to utilize the pre-trained text-to-image diffusion models [45] for better GL2I generation. As discussed in Sec. 3.1.1, a geometric layout L consists of three basic components, including the locations $\{b_i\}$ and the semantic classes $\{c_i\}$ of bounding boxes and the extra geometric conditions v (*e.g.*, camera views).

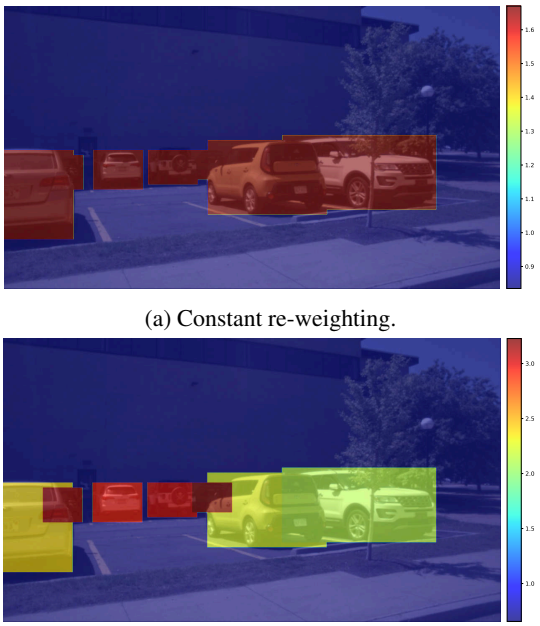
Location "Translation". While classes $\{c_i\}$ and conditions v can be naturally encoded by replacing with the corresponding textual explanations (see Sec. 4.1 for details), locations $\{b_i\}$ can not since the coordinates are continuous. Inspired by [9], we discretize the continuous coordinates by dividing the image into a grid of **location bins**. Each location bin corresponds to a unique **location token** which will be inserted into the text encoder vocabulary of T2I diffusion models. Therefore, each *corner* can be represented by the *location token* corresponding to the *location bin* it belongs to, and the "translation" procedure from the box locations to plain text is accomplished. See Fig. 2 for an illustration.

Specifically, given a grid of size (H_{bin}, W_{bin}), the corner (x_0, y_0) is represented by the $\sigma(x_0, y_0)$ location token as,

$$\sigma(x_0, y_0) = \mathcal{T}[y_{bin} * W_{bin} + x_{bin}], \quad (4)$$

$$x_{bin} = \lfloor x_0/W * W_{bin} \rfloor, \; y_{bin} = \lfloor y_0/H * H_{bin} \rfloor, \quad (5)$$

where $\mathcal{T} = \{< L_i >\}_{i=1}^{H_{bin}*W_{bin}}$ is the set of the location tokens, and $\mathcal{T}[\cdot]$ is the indexing operation. Thus, a bounding box (c_i, b_i) is encoded into a "pharse" with three tokens as $(c_i, \sigma(x_{i,1}, y_{i,1}), \sigma(x_{i,2}, y_{i,2}))$.



(a) Constant re-weighting.

(b) Area re-weighting.

Figure 3: **Foreground prior re-weighting**. (a) Constant re-weighting assigns equal weight to all bounding boxes, while (b) area re-weighting adaptively assigns higher weights to smaller boxes for better small object generation.

Prompt construction. To generate a text prompt, we can serialize multiple box "phrases" into a single sequence. In line with [9], here we utilize a random ordering strategy by randomly shuffling the box sequence each time a layout is presented to the model. Finally, we construct the text prompts with the template, "`An image of {view} camera with {boxes}`". As demonstrated in Tab. 2, this seemingly simple approach can effectively empower the T2I diffusion models for high-fidelity GL2I generation.

3.3. Foreground Prior Re-weighting

The objective function presented in Eqn. 3 is designed under the assumption of a uniform prior distribution across spatial coordinates. However, due to the extreme imbalance between foreground and background, we further introduce an adaptive re-weighting mask, denoted by $m \in \mathcal{R}^{H'\times W'}$, to adjust the prior. This enables the model to focus more thoroughly on foreground generation and better address the challenges posed by the foreground-background imbalance.

