# RareCrafter: Controllable Generative Augmentation for Rare Object Detection in Driving Scenes

Anonymous CVPR submission

Paper ID *****

## Abstract

*Reliable perception of roadway environments is essential for intelligent transportation systems (ITS) and their digital twin frameworks, where object detectors provide a key interface between the physical world and its digital representation. However, existing driving datasets are dominated by common traffic participants, leaving rare but safety-critical objects underrepresented and limiting robustness to uncommon roadway events. We propose RareCrafter, a training-free generative augmentation framework that synthesizes rare objects directly into real driving scenes using controllable generative models. The framework combines structured prompt diversification, depth-conditioned empirical size priors, and region-conditioned verification to ensure that inserted objects remain semantically consistent and maintain realistic, depth-dependent object scales. The generated images are then used to augment training data for object detection. Experiments on the CODA2022 corner-case dataset show that RareCrafter improves rare-object detection by 2.82 mAP on average over real-only training and consistently outperforms a CopyPaste augmentation baseline. These results demonstrate that controllable generative augmentation can effectively mitigate data scarcity for long-tail objects in driving scenes, reducing the need for additional real-world data collection.*

## 1. Introduction

Reliable roadway perception underpins modern transportation systems including intelligent infrastructure, traffic monitoring, and autonomous driving. Increasingly, these systems are integrated into digital representations of the transportation network, often referred to as digital twins, where real-world traffic states are mirrored to support monitoring, planning, and automated decision-making. [11] refers to the perception layer as the ”eye” of a digital twin. Thus, in these environments, perception models form one of the important interfaces between the physical roadway and its digital representation.

The fidelity of digital twin representations depends heavily on the reliability of perception systems that provide observations of the physical environment. Object detection models identify roadway actors and conditions that form the basis for higher-level monitoring and decision-support processes. These downstream components, often described as the ”brain” of a digital twin [11], extract patterns, model system dynamics, and support operational decisions based on the perceived state of the roadway. When perception models fail to detect rare or unexpected objects, the resulting digital representation may omit critical elements of the environment, leading to incomplete situational awareness and potentially reducing the reliability of monitoring, prediction, and decision-support functions built upon that representation.

Contemporary object detection models for street-level scenes are commonly trained and benchmarked on large-scale datasets such as NuScenes [7] and UA-DETRAC [41]. While these datasets have significantly advanced driving-scene research, they primarily represent frequent traffic participants and common roadway conditions, offering limited coverage of rare object categories. As a result, models trained on such datasets often struggle with long-tail events such as wildlife entering the roadway, stalled vehicles, temporary work-zone artifacts, or unexpected obstacles.

Existing approaches to mitigating long-tail imbalance generally fall into three categories: selective data curation, conventional data augmentation, and simulation. Manual curation of rare instances from existing datasets is inherently limited by their sparse occurrence. Traditional augmentation techniques such as copy-paste insertion can increase class frequency but remain constrained to the limited pool of available object instances and may introduce lighting and shadow mismatches or truncated instances inherited from occlusions in the source images. Collecting additional real-world data from sources such as dashcam footage or staged deployments incurs significant logistical and annotation costs while still failing to guarantee coverage of rare events. Simulation platforms such as CARLA provide con-

trolled environments for generating annotated data, but the diversity of scenes and objects depends on available assets and modeling effort, and models trained heavily on simulated imagery often require additional adaptation due to the sim-to-real gap.

Recently, diffusion-based generative models have emerged as powerful tools for producing diverse and high-fidelity images. These models have been successfully applied to tasks such as image editing [5, 14] and inpainting [34, 40], and are increasingly explored for synthetic data generation and augmentation [10]. Unlike traditional augmentation methods that rely on simple transformations such as rotation, flipping, or brightness adjustments, diffusion models can generate entirely new visual instances that extend beyond the configurations present in the original dataset. By learning a generative representation of the data distribution, they can synthesize realistic samples that populate underrepresented regions while maintaining semantic coherence. Furthermore, modern diffusion frameworks support fine-grained control over content and style, allowing augmentation strategies to be tailored to specific failure modes (e.g., rare-object insertion, lighting variations, or domain-shift simulation). These properties collectively make diffusion-based augmentation more expressive, flexible, and aligned with the needs of modern deep vision models than conventional transformation-based techniques.

Despite this potential, the use of diffusion models for safety-critical roadway perception remains limited. Prior work has largely focused on scene stylization or weather adaptation [22], or limited forms of object insertion without fine-grained control over the generated instances [47]. Modern diffusion frameworks support conditioning mechanisms such as text prompts and spatial masks, which enable detailed control over object category, pose, viewpoint, and spatial placement. Such controllability enables the systematic generation of rare traffic objects under diverse contextual configurations, providing a promising mechanism for enriching long-tail object distributions.

To address this gap, we investigate whether controllable generative models can be leveraged to enrich rare-object distributions in driving scenes. Our framework introduces rare traffic objects into real roadway imagery through controlled generative augmentation, enabling the creation of realistic corner cases without costly real-world data collection or manual staging. The proposed approach integrates generative synthesis with verification and filtering mechanisms to ensure that inserted objects remain semantically consistent and spatially plausible for detector training.

Specifically, we propose *RareCrafter*, a training-free generative augmentation framework that synthesizes rare objects directly into real driving scenes using conditional generative models and structured guidance. By enriching training data with rare corner cases, the framework aims to improve the robustness of perception systems when encountering uncommon but foreseeable roadway objects.

The main question explored in this work is: *Can generative models mitigate the costly process of collecting rare-object data in driving scenes?* We empirically evaluate whether generative augmentation can improve object detection performance while reducing reliance on extensive data collection and manual annotation. Our main contributions are:

- *RareCrafter*, a controllable generative augmentation framework for synthesizing rare objects in real driving scene images.
- A structured synthesis pipeline with depth-aware placement, structured prompt diversification, and region-conditioned verification for consistent object insertion.
- Augmenting training data with *RareCrafter*, which improves rare-object detection on a real-world corner-case dataset while preserving performance on common categories.

