Prompt-Free Conditional Diffusion for Multi-object Image Augmentation

Haoyu Wang[1] , **Lei Zhang**[1†] , **Wei Wei**[1] , **Chen Ding**[2] and **Yanning Zhang**[1]

[1]Northwestern Polytechnical University
[2]Xi'an University of Posts & Telecommunications

wanghaoyunwpu@mail.nwpu.edu.cn,
{nwpuzhanglei, weiweinwpu, ynzhang}@nwpu.edu.cn, dingchen@xupt.edu.cn

Abstract

Diffusion models has underpinned much recent advances of dataset augmentation in various computer vision tasks. However, when involving generating multi-object images as real scenarios, most existing methods either rely entirely on text condition, resulting in a deviation between the generated objects and the original data, or rely too much on the original images, resulting in a lack of diversity in the generated images, which is of limited help to downstream tasks. To mitigate both problems with one stone, we propose a prompt-free conditional diffusion framework for multi-object image augmentation. Specifically, we introduce a local-global semantic fusion strategy to extract semantics from images to replace text, and inject knowledge into the diffusion model through LoRA to alleviate the category deviation between the original model and the target dataset. In addition, we design a reward model based counting loss to assist the traditional reconstruction loss for model training. By constraining the object counts of each category instead of pixel-by-pixel constraints, bridging the quantity deviation between the generated data and the original data while improving the diversity of the generated data. Experimental results demonstrate the superiority of the proposed method over several representative state-of-the-art baselines and showcase strong downstream task gain and out-of-domain generalization capabilities. Code is available at here.

1 Introduction

In the past decade, deep neural networks have achieved a surge of success in a wide range of computer vision tasks [He *et al.*, 2016; Dosovitskiy *et al.*, 2020; Radford *et al.*, 2021]. One key premise for such success lies on the collection of large-scale training images. However, in real scenarios even for a specific single task, amassing sufficient images to establish a dataset is often prohibitively costly and laboriously time-intensive, e.g., imageNet [Deng *et al.*, 2009] for image classification. For this problem, a promising solution proves to be image augmentation with generative models [Antoniou *et al.*, 2017], which aims at randomly generating extensive synthetic images based on a few manually collected images to rapidly establish a dataset. Following this idea, various effective image generative models [He *et al.*, 2023; Chen *et al.*, 2023] have been proposed successively. Among them, profiting from the powerful generative capacities, diffusion models have been paid increasing attention to image generation and augmentation. Usually, given some prompts related to the scene content, the diffusion model can directly generate a high-quality image with such a content.

In real-world applications, generating multi-object images with complex spatial relationships is crucial. Although some recent progress have been made for multi-object image generation, due to much increased generation difficulty, most of these methods suffer from obvious limitations. As shown in Fig. 1, several existing methods [Wu *et al.*, 2023b; Nguyen *et al.*, 2023] use category names or image captions as conditions to inputs into the pre-trained diffusion model to generate images, and use attention maps to extract image labels. However, it is difficult to generate a large number of objects using only text prompts, and the quality of labels generated using attention maps is poor when objects overlap. Although some methods [Wang *et al.*, 2024; Wu *et al.*, 2023a] use stronger guidance, such as layout or paragraph, to improve the quality of generated images, these methods are difficult to scale. Some methods [Zhao *et al.*, 2023; Xie *et al.*, 2023] solve this problem by decomposing the multi-object image generation task. They first generate single-object images and their corresponding labels through a pre-trained diffusion model, and then use data augmentation to synthesize multi-object images. Although the number of objects and label quality are increased, artificial facts are often generated, which reduces the reality of the image.

In addition, the above methods tend to pursue training-free and directly use simple text prompts containing category names to generate data, which leads to deviations in style, size, etc. between the generated images and the original data. Furthermore, some methods [Suri *et al.*, 2024; Yang *et al.*, 2024] try to replace or add objects to the original images through image editing. Although this makes the augementated image as realistic as possible, the amount of information added is limited because the layout, background

[†] Corresponding author.

