ReCon: Region-Controllable Data Augmentation with Rectification and Alignment for Object Detection

Haowei Zhu[1], **Tianxiang Pan**[2], **Rui Qin**[1], **Jun-Hai Yong**[1], **Bin Wang** [*1,3]
[1]Tsinghua University [2]Li Auto Inc. [3] BNRist
wangbins@tsinghua.edu.cn

Abstract

The scale and quality of datasets are crucial for training robust perception models. However, obtaining large-scale annotated data is both costly and time-consuming. Generative models have emerged as a powerful tool for data augmentation by synthesizing samples that adhere to desired distributions. However, current generative approaches often rely on complex post-processing or extensive fine-tuning on massive datasets to achieve satisfactory results, and they remain prone to content–position mismatches and semantic leakage. To overcome these limitations, we introduce **ReCon**, a novel augmentation framework that enhances the capacity of structure-controllable generative models for object detection. ReCon integrates region-guided rectification into the diffusion sampling process, using feedback from a pre-trained perception model to rectify misgenerated regions within diffusion sampling process. We further propose region-aligned cross-attention to enforce spatial–semantic alignment between image regions and their textual cues, thereby improving both semantic consistency and overall image fidelity. Extensive experiments demonstrate that ReCon substantially improve the quality and trainability of generated data, achieving consistent performance gains across various datasets, backbone architectures, and data scales. Our code is available at https://github.com/haoweiz23/ReCon.

1 Introduction

Robust object detection and instance segmentation models are essential in modern computer vision (Bochkovskiy et al., 2020; Carion et al., 2020; Zhu et al., 2020; Liu et al., 2024). However, these models are highly dependent on large-scale, meticulously annotated datasets whose creation is expensive and time consuming (Barkai et al., 1993; Cherti et al., 2023). For instance, annotating a single image in the Cityscapes dataset can take up to 60 minutes (Cordts et al., 2016). Consequently, there is a pressing need for efficient and automated methods to synthesize high-quality annotated training data.

Data augmentation has emerged as a vital strategy to alleviate data scarcity by increasing sample diversity and improving model generalization. Traditional augmentation methods (Zhong et al., 2020; DeVries & Taylor, 2017; Yun et al., 2019; Cubuk et al., 2018; Dvornik et al., 2018) typically introduce only minor local variations, falling short of generating truly novel content. Recent advances in generative modeling, especially structurally controllable frameworks, offer a promising alternative by leveraging Canny edges (Zhang et al., 2023; Zavadski et al., 2024), spatial layouts (Chen et al., 2023; Wang et al., 2024b), or instance masks (Wang et al., 2024a; Wu et al., 2023) to maintain fine-grained control during image synthesis.

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Structurally controllable generative models have achieved remarkable progress in geometric manipulation and are now extensively applied to object-detection data augmentation (Fang et al., 2024; Li et al., 2025). One family of methods repaints original images using universal control models such as ControlNet (Zhang et al., 2023; Zavadski et al., 2024) or inpainting models (Rombach et al., 2022; Lugmayr et al., 2022). These pipelines are often complex, requiring extra post-filters to remove noisy outputs (Fang et al., 2024) or multiple sampling processes (Kupyn & Rupprecht, 2024) to generate an image, for example, Kupyn & Rupprecht (2024) generating new samples for a single image containing ten objects via ten separate diffusion samplings.

Moreover, fine-tuning diffusion models conditioned on layouts or masks provides a feasible way for precise end-to-end synthetic data generation for object detection. Recent works demonstrate that such generated datasets yield strong trainability in downstream tasks (Chen et al., 2023; Wang et al., 2024b). However, these approaches typically require fine-tuning on large-scale datasets, which incurs significant computational overhead and remains impractical when data are scarce, which is a common scenario in data augmentation tasks. Furthermore, these approaches often struggle with complex layouts, leading to mis-generated regions and semantic misalignment.

To address these challenges, we propose **Region-Controllable (ReCon)** data augmentation. By integrating region-wise rectification and alignment directly into the diffusion sampling process, ReCon enhances single-pass control over instance synthesis. Without any additional training, ReCon significantly improves consistency between generated content and its annotations. It should be noted that we are not claiming to introduce a novel structural-control generation framework. Instead, our method can, without any additional training, enhance the quality of object detection data produced by existing structural controllable generation models. Specifically, our method first performs Region-Guided Rectification (RGR), in which we detect mis-generated regions by comparing the sampled image against ground-truth annotations using an off-the-shelf grounding model and then rectify those areas by injecting noisy real data points. By applying rectification to areas susceptible to be mis-generated, we boost the accuracy without sacrificing content diversity. Next, we introduce Region-Aligned Cross-Attention (RACA) to mitigate semantic leakage. This mechanism aligns region-specific visual tokens with their corresponding textual descriptions (or other cues) during generation. By enforcing a tight correspondence between image features and text embeddings at each diffusion step, it ensures precise semantic fidelity in the output.

By incorporating these two components into every sampling iteration, ReCon provides fine-grained region control: region-guided rectification preserves spatial agreement with annotations, and region-aligned cross-attention guarantees semantic adherence. The resulting high-fidelity augmented samples significantly enhance downstream object-detection performance. For example, when combined with a Canny-edge conditioned ControlNet model, our method achieves superior performance on the COCO dataset compared to models that have been specifically fine-tuned on COCO. In addition, our method demonstrates high augmentation efficiency: in data-scarce settings, tripling the dataset with our approach outperforms a sevenfold increase achieved by the baseline. Our contributions are as follows:

- We propose ReCon, a novel region-controllable data augmentation method that enhances the regional control capabilities of existing models without requiring additional training.
- We introduce region-guided rectification and region-aligned cross-attention mechanisms to improve control ability during the diffusion sampling process.
- Extensive experiments show that ReCon generates high-quality augmented data and substantially improves detection performance compared to both traditional augmentation techniques and current generative approaches.

2 Related Work

Conditional Generation Models. Recent advances in generative modeling have enabled the synthesis of high-fidelity images. Early research predominantly focused on Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) for image generation. Conditional GANs, for example, were utilized to train classification heads, thereby demonstrating the capability of GANs to model data distributions and generate novel samples in an unsupervised manner (Gurumurthy et al., 2017; Antoniou et al., 2017; Mariani et al., 2018; Zhang et al., 2021; Li et al., 2022a; Zhao & Bilen, 2022;

Xu et al., 2023). However, GANs are often plagued by training instability, mode collapse, and limited controllability, particularly in low-data regimes.

Diffusion models have recently emerged as a robust alternative, offering enhanced controllability and adaptability. These models implement a reverse denoising process that gradually removes noise from an initial Gaussian distribution to approximate the real data distribution (Yang et al., 2023a). Moreover, diffusion models can effectively handle a variety of conditioning inputs, including text, images, layouts, edges, depth maps, points, and masks. This flexibility has enabled their application to a wide range of tasks such as text-to-image synthesis (Podell et al., 2023; Esser et al., 2024), image editing (Meng et al., 2021; Rombach et al., 2022), image inpainting (Lugmayr et al., 2022; Saharia et al., 2022), and data augmentation (Fang et al., 2024; He et al., 2022). For instance, LAMA (Li et al., 2021) proposed a large mask inpainting strategy to enhance image quality, while Taming Transformers (Esser et al., 2020) demonstrated that training in a latent space can outperform more complex baselines. Further innovations include GLIGEN (Li et al., 2023c), which incorporates gated self-attention for improved layout control, and LayoutDiffuse (Cheng et al., 2023), which employs layout-specific attention modules tailored for bounding box guidance. Additionally, methods like GeoDiffusion (Chen et al., 2023) and Instance Diffusion (Wang et al., 2024a) integrate geometry-aware modules to encode spatial features, leading to superior generation outcomes. DetDiffusion (Wang et al., 2024b) introduces a perception-aware loss to effectively bridge the gap between generation and perception.