## 2. Related Works

**Conditional image generation.** Diffusion models have emerged as a powerful class of generative models, achieving state-of-the-art results in high-fidelity image synthesis [15, 28]. They work by training a model to reverse a fixed, gradual process of adding noise; allowing them to start from pure Gaussian noise and iteratively denoise it to a coherent, high-quality image. Unlike unconditional generative models, conditional diffusion models provide a fine-grained control over visual content. Specifically, ControlNet [48] enhances controllability by injecting spatial conditions, such as edge maps or poses, into the diffusion backbone. In addition to diffusion-based architectures, recent hybrid transformer–flow models such as FLUX [19] have demonstrated impressive capabilities in aligning text prompts with high-fidelity visual outputs by integrating large-scale attention modules with rectified flow refinement steps. This design enables Flux to produce photorealistic images with strong semantic consistency and robust controllability. Object-oriented editing leverages conditional generative models to remove [16, 21, 36, 45] an object specified by a mask or synthesize [9, 35, 37, 45] a new object with control over object attributes and appearances through text prompts or reference images. In this work, we take advantage of FLUX-ControlNet-Inpainting [36] to generate rare objects conditioned on a driving background scene and structured prompts.

**Corner case and anomaly driving datasets.** Capturing rare and unexpected events in driving scenes remains challenging due to their infrequent occurrence in real-world data. To facilitate their study, several datasets have been introduced that focus on anomalous objects, unusual traffic situations, and other corner cases in driving scenes.

Table 1. Comparison of methods.

| Method | Training Free | Rare Object Insertion | Fine-Grained Control |
|---|---|---|---|
| RareDiffusion [47] | ✗ | ✓ | ✗ |
| LTDA-Drive [46] | ✓ | ✗ | ✓ |
| RareCrafter (ours) | ✓ | ✓ | ✓ |

One line of research mines or re-annotates existing datasets to highlight anomalous events and corner-case objects [13, 20, 42]. While this improves visibility of rare instances, it remains inherently constrained by the original data distribution. Another direction focuses on curated real-world collections or staged scenarios [25, 27, 30], where unusual objects are deliberately placed in traffic scenes. Although these efforts provide valuable benchmarks, they require substantial manual effort and offer limited diversity of naturally occurring rare-event combinations. Some anomaly-centric datasets [8] further rely on binary anomaly labels without detailed semantic categorization, limiting their usefulness for object-level perception tasks. Alternative approaches attempt to increase rare-event coverage synthetically. Image-level blending methods overlay anomalous objects onto real scenes [3], but are restricted by the available object pool of the source dataset from which the pasted instances are drawn, and often introduce geometric inconsistencies. Simulation-based datasets [4, 6, 13, 17, 25] enable scalable scenario generation under controlled conditions; however, their diversity depends on predefined assets and scene configurations, and models trained primarily on simulated data may struggle to generalize to real-world imagery.

**Diffusion models for image data augmentation.** Diffusion models have emerged as powerful generative tools for producing diverse, high-quality, and semantically coherent augmented data. Traditional augmentation methods expand datasets through simple image transformations such as translation, flipping, cropping, or color jittering, which preserve semantic content while modifying appearance [44]. However, these operations provide limited diversity and are often insufficient for modern vision tasks [1]. Generative augmentation with diffusion models instead samples from an approximate data distribution, enabling the creation of entirely new images or controlled modifications of existing ones. Through conditioning signals such as text prompts, depth maps, bounding boxes, or segmentation masks, these models can generate synthetic object-detection training images [10], generate counterfactual examples for bias mitigation [29], or improve robustness under out-of-distribution shifts [39].

Recent generative models such as FLUX [19] adopt flow-matching as an alternative training paradigm. By learning probability paths grounded in optimal transport, flow-based models enable faster sampling than diffusion models [23], though often at the cost of reduced sample diversity [33]. While typically viewed as a limitation, prior work [2] shows that lower diversity can benefit training-free image editing. Building on this insight, we exploit the structure-preserving behavior of flow-matching models to insert rare objects into driving scenes while minimizing unintended background modifications, enabling targeted augmentation for long-tail object detection.

Existing diffusion-based augmentation methods remain limited for rare-object learning (Table 1). RareDiffusion [47] trains a dedicated diffusion model for a predefined set of categories, restricting scalability to arbitrary classes and offering limited control over viewpoint or structural variations, thereby reducing its effectiveness for targeted augmentation. LTDA-Drive [46] adopts a training-free, text-conditioned inpainting framework; however, it focuses on object categories with abundant training samples, making it less transferable to rare objects with limited representation in existing datasets. Additionally, neither method has publicly released code. In contrast, RareCrafter is training-free, extensible to arbitrary rare categories via pretrained generative models, and enables fine-grained control over object attributes, making it suitable for targeted long-tail data augmentation.

# 3. Methodology

## 3.1. Rare Object Generation

We synthesize rare-object instances using FLUX-ControlNet-Inpainting [36] applied to background images taken from the ONCE dataset [26]. The inpainter receives: (i) the background frame; (ii) a mask defining the insertion region; and, (iii) a text prompt describing the target class. Objects are inserted within predefined bounding boxes to preserve geometric consistency with the scene.

**Structured Prompt Diversification.** A central component of our synthesis pipeline is controlled prompt diversification. Rather than using a single generic description (e.g., "a stroller" or "a motorcycle"), we leverage a large language model to generate a structured set of object-centric prompts. We exploit the LLM's latent knowledge of object taxonomy, materials, structural components, configurations, and viewpoint variations to enumerate semantically meaningful intra-class variations.