Constant re-weighting. To distinguish foreground from background regions, we employ a re-weighting strategy whereby foreground regions are assigned a weight $w(w > 1)$, greater than that assigned to the background regions.

Method	Input Resolution	Epochs	FID↓	Average Precision↑								
				mAP	AP_{50}	AP_{75}	AP^m	AP^l	trailer	const.	ped.	car
Oracle*	-	-	-	48.2	75.0	52.0	46.7	60.5	17.8	30.9	48.5	64.9
LostGAN [51]	256×256	256	59.95	4.4	9.8	3.3	2.1	12.3	0.3	1.3	2.7	12.2
LAMA [33]	256×256	256	63.85	3.2	8.3	1.9	2.0	9.4	1.4	1.0	1.3	8.8
Taming [28]	256×256	64	62.81	4.1	9.3	3.1	1.8	12.2	0.2	2.3	2.9	13.2
Taming [28]	256×256	256	32.84	7.4	19.0	4.8	2.8	18.8	6.0	4.0	3.0	17.3
GeoDiffusion	256×256	64	14.58$^{(+18.26)}$	15.6$^{(+8.2)}$	31.7	13.4	6.3	38.3	13.3	10.8	6.5	26.3
GeoDiffusion	512×512	64	**9.58**$^{(+23.26)}$	31.8$^{(+24.4)}$	62.9	28.7	27.0	53.8	21.2	27.8	18.2	46.0
GeoDiffusion	800×456	64	10.99$^{(+21.85)}$	**34.5**$^{(+27.1)}$	**68.5**	**30.6**	**31.1**	**54.6**	**20.4**	**29.3**	**21.6**	**48.8**

Table 2: **Comparison of generation fidelity on NuImages.** 1) Effectiveness: *GeoDiffusion* surpasses all baseline methods by a significant margin, suggesting the effectiveness of encoding geometric conditions via text prompts. With a larger input resolution, *GeoDiffusion* significantly shortens the gap with real images, making it appealing for training object detectors as in Tab. 3. 2) Efficiency: *GeoDiffusion* generates highly-discriminative images even under annotation-scarce circumstances. The "const." and "ped." suggests construction and pedestrian separately. *: represents the real image *Oracle* baseline.

Area re-weighting. The constant re-weighting strategy assigns equal weight to all foreground boxes, which causes larger boxes to exert a greater influence than smaller ones, thereby hindering the effective generation of small objects. To mitigate this issue, we propose an area re-weighting method to dynamically assign higher weights to smaller boxes. A comparison of this approach can be seen in Fig. 3. Finally, the re-weighting mask m can be represented as,

$$m'_{ij} = \begin{cases} w/c_{ij}^p & (i,j) \in \text{foreground} \\ 1/(H' * W')^p & (i,j) \in \text{background} \end{cases}, \quad (6)$$

$$m_{ij} = H' * W' * m'_{ij} / \sum_{i,j} m'_{ij}, \quad (7)$$

where c_{ij} represents the area of the bounding box to which pixel (i,j) belongs, and p is a tunable parameter. To improve the numerical stability during the fine-tuning process, Eqn. 7 normalizes the re-weighting mask m'. The benefits of this normalization process are demonstrated in Tab. 9.

Objective function. The final objective function of our *GeoDiffusion* model can be formulated as,

$$\mathcal{L}_{\text{GeoDiffusion}} = \mathbb{E}_{\mathcal{E}(x),\epsilon,t} \|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2 \odot m, \quad (8)$$

where y is the encoded layout as discussed in Sec. 3.2.

4. Experiments

4.1. Implementation Details

Dataset. Our experiments primarily utilize the widely used NuImages [3] dataset, which consists of 60K training samples and 15K validation samples with high-quality bounding box annotations from 10 semantic classes. The dataset captures images from 6 camera views (*front, front left, front right, back, back left and back right*), rendering it a suitable choice for our GL2I generation problem. Additionally, to showcase the universality of *GeoDiffusion* for common layout-to-image settings, we present experimental results on the COCO [4, 34] dataset in Sec. 4.3.

