Auto Cherry-Picker: Learning from High-quality Generative Data Driven by Language

Yicheng Chen[1,2], **Xiangtai Li**[1,3], **Yining Li**[1†], **Yanhong Zeng**[1], **Jianzong Wu**[1,4], **Xiangyu Zhao**[1,5], **Kai Chen**[1†]
[1]Shanghai AI Laboratory [2]Fudan University
[3]S-Lab, Nanyang Technological University
[4]Peking University [5]Shanghai Jiao Tong University
Project page: https://yichengchen24.github.io/projects/autocherrypicker

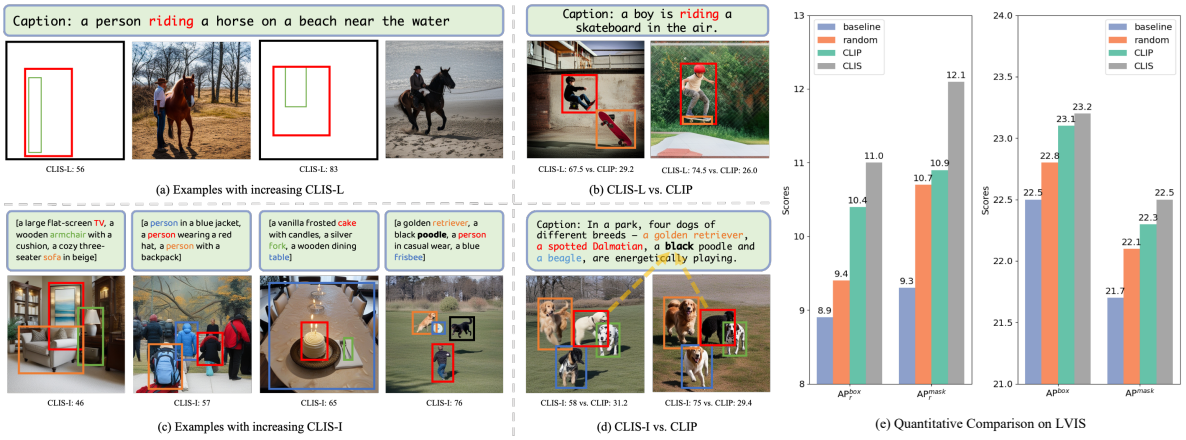

Figure 1. Illustration of quality assessment of generated data samples using CLIS. (a) and (c) compare the quality of samples with different CLIS-L and CLIS-I scores, respectively. Samples with low CLIS fail to align accurately with the condition (e.g., contain extraneous objects or exhibit visual flaws). (b) and (d) compare the preferences of CLIS and CLIP score [27]. (e) compares different selection methods for the same volume of synthetic data used in downstream tasks, reporting AP_r and AP on the LVIS benchmark. See details in Sec. 4.2.

Abstract

*Diffusion models can generate realistic and diverse images, potentially facilitating data availability for data-intensive perception tasks. However, leveraging these models to boost performance on downstream tasks with synthetic data poses several challenges, including aligning with real data distribution, scaling synthetic sample volumes, and ensuring their quality. To bridge these gaps, we present **A**uto **C**herry-**P**icker (ACP), a novel framework that generates high-quality cross-modality training samples at scale to augment perception and multi-modal training. ACP first uses LLMs to sample descriptions and layouts based on object combinations from real data priors, eliminating the need for ground truth image captions or annotations. Next, we use an off-the-shelf controllable diffusion model to generate multiple images. Then, the generated data are refined using a comprehensively designed metric, Composite Layout and Image Score (CLIS), to ensure quality. Our customized synthetic high-quality samples boost performance in various scenarios, especially in addressing challenges associated with long-tailed distribution and imbalanced datasets. Experiment results on downstream tasks demonstrate that ACP can significantly improve the performance of existing models. In addition, we find a positive correlation between CLIS and performance gains in downstream tasks. This finding shows the potential for evaluation metrics as the role for various visual perception and MLLM tasks.*

1. Introduction

Recently, diffusion-based image generation methods [58, 64] have made remarkable progress, enabling various applications, including text-to-image generation(T2I) [20, 21, 57, 60, 71], image editing [5, 26, 53, 59], video generation [22, 29, 30], and more. Compared to previous generative models [13, 34], diffusion models can produce **high-quality** and **high-resolution** examples. Thus, one essential usage of the diffusion-based model is to generate training data for various downstream vision tasks, such as segmentation [42, 80], detection [8, 44], and visual representation learning [38, 69]. Synthetic data alleviates the severe demand for human annotation and provides a more controllable data production process.

† Corresponding Author.

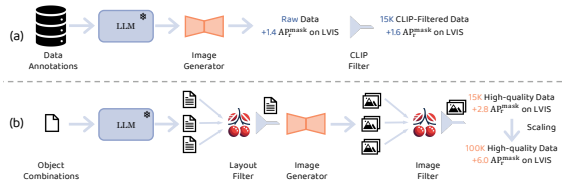

Figure 2. Comparison with previous methods. (a) LMD [43] generates samples conditioned on detailed image descriptions by leveraging LLMs as the layout generator and diffusion-based models as the image generator. Some methods [25] use CLIP filtering to future refine these samples. (b) ACP synthesizes training samples conditioned solely on object combinations in natural language and automatically cherry-picks high-quality ones by evaluating both layouts and images. High-quality training samples are more effective for downstream tasks.