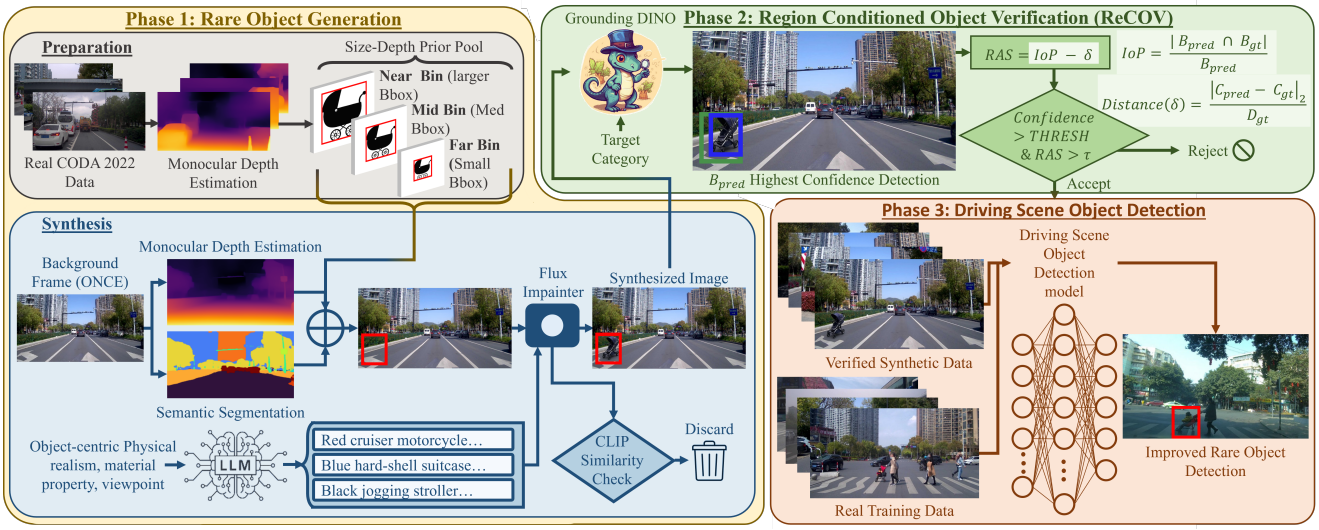


Figure 1. **Overview of the proposed RareCrafter method.** Starting from a real driving-scene image, monocular depth and semantic segmentation are estimated to identify valid ground regions for object insertion. The depth value at the sampled point determines a near, mid, or far bin, which is used to select a bounding-box template from a depth-conditioned empirical pool constructed from real data. Object appearance diversity is introduced through structured prompt diversification, where an LLM generates object-centric prompts describing variations in structure, materials, and configurations, which are used with a FLUX-ControlNet inpainting model to synthesize the rare object at the selected location. Generated samples are filtered using CLIP similarity and further validated through Region-Conditioned Object Verification (ReCOV) with GroundingDINO to ensure the object appears within the intended region. Verified synthetic images are then combined with real training data to train the driving-scene object detector.

Prompts are constrained to describe exactly one object and to emphasize physical realism, viewpoint, material properties, and structural attributes, while deliberately excluding environmental factors such as weather, lighting direction, or scene context. This decouples object specification from background conditions, allowing the inpainting model to adapt seamlessly to the underlying background.

This strategy is particularly beneficial for rare categories, where limited real data can lead detectors to overfit to narrow visual prototypes. By explicitly varying structured object attributes, we approximate a broader intra-class distribution than is available in the training set, mitigating representation sparsity. For example, stroller prompts vary by functional subtype (compact, jogging, double), frame material, fabric type, and structural configuration. Motorcycle prompts vary by vehicle style (sport, cruiser, off-road), paint finish, exposed versus enclosed components, accessories, and usage condition. The system prompt used for prompt construction is included in Supplementary Fig.8.

**Depth-Conditioned Empirical Size Prior.** A critical component of the synthesis pipeline is the construction of insertion masks that are geometrically plausible while remaining stable for generative modeling and detector supervision. Prior work such as LTDA-Drive [46] models object scale by fitting a parametric (e.g., Gaussian) distribution over bounding-box statistics. While effective for frequent categories with abundant samples, such parametric fitting becomes unreliable for rare classes where only a limited number of real instances are available. In this low-data regime, distribution fitting can produce unrealistic aspect ratios and distorted scale configurations (e.g., unusually tall–thin stroller boxes resembling pedestrians), which deviate from the empirical geometry of the category and provide misleading spatial cues to the inpainter.

To avoid distributional artifacts introduced by parametric modeling, we directly leverage the empirical bounding boxes observed in real data. Using the CODA2022 training set, we associate each rare-object instance (suitcase, stroller, motorcycle) with its predicted monocular depth and its normalized bounding-box dimensions. Depth values are discretized into coarse bins (near, mid, far). For each class and depth bin, we construct a non-parametric pool consisting of the observed bounding-box dimensions from real instances. This produces a depth-conditioned empirical prior that preserves realistic aspect ratios and category-specific scale characteristics without requiring distribution fitting.

At synthesis time, candidate insertion locations are sampled from semantically valid regions (e.g., road or sidewalk areas without existing annotations). The depth at the sampled pixel determines the corresponding depth bin, and a bounding-box template is drawn from that bin's empirical pool. This preserves the depth-dependent scale hierarchy observed in real driving scenes.

However, extremely small bounding boxes were em-

pirically found to destabilize diffusion-based inpainting (see Tab. 3): masks with limited spatial support tend to encourage background reconstruction or prompt under-conditioning. To address this, the sampled bounding box is uniformly expanded by a fixed margin before being used both as the inpainting mask and as the detector ground-truth box. This dilation is applied consistently across all depth bins and categories, preserving the relative depth-dependent scale ordering: boxes drawn from the far bin remain smaller than those from the mid and near bins.

