Boosting Few-Shot Detection with Large Language Models and Layout-to-Image Synthesis

Ahmed Abdullah[1], Nikolas Ebert[1], and Oliver Wasenmüller[1]

[1]Mannheim University of Applied Sciences, Germany
{a.abdullah, n.ebert, o.wasenmueller}@hs-mannheim.de

Abstract. Recent advancements in diffusion models have enabled a wide range of works exploiting their ability to generate high-volume, high-quality data for use in various downstream tasks. One subclass of such models, dubbed Layout-to-Image Synthesis (LIS), learns to generate images conditioned on a spatial layout (bounding boxes, masks, poses, etc.) and has shown a promising ability to generate realistic images, albeit with limited layout-adherence. Moreover, the question of how to effectively transfer those models for scalable augmentation of few-shot detection data remains unanswered. Thus, we propose a collaborative framework employing a Large Language Model (LLM) and an LIS model for enhancing few-shot detection beyond state-of-the-art generative augmentation approaches. We leverage LLM's reasoning ability to extrapolate the spatial prior of the annotation space by generating new bounding boxes given only a few example annotations. Additionally, we introduce our novel layout-aware CLIP score for sample ranking, enabling tight coupling between generated layouts and images. Significant improvements on COCO few-shot benchmarks are observed. With our approach, a YOLOX-S baseline is boosted by more than 140%, 50%, 35% in mAP on the COCO 5-,10-, and 30-shot settings, respectively.

Keywords: few-shot detection · layout-to-image synthesis · large language model

1 Introduction

Object detection and image classification models have enjoyed significant improvements over the years, thanks to both architectural improvements [6, 7, 27, 35, 51, 52] and extensively annotated datasets [4, 5, 26, 55]. The advantages that a high volume, appropriately annotated dataset can bring are evident in the rise of performance, applicability, and robustness of state-of-the-art computer vision models across many domains and tasks. However, such datasets are difficult to collect in a transparent fashion [37], and even more difficult to annotate, often requiring extensive manual effort to provide accurate labelling and annotating.

This bottleneck in scaling datasets becomes a limitation that simply training larger models may not be able to overcome. Additionally, the distributions of established datasets may not adequately reflect the full range of real-world
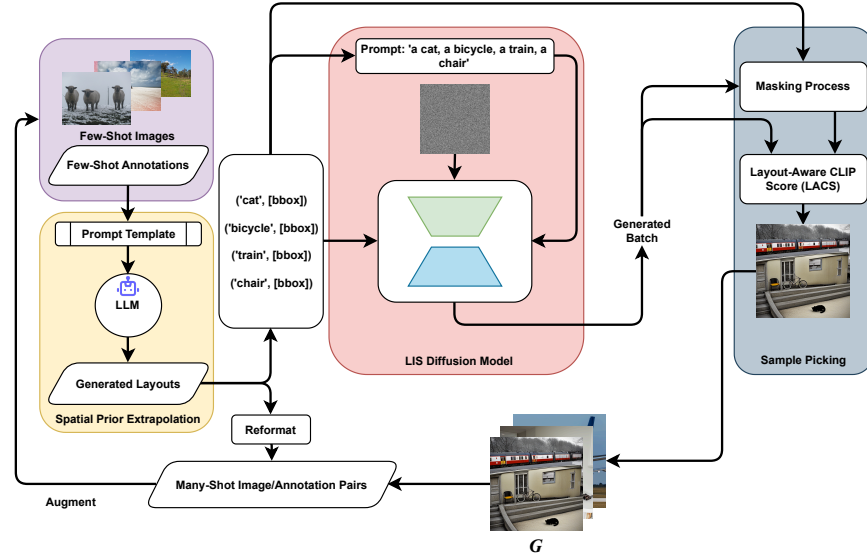
Fig. 1: Overview of proposed framework. First, we employ a pretrained LLM to generate new layouts with the help of a prompt template and the available few-shot annotations. Next, we generate new images using a layout-to-image synthesis diffusion model, conditioned on the newly created layouts. Finally, we employ a layout-aware CLIP score (LACS) to rate the generated samples, and construct a generated set of images G with high layout-adherence by picking the highest scoring images from a given batch. After reformatting the generated layouts into detection annotations, the resulting image-annotation pairs are used to augment the few-shot detection data.

concepts (e.g. specific species of animals). This lack of coverage can result in suboptimal downstream performance of models deployed in scenarios where such concepts are occurring more frequently.

Thus, few-shot learning has seen a noticeable emergence in research [2,56,57] concerned with training models where only a few samples are available, requiring approaches that can generalise to rarely seen concepts. One common approach to account for low-frequency, novel objects is by employing data augmentation pipelines. These range from basic image manipulation operations such as rotation, flipping, scaling, or cropping, to more advanced combinational augmentation strategies, such as CutMix [50], MixUp [53], AugMix [14] and Mosaic [12].

Recently, generative augmentation has emerged as a result of the advancement of image generation methodologies [19,20,36,38]. A plethora of works research the capability of such models to augment existing datasets [1,9,32,34,44]. In the context of image classification, augmentation-by-generation [1,57] is a straightforward process, provided that the generated sample is free of artifacts. This is primarily attributable to the fact that the spatial location of the concept within the synthesized image is not a major concern in this process. However,

generative augmentation for detection and segmentation requires robust conditional generators that can synthesize instances of objects strictly within the conditional spatial boundaries, without compromising the visual quality of the underlying samples.