Previous works have explored generating samples via different types of references, such as captions [43], bounding boxes [74], instance masks [49, 67], and reference images [83]. In particular, InstanceDiffusion [74] generates images with precise instance-level control, including attribute binding and localization, while maintaining high fidelity conditioned on detailed instances. However, these approaches still rely on *manual* annotations, limiting their scalability and diversity in generating large-scale training datasets for downstream tasks. In addition, due to the *inherent randomness* of generative models, the quality of generated data tends to vary, potentially impairing the effectiveness of using this data to train downstream tasks [25]. Consequently, appropriate quality assessment metrics must be employed to rule out low-quality synthetic samples. At the image level, current semantic-based metrics, such as CLIP-Score [25], fall short in precisely assessing dense annotations, while detector-based metrics [65] struggle to capture detailed descriptions and finer relationships between objects. At the layout level, existing methods are constrained by predefined rules [31] and often neglect to consider other reasonable candidates [33]. As illustrated in Fig. 2, current methods face limitations due to costly ground truth annotations and a lack of effective quality assessment metrics. To solve these issues, several essential questions are raised: 1) How can we reduce the reliance on annotations, including caption and layout, while aligning closely with real data distributions? 2) How do we measure the quality for multiple instances, and can we propose a new metric to select good ones? 3) Does the proposed metric reflect the final downstream performance when used for training?

To this end, we propose Auto Cherry-Picker (ACP), a data generation pipeline leveraging pre-trained generative models and LLMs to generate images, detailed descriptions, and layout annotations for cross-modality perception and reasoning tasks. ACP comprises a raw data generator and a comprehensive data filter. The raw data generator redefines the current data synthesis paradigm by minimizing dependence on predefined annotation from original training data and instead using object combinations in natural language. Leveraging these combinations, we first use LLMs to sample fine-grained scene graphs, including object attributes, relations, captions, and spatial layouts. Next, controllable T2I models generate images based on these scene graphs. This pipeline enables scalable synthetic data generation while aligning with real data distributions through object combinations rooted in real-world data priors. It also addresses unbalanced distribution issues by adjusting the category proportion, especially in long-tailed scenarios. To ensure the quality of synthetic data, we introduce a comprehensive metric, **C**omposite **L**ayout and **I**mage **S**core (CLIS), in data filter. CLIS evaluates the reasonableness of generated layouts (CLIS-L) and the quality and alignment of generated images (CLIS-I). CLIS-L assesses the similarity between generated and ground truth layouts from data priors. Fig. 1(a,b) shows that high-quality layouts assessed by CLIS-L mirror real-world layouts and are more likely to yield high-quality images. CLIS-I evaluates visual quality and alignment with scene graphs, using priors from pre-trained large-scale models. Fig. 1(c, d) shows that images with high CLIS-I exhibit superior visual quality and strong alignment with their corresponding scene graph. The filtered scene graphs and corresponding images are used as training samples.

Through comprehensive evaluations, CLIS significantly enhances the performance of the state-of-the-state generation model, InstanceDiffusion, across various generation perspectives, including image fidelity, alignment to text, and layout control. Additionally, we observe a positive correlation between CLIS and performance gains in downstream tasks. Scaling up the training samples generated by ACP results in substantial performance gains in perception and reasoning tasks, particularly in long-tail and open-vocabulary scenarios. On the LVIS dataset [19], we observe a +6.0% improvement in AP_r^{mask} for the long-tail setting using Mask R-CNN [24] and a +1.3% gain in AP_{novel}^{box} for the open-vocabulary setting using Grounding DINO [50]. Additionally, we achieve a score of +80.1 on the MME benchmark and an accuracy improvement of +0.4 on the GQA benchmark with LLaVA-v1.5 [47], validating its efficiency in multi-modal perception and reasoning settings.

We summarize our technical contributions as follows:

- We propose ACP, an innovative training data generator for cross-modality perception and reasoning tasks. It is scalable and aligns with real data distributions.
- We design a comprehensive metric, CLIS, to filter generated data effectively by assessing layouts and images based on priors from real data or pre-trained large-scale models.
- Extensive experiments on visual and cross-modality perception and reasoning benchmarks demonstrate that ACP enhances model performance across various downstream tasks. The correlation between CLIS and performance gain among downstream tasks is well studied.

2. Related Work

Text to Image Generation. Diffusion-based approaches [54, 57, 58, 60] generate images as iterative denoising steps from random noise. Stable Diffusion [58] performs diffusion steps in the latent space of pre-trained autoencoders to achieve efficient training and sampling. Subsequent studies extend text-to-image diffusion models with layout controllability by introducing auxiliary input signals [40, 85] or spatial tokens [82] during training. Another line of works [4, 9, 11, 63, 79] follows a training-free approach by directly intervening the cross-attention layers during the sampling process. However, these controllable T2I methods rely on visual annotations during inference. With recent progress in the field of LLMs, LLM has been introduced in the T2I system to enhance text understanding and alignment [14, 16, 43, 56, 77, 81], further extending methods to condition on detailed text descriptions without the need for manually designing layouts. LMD [43] and LayoutGPT [14] use LLMs as text-guided layout generators through in-context learning. SLD [77] uses LLM to correct the misalignment between the generated images and the user prompt in an iterative closed-loop process. RPG [81] further adopts the recaptioning and planning of MLLM for subregion generation. Compared to previous works, we extend the generation paradigm to be conditioned on simple object combinations in natural language, broadening its applicability in scaling up synthetic data.

Learning from Synthetic Data. Deep learning models, especially for dense prediction tasks, typically require large amounts of data, which can be costly. Therefore, many works use synthetic data to approximate information gathered or measured in the real world [73, 76]. Synthesizing training samples with dense annotations is conditioned on various references. Some works [8, 49, 67] utilize the layout-to-image paradigm to synthesize training samples, conditioning on visual annotations like segmentation masks or bounding boxes. Some [42, 78] utilize an off-the-shelf perception model or adopt a perception head to obtain dense annotations of synthetic images generated based on detailed text descriptions. A series of works [17, 87] can condition on objects via copy-paste synthesis pipeline. However, the reasonableness of layouts is not considered. Among all these studies, no works explore a fully reasonable language-driven pipeline. To fill this gap, our method is driven purely by language and does not require expensive, manually annotated dense labels.