Consequently, the final insertion masks maintain semantic placement consistency and remain anchored to the empirical depth-conditioned size prior, while accommodating the practical stability requirements of generative synthesis.

**Semantic Filtering.** To enforce category consistency, we compute the CLIP similarity between the generated region and the target class name, discarding samples below a predefined threshold. Final acceptance additionally requires passing the region-conditioned verification described in the next section.

### 3.2. Region-Conditioned Object Verification

Initial experiments revealed a failure mode in which the generative model reconstructs the masked region with contextually plausible background content while failing to render the intended object. Such cases may achieve high global semantic similarity (e.g., CLIP score) despite unsuccessful object insertion, motivating an explicit region-level verification procedure.

We therefore introduce a region-conditioned validation stage based on open-vocabulary detection using GroundingDINO [24]. For each generated image, GroundingDINO is prompted with the target category and applied to the full image. Let $B_{\text{gt}}$ denote the inpainting mask (ground-truth insertion region) and $B_{\text{pred}}$ the highest-confidence detection for the target category exceeding a predefined confidence threshold. If no such detection exists, the insertion is rejected.

Because $B_{\text{gt}}$ defines a placement region rather than a tight object box, standard IoU is not suitable. Instead, we measure spatial containment using Intersection-over-Prediction (IoP):

$$\text{IoP}(B_{\text{pred}}, B_{\text{gt}}) = \frac{|B_{\text{pred}} \cap B_{\text{gt}}|}{|B_{\text{pred}}|}$$

IoP quantifies the proportion of the detected object that lies inside the intended insertion region.

To additionally enforce geometric alignment, we introduce a normalized center-distance term. Let $c_{\text{gt}}$ and $c_{\text{pred}}$ denote the box centers, and let $D$ be the diagonal length of the minimal enclosing box covering both $B_{\text{gt}}$ and $B_{\text{pred}}$. We define

$$\delta = \frac{\|c_{\text{pred}} - c_{\text{gt}}\|_2}{D}$$

This term penalizes detections whose centers deviate from the intended insertion location. Center alignment is particularly relevant for our downstream detector, which employs center-based supervision.

We combine containment and alignment into a *Region Alignment Score* (RAS):

$$\text{RAS} = \text{IoP}(B_{\text{pred}}, B_{\text{gt}}) - \delta$$

A generated image is considered successfully verified if (i) the detection confidence exceeds a fixed threshold and (ii) RAS $\geq \tau$, where $\tau$ is a predefined alignment threshold. This region-conditioned verification filters out context-only reconstructions and ensures that the synthesized object is both spatially contained and center-aligned within the designated insertion region.

### 3.3. Driving Scene Object Detection

We use FCOS [38] as the object detection model in all experiments. FCOS is a one-stage, anchor-free architecture that avoids proposal generation and predefined anchor boxes, making it well-suited to driving scenes with substantial scale variation across categories, including rare and small objects.

Our goal is not to optimize detector performance but to measure the effect of generative data augmentation under a controlled setup. Therefore, the detector architecture, training protocol, and hyperparameters remain fixed across all experiments to ensure fair comparison between real-only and augmented training regimes. Evaluation is conducted within the same driving-scene domain as training; we avoid cross-domain settings (e.g., indoor or artistic imagery) so that performance differences can be attributed to augmentation rather than domain shift.

## 4. Experiments

### 4.1. Datasets and Implementation Details

**Datasets**. We use ONCE [26] and CODA [20]. ONCE is a large-scale autonomous driving dataset containing 15k fully annotated scenes across diverse environments, weather conditions, and day/night settings. Unless stated otherwise, we use images from camera 3 only. In our framework, ONCE serves solely as the background source for synthetic rare-object insertion.

CODA is a real-world corner-case dataset with bounding box annotations for 43 object categories. We use the CODA2022 subset, which contains 80,180 annotated objects and is divided into a training set and a test set, each comprising 4,884 images. CODA2022 includes both common driving classes (e.g., car, bus, truck, pedestrian, cyclist) and rare categories.

To prevent data leakage between synthetic training data and evaluation data, we enforce a strict dataset separation.

Synthetic images are generated using ONCE backgrounds that do not overlap with CODA2022, with any shared images removed prior to synthesis. Real rare-object training samples are taken exclusively from the CODA2022 training split, while all evaluations are performed on the CODA2022 test split.

Following CODA's definition of corner cases, which includes (i) novel classes and (ii) novel instances of common classes, our study focuses exclusively on the first type, novel (rare) classes. Accordingly, only rare categories are synthesized, and evaluation is performed on real images containing only common and mentioned rare categories.

For the CopyPaste baseline, object instances are sourced from the LVIS dataset [12], specifically from the *suitcase*, *baby buggy*, and *motorcycle* categories. Since LVIS masks can be noisy, we apply additional mask refinement and filtering before insertion. Dataset splits and their usage are summarized in Tab.5 of Supplementary.

**Implementation Details.** Without loss of generality, we consider three object categories, *suitcase*, *stroller*, and *motorcycle*, to build our framework upon. We selected the FCOS [38] object detection model [38] and trained it with the Adam optimizer [18] for 20 epochs using a base learning rate of $1.6 \times 10^{-5}$ with cosine annealing and a batch size of 16. Images are processed at their original resolution. Data augmentation during epochs 1–18 includes random resizing, cropping, color jitter, and horizontal flipping; the final two epochs use only resizing and horizontal flipping.