In this paper, we exploit these advanced, controllable generative models to produce high-quality synthetic data without extra training, with the goal of enhancing downstream detection tasks.

Generative Data Augmentation. Recent advancements in generative models (Rombach et al., 2022; Esser et al., 2024; Tian et al., 2025) have paved the way for synthesizing high-fidelity images that introduce novel content beyond the capabilities of traditional augmentation techniques (Cubuk et al., 2020, 2018; Yun et al., 2019; Chen et al., 2020). This enhanced data diversity is instrumental in improving the training of perceptual networks for tasks such as object detection.

Initial studies employed GANs (Goodfellow et al., 2020) for data augmentation. However, subsequent research has revealed several limitations of GAN-based approaches. For example, training networks like ResNet50 (He et al., 2016) on data synthesized by models such as BigGAN (Brock et al., 2018) often results in suboptimal performance compared to training with real images. Moreover, the inherent instability in GAN training and the difficulty of generating data under complex conditions (Bansal & Grover, 2023; Gowal et al., 2021; Ravuri & Vinyals, 2019) further constrain their effectiveness.

In contrast, diffusion models offer superior controllability and have gained widespread application in data generation. For image classification, methods such as LECF (He et al., 2022) use GLIDE (Nichol et al., 2021) to generate images and subsequently filter out low-confidence samples to enhance zero-shot and few-shot performance. Similarly, SGID (Li et al., 2023a) leverages BLIP (Li et al., 2022b) to ensure semantic consistency in generated outputs. Feng et al. (2023) filters samples based on feature similarity, while techniques like GIF (Zhang et al., 2022) and DistDiff (Zhu et al., 2024) incorporate additional guidance during the sampling process to refine the quality of generated samples. For object detection, recent methods such as GeoDiffusion (Chen et al., 2023) and DetDiffusion (Wang et al., 2024b) have demonstrated the ability to synthesize high-quality images with precise layout control, specifically designed for training detection models. Additional strategies include using diffusion models with post-filtering based on category-calibrated CLIP scores (Fang et al., 2024) and applying background inpainting to augment training data without extra annotations (Li et al., 2025).

Moreover, synthetic data has shown promise in other domains as well. For instance, MagicDrive (Gao et al., 2023) highlights the benefits of synthetic samples for 3D perception tasks, while TrackDiffusion (Li et al., 2023b) focuses on data generation for multi-object tracking. X-Paste (Zhao et al., 2023) and MosaicFusion (Xie et al., 2024) further contribute by producing samples with clear segmentation boundaries to boost instance segmentation performance.

Despite these advances, most current methods either require additional training of generative models or struggle to balance fidelity and diversity. In this work, we develop the generative data augmentation framework for object detection by leveraging emerging zero-shot recognition models (e.g., GroundedSAM (Ren et al., 2024)) alongside versatile conditional generation models (e.g., Stable Diffusion (Rombach et al., 2022) and ControlNet (Zhang et al., 2023)). Our approach eliminates the

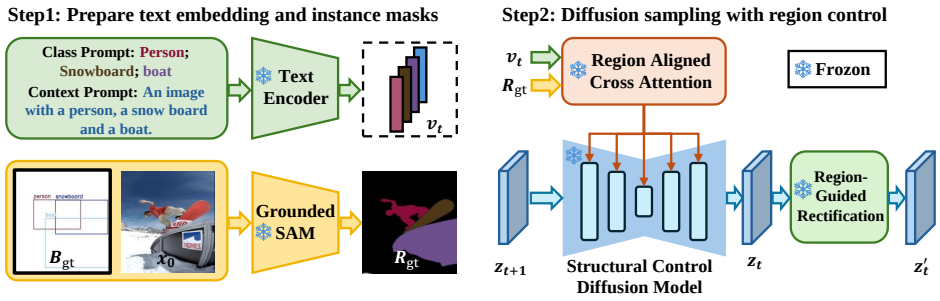


Figure 1: Overview of the ReCon Pipeline. ReCon enhances object-detection data generation by integrating region control into frozen, off-the-shelf models. It first compute the text embedding and instance masks, and then leverages a structural controllable diffusion model as the data generator and introduces region-guided rectification to refine generated results during the sampling process. Additionally, region-aligned cross-attention is incorporated to mitigate semantic leakage. Our method is plug-and-play and can be integrated with existing structure-controllable models.

need to retrain generative models, which is often impractical in data-scarce scenarios, and instead focuses on simplicity and effectiveness in generating task-specific data for target detectors.

3 Method

Task Definition. This study enhances a downstream object detector through data augmentation. By leveraging a pre-existing generator with the original image x, bounding boxes B, and class labels y, we aim to generate a high-fidelity augmented dataset where objects appear within the specified B and carry the correct labels y. The primary challenge is to preserve fidelity to the source while introducing useful novel variations (e.g., new colors, styles, or object poses) to increase content diversity and thereby improve downstream model performance.

Overview. Existing methods often face a trade-off between diversity and fidelity in generating downstream data. Recent works improve data diversity by using in-painting techniques to preserve certain image regions while redrawing others (Li et al., 2025; Ma et al., 2024a). Others improve fidelity by using perceptual models like CLIP (Radford et al., 2021) to filter out low-confidence samples (Fang et al., 2024; Zhao et al., 2023). In this paper, we propose a novel approach that utilizes an off-the-shelf perceptual model to adaptively calibrate image content during sampling, achieving a better balance between diversity and fidelity.

As illustrated in Figure 1, our method builds upon existing structural control models (e.g., ControlNet) to establish an initial layout control. During the sampling process, we employ a region-guided rectification strategy that refines instance-level content by automatically filtering out erroneous or low-confidence samples. Additionally, we introduce a region-aligned cross-attention mechanism to facilitate effective interaction between image content and the corresponding textual features.

3.1 Preliminaries

Stable Diffusion is a generative model that synthesizes high-quality images from textual prompts by operating within a compressed latent space. It comprises forward and reverse diffusion stages.

Forward Process. In the forward process, noise is gradually added to the latent representation $\mathbf{z}_0$ of an image x, turning it into pure Gaussian noise $\mathbf{z}_T$ after T timesteps. This diffusion process is modeled by:

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}\Big(\mathbf{z}_t; \sqrt{\alpha_t}\,\mathbf{z}_{t-1},\, (1-\alpha_t)\,\mathbf{I}\Big), \tag{1}$$

where α_t controls the balance between the previous latent state and the injected noise.

Denoising Process. Starting with $\mathbf{z}_T$ (pure Gaussian noise), the model iteratively predicts and removes noise to generate the clean latent $\mathbf{z}_0$. This reverse sampling process is modeled by:

$$p_\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t) = \mathcal{N}\Big(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \Sigma_\theta(t)\Big), \tag{2}$$

where θ is the denoising network, $\mu_\theta(\mathbf{z}_t, t)$ and $\Sigma_\theta(t)$ denote the predicted mean and covariance. Sequentially applying this reverse process from $t = T$ down to $t = 1$ effectively removes the noise, recovering $\mathbf{z}_0$ for final image generation.

Cross-Attention Mechanism. In text-to-image generation, the text condition $\mathbf{v}_t$ (encoded by a text encoder like CLIP) is integrated into the latent space via cross-attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3}$$

where the query $\mathbf{Q}$ is derived from image features, and $\mathbf{K}$ and $\mathbf{V}$ derived from text embeddings. Here, d_k denotes the key dimension. This mechanism injects semantic text information into the latent features at each denoising step, guiding the image generation to reflect the text prompt.

During sampling, noisy latent representations are progressively denoised while being continuously influenced by the text embeddings. Starting the denoising from different timesteps allows control over the editing intensity, balancing adherence to the original image content. However, since the Stable Diffusion model lacks inherent structural control, additional structure-controllable models are required for generating object detection data.