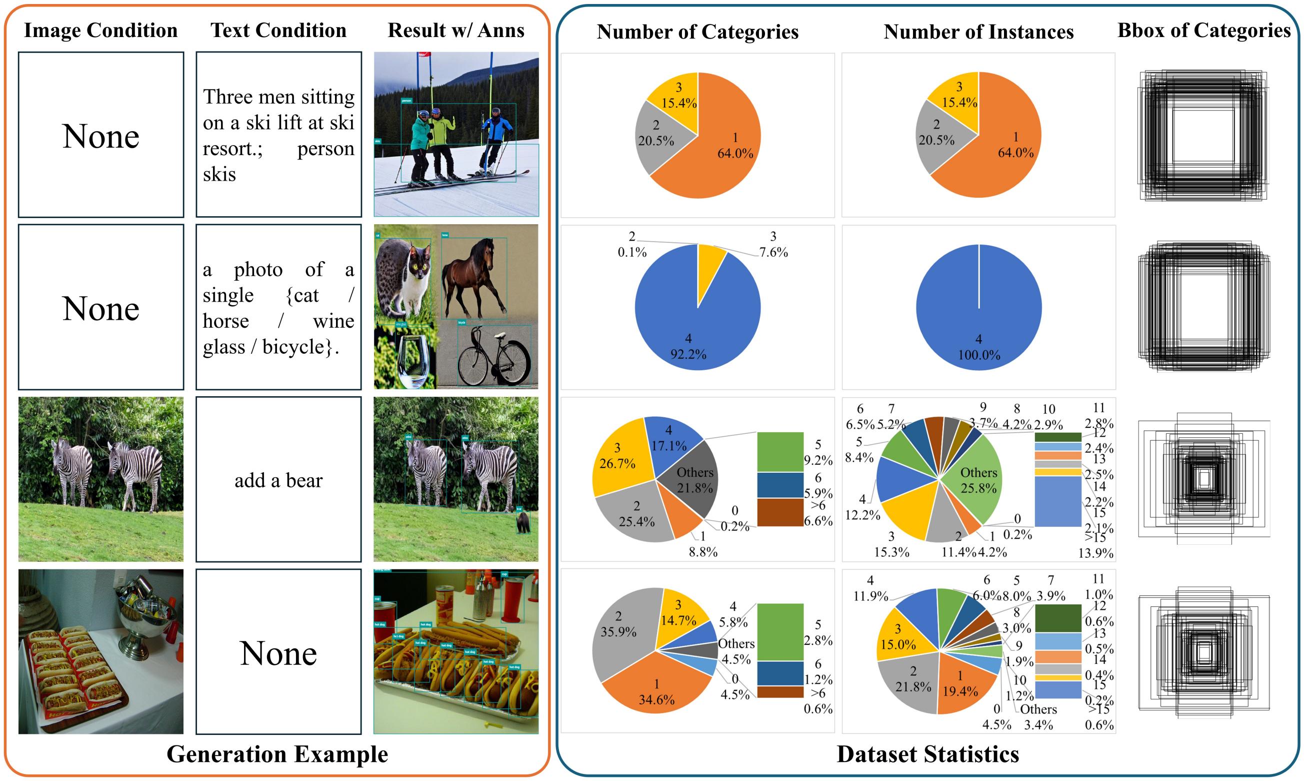

Figure 1: Comparison with state-of-the-art image augmentation methods. Dataset Diffusion decrease in object amount with low annotation quality. MosaicFusion generate counterfactual images and objects are similar in size. Add SD cannot change the background of the image and have low variation in layout. Our method generates a large number of objects while ensuring image diversity.

and most of the objects in the image have not changed.

To fill this gap, we propose a prompt-free conditional diffusion framework, aims to reduce the category and quantity deviations from the original data while improving the diversity of generated images. Inspired by image variation task [Ramesh *et al.*, 2022; Xu *et al.*, 2023; Xu *et al.*, 2024], our framework utilizes a single multi-object image instead of text prompts as the condition of the diffusion model to reduce the category bias brought by text descriptions. More importantly, to better extract and inject the multi-object information into the diffusion procedure, we propose a local-global semantic fusion strategy that utilizes the pre-trained CLIP [Radford *et al.*, 2021] model to separate extract the semantic knowledge within the whole condition image as well as its local crop. On the other hand, to further control the object amount as well as the layout diversity in the generated image, we further propose a reward model based counting loss to explicitly restrict the amount of objects in each category in the generated image, while imposing no any constraint on their spatial layout. By doing this, the proposed model is able to randomly generate high-quality images with the same number of objects in each category as the condition image or even more but showing different layouts, thus guaranteeing the variety of image augmentation. Experimental results demonstrate the superiority of the proposed method over several representative state-of-the-art baselines and showcase good downstream task gains and out-of-domain generalization capabilities.

In summary, this study mainly contributes in four aspects:

1. We propose a prompt-free conditional diffusion framework for multi-object image augmentation. By changing the text condition to a novel local-global semantic fusion strategy, which enables appropriate extracting the multi-object information from the condition image and injecting it into the diffusion model for image generation.
2. We design a reward model based counting loss to constrain the number of objects in each category of generated images, which improves the diversity of images.
3. We contribute new state-of-the-art performance of both downstream tasks and generated quality on MS-COCO dataset in terms of multi-object image augmentation.

2 Related Work

2.1 Text-to-Image Diffusion Models

Driven by multi-modal technology, text-to-image diffusion models exhibit formidable capabilities in image generation. GLIDE [Nichol *et al.*, 2022] uses a cascade architecture and classifier-free guidance [Ho and Salimans, 2022] for image generation based on pre-trained language models. DALL-E2 [Ramesh *et al.*, 2022] adopts a multi-stage model, using CLIP [Radford *et al.*, 2021] text encoder to encode text and images. Imagen [Saharia *et al.*, 2022] uses multiple text encoders to improve sample fidelity and text-image alignment. The latent