For object insertion, we adopt a zoom-in strategy: the target placement box is expanded to include surrounding context, cropped, and resized to $1024 \times 1024$ before being passed to the generative model. This enables consistent inpainting across varying image aspect ratios and object scales. We use the FLUX model [19] with 28 inference steps. Each image is generated up to ten times with different seeds and filtered using $\tau$ of 0.6 and a CLIP Score threshold; images failing the threshold are discarded. Thresholds, determined on the CODA2022 training split, are 0.22 (motorcycle), 0.20 (suitcase), and 0.25 (stroller).

Monocular depth estimation is performed using DPT-Large [32], and ground-region localization using SegFormer [43]. Detection performance is evaluated using the COCO protocol, reporting mAP and mAR averaged over IoU thresholds from 0.50 to 0.95. All experiments are conducted on a single NVIDIA A100 GPU.

## 4.2. Evaluation on Object Detection

In this section, we evaluate the effectiveness of the proposed RareCrafter framework. For each rare category, we generate synthetic data with twice as many images as the real training set and augment it with the original dataset. Fig. 2 presents qualitative results, while Tab. 2 reports quantitative performance.

Table 2. Comparison of object detection performance of FCOS [38] across training configurations for rare categories.

| | Stroller | | Motorcycle | | Suitcase | |
|---|---|---|---|---|---|---|
| **Regime** | mAP | mAR | mAP | mAR | mAP | mAR |
| Real only | 3.66 | 22.60 | 0.13 | 30.50 | 0.60 | 14.60 |
| CopyPaste | 4.26 | **35.03** | 2.16 | 30.26 | 0.63 | 10.76 |
| RareCrafter (ours) | **5.76** | 35.00 | **2.53** | **31.43** | **4.56** | **15.13** |
| RareCrafter vs Real ($\Delta$) | +2.10 | +12.40 | +2.40 | +0.93 | +3.96 | +0.53 |

We compare RareCrafter with a CopyPaste baseline, which inserts segmented instances of the target categories into predefined bounding boxes using alpha blending. In practice, raw instance segmentation masks often contain artifacts such as holes or fragmented regions and may produce objects with unrealistic scales relative to the target bounding box. To mitigate these issues, we refine the masks by filling holes and removing discontinuities, and apply an instance filtering step to discard implausible samples (e.g., extremely small objects or instances requiring excessive rescaling). The remaining instances are resized while preserving their aspect ratio and constrained to fit within the target bounding boxes before compositing them into the scene.

Tab. 2 compares different training regimes using FCOS for rare-category detection. Training on real data only results in very limited performance. Incorporating synthetic data through RareCrafter substantially improves detection performance across all categories. Specifically, RareCrafter increases stroller performance by +2.10 mAP and +12.40 mAR, improves motorcycle mAP by +2.40, and boosts suitcase mAP by +3.96. Compared to the CopyPaste augmentation baseline, RareCrafter consistently achieves higher mAP across all categories. These results demonstrate that RareCrafter effectively mitigates the data scarcity problem for rare objects and significantly improves detection performance. Importantly, performance on common categories remains stable across training regimes (Tab.6 of Supplementary), with an average maximum variation of approximately 1.0 mAP. Qualitative examples of the synthesized training images generated by our RareCrafter framework and the CopyPaste baseline are illustrated in Supplementary Fig.5 and Fig.6, respectively.

## 4.3. Ablation Study

**Ablation on bounding box dilation.** We ablate our mask dilation strategy and perform inpainting directly within bounding boxes sampled from the empirical distribution of the CODA2022 training set. This setting constrains the inpainter to insert rare objects within the original bounding box extents observed in real data. In both settings, the synthesis budget is fixed (100 background images per rare category), and the same empirical bounding-box pool is used; the only difference is whether the sampled box is expanded


Figure 2. **Qualitative object detection results on the CODA2022 dataset.** The detection models are trained under three regimes: real-only data, real data augmented with CopyPaste, and real data augmented with RareCrafter. In the first row, the detector trained on real-only data incorrectly localizes another object as a stroller. In the second row, both the real-only and CopyPaste models fail to detect the motorcycle. In the third row, the real-only model confuses a pedestrian carrying a suitcase with the suitcase itself, while CopyPaste incorrectly localizes another object as a suitcase. In contrast, RareCrafter correctly detects the suitcase without misclassifying the pedestrian.

before inpainting.

We report two synthesis quality metrics. The *CLIP pass rate* measures the proportion of generated samples whose masked region exceeds the CLIP similarity threshold with the target class name. The *ReCOV rate* measures the proportion of CLIP-passed samples that additionally pass the region-conditioned verification described in Sec. 3.2, meaning that GroundingDINO detects the intended object within the designated insertion region with sufficient alignment.

As shown in Tab. 3, removing mask dilation substantially degrades both metrics. Without dilation, the CLIP pass rate drops from 98% to 71%, and the ReCOV rate drops from 97% to 58%. Qualitatively, small insertion regions frequently cause the inpainter to reconstruct plausible background instead of inserting the requested object. Such cases may still achieve moderate semantic similarity, which means that a CLIP-only filter would retain samples where the labeled bounding box contains only background rather than the intended object. This introduces label noise into the synthetic dataset, where the annotated bounding box sometimes contains the intended object and sometimes contains only background.

Applying mask dilation provides the generative model with sufficient spatial support to synthesize the object, reducing prompt-ignoring behavior and improving both semantic consistency and spatial verification. As a result, the CLIP and ReCOV rates both rise, indicating that most samples satisfying the semantic filter also contain a verifiable object instance within the intended region. These results confirm that enlarging the inpainting region is beneficial for stable object synthesis and for preventing noisy synthetic labels.