3.2 Region-Controllable Data Augmentation

Structural Control with ControlNet. ControlNet (Zhang et al., 2023) enhances diffusion models (e.g., Stable Diffusion) by conditioning them on structural cues such as edge, depth, or pose maps. It integrates trainable control layers as follows:

$$\hat{\mathbf{z}}_l = \mathbf{z}_l + \gamma \cdot \text{ControlBlock}(\mathbf{c}_m, \theta_c) \tag{4}$$

where $\mathbf{z}_l$ are the latent features at layer l, $\mathbf{c}_m$ is the structural conditioning map, θ_c are learnable parameters of control blocks, and γ scales the control signal.

In our work, we use ControlNet with an edge canny map to enforce structural constraints during image generation, and we demonstrate that our approach can generalize to other layout-to-image models for diverse guided generation.

Region-Guided Rectification. Existing generative models often encounter issues such as generating an incorrect number of target objects or unintended ones. These challenges significantly affect the quality of the generated data. To address these problems, we propose a region-guided rectification method aimed at perceiving image content during sampling and applying region adjustments. This approach ensures consistency of the content and the layout.

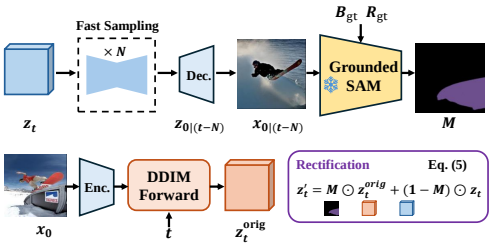


Figure 2: Sampling process with Region-Guided Rectification . We first identify incorrectly generated regions by IoU matching detection results with the original annotations using a grounding model. Next, we derive a rectification mask $\mathbf{M}$ and use $\mathbf{z}_t^{orig}$ sampled from the original data point to correct these errors.

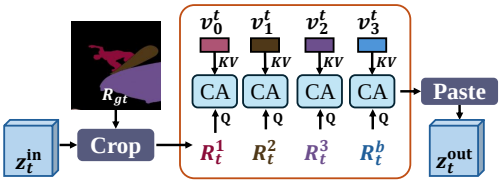


Figure 3: Pipeline of our Region-Aligned Cross-Attention. We first crop region-specific features from $\mathbf{z}_t^{in}$ using predefined regions R_{gt}. For regions belonging to the same category, we perform cross-attention with the corresponding text features. Finally, the interacted region image features are concatenated to produce $\mathbf{z}_t^{out}$.

As shown in Figure 2, given an image annotation comprising multiple bounding boxes B and their corresponding labels y, we employ the Grounded-SAM model (Ren et al., 2024) to detect potential objects in the data point $\mathbf{z}_t$ during the sampling process. We then apply IoU-based matching to identify false positives and false negatives, allowing us to segment regions that are potentially misgenerated.

These out-of-control regions are defined by a binary mask $\mathbf{M}$. The identified regions are then replaced with their corresponding noised versions $\mathbf{z}_t^{\text{orig}}$ sampled from the original image, while the remaining parts of $\mathbf{z}_t$ are preserved. This region-guided rectification process can be formulated as:

$$\mathbf{z}_t' = \mathbf{M} \odot \mathbf{z}_t^{\text{orig}} + (1 - \mathbf{M}) \odot \mathbf{z}_t, \tag{5}$$

where $\odot$ denotes the element-wise multiplication, $\mathbf{z}_t'$ represents the rectified latent data point, and $\mathbf{z}_t^{\text{orig}}$ is the latent point obtained after applying t steps of noise addition to the original image using Equation 1. This method leverages the intrinsic ***overridability*** property of diffusion sampling (Levin & Fried, 2023), allowing regions in intermediate images to be replaced with external content drawn from the same distribution. Such rectifications can influence the final generated image without disrupting the overall inference process.

However, directly detecting the intermediate latent point $\mathbf{z}_t$ with a perception model is impractical due to the ***challenge of finding a pre-trained detector that provides meaningful guidance when the input is noisy***. To address this issue, we leverage recent cache-based diffusion acceleration method (Ma et al., 2024b) to speed up sampling over N steps (with $N = 5$ by default). Furthermore, we utilize the capability of the diffusion model to predict the noise added to $\mathbf{z}_{t-N}$, enabling the prediction of a clean data point $\mathbf{z}_{0|t-N}$. This process is formulated in Equation 6.

$$\mathbf{z}_{0|t-N} = \frac{\mathbf{z}_{t-N} - \sqrt{1-\alpha_t}\theta(\mathbf{z}_{t-N}, t)}{\sqrt{\alpha_t}}, \tag{6}$$

Then we apply region rectification at T_r timesteps ($T_r = 4$), corresponding to the early (0.75 T), middle (0.50 T), latter (0.25 T) and final (0.10 T) stages of the diffusion process. At the early stage, when the overall object layout has begun to emerge, we can correct inaccuracies in the spatial distribution of objects. During the middle stage, as the object starts to take shape, we rectify any incorrect semantic content in the objects. Finally, in the latter and final stages, we refine regions with suboptimal generation quality. More details of the sampling process are presented in Algorithm 1.

Region-Aligned Cross Attention. In text-to-image generation, semantic leakage often occurs, where the content of target regions does not align with the actual textual descriptions. To address this issue, we introduce region-aligned cross attention to mitigate information leakage across regions.

Since attention within the text encoder operates on all prompt tokens, and these tokens may belong to different categories, interference between category-specific features can occur (see the appendix Figure 8). To address this, we individually encode C textual features for C target categories using prompts in the format: `[CLASS]`, as shown in Figure 1. Additionally, we employ a global context description to represent the overall scene, which interacts with the background region. For datasets like COCO, we can directly utilize the provided caption annotations or simply use a custom prompt, such as: "`An image with two cars and three persons`".

Next, as presented in Figure 3, we perform cross-attention interactions between the corresponding object regions and their associated textual features. This step alleviates information leakage in prompt descriptions by ensuring that region-specific textual features influence only their respective regions. An alternative approach to implementing region-aware cross-attention is to use an cross attention mask to suppress text features from unrelated categories, as proposed in (Xue et al., 2023). However, we have observed that, due to the lack of disentanglement during the encoding of textual features from different categories, the masked attention mechanism still suffers from semantic leakage. Besides, Instance Diffusion introduces an instance-masked attention and fusion mechanism to integrate region-specific conditions with corresponding visual tokens. However, it relies on additional region-specific modules and requires retraining to achieve satisfactory performance. In contrast, our approach mitigates the problem of semantic leakage and enhances the fidelity of generated images without the need for additional fine-tuning. Furthermore, we demonstrate that our method can be effectively combined with Instance Diffusion to further boost its performance, as shown in Table 1 and Figure 8.

4 Experiments

4.1 Experiment Settings

We evaluate our method by augmenting downstream object detectors with synthetic samples. We use Stable Diffusion v1.5 (Rombach et al., 2022) with a 25-step DDIM sampler (Song et al., 2020) and

edge-conditioned ControlNet (Zhang et al., 2023) to generate training images. These samples are combined with the original trainset and used to jointly train object detectors. We implement training and evaluation code based on the MMDetection framework (Chen et al., 2019). For consistency with prior work (Wang et al., 2024b), our default detector is Faster R-CNN (Ren et al., 2015) with an R-50-FPN backbone trained for six epochs. We also evaluate our method with diverse detectors including RetinaNet (Lin et al., 2017), ATSS (Zhang et al., 2020), FCOS (Tian et al., 2019), YOLO-X (Ge et al., 2021), and DEIM (Huang et al., 2024). Following previous works, we select images containing 3 to 8 objects for data generation, resulting data set comprising 47,200 images with 227,406 objects. For VOC benchmark, we combine the training sets of VOC 2007 and VOC 2012 for model training, with evaluation performed on the VOC 2007 test set (4,952 images). We use mAP (mean Average Precision), mAR (mean Average Recall), FID to evaluate the performance. Extensive experiments are conducted across various datasets, backbone architectures, and data scales.