Optimization. We initialize the embedding matrix of the location tokens with 2D sine-cosine embeddings [53], while the remaining parameters of *GeoDiffusion* are initialized with Stable Diffusion[1], a pre-trained text-to-image diffusion model constructed on LDM [45]. With VQ-VAE fixed, we fine-tune all parameters of the text encoder and U-Net using AdamW [37] optimizer and a cosine learning rate schedule with a linear warm-up of 3000 iterations. The batch size is set to 64, and learning rates are set to $4e^{-5}$ for U-Net and $3e^{-5}$ for the text encoder. Layer-wise learning rate decay [12] is further adopted for the text encoder, with a decay ratio of 0.95. With a probability of 10%, the text prompt is replaced with a *null* text for unconditional generation. We fine-tune *GeoDiffusion* for 64 epochs, while baseline methods are trained for 256 epochs to maintain a similar training budget with the COCO recipe in [28, 33, 51]. Over-fitting is observed if training baselines longer. During generation, we sample images using the PLMS [35] scheduler for 100 steps with the classifier-free guidance (CFG) set as 5.0.

4.2. Main Results

The quality of object detection data is predicated on three key criteria: **fidelity**, **trainability**, and **generalizability**. Fidelity demands a realistic representation of objects while consistent with geometric layouts. Trainability concerns the usefulness of generated images for the training of object detectors. Generalizability demands the capacity to simulate uncollected, novel scenes in real datasets. In this section, we conduct a comprehensive evaluation of the proposed *GeoDiffusion* with regard to these critical areas.

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

Method	mAP	car	truck	trailer	bus	const.	bicycle	motorcycle	ped.	traffic cone	barrier
Real only	36.9	52.9	40.9	15.5	42.1	24.0	44.7	46.7	**31.3**	**32.5**	38.9
LostGAN [51]	35.6$^{(-1.3)}$	51.7	39.6	12.9	41.3	22.7	42.4	45.6	30.0	31.6	37.9
LAMA [33]	35.6$^{(-1.3)}$	51.7	39.2	14.3	40.5	22.9	43.2	44.9	30.0	31.3	38.3
Taming [28]	35.8$^{(-1.1)}$	51.9	39.3	14.7	41.1	22.4	43.1	45.4	30.4	31.6	38.1
GeoDiffusion	**37.4**$^{(+0.5)}$	**53.0**	**42.4**	**16.5**	**43.5**	**25.7**	**44.8**	**46.7**	30.6	31.7	**39.2**

Table 3: **Comparison of generation trainability on NuImages dataset.** 1) *GeoDiffusion* is **the only** layout-to-image method showing consistent improvements over almost all classes, 2) especially on rare classes, verifying *GeoDiffusion* indeed helps relieve annotation scarcity during detector training. The 800×456 *GeoDiffusion* variant is utilized for trainability evaluation.

4.2.1 Fidelity

Setup. We utilize two primary metrics on the NuImages validation set to evaluate Fidelity. The perceptual quality is measured with the Frechet Inception Distance (FID)[2] [25], while consistency between generated images and geometric layouts is evaluated via reporting the COCO-style average precision [34] using a pre-trained object detector, which is similar to the YOLO score in LAMA [33]. We use a pre-trained Mask R-CNN[3] [21] model trained on the NuImages training set to make predictions on generated images. These predictions are subsequently compared with corresponding ground truth annotations. We further provide the detection results on real validation images as the *Oracle* baseline in Tab. 2 for reference.

Discussion. As shown in Tab. 2, *GeoDiffusion* surpasses all baselines in terms of perceptual quality (FID) and layout consistency (mAP) with 256×256 input, accompanied by a **4× acceleration** in the training procedure, which indicates that encoding geometric conditions into text encoder of T2I diffusion models via text prompts is an effective approach. Moreover, the simplicity of LDM empowers *GeoDiffusion* to generate higher-resolution images with minimal modifications. With 800×456 input, *GeoDiffusion* attains **10.99 FID** and **34.5 mAP**, marking a significant progress towards bridging the gap with real images, especially for the large object generation (54.6 vs. 60.5 AP^l). Fig. 1a provides a qualitative comparison of generated images. *GeoDiffusion* generates images that are more realistic and tightly fitting to the bounding boxes, making it feasible to enhance object detector training, as discussed in Sec. 4.2.2.