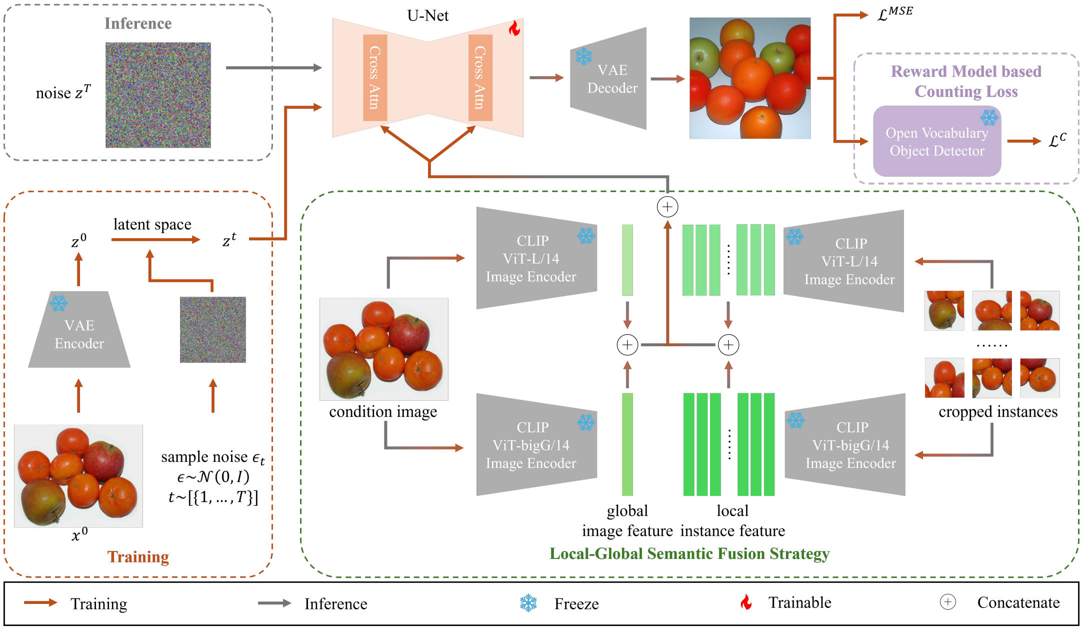

Figure 2: Overview of the proposed prompt-free conditional diffusion framework. We introduce a local-global semantic fusion strategy to generate images with local instance categories and global semantics comparable to the condition image. We also introduce a reward model based counting loss to ensure that the number of objects in each category of the image do not decrease.

diffusion model [Rombach *et al.*, 2022] significantly reduces computational overhead by transferring the diffusion process from the image to a low-dimensional feature space. On its basis, exemplar-based methods [Li *et al.*, 2024] achieve refined control of generated images under the guidance of text by introducing structural information as input, such as mask, edge, pose, etc. Subject-driven image generation methods [Ruiz *et al.*, 2023; Gal *et al.*, 2022] realize the customized generation of specific objects under the guidance of several target images and relevant text prompts. In contrast to the above techniques, the goal of our framework does not require specifying locations or customization of individuals for each instance, but to generate factual images with comparable object amounts and diverse layouts.

2.2 Image Variation

Given an image, image variation aims to generate an image with similar styles or semantics. Currently, there is no unified paradigm for image variation tasks. DALL-E2 [Ramesh *et al.*, 2022] uses the alignment characteristics of the CLIP image and text encoder to encode input images to achieve image variation. ControlNet [Zhang *et al.*, 2023] controls the generation of the diffusion model by adding an additional network structure based on the latent diffusion model. Its reference-only version achieves variation images by splicing the original attention layer of the diffusion model with the attention layer of the control network. Versatile Diffusion [Xu *et al.*, 2023] designs a multi-stream multimodal latent diffusion model framework and supports the diversified generation of a single image stream. Prompt-Free Diffusion [Xu *et al.*, 2024] replaces the text encoder with a semantic context encoder to learn the features of the input image and diversify it. Compared with the method proposed in this paper, the above method performs diversification on the entire image, and its diversified connotation often includes multiple information such as content, style, and color, which cannot guarantee that the instance of the generated image is consistent with the original image.

3 Methodology

3.1 Problem Formulation

Despite the availability of excellent annotation tools such as SAM 2 [Ravi *et al.*, 2024] and Grounding DINO [Liu *et al.*, 2024], the diversified generation of large-scale multi-object images remains a problem that needs to be solved. In multi-object dataset augmentation, consider a collection of N samples, denoted as $\mathcal{D} = \{(x_i, y_i), i = 1, ..., N\}$, where $x_i = \{(o_j, c_j), j = 1, ..., N_i^c\}$, represents an input image containing N_i^c categories, and for each category c_j, it contains o_j objects. $y_i = \{(b_k, c_k), k = 1, ..., N_i^o\}$ denotes the category c_k and the structured box annotations b_k of N_i^o objects, and $\sum_{j=1}^{N_i^c} o_j = N_i^o$. The goal of the task is to generate a set of enhanced images $\mathcal{D}^* = \{x_i^*, i = 1, ..., N\}$ with the same number of input samples, where $x_i^* = \{(o_l, c_l), l = 1, ..., N_i^{c*}\}$, requiring that for each category c_j in the input image x_i, a c_l can be found in the augmented image x_i^* corresponding to it, and the count of it $o_l \geq o_j$.