4.2 Main Results

Compared with State-of-the-art Methods. We compare our approach with state-of-the-art structure-controllable generative diffusion models, as summarized in Table 1. The results demonstrate that our method significantly enhances the effectiveness of structure-guided techniques for object detection data augmentation. Specifically, we evaluate both general-purpose control methods (e.g., ControlNet) and models fine-tuned on COCO (e.g., DetDiffusion). When combined with these methods, our approach further improves their performance and establishes a new state of the art. For instance, integrating ReCon with ControlNet yields a mAP of 35.5, surpassing GeoDiffusion's 34.8. Moreover, our method can act as a plug-and-play enhancement for region-controlled diffusion models without requiring additional training. This is exemplified by the improvement in GLIGEN's mAP from 34.6 to 35.5. These findings validate that our approach enables the generation of higher-quality training samples, resulting in a substantial boost in object detection performance.

Table 1: Comparison with existing generative models on the COCO dataset. ReCon enhances the detector performance by integrating existing methods in a training-free manner. The best results are highlighted in **bold**, while the second-best outcomes are denoted by *underlined italic*.

Method	mAP	AP_{50}	AP_{75}	AP^m	AP^l
Real only	34.5	55.5	37.1	37.9	44.3
▷ *General Control*					
Layout Diffusion (Zheng et al., 2023) [CVPR23]	34.0	54.5	36.5	37.2	43.6
ControlNet (Zhang et al., 2023) [ICCV23]	34.9	55.5	37.7	38.2	45.5
Background-inpainting (Li et al., 2025) [ECCV24]	35.1	55.1	37.7	38.2	45.8
ControlNet-XS (Zavadski et al., 2024) [ECCV24]	35.1	55.8	37.6	38.6	45.0
▷ *Fine tuned on COCO*					
ReCo (Yang et al., 2023b) [CVPR23]	33.6	53.2	36.2	36.7	44.0
GLIGEN (Li et al., 2023c) [CVPR23]	34.6	55.1	37.2	38.1	44.7
GeoDiffusion (Chen et al., 2023) [ICLR24]	34.8	55.3	37.4	38.2	45.4
DetDiffusion (Wang et al., 2024b) [CVPR24]	35.4	55.8	38.3	38.5	**46.6**
Instance Diffusion (Wang et al., 2024a) [CVPR24]	35.0	55.4	37.6	38.4	45.7
ControlNet + **ReCon**	<u>35.5</u>	**56.2**	**38.4**	**39.0**	46.0
GLIGEN + **ReCon**	35.3	<u>56.0</u>	38.1	38.7	45.8
Instance Diffusion + **ReCon**	**35.6**	56.0	<u>38.4</u>	<u>39.0</u>	<u>46.4</u>

Data-Scarce Scenarios. Data augmentation is crucial when training data is limited. To evaluate our approach under such conditions, we conduct experiments in three data-scarce regimes by randomly sampling 1%, 5%, and 10% of the COCO training set and then doubling each subset through augmentation. Our method delivers consistent gains over baseline approaches in all regimes. As shown in Table 2, with only 10% of the data, mAP rises from 18.5% to 21.7%. Training-based generative models often struggle in data-scarce settings due to their dependence on large datasets. In contrast, we employ a generic structure-controlled diffusion model (ControlNet) to produce high-quality object detection samples. We further compare our method to traditional augmentation method RandAugment (Cubuk et al., 2020). Although RandAugment shows noticeable improvement, it

remains inferior to our approach. Moreover, combining our method with RandAugment produces additional improvements, demonstrating compatibility with standard augmentation pipelines.

Few-Shot Scenarios. We also evaluated our method in a 30-shot training setting on YOLOX-S (Ge et al., 2021) using COCO dataset, following the few-shot split protocol of previous work (Wang et al., 2020). Our method performs well even under the few-shot setting, increasing the mAP from 5.4 to 6.7 and AP_{50} from 10.3 to 12.3. More few-shot results are presented in Table 11 in Appendix.

Table 2: Comparison in data-scarce scenarios across varying data proportions.

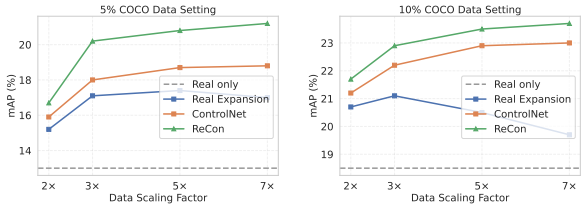
Method	1%	5%	10%
Real only	0.3	13.0	18.5
RandAugment	3.1	16.2	21.4
ControlNet	2.5	15.9	21.2
Recon	3.9	16.7	21.7
Recon + RandAugment	**4.2**	**17.1**	**22.0**

Comparison on Different Dataset. To validate the generalization capability of our method, we conducted additional experiments on PASCAL VOC datasets. We compared our approach against a baseline Faster R-CNN detector trained with $1\times$ schedule. As demonstrated in Table 3, traditional data augmentation methods like RandAugment (Cubuk et al., 2020) show limited effectiveness, while simply duplicate the original dataset leads to overfitting (76.2 vs. 77.1 mAP). Our method achieves superior performance (78.5 mAP) through synthetic generation of diverse high-fidelity images that maintain crucial semantic features.

Table 3: Comparison results on the PASCAL VOC dataset.

Method	Real only	Simple Duplicate	RandAugment	ControlNet	**ReCon**
mAP	77.1	76.2	77.7	77.8	**78.5**

Data Scaling. To assess scalability we measure detection accuracy in low-data regimes (5% and 10%), summarized in Figure 4. Repeating training data (real expansion) provides consistent improvements up to $3\times$: mAP increases from 13.0 to 17.1 on the 5% subset and from 18.5 to 21.1 on the 10% subset. However, further duplication ($5\times$, $7\times$) leads to saturated performance and noticeable degradation, indicating overfitting under extended training. By contrast, our method generates diverse, annotation-consistent samples and continues to yield performance gains without overfitting. As the expansion scale grows, the relative accuracy improvements from our augmented data become more pronounced. Overall, our approach attains comparable or better performance than the ControlNet baseline using substantially fewer augmented examples, highlighting its efficiency for data augmentation.



(a) 5% Data (b) 10% Data

Figure 4: Data scaling on different COCO subsets.

4.3 Ablation Studies

Effectiveness of Each Component. We conducted ablation experiments to assess the contributions of each component in our proposed framework, as presented in Table 4. The results clearly indicate that each component enhances the performance of the downstream model. In particular, the integration of region-guided rectification (RGR) and region-aligned cross attention (RACA) significantly improves the consistency between the generated samples and their corresponding annotations, thereby elevating the quality of the synthesized data. Consequently, our approach increases the baseline mAP from 34.9 to 35.5 and improves the FID from 13.82 to 12.85, demonstrating enhanced trainability and fidelity of the generated samples.

Table 4: Ablation results for different components of our proposed method.

RGR	RACA	FID	mAP	AP_{50}	AP_{75}	AP^m	AP^l
✘	✘	13.82	34.9	55.5	37.7	38.2	45.5
✔	✘	13.21	35.3	56.0	38.1	38.6	45.6
✔	✔	**12.85**	**35.5**	**56.2**	**38.4**	**39.0**	**46.0**

Table 5: Performance comparison of different perception targets.

Target	mAP	AP_{50}	AP_{75}
x_t	35.0	55.6	37.8
$x_{0\|t}$	35.3	55.8	38.2
$x_{0\|(t-N)}$	**35.5**	**56.2**	**38.4**

Table 6: Trainability comparison with DEIM-D-FINE-N (Huang et al., 2024) on COCO.

Method	mAP	AP_{50}	AP_{75}	mAR
Real only	38.5	55.2	41.5	60.4
ControlNet	39.1	55.8	42.1	60.6
ReCon	**39.8**	**56.6**	**42.5**	**61.0**

Table 7: Comparison with different region-guided models.