Hence, Layout-to-Image Synthesis models (LIS) [31, 48, 54] have emerged, giving rise to generative augmentation methods for detection and segmentation tasks [9, 16, 44, 49] that employ diffusion models in their pipelines to provide a performance boost to the downstream detectors and segmentors. Nevertheless, many generative augmentation methods are either not applicable when taking the few-shot detection scenario into account [16, 49], fail to scale on higher augmentation ratios [9], or utilize pseudo-labelling in the pipeline instead of an LIS model [10, 44]. The latter approach leads to subpar image-annotation alignment due to the limitations of pseudo-labelling.

We hypothesize that to effectively reach high quality generative dataset augmentation for few-shot detection tasks, the spatial prior in the annotation space must be expanded in an automated manner, without additional manual effort. This approach ensures that not only the appearance of objects is augmented, but also their location within the scene. Inspired by the progress achieved by Large Language Models (LLM), we employ an LLM-based approach for generating new layouts, effectively extrapolating the spatial prior of the few-shot annotation space. We combine the previous extrapolation step with an LIS model to unlock scalable, high quality generative augmentation for few-shot detection.

In addition, we discover that only a few LIS approaches can generate realistic images while maintaining layout conformity. To ensure close image-layout alignment and realistic objects, we propose our novel Layout-Aware CLIP Score (LACS) for ranking generated images based on both realism and layout-adherence, effectively eliminating noisy samples. CLIP score has been used in prior works to rate the realism of generated images and for calculating the alignment between images and their respective text captions [15, 33, 42]. However, we aim to employ a more advanced scoring scheme to account for hallucinations outside of the conditional regions.

In summary, we propose our novel collaborative LLM-LIS generative augmentation framework (shown in Figure 1) for few-shot object detection that scales well beyond state-of-the-art approaches. We utilize an LLM for spatial prior extrapolation given only a small number of ground truth layouts. Furthermore, we employ our layout-aware CLIP score to enable close image-layout alignment for superior downstream performance. We extensively evaluate our approach on the COCO [26] few-shot detection benchmark and achieve substantial improvements in mean average precision (mAP) on the 5-, 10-, and 30-shot settings.

2   Related Work

2.1   Few-Shot Object Detection

To tackle the challenge of data scarcity and the need to detect novel categories in the wild, few-shot detection aims to train a detector model using only a

small number of instances of those categories in the training data. While many paradigms have been proposed on how to train a detector for few-shot tasks [41, 43], the most common way to handle few-shot detection regimes is pretrain-into-finetune [41]. This is accomplished by pretraining the detection model on high-frequency base categories with ample annotations, and then finetuning on the low-frequency, novel categories of objects while the backbone of the detector model remains frozen.

2.2   Layout-to-Image Synthesis

Regarded as the inverse task of object detection, Layout-to-Image Synthesis (LIS) aims to generate images by providing the model with layout information typically in the form of bounding boxes, masks, edge information, or scribbles. Earlier works [21, 39, 40] utilize Generative Adversarial Networks (GAN) [13] to synthesize images conditioned on a specified semantic or bounding box layout. Inspired by recent advancements in Text-to-Image (T2I) diffusion models [28, 36], many works [42, 45, 48, 54] have emerged to introduce granular location control of concepts synthesized by the diffusion model. This can be done either in a training-free, or a training-based manner.

Training-Free LIS To bypass the computational cost of training auxiliary modules and maintain a wide concept coverage, training-free LIS methods [3, 30, 45] propose leveraging a frozen latent diffusion model [36] for spatially-controllable T2I generation. This is accomplished by guiding the diffusion model features towards the layout condition regions during inference. However, most training-free LIS approaches result in rough layout-adherence, rendering them of limited applicability for few-shot detection.

Training-Based LIS Contrary to training-free LIS, training-based methods [22, 23, 29, 31, 42, 48, 54] integrate auxiliary modules with the frozen encoder-decoder blocks of the diffusion model to learn additional control features by training those modules on detection and segmentation datasets [26, 55]. Gligen [22] uses bounding boxes to formulate grounding tokens, employing an additional gated attention layer to achieve LIS. The recently proposed InstanceDiffusion [42] enables fine-grained instance-level control with flexible layout definitions by tokenizing the location information per-instance and fusing the resulting features with the frozen diffusion model backbone, achieving impressive LIS performance.

2.3   Applications of Generative Augmentation for Object Detection

Fang et al. [9] propose a framework utilizing visual priors (e.g. Holistic Edge [47], semantic segmentation maps) to augment few-shot detection by generating images with ControlNet [54] using the acquired visual priors. InstaGen and DatasetDM [10, 44] incorporate a trainable detector to pseudo-label images generated with a latent diffusion model [36] by exploiting rich attention maps from the diffusion model. Lin et al. [25] implement a copy-paste pipeline to enhance few-shot object detection.
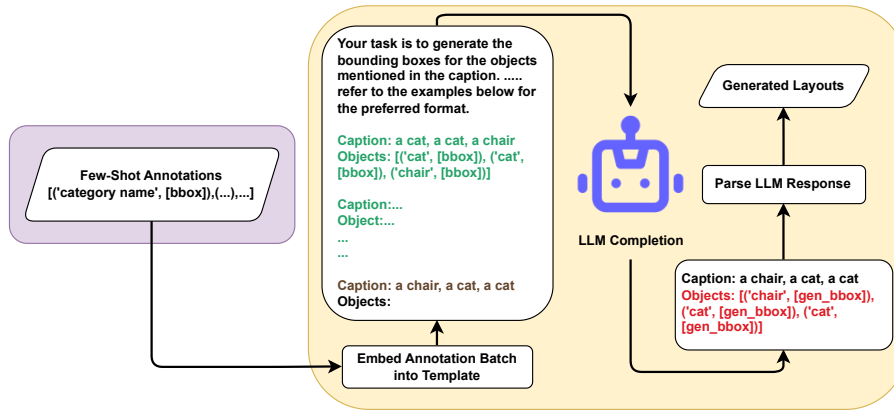
Fig. 2: LLM-based spatial prior extrapolation. First, we embed a batch of few-shot annotations into a prompt template by formatting them as layout descriptions. The embedded layout descriptions serve as in-context examples (in green) to steer the text generation process. We then prompt the LLM to complete a caption of randomly reordered objects (brown) from one of the layout descriptions and obtain a response containing generated bounding boxes (red). Finally, we parse the response to obtain new layouts.