3.2 Overall Architecture

As shown in Fig. 2, the proposed framework consists of two parts: a local-global semantic fusion strategy and a reward model based counting loss.

During the forward diffusion process, the pre-trained latent diffusion model first uses the encoder E to compress the input image $x_i^0 = x_i \in \mathbb{R}^{H\times W\times 3}$ into a latent representation $z_i^0 \in \mathbb{R}^{h\times w\times d}$, while the decoder D can transform the latent representation into pixel space, i.e., $D(z_i^0) \approx x_i^0$, where $\frac{H}{h} = \frac{W}{w} = 8$ and $d = 4$. Then the noise ϵ_t is sampled from the Gaussian distribution and added to it, where t is a time step sampled from the uniform distribution. Finally, a DDPM is trained in the latent space based on the image condition p_i^{img} using the MSE loss and our proposed counting loss to recover z_i^0 from the Gaussian distribution, where the MSE loss $\mathcal{L}_i^{MSE}$ is:

$$\mathcal{L}_i^{MSE} = \mathbb{E}_{z\sim E(x),y,\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(z_i^t, t, \mathcal{C}(p_i^{img}))||_2^2], \tag{1}$$

where z_i^t is the noise feature of time step t, ϵ_θ is the noise prediction network, which takes z_i^t as input and predicts the sampled Gaussian noise guided by the time step t and the conditional feature $\mathcal{C}(p_i^{img})$, where $\mathcal{C}$ is our proposed local-global semantic fusion module.

For the reverse diffusion process, the model directly samples noise in the latent space and uses the trained noise prediction network to gradually denoise it according to the conditional features to obtain the final image.

3.3 Local-Global Semantic Fusion

Numerous papers [Binyamin *et al.*, 2024; Wen *et al.*, 2023; Battash *et al.*, 2024] point out that the text-to-image diffusion model often fails to generate images that accurately match the text prompt, especially when the prompt contains information such as multiple categories or counts. In addition, since most current multi-object generation methods pursue training-free and directly use text prompts to generate images, the generated category distribution is offset from the target dataset distribution due to the inherent bias of the generation model. To address the challenges of category bias introduced by text-based prompts, we replace textual prompts with image-based conditions for diffusion models. Using images as input conditions better captures the category distribution of the target dataset, reducing deviations and improving the fidelity of the generated data.

In order to adapt the latent diffusion model from text-guided image generation to image-guided image generation, we use the image encoder E_{img} pre-trained together with the original text encoder E_{text} using paired text-image data to encode the image condition, so that the obtained conditional features remain in the same feature space without fine-tuning all the parameters of the diffusion model.

The original text encoder uses hidden states of text conditions to capture the semantic relationship between the text context:

$$\mathcal{C}(p_{text}) = E_{text}(T(p_{text})), \tag{2}$$

where $\mathcal{C}(p_{text}) \in \mathbb{R}^{bs\times seq\times emb}$ is the output conditional feature, bs is the batch size of text condition p_{text}, seq is the sequence length, emb is the feature dimension, and T is the tokenizer. In order to further clarify the instance that needs to be enhanced, we crop it from the image, merge it with the original image and input it into the image encoder to extract features:

$$p_i^{img} = \{x_i, Crop(x_i, b_i, pad)\}, \tag{3}$$

where p_i^{img} is the local-global semantic fusion condition, and $Crop(\cdot)$ uses the bounding box information b_i of the global image x_i to crop the local instance to be augmented. In order to better understand the image context, we use hyperparameter pad to control the pixel of outward cropping.

Algorithm 1 Counting Loss

Input: denoised image x_i^*, open vocabulary object detector $\mathcal{D}_{OV}$, number of categories N_i^c, text prompt S_i, class count list L_i^{count}, class index list L_i^{index}, counting loss step γ, counting loss threshold τ

```
if training steps larger than $\gamma$ then
    Let $logits_i \leftarrow \mathcal{D}_{OV}(x_i^*, S_i)$.
    for $j \leftarrow 1, ..., N_i^c$ do
        if $L_i^{index}[j]$ is an integer then
            Let $s_i^j \leftarrow logits_i[L_i^{index}[j]]$
        else
            for idx in $L_i^{index}[j]$ do
                Let $s_i^j \leftarrow$ concatenate all $logits_i[idx]$
            end for
        end if
        Calculate $\mathcal{L}_i^j$ use Eq. 6
    end for
    Calculate $\mathcal{L}_i^C$ use Eq. 7
    return $\mathcal{L}_i^C$
end if
```

Through the above operations, we express the information that is difficult to control with text, such as count and category, through batched local-global image information, highlighting its importance in the condition. To reduce computational complexity, we only need all the features of the original image, and for each cropped image, we only need its $[CLS]$ feature:

$$\mathcal{C}(p_i^{img}) = E_{img}(P(p_i^{img}, M)) \tag{4}$$

where $\mathcal{C}(p_i^{img}) \in \mathbb{R}^{bs\times(1+M)\times emb}$, P is the image processor that processes the image condition for batch training. Specifically, for each image condition p_i^{img}, the image processor fixes its cropped instances to M. When the number of instances is less than M, it is expanded with zero tensors, otherwise M instances are randomly selected for training. We set M to 9, which significantly reduces the computational complexity compared to the text condition while ensuring the semantics of most objects.