We further report the class-wise AP of the top-2 frequent (*i.e.*, car and pedestrian) and rare classes (*e.g.*, trailer and construction, occupying only 1.5% of training annotations) in Tab. 2. We observe that *GeoDiffusion* performs relatively well in annotation-scarce scenarios, achieving higher trailer AP even than the *Oracle* baseline and demonstrating ability to generate highly-discriminative objects even with limited annotations. Instead, similar to previous generative models,

[2]Images are all resize into 256 × 256 before calculation.

[3]https://github.com/open-mmlab/mmdetection3d/tree/master/configs/nuimages

Method	10%	25%	50%	75%
w/o GeoDiffusion	24.6	30.0	33.6	35.8
w/ GeoD (Subset)	26.6$^{(+2.0)}$	31.8$^{(+1.8)}$	34.7$^{(+1.1)}$	36.3$^{(+0.5)}$
w/ GeoD (Full)	**29.5**$^{(+4.9)}$	**32.1**$^{(+2.1)}$	**34.8**$^{(+1.2)}$	**36.4**$^{(+0.6)}$

Table 4: **Necessity of real data.** Our *GeoDiffusion* achieves consistent improvement over various real subsets, which is more significant on more annotation-scarce subset.

GeoDiffusion encounters difficulty with high variance (*e.g.*, pedestrians) and occlusion (*e.g.*, cars) cases, highlighting the areas where further improvements are still needed.

4.2.2 Trainability

In this section, we investigate the potential benefits of *GeoDiffusion*-generated images for object detector training. To this end, we utilize the generated images as augmented samples during the training of an object detector to further evaluate the efficacy of our proposed model.

Setup. Taking data annotations of the NuImages training set as input, we first filter the bounding boxes smaller than 0.2% of the image area, then augment the bounding boxes by randomly flipping with 0.5 probability and shifting no more than 256 pixels. The generated images are considered augmented samples and combined with the real images to expand the training data. A Faster R-CNN [44] initialized with ImageNet pre-trained weights is then trained using the standard 1× schedule and evaluated on the validation set.

Discussion. As reported in Tab. 3, **for the first time**, we demonstrate that the generated images of layout-to-image models can be advantageous to object detector training. Our *GeoDiffusion* is the only method to achieve a consistent improvement for almost all semantic classes, which is much more obvious for rare classes (*e.g.*, **+1.0** for trailer, **+1.7** for construction and **+1.4** for bus, the top-3 most rare classes occupying only 7.2% of the training annotations), revealing that *GeoDiffusion* indeed contributes by relieving the annotation scarcity of rare classes, thanks to the data efficiency of *GeoDiffusion* as discussed in Sec. 4.2.1.

Necessity of real data. We further demonstrate the ease of real data necessity brought by *GeoDiffusion* by varying



Figure 4: **Visualization of generation generalizability**. From left to right, we present the query layout, the generated images conditional on the original, the flipped and shifted layouts. *GeoDiffusion* performs surprisingly well on the real-life collected geometric layouts (2nd & 3rd columns), while revealing superior robustness for out-of-distribution situations (4th column).

Method	Epoch	FID↓	mAP↑	AP_{50}↑	AP_{75}↑
LostGAN [51]	200	42.55	9.1	15.3	9.8
LAMA [33]	200	31.12	13.4	19.7	14.9
CAL2IM [23]	200	29.56	10.0	14.9	11.1
Taming [28]	68+60	33.68	-	-	-
TwFA [55]	300	22.15	-	28.2	20.1
Frido [15]	200	37.14	17.2	-	-
GLIGEN [32]	86	21.04	22.4	36.5	24.1
LayoutDiffuse [11]	60	20.27	-	-	-
GeoDiffusion	**60**	**20.16**	**29.1**	**38.9**	**33.6**