While both Fang et al. [9] and InstaGen [10] offer promising results in object detection performance, they fall short due to either non-scalability with higher augmentation ratios [9] or insufficient image-annotation alignment of generated data due to noisy pseudo-labelling [10]. Our approach seeks to solve those shortcomings by extrapolating the spatial prior of bounding box distributions to achieve scalable, high quality generative augmentation. Instead of relying on post-generation pseudo-labelling, we deploy a pretrained LIS diffusion model coupled with our novel Layout-Aware CLIP Score (LACS) to ensure close image-layout alignment.

3 Method

In this section, we describe the key modules of our generative augmentation framework. As seen in Figure 1, we employ three modules to obtain high quality generated data. First, an LLM-based module extrapolates the spatial prior of the few-shot annotation data by generating bounding box layouts. The generated layouts are then used to guide an LIS model for image synthesis. This is followed by our novel layout-aware CLIP score, which is used to rate the generated batch of images for sample picking.

3.1   LLM-based Spatial Prior Extrapolation

Motivated by the emerging application of language models as layout generators [11, 24, 46], we incorporate a pretrained large language model for spatial prior extrapolation (SPE) via generation of novel layouts.

To elaborate, the spatial prior encompasses the existing layout annotations in the few-shot data, and is presented to the generator model (in our case, the LLM) as in-context examples. The task of the generator model is to *extrapolate* from the presented examples, and generate plausible layouts to be used for image generation via LIS, hence the term *Spatial Prior Extrapolation*.

We employ LLM auto-completion with a pretrained Mixtral-8x7B-Instruct-v0.1. [17]. We choose Mixtral as our LLM due to its' competitive performance in layout generation when compared to closed-source models such as GPT-3.5 and GPT-4 in spatial and numeracy reasoning tasks, as tested by LMD [24]. Similar to LMD [24], we first compose a prompt template. The template consists of detailed bounding box layout generation instructions for the LLM to follow, in addition to spatial restrictions with respect to box placement and image dimensions. We then embed a batch of per-image ground truth annotations A_b as context examples into the prompt template by formatting them as layout descriptions (e.g. caption:'a cat, a car, a person', objects: ['cat',[bbox1], 'car', [bbox2], 'person', [bbox3]]), as seen in Figure 2.

We process the batch A_b by concatenating the captions of the ground truth layout descriptions (one caption per auto-completion) after randomizing the occurrence-order of the objects within the captions. The LLM is then used to auto-complete the prompt, generating object bounding boxes corresponding to the concatenated caption. To induce generation diversity, we also randomize the order of examples in the prompt template for each auto-completion.

We repeat the auto-completion for α times per batch element, where α is the augmentation ratio (number of synthetic layouts over real layouts). The bounding box layouts generated by the LLM are subsequently parsed and utilized in the LIS process and as annotations for detector training.

3.2   Layout-to-Image Synthesis Diffusion Model

At the core of our framework, we employ a pretrained LIS diffusion model for image generation. In contrast to typical text-to-image latent diffusion models [36], an LIS diffusion model accepts spatial layouts as conditional signals to guide the results of text-to-image generation.

For generating images from a layout and a text condition, a batch of Gaussian noise latent codes $Z \in \mathbb{R}^{b \times h \times w \times c}$ is initialized, where b, h, w, c correspond to the batch size, height, width, and number of channels of the latent code, respectively. The latent denoising network **LDN** of the LIS model produces the denoised latent codes $\tilde{Z}$ as

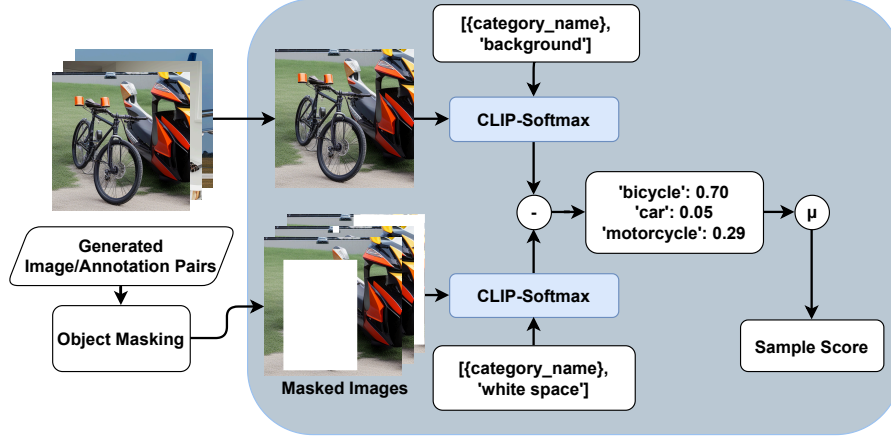$$\tilde{Z} = \mathbf{LDN}(Z, p, l, t), \tag{1}$$