3.4 Reward Model Based Counting Loss

With the proposed local-global semantic fusion strategy, we can improve the fidelity of the generated image. To further ensure that the object amounts do not degrade, we propose a reward model based counting loss. Specifically, for the input image x_i, we obtain the image x_i^* by one-step denoising during training:

$$x_i^* = \frac{1}{\sqrt{\alpha_t}}(x_i^t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_i^t, t)) + \sigma_t \mathbf{z}, \tag{5}$$

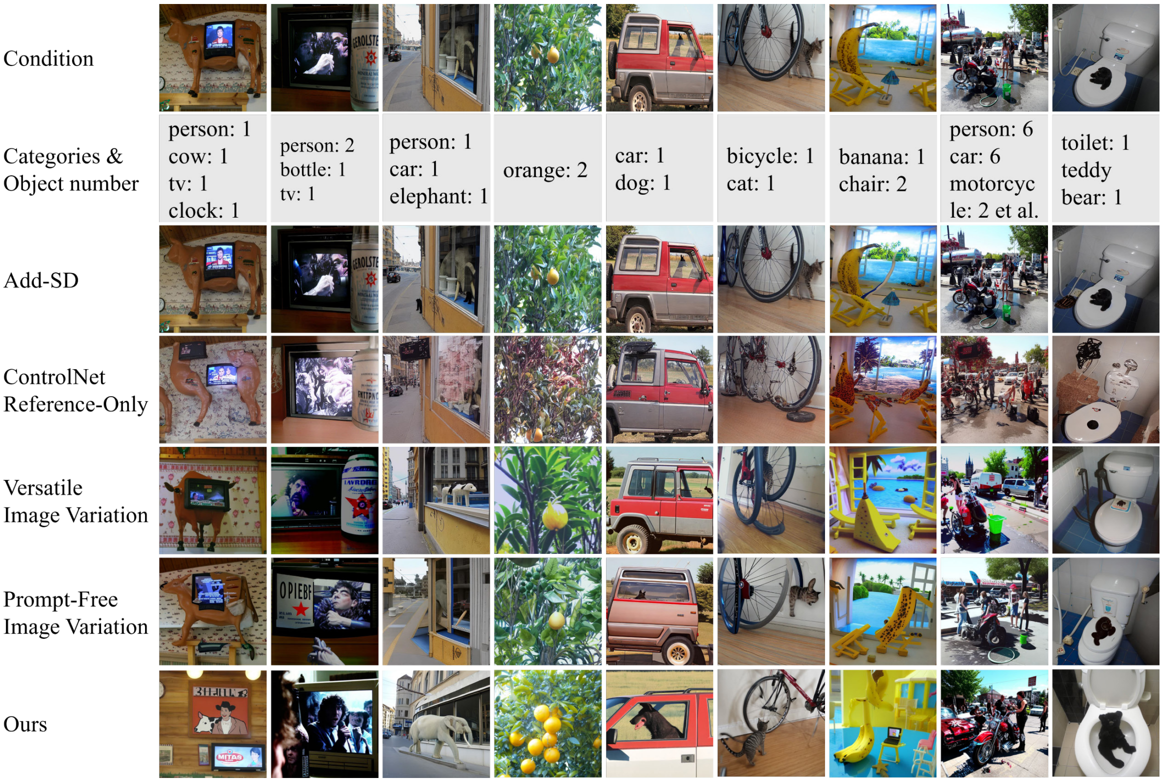
Condition
Categories & Object number
person: 1 cow: 1 tv: 1 clock: 1
person: 2 bottle: 1 tv: 1
person: 1 car: 1 elephant: 1
orange: 2
car: 1 dog: 1
bicycle: 1 cat: 1
banana: 1 chair: 2
person: 6 car: 6 motorcycle: 2 et al.
toilet: 1 teddy bear: 1
Add-SD
ControlNet Reference-Only
Versatile Image Variation
Prompt-Free Image Variation
Ours

Figure 3: Qualitative comparison. We compare with Dataset Diffusion w/SDXL and SDXL img2img, ControlNet Reference-Only, Versatile Image Variation and Prompt-Free Image Variation on COCO 2017 validation set. Our method is superior to other methods in terms of sufficient objects, realism, and layout diversity. Better viewed with zoom-in.

where ϵ_t is the noise prediction network, α_t, $\overline{\alpha}_t$, σ_t are the hyperparameters defined by DDPM, and $\mathbf{z} \sim \mathcal{N}(0,1)$ is used to adjust the signal-to-noise ratio.