Table 5: **Comparison of generation fidelity on COCO.** *GeoDiffusion* obtains SoTA results on 256×256 inputs with superior efficiency, revealing universality of *GeoDiffusion*.

the amount of real data usage. We randomly sample 10%, 25%, 50%, and 75% of the real training set, and each subset is utilized to train a Faster R-CNN together with generated images separately. We consider two augmentation modes, 1) *Full*: the *GeoDiffusion* trained on the full training set is utilized to augment each subset as in Tab. 3. 2) *Subset*: we re-train *GeoDiffusion* with each real data subset separately, which are then used to augment the corresponding subset. The number of gradient steps are maintained unchanged for each pair experiment with the same amount of real data by enlarging the batch size adaptively.

As shown in Fig. 1b, *GeoDiffusion* achieves consistent improvement over different real training data budgets. The more scarce real data is, the more significant improvement *GeoDiffusion* achieves, as in Tab. 4, sufficiently revealing that generated images do help ease the data necessity. With only 75% of real data, the detector can perform comparably with the full real dataset by augmenting with *GeoDiffusion*.

Method	mAP	AP_{50}	AP_{75}	AP^s	AP^m	AP^l
Real only	37.3	**58.2**	40.8	**21.1**	40.7	48.2
LAMA [33]	35.3	54.8	38.2	19.3	38.6	45.7
GeoDiffusion	**37.6**	57.7	**40.9**	20.3	**41.1**	**49.9**

Table 6: **Comparison of trainability on COCO.** Our *GeoDiffusion* can indeed benefit regardless of domains.

4.2.3 Generalizability

In this section, we evaluate generalizability and robustness of *GeoDiffusion* on novel layouts unseen during fine-tuning.

Setup. To guarantee that the input geometric layouts are reasonable (*e.g.*, no violation of the basic physical laws like objects closer to the camera seem larger), we first randomly sample a query layout from the validation set, based on which we further disturb the query bounding boxes with 1) *flipping* and 2) *random shifting* for no more than 256 pixels along each spatial dimension, the exact same recipe we use in Sec. 4.2.2 for bounding box augmentation. Check more generalizability analysis for OoD circumstances in Sec. B.

Discussion. We visualize the generated results in Fig. 4. Our *GeoDiffusion* demonstrates superior generalizability to conduct generation on novel unseen layouts. Specifically, *GeoDiffusion* performs surprisingly well given geometric layouts collected in real-world scenarios and its corresponding flip variant (2nd & 3rd columns in Fig. 4). Furthermore, we observe that *GeoDiffusion* demonstrates strong robustness to layout augmentation even if the resulting layouts are out-of-distribution. For example, our *GeoDiffusion* learns to generate a downhill for boxes lower than current grounding plane (*e.g.*, the *shift* column of the 1st row), or an uphill



Figure 5: **Visualization of image inpainting on COCO.**

Method	mAP	AP_{50}	AP_{75}
Stable Diffusion	17.6	23.2	20.0
GeoDiffusion	**19.0**$^{(+1.4)}$	**26.2**$^{(+3.0)}$	**21.6**$^{(+1.6)}$

Table 7: **Comparison of image inpainting on COCO.**

for bounding boxes higher than the current grounding plane (*e.g.*, *shift* of 2nd & 3rd rows) to maintain consistent with given geometric layouts. The remarkable robustness also convinces us to adopt bounding box augmentation for the object detector training in Sec. 4.2.2 to further increase the annotation diversity of the augmented training dataset.

4.3. Universality

Setup. To demonstrate the universality of *GeoDiffusion*, we further conduct experiments on COCO-Stuff dataset [4] following common practices [23, 28, 33]. We keep hyper-parameters consistent with Sec. 4.1, except we utilize the DPM-Solver [38] scheduler for 50-step denoising with the classifier-free guidance ratio set as 4.0 during generation.

Fidelity. We ignore object annotations covering less than 2% of the image area, and only images with 3 to 8 objects are utilized during validation following [33]. Similarly with Sec. 4.2.1, we report FID and YOLO score [33] to evaluate the perceptual quality and layout consistency respectively. As shown in Tab. 5, *GeoDiffusion* outperforms all baselines in terms of both the FID and YOLO score with significant efficiency, consistent with our observation on NuImages in Tab. 2, revealing the universality of our *GeoDiffusion*. We further provide the qualitative comparison in Fig. 9.