Fig. 3: Overview of LACS. First, we create **n** masked images from a generated image, where **n** is the number of object categories in the image. Next, for each category, we perform zero-shot classification on both the generated image and the masked image and obtain a per-category layout-adherence score by subtracting the two classification scores. We average over all categories to arrive at the final sample score.

where p corresponds to the text prompt, l the spatial layout condition, and t the number of denoising time steps.

The denoised latents $\tilde{Z}$ are then decoded by the latent decoder $\mathcal{D}$ to generate a batch of images $\tilde{X}$ as

$$\tilde{X} = \mathcal{D}(\tilde{Z}). \tag{2}$$

In Section 4.3, we ablate over different pretrained State-of-the-Art LIS models with both masks and bounding boxes as conditional layouts. We discover that InstanceDiffusion [42] with bounding box conditions offers the best synthesis in terms of layout-adherence and overall image quality.

3.3   Layout-Aware CLIP Score

To ensure a close image-layout alignment, we integrate a novel layout-aware CLIP score into our framework (shown in Figure 3). For each image $\tilde{x}$ in a batch of LIS generated images $\tilde{X}$, we first obtain the CLIP [33] softmax score CS for each category present in the layout

$$CS(text_n, \tilde{x}) = S_0(C_{logits}(\tilde{x}, \{text_n, 'background'\})), \tag{3}$$

where $text_n$ corresponds to the text of the n'th category in the layout, C_{logits} the CLIP cosine similarity logits, and S_0 denotes the softmax output of the first logit from the CLIP inference.

Next, we produce a masked image $\tilde{x}_{masked_n}$ from the generated image by masking the bounding boxes of the respective category with white pixels. We

Fig. 4: Samples generated with InstanceDiffusion [42] and their LACS score. Green boxes highlight the conditional layout, while red regions highlight out-of-layout hallucinations.

obtain the masked CLIP softmax score CS_{mask} by inferring on the masked image and switching the second text prompt from $'background'$ to $'whitespace'$ as

$$CS_{mask}(text_n, \tilde{x}_{masked_n}) = S_0(C_{logits}(\tilde{x}_{masked_n}, \{text_n, 'whitespace'\})). \quad (4)$$

The text prompt $'background'$ is introduced as a placeholder to quantify the presence of the object category in the image for the first score CS. In the masked score CS_{mask}, the text prompt $'whitespace'$ is used for inducing a lower value in the logit corresponding to the object category text $text_n$, when a generated image $\tilde{x}$ contains little to no out-of-layout hallucinations. The opposite occurs when said hallucinations are present in the generated image, leading to a higher logit value for $text_n$. Finally, we obtain the final Layout-Aware CLIP Score (LACS) for image $\tilde{x}$ by subtracting CS_{mask} from CS for all n categories and calculating the mean score as

$$LACS(\tilde{x}) = \frac{1}{n} \sum_{1}^{n} CS(text_n, \tilde{x}) - CS_{mask}(text_n, \tilde{x}_{masked_n}). \quad (5)$$

The score above is used to sort the individual images $\tilde{x}$ in the generated batch $\tilde{X}$, where the best ranking images contain the least out-of-layout hallucinations, and vice-versa. We utilize the score for sample picking in our experiments (LACS-SP), and for calculating the mean layout-adherence of generated images (mLACS) for quality comparisons. Figure 4 shows the LACS score for images generated with InstanceDiffusion [42] and its' correlation with out-of-layout hallucinations.

Table 1: Main results over novel COCO [26] 5-,10-, and 30-shot settings. We compare our approach against state-of-the-art generative augmentation methods [9, 10] in few-shot detection settings and showcase the effect of spatial prior extrapolation (SPE) and top-1 sample picking w.r.t. our layout-aware CLIP Score (LACS-SP). We utilize InstanceDiffusion [42] in our framework and compare downstream performance with and without our contributions (SPE and LACS-SP). Results for Fang et al. [9] correspond to their best augmentation ratio α of 1 as reported in their findings. For comparison with InstaGen [10], we synthesize with an augmentation ratio α of 4 with both their method and ours.

Detector		Features		5-Shot		10-Shot		30-Shot	
		SPE	LACS-SP	mAP	AP^{50}	mAP	AP^{50}	mAP	AP^{50}
YOLOX-S [12]	Baseline			5.0	10.1	9.6	18.1	14.2	26.7
	w/ Fang et al. [9]			5.9	11.4	11.1	20.6	15.9	27.8
	w/ InstaGen [10]			11.5	19.7	15.1	25.3	18.7	30.9
	w/ InstDiff [42]			11.0	19.2	14.9	25.6	19.5	32.6
	w/ Ours	✓		12.4	21.4	15.6	26.8	20.1	33.5
			✓	11.4	19.7	15.7	26.5	19.7	33.0
		✓	✓	**12.8**	**21.9**	**16.1**	**27.1**	**20.4**	**34.1**
DINO-Swin-L [52]	Baseline			18.6	26.0	24.3	33.7	35.8	49.5
	w/ Fang et al. [9]			20.3	28.1	26.0	36.8	35.0	48.8
	w/ InstaGen [10]			26.4	36.0	30.4	41.3	36.4	50.0
	w/ Ours	✓	✓	**27.4**	**38.9**	**30.9**	**43.8**	**37.1**	**53.2**