Method	mAP	AP_{50}	AP_{75}
Swin-Tiny	35.5	56.2	38.4
Swin-Base	**35.6**	**56.2**	**38.7**

Perception Target. Our method leverages cache-based fast sampling (Ma et al., 2024b) to recover a clean x_0, providing a more accurate control signal for region-guided rectification. we compare different perception targets: x_t, $x_{0|t}$, and $x_{0|(t-N)}$. As shown in Table 5. While x_t yields modest gains due to low recall which in turn causes the model to favor a lower overall editing strength. In contrast, employing $x_{0|t}$ further enhances performance, and the best results are achieved when using $x_{0|(t-N)}$ obtained via the fast sampling method.

Different Detection Backbone. We evaluate multiple object detectors and report results on the state-of-the-art DEIM (CVPR25) method in Table 6. Additional detector comparisons are provided in Table 10 in the Appendix. Our experiments show that our method consistently improve performance across different detectors, demonstrating its robustness and effectiveness.

Figure 5: Visualization comparison of generated samples. Our methods show better image fidelity and content-annotation consistency.

Perception Model. Our approach employs the Grounded-SAM (Ren et al., 2024) as region-guided model to provide region-aware guidance. Additionally, we compare different backbone models for the detector within Grounded-SAM, as detailed in Table 7. The experimental results indicate that better perception leads to improved performance, suggesting that our method stands to benefit from stronger foundation models.

4.4 Qualitative Results

Figure 5 shows that our method substantially improves both the fidelity and the localization accuracy of generated samples. Unlike prior structure-control methods such as GLIGEN and ControlNet, which lack mechanisms for fine-grained region rectification and hence can exhibit imprecise localization and semantic leakage, our approach enforces strict consistency with the provided annotations. For

example, it removes an extraneous zebra produced by GLIGEN outside the target bounding box (row 1) and a superfluous sheep outside the region of interest (row 3), and it correctly restores a person that ControlNet fails to generate (row 2). By aligning generated content with the original annotations, our method improves overall generation quality while maintaining high fidelity and sample diversity. More visualization results are provided in the Appendix.

5 Conclusion

This paper presents **Re**gion-**Con**trollable data augmentation (ReCon), a training-free, diffusion-based method developed to generate high-quality, content-label-aligned synthetic data for enhancing object detection models. Extensive experimental evaluations demonstrate that ReCon outperforms traditional augmentation and generative methods, ultimately leading to superior detection performance.

6 Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62021002.

References

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*, 2023.

N Barkai, Hyunjune Sebastian Seung, and Haim Sompolinsky. Scaling laws in learning of classification tasks. *Physical review letters*, 70(20):3167, 1993.

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

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. *arXiv preprint arXiv:2306.04607*, 2023.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020.

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.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, June 2023.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3043–3054, 2023.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout, 2017.

Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 364–380, 2018.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1257–1266, 2024.

Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2704–2714, 2023.

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.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233, 2021.

Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 166–174, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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.

Shihua Huang, Zhichao Lu, Xiaodong Cun, Yongjun Yu, Xiao Zhou, and Xi Shen. Deim: Detr with improved matching for fast convergence. *arXiv preprint arXiv:2412.04234*, 2024.

Orest Kupyn and Christian Rupprecht. Dataset enhancement with instance-level augmentations. In *European Conference on Computer Vision*, pp. 384–402. Springer, 2024.

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, Daniel Haziza, Luca Wehrstedt, Jeremy Reizenstein, and Grigory Sizov. xformers: A modular and hackable transformer modelling library. `https://github.com/facebookresearch/xformers`, 2022.

Eran Levin and Ohad Fried. Differential diffusion: Giving each pixel its strength. *arXiv preprint arXiv:2306.00950*, 2023.

Bohan Li, Xinghao Wang, Xiao Xu, Yutai Hou, Yunlong Feng, Feng Wang, and Wanxiang Che. Semantic-guided image augmentation with pre-trained models. *arXiv preprint arXiv:2302.02070*, 2023a.

Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21330–21340, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022b.

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*, 2023b.

Yuhang Li, Xin Dong, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. A simple background augmentation method for object detection with diffusion model. In *European Conference on Computer Vision*, pp. 462–479. Springer, 2025.

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*, 2023c.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Siwei Lyu. Deepfake detection: Current challenges and next steps. pp. 1–6, 2020.

Fulong Ma, Weiqing Qi, Guoyang Zhao, Ming Liu, and Jun Ma. Erase, then redraw: A novel data augmentation approach for free space detection using diffusion model. *arXiv preprint arXiv:2409.20164*, 2024a.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15762–15772, 2024b.

Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 786–808, 2023.

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.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.

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

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. *arXiv preprint arXiv:2002.08911*, 2020.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*. Association for Computing Machinery, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2025.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.

Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *International Conference on Machine Learning*, pp. 9919–9928. PMLR, 2020.

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6232–6242, 2024a.

Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, et al. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7246–7255, 2024b.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1206–1217, 2023.

Jiahao Xie, Wei Li, Xiangtai Li, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *International Journal of Computer Vision*, pp. 1–20, 2024.

Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16111–16121, 2024.

Austin Xu, Mariya I Vasileva, Achal Dave, and Arjun Seshadri. Handsoff: Labeled dataset generation with no additional human annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7991–8000, 2023.

Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14256–14266, 2023.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023a.

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*, 2023b.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Denis Zavadski, Johann-Friedrich Feiden, and Carsten Rother. Controlnet-xs: Rethinking the control of text-to-image diffusion models as feedback-control systems. In *European Conference on Computer Vision*, pp. 343–362. Springer, 2024.

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

Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9759–9768, 2020.

Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. *arXiv preprint arXiv:2211.13976*, 2022.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10145–10155, 2021.

Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. *arXiv preprint arXiv:2204.07513*, 2022.

Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In *International Conference on Machine Learning*, pp. 42098–42109. PMLR, 2023.

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.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Haowei Zhu, Ling Yang, Jun-Hai Yong, Hongzhi Yin, Jiawei Jiang, Meng Xiao, Wentao Zhang, and Bin Wang. Distribution-aware data expansion with diffusion models. *Advances in Neural Information Processing Systems*, 37:102768–102795, 2024.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

A Limitations and Societal Impacts

Limitations. Although ReCon produces better FID scores and improves detector mAP without additional training, it may increase computation time as data volume grows. The requirement for an extra perception model also raises development costs. We have introduced acceleration techniques to lower these costs. Integrating a fast sampler (Luo et al., 2023) and a lightweight perception model offers a promising path to maximize the efficiency of diffusion based data augmentation for detector training.

Societal Impacts. Generative models (Podell et al., 2023; Zavadski et al., 2024) offer a cost-effective alternative to manual data collection and annotation for object detection. Our method enhances these models' ability to produce annotation-consistent content without additional training and improves downstream detector performance. This efficiency can benefit organizations and researchers who face limited data resources.

However, because generative models are pre-trained on large, uncurated vision–language datasets from the internet, they may inherit social biases and stereotypes (Cho et al., 2023; Naik & Nushi, 2023; Ross et al., 2020) and produce discriminatory outputs. It is therefore essential to integrate bias detection and mitigation mechanisms. By applying region-wise rectification and alignment, our approach improves content accuracy and reduces bias.

Another concern is the potential misuse of synthetic imagery for purposes such as deepfakes (Lyu, 2020), which can spread misinformation and undermine societal trust. To address this risk, the community must establish regulations and best practices that ensure responsible creation and use of synthetic data models.

B Pseudo Algorithm

We present the pseudo algorithm of our Region-Guided Rectification Sampling in Algorithm 1. In **Stage 1**, we use a pre-trained perception model to identify instance masks. If annotated masks are already available, this step can be skipped. Next, we apply *exclusive dilation* to mitigate boundary segmentation issues and obtain a refined mask region R_{gt}, which serves as the initial ground-truth region. If visual priors such as Canny edge maps are provided, we further suppress false-positive regions in these priors, as they are prone to misidentification and may introduce noise in the corresponding visual condition maps.