Trainability. We then utilize *GeoDiffusion* to augment COCO detection training set following the exact same box augmentation pipeline in Sec. 4.2.2. As reported in Tab. 6, *GeoDiffusion* achieves significant improvement on COCO validation set also, suggesting that *GeoDiffusion* can indeed generate high-quality detection data regardless of domains.

Inpainting. We further explore applicability of inpainting for *GeoDiffusion*. Similarly with the fidelity evaluation, we randomly mask one object for each of COCO detection validation set, and then ask models to inpaint missing areas. A YOLO detector is utilized to evaluate the recognizability of inpainted results similarly with YOLO score [33] following [32]. As in Tab. 7, *GeoDiffusion* surpasses SD baseline remarkably. We further provide a qualitative comparison in Fig. 5, where *GeoDiffusion* can successfully deal with the complicated image synthesis requirement.

Pre-trained Text encoder	Grid size (H_{bin}, W_{bin})	# Pixels / bin	FID↓	mAP↑
✗	400×228	2×2	16.45	9.5
✓	**400×228**	**2×2**	**11.63**	**23.7**
✓	200×114	4×4	11.83	21.4
✓	100×57	8×8	14.94	20.8

Table 8: **Ablations on text encoder and location grid size.** The foreground re-weighting is not adopted in this ablation. The default settings are marked in gray.

4.4. Ablation study

Setup. We conduct ablation studies mainly with respect to fidelity and report the FID and COCO-style mAP following the exact same setting in Sec. 4.2.1. Specifically, the input resolution is set to be 800×456, and our *GeoDiffusion* is fine-tuned for 64 epochs on the NuImages training set. The optimization recipe is maintained the same with Sec. 4.1. Check more ablations and discussions in Sec. A and B.

Pre-trained text encoder. To verify the necessity of using pre-trained text encoder, we only initialize the VQ-VAE and U-Net with Stable Diffusion, while randomly initializing the parameters of the text encoder for comparison following the official implementation of LDM. As demonstrated in Tab. 8, the default *GeoDiffusion* significantly surpasses the variant without the pre-trained text encoder by 4.82 FID and 14.2 mAP, suggesting that with a proper "translation", the pre-trained text encoder indeed possesses transferability to encode geometric conditions and enable T2I diffusion models for high-quality object detection data generation.

Location grid size. We further ablate the location grid size (H_{bin}, W_{bin}) in Tab. 8. A larger grid size can achieve consistent improvement for both the perceptual quality and layout consistency. Indeed, a larger grid size suggests a smaller bin size and less coordinate discretization error, and therefore, a more accurate encoding of geometric layouts. Due to the restriction of hardware resources, the most fine-grained grid size we adopt is 2×2 pixels per location bin.

5. Conclusion

This paper proposes *GeoDiffusion*, an embarrassing simple architecture to encode geometric conditions via the text prompts and empower pre-trained text-to-image diffusion models for high-quality object detection data generation. The proposed model is shown to be effective in generating realistic images that conform to specified geometric layouts, as evidenced by the extensive experiments that demonstrate high fidelity, enhanced trainability in annotation-scarce scenarios, and improved generalizability to novel scenes.

Acknowledgments. We gratefully acknowledge support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

[1] 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

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 2

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 5, 13

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 5, 8, 14

[5] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *ICCV*, 2021. 11

[6] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *CVPR*, 2023. 11

[7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 2

[8] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 11

[9] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 4

[10] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *ICLR*, 2023. 2

[11] 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. 7, 11

[12] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020. 5

[13] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *ECCV*, 2018. 2, 3

[14] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *ICCV*, 2017. 2, 3

[15] 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. *arXiv preprint arXiv:2208.13753*, 2022. 7

[16] Hao-Shu Fang, Jianhua Sun, Runzhong Wang, Minghao Gou, Yong-Lu Li, and Cewu Lu. Instaboost: Boosting instance segmentation via probability map guided copy-pasting. In *ICCV*, 2019. 2, 3