4 Experiments

4.1 Experiment Setup

For the evaluation of our method, we conduct our experiments using the COCO [26] detection dataset. We evaluate under two settings: the established few-shot novel category finetuning [9, 18, 41], and all-category training using a small synthetic dataset. Unless otherwise stated, we use Mixtral-8x7B-Instruct-v0.1 [17] for bounding box layout generation and embed its prompt template (outlined in Section 3.1) with batches A_b of five example layouts from the ground truth annotations. Here, we set the augmentation ratio α to 4. When not utilizing spatial prior extrapolation with Mixtral, we simply oversample the ground truth annotations with x-axis flipping to reach the desired augmentation ratio. For layout-to-image synthesis, we utilize a pretrained InstanceDiffusion [42] with bounding boxes as layouts and set the number of denoising steps t to 50, the guidance scale to 7.5, grounding-alpha (percentage of timesteps using grounding inputs) to 0.8, and MIS (percentage of timesteps using the multi-instance sampler) to 0.36. We generate a batch $\tilde{X}$ of five images, rank them with our novel LACS score, and pick the top-1 image from the batch. Unless otherwise stated, we augment the few-shot real data with the resulting generated data. All detectors are set to a fixed seed for reproducibility when finetuning.

4.2 Evaluation on COCO Detection

Few-Shot Novel Category Finetuning In the standard few-shot settings on COCO [26], the 80 dataset categories are split into two sets: 60 base-categories

Table 2: Comparison with InstaGen [10] on finetuning Faster RCNN [35] using all COCO [26] categories. We generate 1284 synthetic image-annotation pairs with both our approach and InstaGen [10] and finetune using synthetic data only.

	mAP	AP^{50}	AP^{75}	mAP^s	mAP^m	mAP^l	mLACS
InstaGen [10]	3.0	7.8	1.8	0.1	2.3	6.9	0.59
Ours	**4.3**	**10.0**	**3.3**	**0.7**	**5.2**	**7.7**	**0.65**

and 20 novel-categories. The base-category images are used to pretrain a detector, while the novel-category images are used to finetune the detector after the initial pretraining. We test our proposed framework on 5-, 10-, and 30-shot settings with YOLOX-S [12] and DINO-Swin-L [52] and report the standard COCO mAP and AP^{50} metrics [8] on the novel categories. For fair comparison, we use identical hyperparameters for both detector networks as Fang et al. [9]. We observe improvements over both Fang et al. [9] and state-of-the-art detection dataset generator InstaGen [10] in Table 1 in all scenarios. When utilizing spatial prior extrapolation (SPE) via Mixtral [17], we notice a considerable improvement in mAP due to more diverse box locations, as opposed to simply oversampling the ground truth annotations. When combining SPE with top-1 sample picking via our layout-aware CLIP score, we are able to improve the YOLOX-S [12] baseline by 156%, 67%, and 43% on the 5-, 10-, and 30-shot settings, respectively. Interestingly, running only InstanceDiffusion [42] in the framework brings a noticeable performance boost over the baseline, owing to the benefit of instance-awareness of the LIS model.

All-Categories Training on Synthetic Data Only To test the generalization of our approach compared to InstaGen [10], we train a standard Faster RCNN [35] detector for 12 epochs using synthetic data only from both our proposed framework and InstaGen. Here, we utilize all 80 COCO categories and generate 1284 images using 5-shot layouts as the only available data. In Table 2, we report additional metrics AP^{75}, mAP for small, medium and large objects (mAP^s, mAP^m, mAP^l), as well as the mean layout-aware CLIP score (mLACS) for both generated datasets.

The performance improvements of our method stem from the use of controllable, instance-aware layout-to-image synthesis, in addition to appropriate layout adherence via sample picking.

4.3   Ablation Studies

In this section, we ablate over several design choices for our framework. We analyse different LIS models, masks or bounding boxes as conditional layouts, quality vs. quantity in LIS generation, an alternative approach for spatial prior extrapolation, and scalability of our framework with respect to higher augmentation

Table 3: Ablation over different LIS models [29, 42, 48] and conditioning signals (Cond.), with and without top-1 sample picking w.r.t. our layout-aware CLIP score (LACS-SP) over COCO [26] 10-shot with YOLOX-S [12]. We record mAP and AP^{50} bounding box metrics as well as dataset quality (Qual.) metrics mean LACS (mLACS) and average CLIP classification score of cropped bounding boxes (CS-Crop). For comparison, we include finetuning results without the use of generated images as baseline.

Method	LACS-SP	Cond.		Box-Metric		Qual.-Metric	
		Box	Mask	mAP	AP^{50}	mLACS	CS-Crop
Baseline [12]		N/A	N/A	9.6	18.1	-	-
Freestyle [48]		N/A	✓	10.1	18.8	0.483	0.705
	✓	N/A	✓	10.7	19.6	0.767	0.713
PLACE [29]		N/A	✓	10.3	19.3	0.447	0.774
	✓	N/A	✓	11.2	20.4	0.687	0.776
Instance-Diffusion [42]		✓	✓	13.3	22.9	0.587	0.787
	✓	✓	✓	13.2	22.8	**0.804**	0.784
			✓	9.1	16.9	0.317	0.823
	✓		✓	9.6	17.8	0.563	**0.824**
		✓		14.8	25.6	0.522	0.817
	✓	✓		**15.6**	**26.7**	0.735	0.819

ratios. In all of the ablation studies, we finetune a YOLOX-S [12] in standard COCO [26] few-shot testing with generated data and evaluate on COCO-val17.