Then we use the image annotations to construct supervision information. The proposed counting loss focuses solely on category counts, rather than bounding box positions, to promote diverse layout generation. This design ensures that object counts match the desired distribution without imposing rigid spatial constraints, enhancing both flexibility and diversity in the generated images. For the N_i^c categories contained in the image, we first count the number of objects in each category and obtain a one-to-one corresponding category name list $L_i^{class} = \{name(c_j), j = 1, ..., N_i^c\}$ and count list $L_i^{count} = \{len(o_j), j = 1, ..., N_i^c\}$, where $name(\cdot)$ is used to get the category name, and $len(\cdot)$ is a function of counting numbers. Then we connect each name in L_i^{class} with a period to construct the text prompt S_i of the reward model. Since some category names have more than one word, we also record the index list L_i^{index} of each category in S_i to obtain the result of the reward model.

Finally, we use the pre-trained open vocabulary object detector as the reward model to detect the categories in the image according to S_i. For detection result of c_j, we take the highest confidence sample based on input image and calculate the loss according to the threshold hyperparameter τ:

$$\mathcal{L}_i^j = \sum_{L_i^{count}[j]} ReLU(\tau - topk(s_i^j, k = L_i^{count}[j])), \qquad (6)$$

where s_i^j is the confidence result of in category c_j of image x_i detected by the reward model. And the final counting loss $\mathcal{L}_i^C$ of image x_i is:

$$\mathcal{L}_i^C = \frac{\sum_{j=1}^{N_i^c}(\mathcal{L}_i^j)}{\sum_{j=1}^{N_i^c}(L_i^{count}[j])}. \qquad (7)$$

The hyperparameter γ determines the training step at which counting loss begins to take effect. This avoids noisy gradients during early training stages when the denoised images may still contain significant noise. The calculation method of counting loss is outlined in Algorithm 1. The overall training loss of the proposed framework can be formulated as:

$$\mathcal{L} = \sum_i (\mathcal{L}_i^{MSE} + \lambda \mathcal{L}_i^C), \qquad (8)$$

where λ is a hyperparameter for adjusting the loss weight.

4 Experiments

4.1 Experimental Setups

Datasets

We validate the proposed framework and comparison methods on the MS-COCO [Lin *et al.*, 2014] dataset, a relatively complex object detection dataset containing 80 categories, with an average of 7.7 objects per image. We use $train2017$ containing 118K images to train the proposed method and generate images for downstream task evaluation, and use the COCO validation set val2017 consisting of 5K images for generation quality evaluation.

Implementation Details

We use Stable Diffusion XL [Podell *et al.*, 2023] and Grounding DINO [Liu *et al.*, 2024] as LDM and reward model respectively. We fine-tune the model using LoRA [Hu *et al.*, 2021] at 512 × 512 resolution, we set the learning rate to 1e-4, total batch size to 32, and train on two RTX 3090 GPUs using the AdamW [Loshchilov and Hutter, 2019] optimizer with constant scheduler. For training images, we use center crop and random flip as data augmentation. In the inference stage, we use the Euler scheduler with 50 steps for generation.

Metrics

In addition to the qualitative results, we use multiple quantitative indicators to evaluate our proposed method from various dimensions. For downstream task evaluation, we use mAP (mean Average Precision) and AP50 to evaluate the generated data, and for generation quality evaluation, we use the widely used Frechet Inception Distance (FID) [Heusel *et al.*, 2017] to evaluate the fidelity of the generated images. In addition, to evaluate the diversity of the generated images, we calculate the diversity score (DS) by comparing the LPIPS [Zhang *et al.*, 2018] metric of paired images. Finally, to evaluate the object amounts of the generated images, we designed an instance quantity score (IQS) that detects the instance quantity of each category under multiple confidence settings using the pre-trained YOLOv8m [Jocher *et al.*, 2023] and compares it with the original images. Algorithm is shown in Appendix.

4.2 Comparison Methods

We compare the proposed method with the state-of-the-art multi-object image augmentation methods Dataset Diffusion [Nguyen *et al.*, 2023], Mosaic Fusion [Xie *et al.*, 2023], Add SD [Yang *et al.*, 2024], and image variation methods ControlNet Reference Only [Zhang *et al.*, 2023], Versatile Diffusion [Xu *et al.*, 2023], and Prompt-Free Diffusion [Xu *et al.*, 2024]. We use the pre-trained model of the above methods to generate images under its default parameters.

4.3 Downstream Task Evaluation

To verify the effectiveness of the proposed method, we use the generated data to train downstream detection and segmentation models and compare the metrics of the validation set to test the ability of image augmentation. Specifically, we use the training set of the COCO dataset to generate 10k data for each method and mix it with the original training set to train Mask RCNN[He *et al.*, 2017]. For Dataset Diffusion, we use the pre-trained model to perform instance segmentation on its semantic labels. For Add SD, image variation methods, and our method, we use Grounding DINO[Liu *et al.*, 2024] and SAM[Kirillov *et al.*, 2023] to generate annotations.

Tab. 1 shows the performance indicators of all methods on the validation set. Our method achieves the state-of-the-art performance on both detection models. These results demonstrate the effectiveness of the proposed method and provide promising results for the further application of generative models in detection tasks.