[17] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Laurent Itti, and Vibhav Vineet. Dall-e for detection: Language-driven context image synthesis for object detection. *arXiv preprint arXiv:2206.09592*, 2022. 2, 3

[18] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 2, 3

[19] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 2, 4

[20] 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. 2

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 6

[22] 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. 2

[23] Sen He, Wentong Liao, Michael Ying Yang, Yongxin Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Context-aware layout to image generation with enhanced object appearance. In *CVPR*, 2021. 2, 3, 7, 8

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

[25] 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. 6

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3, 4

[27] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. In *JMLR*, 2022. 2

[28] Manuel Jahn, Robin Rombach, and Björn Ommer. High-resolution complex scene synthesis with transformers. *arXiv preprint arXiv:2105.06458*, 2021. 1, 3, 5, 6, 7, 8

[29] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. *arXiv preprint arXiv:2303.17559*, 2023. 11

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 4

[31] 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. 2

[32] 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. *arXiv preprint arXiv:2301.07093*, 2023. 7, 8, 11

[33] Zejian Li, Jingyu Wu, Immanuel Koh, Yongchuan Tang, and Lingyun Sun. Image synthesis from layout with locality-aware mask adaption. In *ICCV*, 2021. 2, 3, 5, 6, 7, 8

[34] 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, 6

[35] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 5

[36] Zhili Liu, Jianhua Han, Kai Chen, Lanqing Hong, Hang Xu, Chunjing Xu, and Zhenguo Li. Task-customized self-supervised pre-training with scalable dynamic routing. In *AAAI*, 2022. 11

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[38] 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. *arXiv preprint arXiv:2206.00927*, 2022. 8

[39] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[40] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[41] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 2

[42] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017. 1

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

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 6

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 4, 5

[46] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022. 2

[47] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. In *PAMI*, 2022. 2

[48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[51] Wei Sun and Tianfu Wu. Image synthesis from reconfigurable layout and style. In *ICCV*, 2019. 2, 3, 5, 6, 7

[52] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 4

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 5

[54] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *arXiv preprint arXiv:2212.06909*, 2022. 2

[55] Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. Modeling image composition for complex scene generation. In *CVPR*, 2022. 7

[56] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *CVPR*, 2019. 2, 3

[57] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: Revisit copy-paste at scale with clip and stablediffusion. *arXiv preprint arXiv:2212.03863*, 2022. 2, 3

[58] LIU Zhili, Kai Chen, Jianhua Han, HONG Lanqing, Hang Xu, Zhenguo Li, and James Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. In *ICLR*, 2023. 11

Appendix

A. More Ablation Study

Foreground prior re-weighting. In Tab. 9, we study the effect of foreground re-weighting. Adopting the constant re-weighting obtains a significant +3.4 mAP improvement (27.1 *vs.* 23.7), which further increases to 30.1 mAP with the help of area re-weighting, revealing that the foreground modeling is essential for object detection data generation. Note that the mAP improvement comes at a cost of a minor FID increase (11.99 *vs.* 11.63) since we manually adjust the prior distribution over spatial locations. We further verify the effectiveness of mask normalization in Eqn. 7, which can significantly decrease FID while maintaining the mAP value almost unchanged (4th & 6th rows), suggesting that mask normalization is mainly beneficial for fine-tuning the diffusion models after foreground re-weighting.

Importance of camera views. In this work, camera views are considered as an example to demonstrate that text can be indeed used as a universal encoding for various geometric conditions. A toy example is provided in Fig. 1a, fully proving text indeed has potential to decouple various conditions with a unified representation. Moreover, we further build a *GeoDiffusion w/o camera views*, significantly worse than the default *GeoDiffusion* as shown in Tab. 10, revealing the importance of adopting camera views.

Location tokens. As stated in Sec. 4.1, location tokens are first initialized with 2D sine-cosine embeddings and then fine-tuned together with the whole model. We further train a *GeoDiffusion with fixed location tokens*, which performs worse than the default *GeoDiffusion* as in Tab. 10.