Generative Model Evaluation. Assessment of AI-generated content is challenging due to its subjective nature and the complexity of factors contributing to the generation quality. Metrics like Inception Score (IS) [61], Fréchet Inception Distance (FID) [28], and LPIPS [86] are commonly used for visual quality and diversity assessment. Some methods [7, 27] focus on the alignment between text and generated image. CLIPScore [27] computes the cosine similarity between text features and generated-image features extracted by CLIP. BLIP-CLIP [7] applies BLIP [39] to generate captions, then calculates the CLIP text-text cosine similarity between the generated captions and text prompts. For layout quality assessment, LayoutDM [33] proposes Maximum IoU [36] to measure the similarity between generated and real layouts. They compute an optimal match between generated and real layouts, maximizing the average IoU for corresponding categories. HRS [2] and T2I-CompBench [31] evaluate fixed spatial relationships, such as right, bottom, near, etc., of layouts using simple rules. Compared to these metrics, CLIS evaluates instance-level results in complex scenes and combines the reasonableness of layout and visual quality in one shot, making it a suitable metric to generate high-quality data for downstream tasks.

3. Method

Auto Cherry-Picker is a training-free cross-modality perception and reasoning dataset generation pipeline. It can produce pairs of images and scene graphs conditioned on object combinations while automatically selecting high-quality ones for training downstream models. We first introduce the preliminary in Sec. 3.1. Then, we detail our framework, including the raw data generator and the data filter in Sec. 3.2. Finally, we explain the deployment on various downstream tasks in Sec. 3.3.

3.1. Preliminary

Task Formation. Given a data pool $D_P = \{D_t\}_{t=1}^{T}$, where D_t denotes the training set for downstream task t, the objective is to generate a high-quality synthetic dataset $\mathcal{D}_t$ that, when combined with original D_t, enhances model performance on the corresponding downstream task t.

Data Priors. Data priors are defined as

$$P = \{p_i = ((s_i, o_i), r_i, L_i) | L_i = \{(l_s, l_o)_k^i\}_{k=1}^{M_i}\}_{i=1}^{N} \quad (1)$$

where s_i denotes the subject list, o_i denotes the object list, (s_i, o_i) denotes the object combination, r_i represents the relationship between them, L_i is a set of ground truth layouts corresponding to (s_i, o_i) and r_i. We use LLMs to extract layouts, corresponding categories, and relationships from open-source datasets D_P, resulting in data priors P.

Synthetic Data. A single item d_i in synthetic dataset $\mathcal{D}$ is defined as

$$d_i = \{(G_i, I_i) \mid G_i = (c_i, \{(\mathcal{S}_k, \mathcal{O}_k)\}_{k=1}^{K_i})\} \quad (2)$$

$$(\mathcal{S}_k, \mathcal{O}_k) = \{(s_k, o_k), (a_s, a_o)_k, (l_s, l_o)_k, r_k\} \quad (3)$$

where I_i denotes the image, and G_i denotes the scene graph. G_i includes an overall caption c_i and a set of detailed instance-level annotations $(\mathcal{S}_k, \mathcal{O}_k)$. Each pair includes original labels (s_k, o_k), attributes $(a_s, a_o)_k$, layouts $(l_s, l_o)_k$, and their relationship r_k.

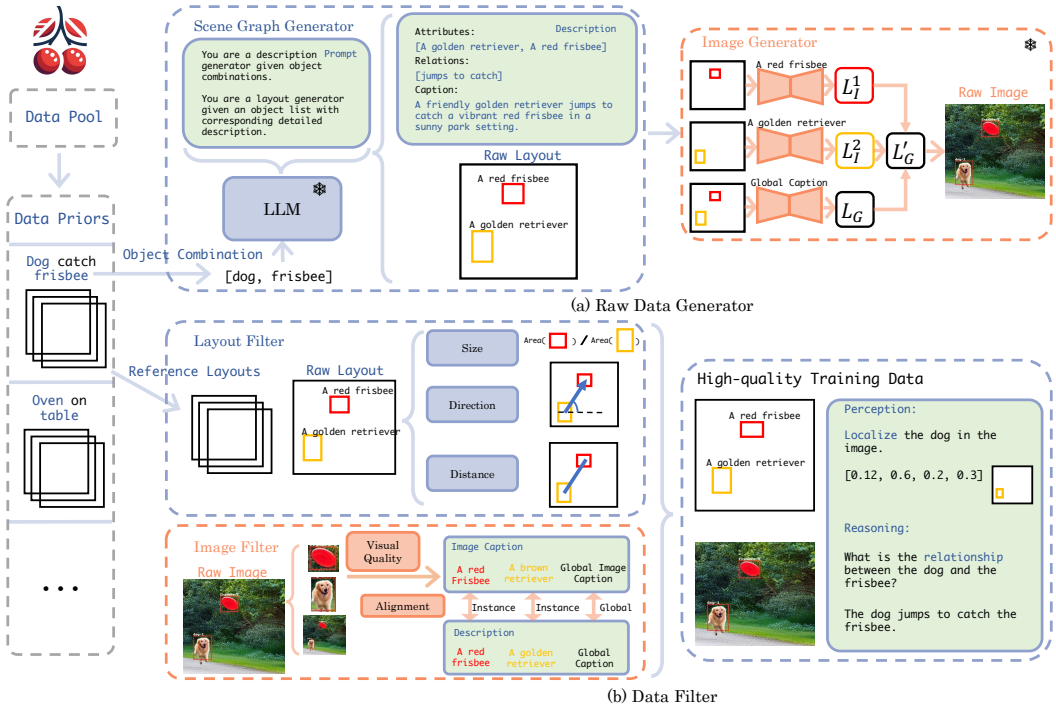

Figure 3. Illustration of Auto Cherry-Picker pipeline. It contains a (a) raw data generator and a (b) data filter using CLIS. Conditioned on input object combination sampled from data priors, Scene Graph Generator generates detailed attributes, relations, captions, and corresponding layouts. Subsequently, the Image Generator produces images based on the scene graph. These raw layouts and images are refined through filters using CLIS-L and CLIS-I, respectively, to produce high-quality training data.

3.2. ACP Framework

As depicted in Fig. 3, we aim to design a high-quality cross-modality training data generator comprising two key components: a raw data generator and a data filter. The former seeks to generate data, while the latter selects good ones.