Task	**bbox**		**mask**	
	mAP	AP50	mAP	AP50
train2017	38.65	59.48	35.24	56.32
Dataset Diffusion	38.34	59.09	35.12	56.12
MosaicFusion	38.74	59.60	35.15	56.45
Add SD	37.88	58.67	-	-
ControlNet	38.75	59.62	35.33	56.39
Versatile	38.80	59.50	35.36	56.50
Prompt-Free	38.68	59.23	35.15	56.21
Ours	**39.04**	**59.86**	**35.43**	**56.73**

Table 1: Performance comparison of downstream task evaluations across state-of-the-art methods.

4.4 Generation Quality Evaluation

To further verify the generation quality of the model, we use the validation set of the COCO dataset to evaluate the generated images. Specifically, we use each image in the validation set as a condition for image augmentation and calculate the fidelity, diversity score, and instance quantity score of all images. Since Dataset Diffusion [Nguyen *et al.*, 2023] and MosaciFusion [Xie *et al.*, 2023] do not support image-based augmentation, we do not compare with them here.

Qualitative Evaluation

Fig. 3 shows the visualization results on some challenging samples. Add SD[Yang *et al.*, 2024] does not always successfully add targets. ControlNet[Zhang *et al.*, 2023] cannot understand the semantics of the input image and loses the object of interest after image augmentation. The diversity of the images generated by Versatile Diffusion[Xu *et al.*, 2023] and Prompt Free Diffusion[Xu *et al.*, 2024] is not good. The layout of the image is almost the same as the original image, and there will be problems with missing targets and even counterfactual images. Compared with the above methods, our method achieves the best results in the balance of layout diversity, number of generated objects, and consistency with facts.

Quantitative Evaluation

As demonstrated in Tab. 2, val2017 represents the results of the original val dataset. The proposed method achieves the best or suboptimal results across FID, DS and IQS metrics, which proves the effectiveness of our method. It is worth not-

Methods	**FID ↓**	**DS ↑**	**IQS ↑**
val2017	-	-	45.02
Add SD[Yang *et al.*, 2024]	**6.90**	0.19	**32.55**
ControlNet[Zhang *et al.*, 2023]	25.50	0.64	15.91
Versatile[Xu *et al.*, 2023]	19.01	0.65	24.64
PFD[Xu *et al.*, 2024]	22.39	0.62	20.23
Ours	18.59	**0.71**	29.17

Table 2: Quantitative comparison with state-of-the-art methods. ↑ means higher is better, ↓ means lower is better. All generated images are evaluated at 512 × 512 resolution.

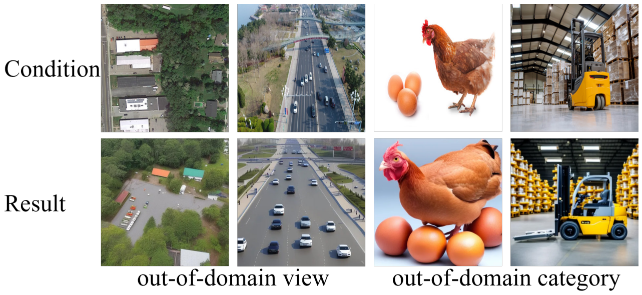

Figure 4: Out-of-domain experimental results under two settings.

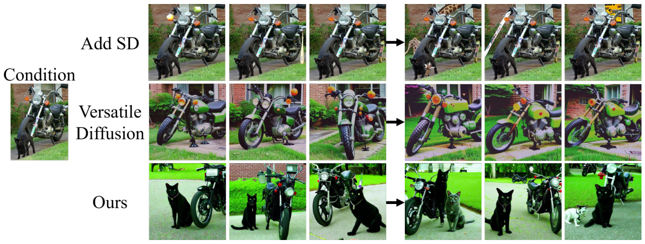

Figure 5: Recurrent generation for a given condition

ing that although Add SD [Yang *et al.*, 2024] is superior to the proposed method in terms of fidelity through image editing, its diversity score is greatly reduced, and the instance quantity score is even lower than the original dataset. The proposed method has optimal performance in terms of the balance of fidelity, diversity and instance quantity.

Out-of-domain Evaluation

We also conducted experiments on the out-of-domain generalization ability of the model, including two settings. The cross-view setting is shown in the first two columns of Fig. 4. We experiment with satellite and drone remote sensing images. Our method can correctly understand the semantics in the input image and perform cross-view image augmentation on it. The cross-category setting is shown in the last two columns of Fig. 4. We test categories such as chickens, eggs, forklift, et al. that are not in the COCO dataset. Our method can understand the semantics of unseen categories through only images and generate diverse images.

Multiple Random Generation & Recurrent Generation

As shown in Fig. 5, we verified the effects of different methods on multiple augmentations of a single image and further augmentations of the augmented image. Our method achieved the best layout diversity. More results can be found in the supplementary materials.

4.5 Ablation Study

Effectiveness of Each Component

To further verify the effectiveness of our proposed components, we conducted a series of ablation studies. These studies mainly focus on two key components of our model: the local-global semantic fusion strategy and the reward model based counting loss. We use SDXL img2img as our baseline. As shown in Tab. 3, after using semantic fusion module to replace the original text condition module, the fidelity and diversity of the model have been significantly improved. Similarly, after adding counting loss, the instance quantity score of the model was further improved.