Comparing LIS Methods and Necessity of Layout-Adherence To generate useful synthetic data for downstream detection tasks, an LIS model boasting both realistic generation capability and layout adherence is of utmost importance. We benchmark three state-of-the-art pretrained LIS models with and without our sample picking: Freestyle [48], PLACE [29], and InstanceDiffusion [42]. In this experiment, we oversample layouts from the 10-shot annotations to reach an augmentation ratio of 4. For generating with mask layouts, we fill bounding boxes with random instance segmentation masks from the 10-shot annotations. We assess downstream detection performance along with the layout adherence and quality of the generated instances. To quantify layout adherence and instance quality, we take the mean layout-aware CLIP score (mLACS) over all generated images to measure the former, and calculate the average CLIP [33] classification scores on cropped bounding boxes (CS-Crop) for the latter.

As evident from the results from Table 3, top-1 sample picking with our proposed LACS metric yields a favorable boost in detection performance in almost all cases. This improvement results from discarding noisy images, effectively improving the image-layout alignment in generated images. All three LIS models benchmarked offer an improvement in downstream detection, with InstanceDiffusion [42] showing the most notable gains. We notice that conditioning InstanceDiffusion with only bounding boxes results in the best improvement, while introducing masks leads to either out-of-layout hallucination (in case of using
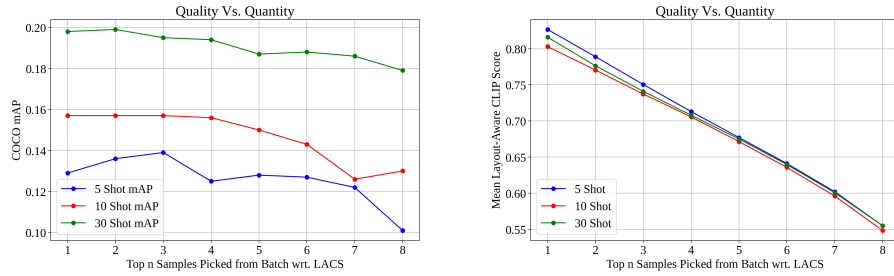
Fig. 5: Quality vs. quantity analysis. We analyse the effect of picking the top n samples from a generated batch of eight images sorted by the layout-aware clip-score (LACS).

masks only) or increased instance artifacts (in case of using boxes and masks). One may speculate that the cause could be inaccurate mask pseudo-labelling used in the creation of its training data.

Quality vs. Quantity of Generated Images One may question whether or not including more samples from the generated batch is beneficial to detection due to scene variance, regardless if the batch contains images not adhering well to the layout. To test this, we generate annotations for 5-, 10-, and 30-shot settings with an augmentation ratio of 4 and compare different top-n sample picking strategies: given a conditional layout, we sort a generated batch of eight images according to our LACS metric and pick the top-n scoring samples from the sorted batch. To counter-act the effect of longer finetuning, we scale back the number of epochs with respect to the number of samples picked, such that each strategy is trained for the same duration. Figure 5 shows a degradation occurring in detector performance when more than 50% of samples are picked from the generated batch. This results in overall lower layout-adherence, which is critical for training a detector on LIS synthetic data.

Alternative Approach for Spatial Prior Extrapolation In this ablation study, we explore an alternative approach to spatial prior extrapolation with Gaussian Mixture Models (GMM). Specifically, we model bounding boxes of shape $bbox \in \mathbb{R}^{x \times y \times w \times h}$ as a random distribution and fit an ensemble of $\boldsymbol{N}$ GMMs on the bounding boxes of novel categories (one category per GMM). After fitting, we sample valid bounding boxes from the fitted ensemble of GMMs according to object co-occurrences from a ground truth layout (e.g. how many objects of category 'person', 'car', 'cat', etc). We also explore fitting an additional GMM to model object co-occurrences as a random distribution as well. We benchmark both GMM approaches (with and without object co-occurrence modelling) by fitting them on COCO [26] 10-shot annotations and generating bounding box layouts with the default augmentation ratio α of 4. We compare this approach against using Mixtral [17] for layout generation as outlined

Table 4: Ablation over different spatial prior extrapolation methods. We compare an LLM approach discussed in Section 3.1 with Mixtral [17], ground truth 2× oversampling with x-axis flipping, and N-Ensemble of Gaussian mixture models (GMM). For the GMM approach, we consider drafting object co-occurrences directly from ground truth data (1), or from an auxiliary GMM fitted on the ground-truth object co-occurrence data (2). We also monitor the resulting layout-adherence via mean layout-aware CLIP score (mLACS) discussed in Section 3.3.