Raw Data Generator. ACP generates data samples by harnessing information from data priors P to 1) closely align with real data distributions and 2) enable scalable generation. As shown in Fig. 3(a), to generate d_i, the generator first samples object combinations $\{(s_k, o_k)\}_{k=1}^{K_i}$ from P. Next, LLMs generate descriptions from these initial combinations, leveraging their in-context learning capability [6]. The description contains detailed attributes $\{(a_s, a_o)_k\}_{k=1}^{K_i}$, relationships $\{r_k\}_{k=1}^{K_i}$ between different objects, and an overall dense caption c_i. Then, we utilize the spatial reasoning ability of LLMs to plan layouts based on the relationships and descriptions of objects. The LLM involved in this process is referred to as the scene graph generator. Using the synthetic detailed description and spatial layouts, we adopt an off-the-shelf diffusion-based image generator to generate images by initiating the reverse diffusion process with different random noise.

Data Filter. For the raw generated data from stage one, a data filter is utilized to cherry-pick high-quality training data. As depicted in Fig. 3(b), the Layout Filter with CLIS-L and the Image Filter with CLIS-I assess the quality of layouts and image contents separately.

(a) Layout Filter. Precise relationships between objects in detailed image captions are essential for reducing hallucinations [84] and enhancing the reasoning abilities of MLLMs. Layout attributes provide important relational information, serving as valuable supplements for image-level evaluation. High-quality layouts are also crucial for improving the likelihood of image generators generating high-quality images. Since synthetic layouts lack ground truth, data priors are used to assess the reasonableness of layout pairs. Intuitively, we assume that a layout pair similar to the existing high-quality layout pairs is itself of high quality. CLIS-L evaluates this similarity from three perspectives: size, distance, and direction. Formally, given an object combination (s, o), their relationship r, and corresponding layout pair $\mathcal{L}_1 = (l_s, l_o)$, the reference layouts from data priors can be denoted as $L_P = P((s, o), r)$, CLIS-L is defined as:

$$\begin{aligned}\text{CLIS-L}(s, o, r, \mathcal{L}_1) = F_p(\{S_{size}(\mathcal{L}_1, \mathcal{L}_2) + \\ S_{dist}(\mathcal{L}_1, \mathcal{L}_2) + S_{dir}(\mathcal{L}_1, \mathcal{L}_2)\} | \mathcal{L}_2 \in L_P)\end{aligned} \tag{4}$$

Here, F_p refers to the operation of obtaining the p-th percentile to make CLIS-L robust against potential errors in P. S_{size} and S_{dist} are computed using the following similarity function S_{sim}:

$$S_{sim}(\mathcal{L}_1, \mathcal{L}_2, t) = 1 - \frac{|f_t(\mathcal{L}_1) - f_t(\mathcal{L}_2)|}{max(f_t(\mathcal{L}_1), f_t(\mathcal{L}_2))} \tag{5}$$

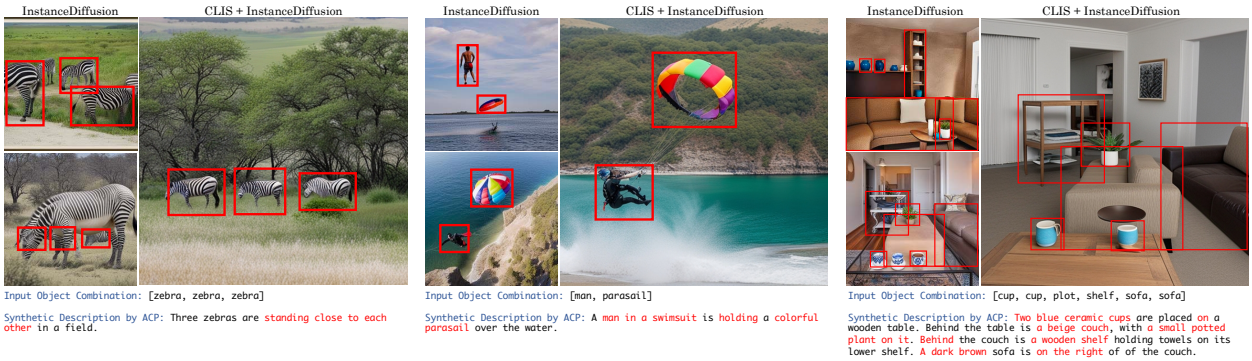

Figure 4. Comparison of generation results based on the same input object combinations and synthetic descriptions with and without CLIS. More generation results can be found in Appx. 13.

Model	FID↓	CLIP score↑	YOLO score↑
Stable Diffusion [58]	56.8	26.6	N/A
BoxDiff-SD [79]	60.0	26.4	4.4
BoxDiff-GLIGEN [79]	61.0	25.9	21.5
GLIGEN [40]	63.5	25.4	35.1
InstanceDiffusion [74]	53.5	25.2	45.6
GLIGEN w. CLIS	59.9 (-3.6)	25.8 (+0.4)	36.8 (+1.7)
InstanceDiffusion w. CLIS	48.9 (-4.6)	25.8 (+0.6)	47.9 (+2.3)

Table 1. Generation results of CLIS on the COCO val set.

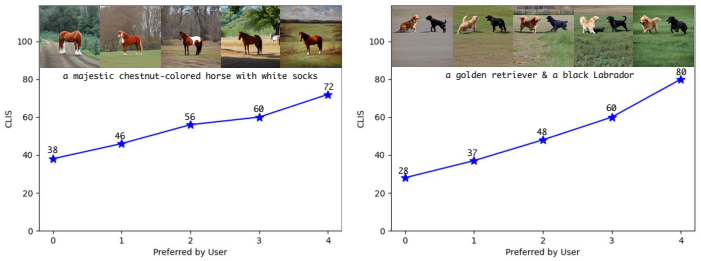

Figure 5. Consistent with human judgement. See details in Appx. 10.