Method	**FID ↓**	**DS ↑**	**IQS ↑**
Baseline	29.49	0.36	25.67
+ Semantic Fusion	20.43	0.68	27.98
+ Counting Loss	**18.59**	**0.71**	**29.17**

Table 3: Ablations of local-global semantic fusion strategy (SF) and reward model based counting loss (CL).

Method	**FID ↓**	**DS ↑**	**IQS ↑**
Image Only	19.27	0.70	27.34
w/ Category Name	19.02	0.70	28.22
w/ Content Image	**18.59**	**0.71**	**29.17**
w/ Both	18.85	**0.71**	28.90

Table 4: Ablations of different conditions.

Method	**FID ↓**	**DS ↑**	**IQS ↑**
Random Crop	**18.42**	0.71	25.99
Grounding DINO	19.24	**0.72**	27.22
Ground Truth	18.59	0.71	**29.17**

Table 5: Ablations of different inference methods.

Analysis of Different Conditions

We also conducted ablation experiments under different conditions. As shown in Tab. 4, although both category name and content conditions can improve the performance of the model, the category name is a subset of the content, and using only the content can enable the model to achieve higher performance.

Analysis of Different Inference Methods

In practical applications, we cannot always obtain the ground truth of image annotations. So we use random crop and Grounding DINO detection results as contents for inference. As shown in Tab. 5, the model can achieve similar fidelity and diversity with a slight decrease in instance quantity score.

5 Conclusion

This paper introduces a prompt-free conditional diffusion framework for multi-object image augmentation. Through the proposed local-global semantic fusion strategy and the reward model based counting loss, the model can augment the images in a large-scale and diverse manner that conforms to the original category distribution. Qualitative and quantitative experimental evaluations substantiate the efficacy and superiority of our proposed methodology. At the same time, both out-of-domain generalization ability and recurrent augmentation ability of the model provide more possibilities for its application.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grand 62372379, Grant 62472359, and Grant 62472350; in part by the Xi'an's Key Industrial Chain Core Technology Breakthrough Project: AI Core Technology Breakthrough under Grand 23ZDCYJSGG0003-2023; in part by National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing fundation under Grant TJ-04-23-04; in part by Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grant CX2025092.

References

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

[Battash *et al.*, 2024] Barak Battash, Amit Rozner, Lior Wolf, and Ofir Lindenbaum. Obtaining favorable layouts for multiple object generation, 2024.

[Binyamin *et al.*, 2024] Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects, 2024.

[Chen *et al.*, 2023] Weijie Chen, Haoyu Wang, Shicai Yang, Lei Zhang, Wei Wei, Yanning Zhang, Luojun Lin, Di Xie, and Yueting Zhuang. Adapt anything: Tailor any image classifiers across domains and categories using text-to-image diffusion models, 2023.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Gal *et al.*, 2022] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[He *et al.*, 2016] 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*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[He *et al.*, 2023] 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? In *The Eleventh International Conference on Learning Representations*, 2023.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[Hu *et al.*, 2021] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[Jocher *et al.*, 2023] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics, 2023. Accessed: 2024-09-19.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[Li *et al.*, 2024] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback, 2024.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2024] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[Nguyen *et al.*, 2023] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 76872–76892. Curran Associates, Inc., 2023.

[Nichol *et al.*, 2022] 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, 2022.

[Podell *et al.*, 2023] 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, 2023.

[Radford *et al.*, 2021] 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*, pages 8748–8763. PMLR, 2021.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024.

[Rombach *et al.*, 2022] 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*, pages 10684–10695, 2022.

[Ruiz *et al.*, 2023] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[Saharia *et al.*, 2022] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[Suri *et al.*, 2024] Saksham Suri, Fanyi Xiao, Animesh Sinha, Sean Culatana, Raghuraman Krishnamoorthi, Chenchen Zhu, and Abhinav Shrivastava. Gen2det: Generate to detect. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024.

[Wang *et al.*, 2024] 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*, pages 6232–6242, 2024.

[Wen *et al.*, 2023] Song Wen, Guian Fang, Renrui Zhang, Peng Gao, Hao Dong, and Dimitris Metaxas. Improving compositional text-to-image generation with large vision-language models, 2023.

[Wu *et al.*, 2023a] Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, Di Zhang, and Zhongyuan Wang. Paragraph-to-image generation with information-enriched diffusion model, 2023.

[Wu *et al.*, 2023b] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36:54683–54695, 2023.

[Xie *et al.*, 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, 2023.

[Xu *et al.*, 2023] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023.

[Xu *et al.*, 2024] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8682–8692, 2024.

[Yang *et al.*, 2024] Lingfeng Yang, Xinyu Zhang, Xiang Li, Jinwen Chen, Kun Yao, Gang Zhang, Errui Ding, Lingqiao Liu, Jingdong Wang, and Jian Yang. Add-sd: Rational generation without manual reference. *arXiv preprint arXiv:2407.21016*, 2024.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhao *et al.*, 2023] 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*, pages 42098–42109. PMLR, 2023.