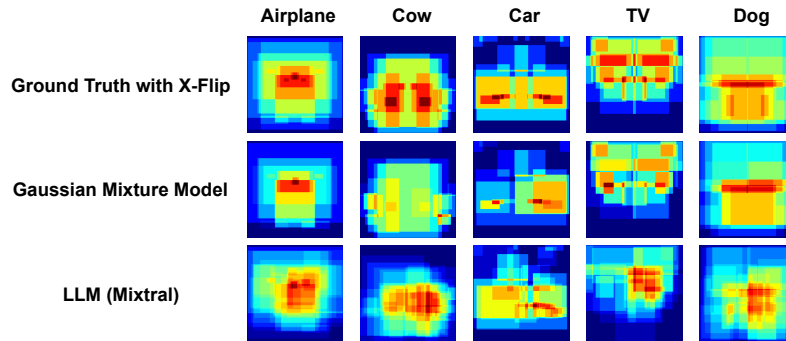| Method	mAP	AP^{50}	AP^{75}	mAP^s	mAP^m	mAP^l	mLACS
GT Oversampling + X-flip	15.6	26.7	15.9	**5.5**	14.4	**23.2**	0.735
N-Ensemble of GMM (1)	14.4	25.3	14.7	4.3	13.7	21.6	0.742
N-Ensemble of GMM (2)	14.6	25.4	15.1	4.5	12.8	22.4	0.732
LLM (Mixtral [17])	**15.9**	**27.0**	**16.7**	**5.5**	**14.9**	23.1	**0.780**

Fig. 6: Bounding box heatmaps for ground truth boxes, Gaussian mixture model generated boxes, and Mixtral [17] generated boxes on 5 novel COCO [26] categories.

in Section 3.1 and a simple strategy of oversampling the ground truth annotations with x-axis flipping. Results shown in Table 4 indicate that modelling layouts as random distributions via GMMs does not yield improvements over ground truth oversampling, which can be attributed to layout data scarcity in few-shot settings. Interestingly, Mixtral [17] is able to extrapolate reasonably from the provided examples and produces plausible and diverse bounding box layouts. Figure 6 shows the bounding box heatmaps for ground truth boxes, GMM-sampled boxes, and Mixtral-sampled boxes.

Effect of Higher Augmentation Ratios To showcase the full extent of our generative augmentation framework, we incrementally increase the layout augmentation ratio α and monitor the downstream YOLOX-S [12] performance. We compare ground truth layout oversampling (GTOS) against our LLM-based spatial prior extrapolation (SPE). Figure 7 shows superior scalability when generating with SPE than with GTOS. This effect is more pronounced on augmenta-
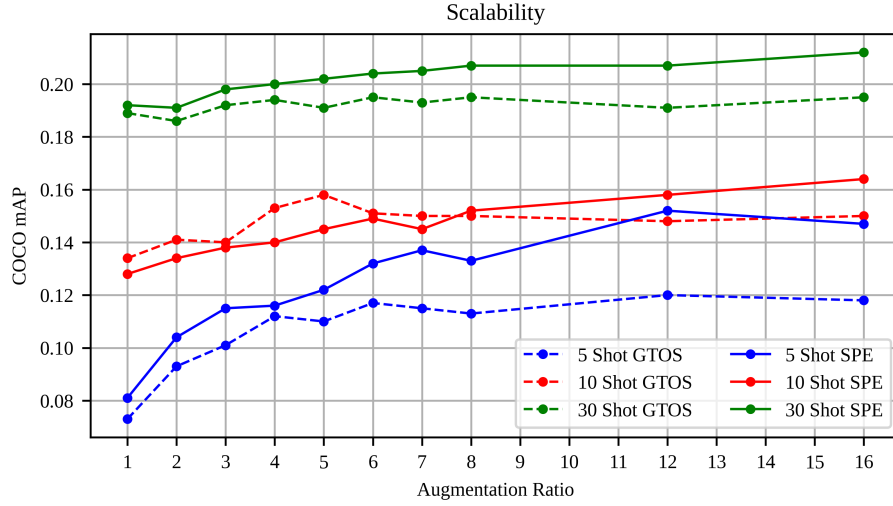
Fig. 7: Scalability of our method with respect to higher augmentation ratios α. We generate bounding box annotations with our LLM-based spatial prior extrapolation (SPE) method and compare with simple ground truth oversampling with x-axis flipping (GTOS). We synthesize images conditioned on the generated box layouts with InstanceDiffusion [42] and subsequently evaluate on YOLOX-S [12] finetuning.

tion ratios of 8 and higher. In 10-shot settings, SPE lags slightly behind GTOS, but begins to outperform GTOS on higher augmentation ratios. In 5-shot and 10-shot settings, SPE remains superior across all augmentation ratios.

5   Conclusion

We propose a generative augmentation framework for few-shot object detection, combining a large language model with a layout-to-image synthesis model to generate high-quality image-annotation pairs. We introduce the Layout-Aware CLIP Score (LACS), a CLIP-based [33] metric for evaluating layout adherence in generated images. Our framework outperforms state-of-the-art generative methods [9,10] on few-shot COCO [26] benchmarks, and ablation studies demonstrate the benefits of LLM-based spatial prior extrapolation and improving layout adherence through sample selection using LACS.

Acknowledgement. We would like to thank Haoyang Fang [9] for their help in reproducing their experiment baselines.