where $t \in \{\text{Area}, \text{IoU}, \text{RD}\}$ denotes the attribute f_t focuses on. Specifically, f_{Area} computes the area ratio of layout pair, while f_{IoU} and f_{RD} compute the IoU and relative distance, respectively. Thus, we define S_{size} and S_{dist}:

$$S_{size}(\mathcal{L}_1, \mathcal{L}_2) = S_{sim}(\mathcal{L}_1, \mathcal{L}_2, \text{Area}) \quad (6)$$

$$S_{dist}(\mathcal{L}_1, \mathcal{L}_2) = S_{sim}(\mathcal{L}_1, \mathcal{L}_2, \text{IoU}) + S_{sim}(\mathcal{L}_1, \mathcal{L}_2, \text{RD}) \quad (7)$$

For S_{dir}, we compute the cosine similarity between direction vectors of $\mathcal{L}_1$ and $\mathcal{L}_2$:

$$S_{dir}(\mathcal{L}_1, \mathcal{L}_2) = cos(f_{\text{Dir}}(\mathcal{L}_1), f_{\text{Dir}}(\mathcal{L}_2)) \quad (8)$$

where f_{Dir} computes the direction vector between two layouts.

(b) Image Filter. To enable models with strong perception and complex reasoning abilities, we emphasize the visual quality of the image and its alignment with corresponding detailed descriptions. A pre-trained image captioning model generates global and local descriptions for synthetic images, respectively, where clear descriptions indicate high image quality and fidelity. Local descriptions are generated by guiding the model to focus on the corresponding layouts. An LLM then compares these descriptions with those in the generated scene graph to evaluate alignment. Thus, CLIS-I is formulated as:

$$\text{CLIS-I}(G, I) = F_{sim}(G'^T, G^T) = F_{sim}(F_C(I, G^L), G^T) \quad (9)$$

where G^T represents the textual part of the scene graph G (global caption and local objects attributes), G^L represents the layouts within G, I is the synthetic image corresponding to G, and F_C, F_{sim} denote the captioning and similarity functions, respectively. Notably, CLIS-I includes instance-level scores without additional calculations, enabling instance-level filtering. Please refer to Appx. 6 for details of foundation models and prompts utilized in ACP.

3.3. Deployment on Downstream Tasks

Cross-modality data generated by ACP can be readily transformed into training samples for various downstream tasks. In this work, we validate its generalizability in two primary settings: visual perception and multi-modal perception and reasoning.

Visual Perception Tasks. Layouts naturally serve as bounding boxes for detection tasks in synthetic training samples. To further deploy these samples for segmentation tasks, we adopt SAM [37] to obtain masks within the layout. Combining these segmentation masks with layout annotations and images prepares synthetic samples for visual perception tasks, especially in unbalanced scenarios such as long-tailed instance segmentation and open-vocabulary object detection.

Multi-modal Perception and Reasoning Tasks. For instruction fine-tuning of MLLMs, we construct question-answer pairs using a predefined question template. Details can be found in Appx. 7.1. The synthetic instruction data aims to enhance the perception and reasoning abilities of MLLMs. The question template includes prompts that require models to provide detailed descriptions of objects based on location or category, localize objects from descriptions, distinguish relationships between objects, and more.

CLIS Setting. CLIS selects high-quality samples through

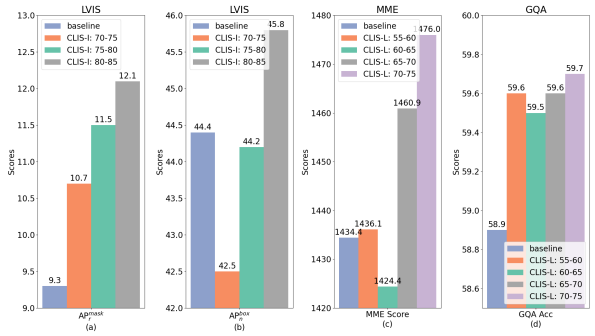

Figure 6. Correlation between CLIS and performance gains on downstream tasks. (a,b): Synthetic data with different ranges of CLIS-I on long-tailed instance segmentation and open-vocabulary detection scenarios of LVIS benchmarks using Mask R-CNN and Grounding-DINO as baseline, respectively. (c,d): Synthetic data with different ranges of CLIS-L on multi-modal perception and reasoning MME and GQA benchmarks based on LLaVA-v1.5.

two approaches: 1) it chooses the highest-scoring image and layouts from a set generated from the same object combinations, and 2) further to refine high-quality training samples for different downstream tasks, it applies independent score thresholds for layouts (CLIS-L), instances, and images (CLIS-I). Please see Appx. 7.2 for more details about CLIS computation, score distribution, and filtering ratio.

4. Experiments

We first introduce experiments set up in Sec. 4.1. We then validate CLIS in Sec. 4.2 from two perspectives: 1) its correlation with image fidelity of generated samples and 2) its correlation with the performance gain in downstream tasks. Next, we demonstrate ACP's effectiveness on several downstream tasks and demonstrate its potential for continuous scaling up of data size in Sec. 4.3. We validate the designed modules in ACP via a series of ablation studies in Sec. 4.4. More results are available in the supplementary material. All the source code and datasets will be available to the public.

4.1. Implementation Details

Datasets. We evaluate generation quality using COCO [45] following prior works [11, 12]. We randomly sample 1181 images from the COCO validation set, each paired with a fixed caption. For downstream tasks, we conduct object detection and instance segmentation experiments on COCO and LVIS v1.0 [19]. Additionally, we evaluate image-based visual question answering (VQA) using the MME [15] and GQA [32] benchmarks. The MME Perception benchmark is widely used to evaluate the perception abilities of MLLMs. GQA is a comprehensive dataset for assessing visual reasoning abilities.

Baselines. For generation models, we primarily select existing controllable diffusion-based T2I models, including GLIGEN [40], BoxDiff [79] and InstanceDiffusion [74],

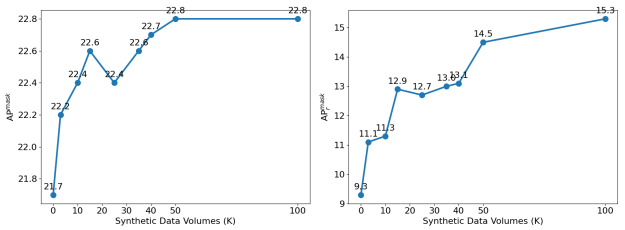

Figure 7. Data scaling on the LVIS benchmark using Mask R-CNN.

following their official settings. We also include Stable Diffusion [58] as a baseline T2I model. For downstream tasks, we adopt Mask R-CNN [24] and CenterNet2 [89] as baselines for long-tailed instance segmentation. For open-vocabulary object detection, we use Grounding-DINO [50, 88], and for VQA, we employ LLaVA-v1.5 [46, 47].