Table 3. Ablation on mask dilation for the inpainting region. Enlarging the bounding box significantly improves both semantic consistency (CLIP pass rate) and successful object insertion verified by ReCOV.

| Design Choice | CLIP Pass Rate (%) ↑ | ReCOV Rate (%) ↑ |
|---|---|---|
| Without dilation | 71 | 58 |
| With dilation | **98** | **97** |

**Ablation on prompt diversification.** We ablate our prompt generation pipeline by replacing the structured prompts with a simple category-only prompt (e.g., "a stroller") during inpainting. As shown in Tab. 4, structured prompt diversification consistently improves detection performance across all rare categories. The gains are particularly pronounced for motorcycle and suitcase, where mAP increases from 0.30 to 2.53 and from 1.8 to 4.56, respectively. These results suggest that object-centric prompt diversification introduces greater intra-class appearance variation in the synthesized data, helping the detector learn more robust representations for rare categories.

**Ablation on synthetic data ratio.** To study the effect of synthetic data scaling, we vary the ratio of RareCrafter-generated images added to the training set from $0\times$ (real-only) to $4\times$ the size of the real dataset. As shown in Fig. 3, adding synthetic samples consistently improves performance over the real-only baseline across all rare categories. For stroller and suitcase, performance peaks around

Table 4. Ablation study on the effect of the prompt generation pipeline for RareCrafter across rare categories using mAP metric for FCOS.

| | Stroller | | Motorcycle | | Suitcase | |
|---|---|---|---|---|---|---|
| Regime | Simple prompt | Structured prompt | Simple prompt | Structured prompt | Simple prompt | Structured prompt |
| RareCrafter | 4.16 | 5.76 | 0.30 | 2.53 | 1.8 | 4.56 |

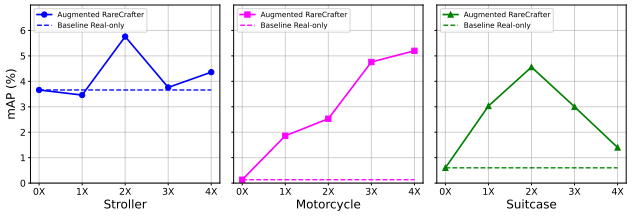


Figure 3. Ablation study on the impact of synthetic data scaling using RareCrafter.

a 2× augmentation ratio and slightly decreases with further scaling, indicating diminishing returns. In contrast, motorcycle performance continues to improve as more synthetic data is introduced.

**Ablation on the inpainting model.** To evaluate the impact of the inpainting backbone, we replace Flux Inpainting [36] with the SDXL Inpainting model [31] while keeping all other components unchanged. Both models receive the same text prompt and inpainting mask. Figure 4 shows a qualitative comparison.

Flux demonstrates stronger prompt adherence and better captures fine-grained attributes described in the text. The generated objects exhibit richer high-frequency details and more realistic textures, whereas SDXL outputs often appear simplified or stylized. This difference is particularly visible in object geometry and material appearance, where Flux produces more realistic structure and shading.

We also observe that SDXL occasionally introduces unintended elements (e.g., an extra person in the stroller example), while Flux typically restricts generation to the intended object. In addition, Flux achieves better integration with the surrounding scene, producing inpainted regions with lighting, color, and texture that are more consistent with the background. Additional qualitative comparisons between Flux and SDXL are provided in Supp. Fig7.

## 5. Conclusion

This work introduced *RareCrafter*, a training-free generative augmentation framework for mitigating the long-tail distribution of object categories in driving-scene datasets. By inserting rare traffic objects into real roadway imagery using controllable flow matching-based generation, RareCrafter enables systematic creation of rare training instances without additional real-world data collection.

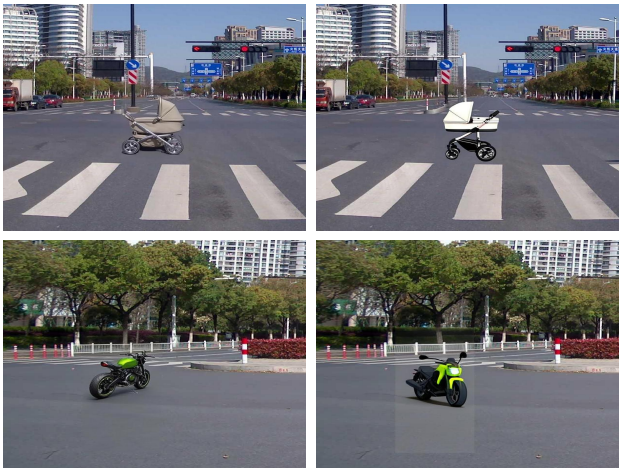

Figure 4. **Ablation on the inpainting model.** Comparison between Flux and SDXL Inpainting using the same prompt and mask. Flux (left) generates objects with stronger prompt adherence, richer details, and better scene integration, while SDXL (right) often produces simplified or stylized outputs with a cartoon-like appearance.

The framework combines prompt diversification, depth-conditioned size priors, and region-based verification to guide synthesis and maintain semantic consistency. Experiments on the CODA2022 corner-case dataset show augmenting training data with synthesized rare-object instances improves detection performance and outperforms a Copy-Paste baseline, demonstrating the potential of controllable generative models to enrich underrepresented driving-scene data and reduce the burden of collecting rare safety-critical scenarios.

During our experiments, we observed that approximating object scale using empirical depth-conditioned size statistics provides stable synthesis but does not explicitly model precise geometric relationships between objects and the scene. While the approach preserves the depth-dependent scale hierarchy observed in real driving data, stronger geometric cues could improve scale consistency. Future work may explore additional geometric supervision, such as depth annotations, 3D structure, or LiDAR measurements. Improving the representation of rare roadway objects in perception systems may ultimately enhance the fidelity of digital roadway representations and digital twin environments used for monitoring and decision support.