References

1. Bansal, H., Grover, A.: Leaving reality to imagination: Robust classification via generated datasets. arXiv preprint 2302.02503 (2023)

2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: Neural Information Processing Systems (NeurIPS) (2020)
3. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Winter Conference on Applications of Computer Vision (WACV) (2024)
4. Cruz, S.D.D., Wasenmüller, O., Beise, H.P., Stifter, T., Stricker, D.: SVIRO: Synthetic vehicle interior rear seat occupancy dataset and benchmark. In: Winter Conference on Applications of Computer Vision (WACV). pp. 973–982 (2020)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
6. Ebert, N., Stricker, D., Wasenmüller, O.: PLG-ViT: Vision transformer with parallel local and global self-attention. Sensors (2023)
7. Ebert, N., Stricker, D., Wasenmüller, O.: Transformer-based detection of microorganisms on high-resolution petri dish images. In: International Conference on Computer Vision Workshops (ICCVW) (2023)
8. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision (IJCV) (2010)
9. Fang, H., Han, B., Zhang, S., Zhou, S., Hu, C., Ye, W.M.: Data augmentation for object detection via controllable diffusion models. In: Winter Conference on Applications of Computer Vision (WACV) (2024)
10. Feng, C., Zhong, Y., Jie, Z., Xie, W., Ma, L.: Instagen: Enhancing object detection by training on synthetic dataset. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
11. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: Layoutgpt: Compositional visual planning and generation with large language models. In: Neural Information Processing Systems (NeurIPS) (2024)
12. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint 2107.08430 (2021)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM (2020)
14. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. In: International Conference on Learning Representations (ICLR) (2020)
15. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021)
16. Jia, Y., Hoyer, L., Huang, S., Wang, T., Van Gool, L., Schindler, K., Obukhov, A.: Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2024)
17. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al.: Mixtral of experts. arXiv preprint 2401.04088 (2024)
18. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: International Conference on Computer Vision (ICCV) (2019)

19. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Neural Information Processing Systems (NeurIPS) (2020)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
21. Li, Y., Cheng, Y., Gan, Z., Yu, L., Wang, L., Liu, J.: Bachgan: High-resolution image synthesis from salient object layout. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
22. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
23. Li, Y., Keuper, M., Zhang, D., Khoreva, A.: Adversarial supervision makes layout-to-image diffusion models thrive. In: International Conference on Learning Representations (ICLR) (2024)
24. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. arXiv preprint 2305.13655 (2023)
25. Lin, S., Wang, K., Zeng, X., Zhao, R.: Explore the power of synthetic data on few-shot object detection. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2023)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (ECCV) (2014)
27. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
28. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint 2310.04378 (2023)
29. Lv, Z., Wei, Y., Zuo, W., Wong, K.Y.K.: Place: Adaptive layout-semantic fusion for semantic image synthesis. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
30. Mo, S., Mu, F., Lin, K.H., Liu, Y., Guan, B., Li, Y., Zhou, B.: Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
31. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In: AAAI conference on Artificial Intelligence (2024)
32. Oehri, S., Ebert, N., Abdullah, A., Stricker, D., Wasenmüller, O.: GenFormer – generated images are all you need to improve robustness of transformers on small datasets. In: International Conference on Pattern Recognition (ICPR) (2024)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
34. Reichardt, L., Uhr, L., Wasenmüller, O.: Text3DAug – prompted instance augmentation for lidar perception. In: International Conference on Intelligent Robots and Systems (IROS) (2024)

35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems (NeurIPS) (2015)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
37. Scheuerman, M.K., Hanna, A., Denton, E.: Do datasets have politics? disciplinary values in computer vision dataset development. Proceedings of the ACM on Human-Computer Interaction (2021)
38. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint 2010.02502 (2020)
39. Sun, W., Wu, T.: Image synthesis from reconfigurable layout and style. In: International Conference on Computer Vision (ICCV) (2019)
40. Tang, H., Qi, X., Sun, G., Xu, D., Sebe, N., Timofte, R., Van Gool, L.: Edge guided gans with contrastive learning for semantic image synthesis. arXiv preprint 2003.13898 (2020)
41. Wang, X., Huang, T., Gonzalez, J., Darrell, T., Yu, F.: Frustratingly simple few-shot object detection. In: International Conference on Machine Learning (ICML) (2020)
42. Wang, X., Darrell, T., Rambhatla, S.S., Girdhar, R., Misra, I.: Instancediffusion: Instance-level control for image generation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
43. Wang, Y.X., Ramanan, D., Hebert, M.: Meta-learning to detect rare objects. In: International Conference on Computer Vision (ICCV) (2019)
44. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. In: Neural Information Processing Systems (NeurIPS) (2023)
45. Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: International Conference on Computer Vision (ICCV) (2023)
46. Xie, J., Ye, K., Li, Y., Li, Y., Lin, K.Q., Zheng, Y., Shen, L., Shou, M.Z.: Visorgpt: Learning visual prior via generative pre-training. arXiv preprint 2305.13777 (2023)
47. Xie, S., Tu, Z.: Holistically-nested edge detection. In: International Conference on Computer Vision (ICCV) (2015)
48. Xue, H., Huang, Z., Sun, Q., Song, L., Zhang, W.: Freestyle layout-to-image synthesis. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
49. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. In: Neural Information Processing Systems (NeurIPS) (2024)
50. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: International Conference on Computer Vision (ICCV) (2019)
51. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
52. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint 2203.03605 (2022)
53. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint 1710.09412 (2017)
54. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: International Conference on Computer Vision (ICCV) (2023)

55. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
56. Zhou, J., Zheng, Y., Tang, J., Jian, L., Yang, Z.: Flipda: Effective and robust data augmentation for few-shot learning. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2022)
57. Zhou, Y., Sahak, H., Ba, J.: Training on thin air: Improve image classification with generated data. arXiv preprint 2305.15316 (2023)