Evaluation Protocols. To assess image-level quality, we employ the Fréchet Inception Distance (FID) [28] with the COCO validation set as the reference dataset. We also use the CLIP score and YOLO score [41] to measure alignment and layout accuracy, respectively. For segmentation and object detection, we use Average Precision (AP) as the primary evaluation metric and report AP for novel and rare categories in open-vocabulary and long-tailed scenarios, respectively. For VQA benchmarks, we report the averaged score for MME and accuracy for GQA. Please refer to Appx. 8 for more details about the setup of our experiments.

4.2. Study Efficacy of CLIS

Generation Results. We first evaluate the efficacy of CLIS from a conventional generative perspective. Table 1 shows that CLIS enhances generation quality, as evidenced by a decrease in FID scores and increases in both CLIP and YOLO scores when applied to both GLIGEN [40] and InstanceDiffusion [74], demonstrating its effectiveness and generalizability. We also present visualization results in Fig. 4, comparing samples generated with and without CLIS, using the same input object combinations and synthetic descriptions. CLIS enhances layout quality in both action-based and complex multi-object scenarios. Furthermore, it ensures better visual quality and alignment between images and textual descriptions. We additionally evaluate its consistency with human preference in Appx. 10.

Correlation with Performance Gains on Downstream Tasks. We first compare CLIS with other widely used selection metrics. ACP generates 30K raw training samples for long-tailed instance segmentation. We then apply three different selection methods (random, CLIP, and CLIS) to pick the top 50% of synthetic data to train Mask R-CNN. Fig. 1(e) shows that CLIS achieves the highest score on LVIS, demonstrating its superior correlation compared to other metrics.

We further analyze the separate correlation of CLIS-I on

Table 2. Results on visual perception downstream tasks. (left): LVIS long-tailed instance segmentation benchmarks. (right): Open-vocabulary object detection benchmarks.

Method	Backbone	AP_r^{mask}	AP^{mask}
Mask R-CNN [24]	ResNet-50	9.3	21.7
w. ACP	ResNet-50	14.5 (+5.2)	22.8 (+1.1)
CenterNet2 w. Copy-Paste [18]	Swin-B	29.3	39.3
w. ACP	Swin-B	30.7 (+1.4)	39.6 (+0.3)

Dataset	Method	Backbone	AP_{novel}^{box}	AP^{box}
LVIS	Grounding-DINO	Swin-T	31.7	48.7
	w.ACP	Swin-T	33.0 (+1.3)	49.2
COCO	Grounding-DINO	Swin-T	60.4	57.1
	w.ACP	Swin-T	60.8 (+0.4)	56.9

Table 3. ACP boosting the results on the multi-modal MME and GQA benchmarks.

Method	LM Backbone	MME	GQA
LLaVA-1.5	Vicuna-7B	1434.4	58.9
LLaVA-1.5	Vicuna-13B	1438.3	60.7
LLaVA-1.5	LLama-3-8B	1445.3	60.1
LLaVA-1.5 w. ACP	Vicuna-7B	1514.5 (+80.1)	59.3 (+0.4)

Table 4. Compared with existing data generation methods on the LVIS benchmark. MosaicFusion defaults to using 4K synthetic images of rare categories. For a fair comparison, we extend it to all categories, denoted as MosaicFusion†. Additionally, we adjust the rare ratio in ACP by incorporating 2K rare synthetic images while keeping the total number unchanged, denoted as ACP‡.

Method	Backbone	AP_r^{mask}	AP^{mask}	Backbone	AP_r^{mask}	AP^{mask}
CenterNet2 (baseline)	ResNet-50	17.8	26.1	SwinB	27.3	34.1
w. X-Paste [87]	ResNet-50	17.9	<u>28.0</u>	SwinB	26.8	<u>35.1</u>
w. MosaicFusion [80]	ResNet-50	<u>19.6</u>	26.7	SwinB	<u>29.8</u>	34.3
w. MosaicFusion†	ResNet-50	18.6	27.0	SwinB	29.2	34.6
w.ACP	ResNet-50	19.2	<u>28.0</u>	SwinB	29.6	**35.2**
w.ACP‡	ResNet-50	**21.8**	**28.1**	SwinB	**30.6**	<u>35.1</u>

visual perception task performance and CLIS-L on multi-modal perception and reasoning tasks. We sample an equal number of synthetic samples across different CLIS-I and CLIS-L ranges to train baseline models, using 10K samples for CLIS-I and 1.3K for CLIS-L. Fig. 6 shows that higher CLIS-I correlates with improved performance in both long-tailed and open-vocabulary settings. Similarly, a positive correlation is observed between CLIS-L and performance on the MME and GQA benchmarks. These results validate the effectiveness of using CLIS in data filter across various downstream tasks.

4.3. Synthetic Dataset Scale Up

Data Scaling. To investigate the data scaling effects, we conduct experiments on LVIS using Mask R-CNN as the baseline, varying the size of synthetic data. Fig. 7 illustrates that ACP consistently boosts performance with increasing data volumes. Remarkably, ACP achieves its best results with the largest data size (100K), reaching an AP_r^{mask} of 15.3 (+6.0) and an AP^{mask} of 22.8 (+1.1), which shows potential for continuous scaling up. Interestingly, we find that as the volume of selected data grows, performance initially rises rapidly while plateaus after roughly 50K synthetic examples. To balance computation efficiency with performance gains, we choose 50K as our default synthetic data volume for subsequent experiments.