In **Stage 2**, we define a rectification interval of T_r steps within the sampling process. First, we perform N steps of fast sampling to infer the latent $z'_{0|t-N}$, which is then decoded into an image. This image is passed through an object detector to identify false positives and false negatives. We then segment the corresponding regions: false positives are segmented on x'_{t-N}, and false negatives on the reconstructed image x_0. These regions are merged to produce a region-guided correction mask M. To rectify errors, we generate a noisy latent z_t^{orig} by adding t steps of noise to the original image. Finally, we mix z_t and z_t^{orig} using the region-guided mask to correct misgenerated areas.

C Experimental Setting

C.1 Metrics Definition

We provide a detailed explanation of the metrics used to evaluate the effectiveness of our method below:

mAP: The mean of the average precision values computed over multiple IoU thresholds (typically from 0.50 to 0.95 in steps of 0.05) and across all classes, reflecting overall detection accuracy under varying overlap requirements.

AP$_{50}$: The average precision at a single IoU threshold of 0.50, measuring detection quality under a more lenient overlap criterion.

AP$_{75}$**:** The average precision at a single IoU threshold of 0.75, measuring detection quality under a stricter overlap criterion.

Algorithm 1 Region-Guided Rectification Sampling Process

Require: VAE decoder θ_d, Object detector $\mathcal{D}$, segmentation model $\mathcal{S}$
Require: Original image x_0, initial latent z_0, control map c, text prompt p, initial noise ϵ
Require: Ground-truth boxes B_{gt} with labels y_{gt}
Require: Refinement time steps T_r, fast-sampling step count N
Ensure: Final output image $\hat{x}_0$

Stage 1: Initial Mask preparation
$B_{\text{pred}}, y_{\text{pred}} \leftarrow \mathcal{D}(x_0)$ ▷ Detect candidate boxes in original image
$B_{fp} \leftarrow \text{match\_by\_IoU}(B_{\text{pred}}, B_{\text{gt}}, y_{\text{pred}}, y_{\text{gt}}, \tau = 0.5)$ ▷ Select false positives
$(R_{\text{gt}}, R_{fp}) \leftarrow \mathcal{S}\big(x_0,\, [B_{\text{gt}}, B_{fp}]\big)$ ▷ Segment GT and FP regions
$(R_{\text{gt}}, R_{fp}) \leftarrow \text{ExclusiveDilate}(R_{\text{gt}}, R_{fp},\, \text{kernel} = 7)$ ▷ Dilate masks and avoid overlap
$c \leftarrow c \,\odot\, (1 - R_{fp})$ ▷ Keep control without FP regions
Stage 2: Sampling with region-guided rectification
for $t = T-1, T-2, \ldots, 0$ **do**
    $z_t = \sqrt{\alpha_t}\,\dfrac{z_{t+1} - \sqrt{1-\alpha_{t+1}}\,\epsilon_\theta(z_{t+1},\, t+1)}{\sqrt{\alpha_{t+1}}} \;+\; \sqrt{1-\alpha_t}\,\epsilon_\theta(z_{t+1},\, t+1)$ ▷ Denoise
    if $t \in T_r$ **then**
        $z'_{t-M} \leftarrow \text{FastSample}(z_t,\, N)$ ▷ Accelerated N-step sampling
        $z'_{0|\,t-N} \leftarrow \dfrac{z'_{t-N} - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}}$
        $x'_{0|\,t-N} \leftarrow \theta_d(z'_{0|\,t-N})$
        $B_{\text{pred}}, y_{\text{pred}} \leftarrow \mathcal{D}(x'_{0|\,t-N})$
        $(B_{fp}, B_{fn}) \leftarrow \text{match\_by\_IoU}(B_{\text{pred}},\, B_{\text{gt}},\, y_{\text{pred}},\, y_{\text{gt}},\, \tau = 0.5)$ ▷ Select false positives and false negatives bboxes
        $M \leftarrow \text{merge}(\mathcal{S}\big([x'_{0|\,t-N}, x_0],\, [B_{fp}, B_{fn}]\big))$ ▷ Merge false positive and false negative region to build region-rectification mask
        $z_t^{\text{orig}} = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon$
        $z_t \leftarrow z_{t|0} \,\odot\, M \;+\; z_t \,\odot\, (1 - M)$ ▷ Region-Guided Rectification
    end if
end for
$\hat{x}_0 \leftarrow \theta_d(\hat{z}_0)$
return $\hat{x}_0$

AP^m : The average precision for medium-sized objects (area $\in$ [32², 96²] pixels), indicating performance on objects of moderate scale.

AP^l : The average precision for large objects (area > 96² pixels), indicating performance on larger targets that are generally easier to localize.

FID: Fréchet Inception Distance (FID) measures the similarity between real and generated images. We compute FID using 5,000 samples. Lower FID indicates higher image quality and diversity.

C.2 More Training Details

In this study, we train several object detection models on downstream detection datasets, including Faster R-CNN (Ren et al., 2015) with an R50 FPN backbone, ATSS (Zhang et al., 2020) with an R50 FPN backbone, FCOS (Tian et al., 2019) with an R50 FPN backbone, YOLOX-S (Ge et al., 2021), RetinaNet with a Swin-Tiny (Liu et al., 2021) FPN backbone and DEIM-D-FINE -N (Huang et al., 2024) model. For Faster R-CNN, ATSS, FCOS, and RetinaNet, we follow the standard $1\times$ training schedule, running 12 epochs for all experiments, except for Faster R-CNN on the COCO dataset, where the training is reduced to 6 epochs. We use random flipping as the default data augmentation strategy. For YOLOX-S, we follow the official training setup with 300 epochs and apply a stronger augmentation pipeline, including mosaic, random affine transformations, mixup, random flipping, and HSV-based random augmentation. For DEIM, we follow the official training configuration to train model for 40 epochs.

D More Visualization Results

D.1 Visualization Samples Analysis

As shown in Figure 6, we present additional samples generated by our method, which illustrate its ability to produce novel content with high image-annotation consistency and strong visual fidelity. Our method remains robust even in challenging scenarios involving tiny object boxes and densely distributed objects, ultimately improving the effectiveness of downstream detection models.

In Figure 6, we notice that some of the smaller objects in our generated images remain strikingly similar to those in the originals (the plant in the second row). This is partly due to the Canny-conditioned ControlNet, which helps preserve source-like structures, a result we see much less often with GLIGEN. More fundamentally, when a region proves difficult to synthesize accurately, our method continuously applies region-guided corrections to bring it closer to the original image. We lower the editing intensity in these challenging areas in order to maintain their structural and semantic integrity, so that any differences from the source appear only in texture and lighting.

Texture-level perturbations on hard-to-generate samples also help us produce valuable corner cases. For example, Figure 7 shows a truck on the left that is heavily occluded and has an implausible aspect ratio. By introducing controlled texture disturbances into the real image, we can create new corner cases that further enhance the model's robustness.

Additionally, a case study is provided in Figure 8, which demonstrates that although Instance Diffusion incorporates instance-level descriptions with specific visual tokens, it still suffers from semantic leakage. Our method can further mitigate this issue and enhance performance in such cases.

D.2 Rectification Mask Analysis

Our method introduces region-guided rectification to correct misgenerated areas during the diffusion sampling process. We quantify the rectification mask's area at different sampling stages to analyze how much of the image each correction step influences. Across 1,000 samples, we measured the ratio of mask area to total image area at the 75%, 50%, 25%, and 10% sampling steps, obtaining 12.16%, 8.87%, 7.12%, and 6.77%, respectively. This trend demonstrates that our rectification procedure progressively reduces misgenerated regions as sampling proceeds.