# References

[1] Panagiotis Alimisis, Ioannis Mademlis, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, and Georgios Th. Papadopoulos. Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions, 2025. 3

[2] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 7877–7888, 2025. 3

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 3

[4] Daniel Bogdoll, Iramm Hamdard, Lukas Namgyu Rößler, Felix Geisler, Muhammed Bayram, Felix Wang, Jan Imhof, Miguel de Campos, Anushervon Tabarov, Yitian Yang, Martin Gontscharow, Hanno Gottschalk, and J. Marius Zöllner. *AnoVox: A Benchmark for Multimodal Anomaly Detection in Autonomous Driving*, page 206–223. Springer Nature Switzerland, 2025. 3

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 2

[6] Tom Bu, Xinhe Zhang, Christoph Mertz, and John M. Dolan. Carla simulated data for rare road object detection. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2794–2801, 2021. 3

[7] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving, 2020. 1

[8] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation, 2021. 3

[9] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2

[10] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. 2024. 2, 3

[11] Yongjie Fu, Mehmet Kerem Turkcan, Mahshid Ghasemi, Zhaobin Mo, Chengbo Zang, Abhishek Adhikari, Zoran Kostic, Gil Zussman, and Xuan Di. Ai-powered cps-enabled vulnerable-user-aware urban transportation digital twin: Methods and applications. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–18, 2026. 1

[12] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. 6

[13] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings, 2022. 3

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

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[16] Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance, 2025. 2

[17] Hoon Kim, Kangwook Lee, Gyeongjo Hwang, and Changho Suh. Crash to not crash: Learn to identify dangerous vehicles using a simulator. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):978–985, 2019. 3

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[19] Black Forest Labs. Flux: A hybrid transformer–flow architecture for high-fidelity text-to-image synthesis. Model Card, 2024. https://blackforestlabs.ai/. 2, 3, 6

[20] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, Xiaodan Liang, Zhenguo Li, and Hang Xu. Coda: A real-world road corner case dataset for object detection in autonomous driving, 2022. 3, 5

[21] Ruibin Li, Tao Yang, Song Guo, and Lei Zhang. Rorem: Training a robust object remover with human-in-the-loop, 2025. 2

[22] Hongbin Lin, Zilu Guo, Yifan Zhang, Shuaicheng Niu, Yafeng Li, Ruimao Zhang, Shuguang Cui, and Zhen Li. Drivegen: Generalized and robust 3d detection in driving via controllable text-to-image diffusion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27497–27507, 2025. 2

[23] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. 3

[24] 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. *arXiv preprint arXiv:2303.05499*, 2023. 5

[25] Kira Maag, Robin Chan, Svenja Uhlemeyer, Kamil Kowol, and Hanno Gottschalk. Two video data sets for tracking and retrieval of out of distribution objects, 2022. 3

[26] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Chunjing Xu, et al. One million scenes for autonomous driving: Once dataset. 2021. 3, 5

[27] Alexey Nekrasov, Malcolm Burdorf, Stewart Worrall, Bastian Leibe, and Julie Stephany Berrio Perez. Spotting the Unexpected (STU): A 3D LiDAR Dataset for Anomaly Segmentation in Autonomous Driving. In *"Conference on Computer Vision and Pattern Recognition (CVPR)"*, 2025. 3

[28] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 2

[29] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R. Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models, 2025. 3

[30] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: Detecting small road hazards for self-driving vehicles, 2016. 3

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 8, 12

[32] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *CoRR*, abs/2103.13413, 2021. 6

[33] Johannes Schusterbauer, Ming Gui, Pingchuan Ma, Nick Stracke, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. *FMBoost: Boosting Latent Diffusion with Flow Matching*, page 338–355. Springer Nature Switzerland, 2024. 3

[34] Chaehun Shin, Jooyoung Choi, Heeseung Kim, and Sungroh Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator. *arXiv preprint arXiv:2411.15466*, 2024. 2

[35] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Generative object compositing, 2022. 2

[36] Alimama Creative Team. Flux-controlnet-inpainting. https://github.com/alimama-creative/FLUX-Controlnet-Inpainting, 2024. 2, 3, 8, 12

[37] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proc. Int. Conf. Computer Vision (ICCV)*, 2019. 5, 6, 13

[39] Joshua Vendrow, Saachi Jain, Logan Engstrom, and Aleksander Madry. Dataset interfaces: Diagnosing model failures using controllable counterfactual generation, 2023. 3

[40] Yikai Wang, Chenjie Cao, Junqiu Yu, Ke Fan, Xiangyang Xue, and Yanwei Fu. Towards enhanced image inpainting: Mitigating unwanted object insertion and preserving color consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025. 2

[41] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking, 2020. 1

[42] Zhongyu Xia, Jishuo Li, Zhiwei Lin, Xinhao Wang, Yongtao Wang, and Ming-Hsuan Yang. Openad: Open-world autonomous driving benchmark for 3d object detection, 2025. 3

[43] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021. 6

[44] Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, 137:109347, 2023. 3

[45] Yongsheng Yu, Ziyun Zeng, Haitian Zheng, and Jiebo Luo. Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting, 2025. 2

[46] Mahmut Yurt, Xin Ye, Yunsheng Ma, Jingru Luo, Abhirup Mallik, John Pauly, Burhaneddin Yaman, and Liu Ren. Ltda-drive: Llms-guided generative models based long-tail data augmentation for autonomous driving, 2025. 3, 4

[47] Hancheng Zhang, Yuanyuan Hu, Zhendong Qian, Jirui Sha, Min Xie, Yuyang Wan, and Pengfei Liu. Enhancing rare object detection on roadways through conditional diffusion models for data augmentation. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):19018–19029, 2024. 2, 3

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

## 6. Supplementary

This section provides additional material supporting the experiments presented in the main paper. First, we present qualitative examples of synthetic training images generated with the proposed RareCrafter pipeline and with the CopyPaste baseline, illustrating the visual characteristics of each synthesis approach when inserting rare objects into real ONCE driving scenes. We further include additional qualitative comparisons between different inpainting models used for generative synthesis, highlighting differences in prompt adherence and scene integration.