Long-tailed Instance Segmentation Benchmark. Table 2 (left) presents our results on the long-tailed instance segmentation settings of LIVS. ACP demonstrates significant performance gains over the commonly-used Mask R-CNN baseline, with an improvement of 1.1% in AP^{mask}, and the most notable improvement in rare categories (+5.2% AP_r^{mask}). We also observe consistent performance improvements with a stronger CenterNet2 baseline, which employs Swin-B as the backbone and copy-paste [18] for data augmentation. With this setup, ACP achieves a 1.4% higher AP_r^{mask}. This underscores ACP's strong generalization ability across different detector architectures and its effectiveness in conjunction with existing data augmentation methods.

Open-vocabulary Object Detection Benchmark. We further demonstrate the effectiveness of ACP in the challenging open-vocabulary detection setting. We use Grounding-DINO [50] as our baseline, which is pre-trained on large-scale data corpus, including Objects365 [62], GoldG [35], GRIT [55], and V3Det [72], totaling 61.8M images, following [88]. Table 2 (right) shows that ACP performs favorably against Grounding-DINO by 1.3% in LVIS AP_n^{box} and 0.4% in COCO AP_n^{box}, despite using a limited volume of generated training samples compared to the size of pre-train real data. This validates how high-quality synthetic data can complement real data effectively.

Multi-modal Benchmarks. We further evaluate the effectiveness of ACP on multi-modal perception and reasoning tasks. We adopt LLaVA-v1.5 with Vicuna-7B as our baseline. Table 3 indicates that ACP significantly enhances the model's perception ability on the MME benchmark, achieving an improvement of 80.1, which exceeds the performance of LLaVA-v1.5 even with stronger language model backbones such as Vicuna-13B and LLama-3-8B. Additionally, ACP improves performance on the widely recognized GQA reasoning benchmark. These results validate the effectiveness of our method in cross-modality settings.

Comparison with Previous Methods. We conduct a quantitative comparison with X-Paste [87] and MosaicFusion [80] using CenterNet2 with two backbones, ResNet-50 and Swin-B. Table 4 shows that ACP consistently achieves the highest AP^{mask} on both backbones and strikes a strong balance between AP^{mask} and AP_r^{mask}. When adjusting the rare ratio, ACP‡ can further enhance AP_r^{mask} significantly.

Table 5. Ablation study on Layout Generator. We compare our method to ACP with ground truth layouts on the LVIS using Mask R-CNN as our baseline.

Model	Layout Generator	AP_r^{mask} ↑	AP^{mask} ↑
Baseline	×	9.3	21.7
ACP	×	11.3	22.9
ACP	✓	12.2	23.0

Table 6. Ablation study on CLIS-I. We compare three variants of CLIS-I by filtering images generated from the same scene graph.

Model	F_C	F_{sim}	FID↓	AP_r^{mask}↑	AP^{mask}↑
Baseline	×	×	N/A	9.3	21.7
CLIP	×	✓	40.8	10.3	21.9
CLIP-text	✓	×	41.2	11.3	22.2
CLIS-I	✓	✓	41.3	11.5	22.3

In contrast, X-Paste yields high AP^{mask} but provides no additional gains on AP_r^{mask} compared to the baseline. Meanwhile, MosaicFusion boosts AP_r^{mask} but offers only modest AP^{mask} improvements, and its advantage in rare categories diminishes when extended to all categories. While ACP's AP^{mask} improvements over X-Paste are relatively modest, ACP achieves significant gains in AP_r^{mask}, enhancing the robustness and practicality of trained detectors. These results demonstrate the superiority of synthesizing images with reasonable layouts compared to composing training examples by pasting multiple synthesized instances onto a background. Efficiency analysis can be found in Appx. 11.

4.4. Ablation and Analysis

Layout Generator. We perform an ablation study on the layout generator by comparing with ACP utilizing ground truth layouts derived from training annotations. First, we conduct a qualitative comparison by showcasing images generated from both ground truth layouts and synthetic layouts in Fig. 8. The layout generator is capable of producing high-quality, reasonable layouts and helps mitigate issues such as overlap and small object sizes in ground truth layouts, which can lead to sub-optimal results in diffusion models. Next, we perform a quantitative comparison using 5K synthetic samples on LVIS with Mask R-CNN as our baseline. Table 5 shows that ACP with synthetic layouts achieves a +0.9% increase in AP_r^{mask} compared to ACP with ground truth layouts, demonstrating the effectiveness of generated layouts. We hypothesize that this improvement is due to the layout generator introducing data augmentation at the layout level rather than solely relying on the image generator's image-level augmentation while maintaining sufficient quality for the image generator. Further examples can be found in Appx. 13.1.

Components of CLIS-I. We conduct an ablation study on two operations in CLIS-I: F_{sim} and F_C in Eq 9. Specifically, we replace LLM alignment function F_{sim} with CLIP text alignment score and replace image caption function F_C with CLIP score to directly measure alignment between image and text. For the same set of scene graphs, the Image Generator produces 4 images for each scene graph, and these image filtering methods independently select the highest-scoring image. Each method generates the same volume of data (4K) and uses the same scene graph annotations. We evaluate these methods from both generation perspectives and performance gains in the downstream task. In particular, we use the LVIS benchmark and Mask R-CNN as the baseline detector. Table 6 shows that CLIS-I demonstrates the most significant performance gain in the downstream task, with a +2.2% increase in AP_r^{mask}, aligning with our initial motivation. Interestingly, while CLIP selection excels in generation evaluation, it yields suboptimal results in the downstream tasks. This highlights that CLIS-I is more strongly correlated with downstream task performance gains than conventional generation metrics.

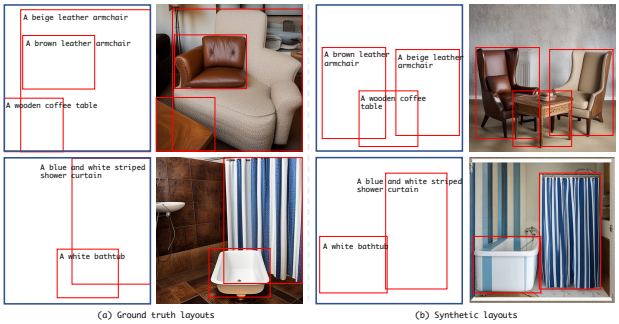

Figure 8. Visualization of the comparison between ground truth layouts and synthetic layouts, along with their corresponding synthetic images.