E More Experimental Results

E.1 Comparison of Rectification at Different Diffusion Timesteps

We study the sensitivity of our region-rectification schedule to the choice of diffusion timesteps. Concretely, we apply rectification at single timesteps as well as at several multi-stage schedules and report mean average precision (mAP), AP_{50} and AP_{75} in Table 8. Applying rectification only at an early stage ($0.75T$) produces a modest improvement over the ControlNet baseline (mAP 35.2 vs. 34.9), but is not optimal. We attribute this to two factors: (1) early-stage rectification can create shortcuts that allow spurious regions to be propagated and "fixed" during later synthesis steps, and (2) features at $0.75T$ remain highly diverse and only partially developed, so a single early rectification cannot fully suppress all errors.

When rectification is applied at multiple stages we observe consistent improvements. A two-stage schedule $[0.5T,\ 0.25T]$ raises mAP to 35.3, and a three-stage schedule $[0.5T,\ 0.25T,\ 0.1T]$ further increases mAP to 35.4. Our four-stage schedule $[0.75T,\ 0.5T,\ 0.25T,\ 0.1T]$ achieves the best balance across metrics (mAP 35.5, AP_{50} 56.2, AP_{75} 38.4). Extending the schedule to six stages yields only marginal additional benefit (mAP 35.5, AP_{75} 38.5), indicating that performance gains saturate beyond four rectification stages. These results justify our chosen four-stage schedule as an effective and parsimonious design choice.

E.2 Comparison with Image Editing Method

Existing methods suffer from semantic leakage and discrepancies between generated content and the original annotations. To address these issues, our approach calibrates intermediate sampling results using latent point sampled within the diffusion sampling process. We also explore the effectiveness

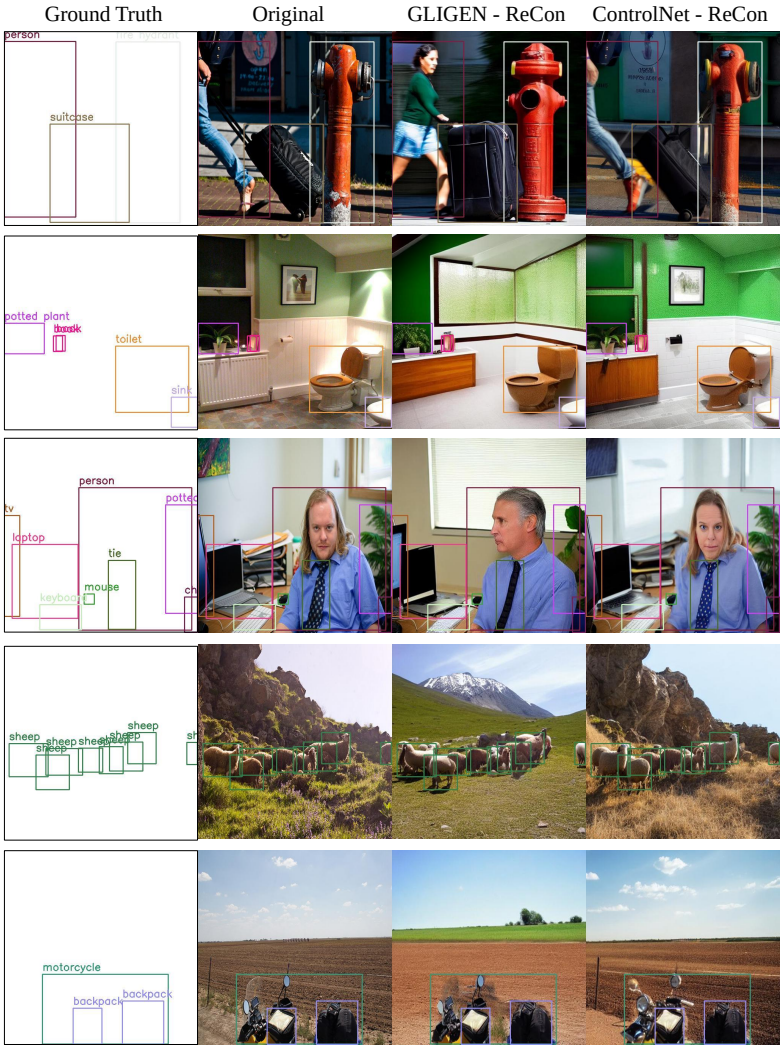
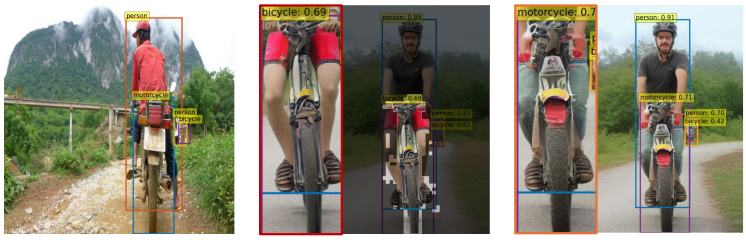
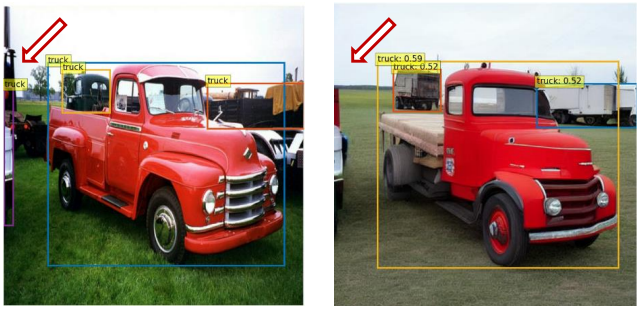


Figure 6: Visualization results of samples generated with our methods.



Original Image Instance Diffusion Ours

Figure 7: Successful case analysis. Existing state-of-the-art methods suffer from semantic leakage, while our approach effectively mitigates this issue.



Original Image Ours

Figure 8: Corner case analysis. For corner cases with severe occlusion and unrealistic aspect ratios (truck on the left side of image), our method can augment them to generate new corner-case examples.

Table 8: Ablation study of region rectification at different diffusion timesteps.

Method	mAP	AP_{50}	AP_{75}
ControlNet	34.9	55.5	37.7
$0.75T$	35.2	55.7	38.0
$0.5T$	35.3	56.0	38.0
$0.25T$	35.3	55.8	38.1
$0.1T$	35.2	55.8	38.0
$[0.5, 0.25]T$	35.3	55.8	38.2
$[0.5, 0.25, 0.1]T$	35.4	56.0	38.4
$[0.75, 0.5, 0.25, 0.1]T$	**35.5**	**56.2**	38.4
$[0.75, 0.625, 0.5, 0.375, 0.25, 0.1]T$	35.5	56.2	**38.5**

of image-editing methods which modify specific regions of the original image while preserving the remainder. We compare our method against an image editing method: SDEdit (Meng et al., 2021). Notably, SDEdit applies a uniform editing strength across the entire image, which may result in certain regions being either over-enhanced or under-enhanced. As demonstrated in Table 9, our method, which leverages region-controllable data augmentation guided by a perceptual model, achieves superior performance.

Table 9: Comparison with image editing method.

Method	mAP	AP_{50}	AP_{75}
Real only	34.5	55.5	37.1
SDEdit	35.2	55.9	38.1
ReCon	**35.5**	**56.2**	**38.4**

E.3 Comparison with More Detectors

As shown in Table 10, we compare additional object detectors, and the results demonstrate that our method consistently improves the performance of each.

E.4 More Data Scaling

In the main text, we conducted data-scaling experiments under data-scarce settings. We further validated this phenomenon in few-shot scenarios. As shown in Table 11, we compare ReCon with simple duplication-based augmentation of original images under different expansion ratios. Our findings indicate that as the scale of data expansion increases, the corresponding performance improvement becomes more evident. Moreover, we observe that augmenting with duplicated original

Table 10: Trainability comparison for more detectors on COCO.