B. More Discussions

More generalizability analysis. We first provide more visualization for augmented bounding boxes similarly with Sec. 4.2.3. As shown in Fig. 6, *GeoDiffusion* demonstrates superior robustness towards the real-life collected and augmented layouts, convincing us to flexibly perform bounding box augmentation for more diverse augmented set.

We further explore the generalizability for totally out-of-distribution (OoD) layouts (*i.e.*, unseen boxes and classes) in Fig. 7. *GeoDiffusion* performs surprisingly well for OoD bounding boxes (*e.g.*, unusually large bounding boxes with only one object in an entire image) as in Fig. 7a, but still suffers from unseen classes as in Fig. 7b, probably due to the inevitable forgetting during fine-tuning. Parameter-efficient fine-tuning (PEFT) [11, 32] might ease the problem but at a cost of generation quality as shown in Tab. 5. Considering our focus is to generate high-quality detection data to augment real data, fidelity and trainability are considered as the primary criteria in this work.

Constant w	Area p	Norm	FID↓	mAP ↑
1.0 (no re-weight)	0	✓	**11.63**	23.7
2.0	0	✓	11.77	27.1
4.0	0	✓	12.90	26.2
2.0	**0.2**	✓	11.99	**30.1**
2.0	0.4	✓	14.99	28.3
2.0	0.2	✗	14.76	29.8

Table 9: **Ablations on foreground re-weighting.** The best result is achieved when both constant and area re-weighting are adopted. Default settings are marked in gray.

Camera views	Learned Location Token	FID↓	mAP ↑
✓	✓	**10.99**	**34.5**
✗	✓	11.95	31.0
✓	✗	14.36	27.7

Table 10: **Ablations on the camera views and locatuon tokens.** Default settings are marked in gray.

Location translation via text. Although seeming cumbersome, the main purpose is to adopt **text as a universal encoding** for various geometric conditions and empower pre-trained T2I DMs for detection data generation, which, however, might not support the dense pixel-level semantic control currently (*e.g.*, mask-to-image generation).

Extendibility. Thus, *GeoDiffusion* can be extended to other descriptions as long as they can be discretized (*e.g.*, locations) or represented by text prompts (*e.g.*, car colors).

Adaptation of the Foreground Re-weighting with existing methods can be beneficial if applicable. However, existing methods utilize specific modules to encoder layouts (*e.g.*, RoI Align to take foreground features only in LAMA), suggesting that specific designs might still be required for the adaptation, which is beyond the scope of this work.

Limitation. We notice that the *GeoDiffusion*-generated images by now can only contribute as augmented samples to train object detectors together with the real images. It is appealing to train detectors with generated images solely, and we will explore it in the future work. We hope that our simple yet effective method can bring researchers' attention to large-scale object detection *dataset* generation with more flexible and controllable generative models.

Meanwhile, more flexible usage of the generated images beyond data augmentation, especially incorporation with generative pre-training [6, 58], contrastive learning [5, 36] and DM-based perception models [8, 29], is an appealing future research direction.



Figure 6: **More visualization of generation generalizability for augmented bounding boxes**. *GeoDiffusion* demonstrates superior performance for real-life collected and augmented layouts, consistently with what we have observed in Fig. 4.



(a) **Out-of-distribution bounding boxes.**



(b) **Out-of-distribution classes.**

Figure 7: **More visualization of generation generalizability for totally out-of-distribution layouts.** Our *GeoDiffusion* demonstrates strong robustness towards (a) OoD bounding boxes, but suffers from (b) unseen classes during fine-tuning.

C. More Qualitative Comparison

We provide more qualitative comparison on NuImages and COCO-Stuff datasets in Fig. 8 and 9 respectively.



Figure 8: **More qualitative comparison on the NuImages [3] dataset**.



Figure 9: **More qualitative comparison on the COCO-Stuff [4] dataset**. Our *GeoDiffusion* can successfully deal with both outdoor (1st-4th rows) and indoor (5th-7th rows) scenes, while demonstrating significant fidelity and diversity (4th-6th columns are generated images by *GeoDiffusion* under three different random seeds).