5. Conclusion

In this paper, we propose Auto Cherry-Picker, a cross-modality training data generator conditioned on object combinations with a comprehensively designed CLIS metric to ensure the quality of generated data. ACP is effective in various downstream tasks, including perception and reasoning tasks, particularly in improving the performance in annotation-scarce scenarios. CLIS can be used to pick high-quality generation samples, where we also find the generated data with higher CLIS can lead to better performance for perception tasks. Moreover, our method can be easily adapted to stronger LLMs and image generation models. Our research bridges the gap between generation data and downstream performance. We hope our results can inspire generation metric design in the future.

Limitations. Scaling the synthetic data size is resource-intensive. Currently, we only utilize a limited volume of synthetic data to ensure quality. Exploring methods that leverage low-quality data would be beneficial. We present more details on limitations and future work in the Appx. 9.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[2] Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *ICCV*, 2023. 3, 6

[3] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018. 5

[4] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023. 3

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 1

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 4

[7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *TOG*, 2023. 3

[8] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric con-trol for object detection data generation. In *ICLR*, 2024. 1, 3

[9] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *WACV*, 2023. 3

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 4

[11] G. Couairon, M. Careil, M. Cord, S. Lathuiliere, and J. Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *ICCV*, 2023. 3, 6

[12] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024. 6

[13] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 1

[14] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. In *NeurIPS*, 2024. 3

[15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2024. 6

[16] Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. Llm blueprint: Enabling text-to-image generation with complex and detailed prompts. In *ICLR*, 2024. 3

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

[18] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 7

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2, 6, 4

[20] Yue Han, Jiangning Zhang, Junwei Zhu, Xiangtai Li, Yanhao Ge, Wei Li, Chengjie Wang, Yong Liu, Xiaoming Liu, and Ying Tai. A generalist facex via learning unified facial representation. *arXiv preprint arXiv:2401.00551*, 2023. 1

[21] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face adapter for pre-trained diffusion models with fine-grained id and attribute control. In *ECCV*, 2024. 1

[22] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *NeurIPS*, 2022. 1

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 6, 7

[25] 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? *ICLR*, 2023. 2

[26] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[27] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 1, 3

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 3, 6

[29] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1

[30] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 1

[31] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023. 2, 3, 6

[32] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 6

[33] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *CVPR*, 2023. 2, 3

[34] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[35] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 7, 4

[36] Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Constrained graphic layout generation via latent optimization. In *ACM MM*, 2021. 3

[37] 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 *ICCV*, 2023. 5

[38] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Dreamteacher: Pretraining image backbones with deep generative models. In *ICCV*, 2023. 1

[39] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[40] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 3, 5, 6

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

[42] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, 2023. 1, 3

[43] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *TMLR*, 2024. 2, 3

[44] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *CVPR*, 2023. 1

[45] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 6, 4, 5

[47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2024. 2, 6

[48] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, 2020. 4

[49] Shuangting Liu, Jiaqi Zhang, Yuxin Chen, Yifan Liu, Zengchang Qin, and Tao Wan. Pixel level data augmentation for semantic image segmentation using generative adversarial networks. In *ICASSP*, 2019. 2, 3

[50] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 2, 6, 7

[51] 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 *ICCV*, 2021. 4

[52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[53] 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. 1

[54] 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. 3

[55] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *ICLR*, 2024. 7, 4

[56] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *ACMMM*, 2023. 3

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 5, 6

[59] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, 2022. 1

[60] 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, 2022. 1, 3

[61] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 3

[62] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 7, 4

[63] Jaskirat Singh, Stephen Gould, and Liang Zheng. High-fidelity guided image synthesis with latent diffusion models. In *CVPR*, 2023. 3

[64] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. In *NeurIPS*, 2021. 1

[65] Saksham Suri, Fanyi Xiao, Animesh Sinha, Sean Chang Culatana, Raghuraman Krishnamoorthi, Chenchen Zhu, and Abhinav Shrivastava. Gen2det: Generate to detect. *arXiv preprint arXiv:2312.04566*, 2023. 2

[66] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 5

[67] Weimin Tan, Siyuan Chen, and Bo Yan. Diffss: Diffusion model for few-shot semantic segmentation. *arXiv preprint arXiv:2307.00773*, 2023. 2, 3

[68] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *CVPR*, 2024. 2

[69] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In *NeurIPS*, 2024. 1, 2

[70] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[71] Chaoyang Wang, Xiangtai Li, Lu Qi, Henghui Ding, Yunhai Tong, and Ming-Hsuan Yang. Semflow: Binding semantic segmentation and image synthesis via rectified flow. In *NIPS*, 2024. 1

[72] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *ICCV*, 2023. 7, 4

[73] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *CVPR*, 2019. 3

[74] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *CVPR*, 2024. 2, 5, 6

[75] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 4

[76] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *ICCV*, 2021. 3

[77] Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *CVPR*, 2024. 3

[78] 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. *NeurIPS*, 2023. 3

[79] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, 2023. 3, 5, 6

[80] 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. In *IJCV*, 2024. 1, 7

[81] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *ICML*, 2024. 3

[82] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Reco: Region-controlled text-to-image generation. In *CVPR*, 2023. 3

[83] Hanrong Ye, Jason Kuen, Qing Liu, Zhe Lin, Brian Price, and Dan Xu. Seggen: Supercharging segmentation models with text2mask and mask2img synthesis. *arXiv preprint arXiv:2311.03355*, 2023. 2

[84] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. Halle-control: Controlling object hallucination in large multimodal models. *arXiv preprint arXiv:2310.01779*, 2024. 4

[85] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[86] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

[87] 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 *ICML*, 2023. 3, 7, 4

[88] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 6, 7, 4

[89] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 6