Detector	Method	mAP	AP_{50}	AP_{75}
RetinaNet - Swin-T (Liu et al., 2021)	Real only	34.0	54.0	35.8
	ReCon	**35.1**	**54.9**	**37.0**
FCOS - R50 (Tian et al., 2019)	Real only	36.6	56.0	39.1
	ReCon	**37.3**	**56.4**	**39.7**
ATSS - R50 (Zhang et al., 2020)	Real only	39.4	57.5	42.6
	ReCon	**40.0**	**58.3**	**43.2**

images provides only marginal gains, whereas our method achieves comparable results to 10× duplication by using only 2× generated data.

Table 11: Few-shot data augmentation results across different data scales.

Method	mAP	AP_{50}	AP_{75}	AP^m	AP^l
Real only	5.4	10.3	5.0	5.2	8.9
Expanded 2×					
Real Expansion	5.7	10.9	5.5	5.8	9.7
ReCon	**6.7**	**12.3**	**6.5**	**6.7**	**11.2**
Expanded 5×					
Real Expansion	6.1	11.3	6.1	6.3	10.2
ReCon	**7.7**	**14.1**	**7.6**	**7.7**	**12.5**
Expanded 10×					
Real Expansion	6.4	11.5	6.5	6.3	10.2
ReCon	**8.0**	**14.4**	**7.9**	**7.7**	**13.2**

E.5 Robustness Evaluation

We perform three independent runs with different random seeds and report the mean and standard deviation in Table 12. The results show that our method consistently outperforms the baseline and demonstrates lower variance, indicating improved stability and robustness.

Table 12: Performance stability across three runs with different random seeds. We report the mean and standard deviation of mAP and AP_{50}.

Method	mAP (%)	AP_{50} (%)
Baseline	34.9 ± 0.08	55.3 ± 0.12
ReCon	$\mathbf{35.5 \pm 0.05}$	$\mathbf{56.3 \pm 0.14}$

E.6 Runtime and Inference Efficiency

Table 13 reports per-sample inference times on the COCO dataset for baseline control methods and for those methods augmented with our method. All runtime measurements were collected on a single NVIDIA RTX 3090 GPU. On average, our approach introduces only an additional 0.79–1.04 seconds per sample compared to the corresponding baseline control model. Compared to the training-free layout-to-image control method LayoutGuidance, our method does not rely on backward guidance and therefore avoids the substantial time cost incurred by its optimization-based procedure. In contrast to training-based alternatives that require additional fine-tuning on downstream datasets, our framework can be applied directly at inference time without introducing extra training overhead. Finally, our method is fully compatible with structural controllable models such as ControlNet, yielding improved performance while incurring less than one second of additional inference time per sample.

We note that further acceleration is possible: replacing the grounding component with a more efficient model (e.g., EfficientSAM (Xiong et al., 2024)) or integrating memory- and compute-efficient attention libraries (e.g., xFormers (Lefaudeux et al., 2022)) should reduce the added overhead. We will include these quantitative runtime results and the above discussion in the revised manuscript to clarify the practical efficiency of our approach.

Table 13: Inference time comparison (seconds per sample) on COCO measured on an NVIDIA RTX 3090. Values in parentheses indicate the baseline time plus the additional overhead introduced by our method.

Method	Inference time (s)
ControlNet	2.55
Layout Guidance	12.58
Instance Diffusion	8.98
ControlNet + ReCon	3.34 (2.55 + 0.79)
Instance Diffusion + ReCon	10.02 (8.98 + 1.04)

E.7 More Ablation Experiments

To evaluate the contribution of region-aligned cross-attention (RACA) independently from the full rectification pipeline, we augmented the InstanceDiffusion baseline with RACA while keeping all other components unchanged. The RACA layer replaces the standard cross-attention with a region-aligned cross-attention mechanism and introduces *no additional trainable parameters*, allowing direct deployment on pretrained generators without fine-tuning. Table 14 summarizes the results: adding RACA yields consistent improvements over the InstanceDiffusion baseline, confirming that region-aligned attention alone contributes meaningfully to region alignment and downstream detection performance.

Table 14: Effect of region-aligned cross-attention (RACA) within InstanceDiffusion.

Method	mAP	AP_{50}	AP_{75}
InstanceDiffusion	35.0	55.4	37.6
InstanceDiffusion w/ RACA	35.2	55.7	38.0

E.8 Rationale Analysis of Our Method

Prior work has applied perception models to filter and then re-generate low-quality synthetic samples in two-step generate-then-filter pipelines (Fang et al., 2024). While effective in improving sample fidelity, such repeated generate-and-filter cycles incur substantial computational overhead. In contrast, our approach integrates grounding feedback directly into the diffusion trajectory, enabling on-the-fly rectification of misaligned regions within a single forward pass. This integrated rectification avoids repeated re-sampling and produces high-quality, aligned image-label pairs far more efficiently than multi-round generate-and-filter procedures.

Our rectification mechanism (RGR) identifies mismatched regions via grounding-based IoU matching between generated regions and target annotations, and selectively injects controlled perturbations into those mismatched regions during the diffusion trajectory. This targeted intervention encourages subsequent denoising steps to correct region appearance and spatial extent while leaving well-aligned regions largely untouched.

The proposed design achieves three complementary properties that are critical for large-scale synthetic data pipelines:

- **Usability.** The method does not require model retraining, multi-round generation, or post-hoc filtering; it operates during a single forward pass and can be applied to pretrained generators out-of-the-box.
- **Effectiveness.** Experiments across multiple datasets and generator architectures show consistent improvements in both image quality and detection metrics.
- **Efficiency.** The rectification strategy introduces limited additional computational overhead to inference, making it suitable for large-scale synthetic data generation.

These benefits stem from two lightweight, jointly designed modules (region-aligned attention and grounding-guided rectification) that together ensure improved alignment without compromising generation speed or flexibility. To reach the final design we explored and quantitatively compared a variety of correction and alignment strategies; the selected combination offers a favorable trade-off between alignment quality, computational cost, and ease of integration.

In summary, the combination of on-the-fly grounding feedback, parameter-free region-aligned attention, and targeted rectification yields a practical and effective solution for producing well-aligned image-label pairs. The plug-and-play compatibility with pretrained generators and consistent empirical gains across baselines and datasets underscore the substantive contributions of this work.

E.9 Further Analysis of Rectified Regions

We analyzed the relationship between object area and the likelihood of rectification under our proposed Region-Guided Rectification strategy. Specifically, we examined 500 objects identified as requiring rectification and plotted their area distribution using kernel density estimation, as shown in Figure 9. The results reveal that smaller regions are more likely to be rectified. This is because small objects are generally more difficult to synthesize accurately, diffusion models tend to generate artifacts or errors in such regions.

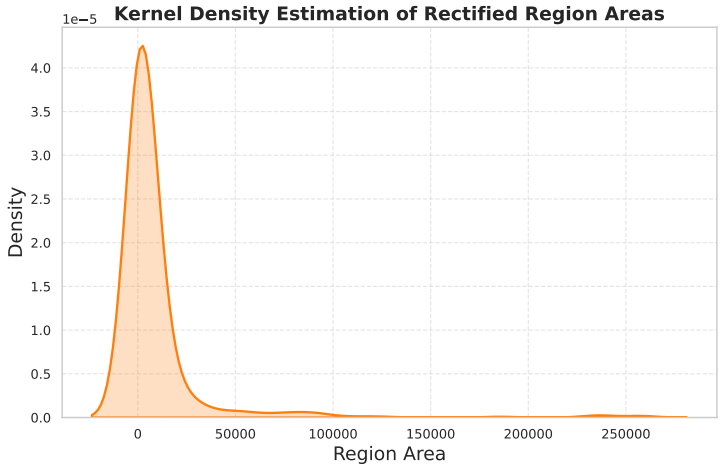


Figure 9: Kernel density estimation of rectified region areas.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our contributions have been claimed in the abstract and introduction accurately.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a limitation discussion in Appendix A

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the detailed information in the Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The implementation code is provided in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the detailed information of dataset, training and testing setting in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes. The paper reports error bars in the form of mean and standard deviation in Section E.5 in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the introduction of compute resources in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impacts of the work in Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We only use LLM for writing improvement.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.