Next, we summarize the datasets, splits, and usage protocols employed in our pipeline, including the ONCE background pool, CODA2022 training and evaluation splits, LVIS foreground sources used for CopyPaste, and the construction of synthetic datasets. This summary also clarifies the experimental regimes used for training detectors with real-only and real+synthetic data.

To complement the rare-class experiments reported in the main paper, we additionally report detection performance on common object categories to verify that augmenting training with synthetic rare objects does not degrade performance on frequent classes.

Finally, we provide the full system prompt and prompt template used to generate structured object descriptions for the generative synthesis pipeline, which are used to produce diverse object prompts for the inpainting model.


Figure 5. **Synthetic training images generated using the RareCrafter framework.** Rare object instances are synthesized directly into real ONCE driving scenes using FLUX-ControlNet-Inpainting conditioned on structured text prompts and predefined insertion regions. The generative model adapts the inserted object to the surrounding scene context, allowing object appearance, shading, and texture to be consistent with the background. Each column corresponds to one target category (stroller, suitcase, and motorcycle), illustrating examples of generated objects placed in different scene configurations.

Table 5. Datasets, splits, and usage in our pipeline. ONCE provides backgrounds, CODA2022 provides real supervision/evaluation, and Synthetic data supplements rare classes. **Common classes (5):** bus, car, cyclist, truck, pedestrian. **Rare classes (3):** motorcycle, stroller, suitcase.

| Dataset | Split / Subset | Size & Labels | Purpose | Protocol & Key Notes |
|---|---|---|---|---|
| **ONCE** | **Background Pool** ($D_{\text{bg}}$)<br>(Training scenes, Cam 3) | **6,002 images**<br>5 Classes (Common) | **Synthesis Backgrounds:**<br>Canvas for inserting rare objects (via FLUX or Copy-Paste). | **Leakage Prevention:**<br>We strictly remove ONCE images that overlap with CODA. Overlapping frames are discarded *before* synthesis. |
| **CODA2022** | **Validation** ($D_{\text{val}}$)<br>Subsets:<br>– Full ($D_{\text{val}}^{\text{full}}$) | **4,884 images**<br>8 Classes (5 Common, 3 Rare) | **Real Training Data:**<br>Used for detector training, hyperparameter selection (e.g., CLIP threshold), and bbox size priors. | **Subset Definitions:**<br>• $D_{\text{val}}^{\text{full}}$: Images with **all** 8 classes. |
| | **Test Set** ($D_{\text{test}}$) | **4,884 images**<br>8 Classes | **Evaluation Only:**<br>Final reporting on real-world rare objects. | **Strict Separation:**<br>$D_{\text{test}}$ has no overlap with $D_{\text{bg}}$ (ONCE) or synthetic data ($D_{\text{syn}}$). |
| **LVIS** | Source Foreground | **160k images**<br>Instance masks | **Copy-Paste Source:**<br>Provides cutouts for motorcycle, stroller, and suitcase that are pasted onto ONCE backgrounds. | Used **only** for the CopyPaste baseline; masks are refined/filtered due to noisy annotations before compositing. |
| **Synthetic** | **1. Copy-Paste** ($D_{\text{syn}}^{\text{CP}}$)<br>(LVIS + ONCE)<br><br>**2. Generative** ($D_{\text{syn}}^{\text{RC}}$)<br>(FLUX + ONCE) | **1 image / class**<br>8 Classes (5 Common, 3 Rare) | **Rare Class Training:**<br>Supplements rare classes in *Real+Synthetic* experiments. | **Regime Definitions (Source of Rare Classes):**<br>• **Real-only:** Train on $D_{\text{val}}^{\text{full}}$. Rare classes are **real only**.<br>• **Real+Syn:** Train on $D_{\text{val}}^{\text{full}}$ + $D_{\text{syn}}$. Real rare classes are **augmented** with synthetic data. |


Figure 6. **Synthetic training images generated using the CopyPaste baseline.** Object instances from the LVIS dataset are composited into ONCE driving scenes using alpha blending within predefined bounding boxes. Because pasted instances are taken directly from existing images, they may exhibit artifacts originating from the source dataset, such as incomplete objects due to occlusion in the original image, fragmented or noisy segmentation masks, and inconsistencies in lighting, texture, or shadows relative to the target scene. Each column corresponds to one target category (stroller, suitcase, and motorcycle).


Figure 7. **Additional qualitative comparison between Flux Inpainting and SDXL Inpainting.** Top row: results generated by Flux Inpainting. Bottom row: results generated by SDXL Inpainting using the same prompts and inpainting masks. Flux generally produces objects with stronger prompt adherence, richer details, and better integration with the surrounding scene. In contrast, SDXL often generates simplified or stylized, toy-like objects, may miss fine-grained attributes, and can introduce unintended elements (e.g., a person pushing the stroller or a rider on the motorcycle).

Table 6. Comparison of object detection performance of FCOS [38] across training configurations for common categories.

| Regime | Bus | Car | Cyclist | Pedestrian | Truck |
|---|---|---|---|---|---|
| Real only | 34.63 | 53.76 | 36.26 | 30.40 | 39.10 |
| + CopyPaste | 33.46 | 53.13 | 35.30 | 29.66 | 38.43 |
| + RareCrafter (ours) | 33.56 | 53.10 | 34.76 | 29.36 | 38.46 |

**System Prompt**

You are generating structured prompts for an image inpainting model (Flux Inpainting Alimama).
Your task is to produce a list of independent prompts for inserting ONE motorcycle into an existing scene.
IMPORTANT RULES:

- Each prompt must describe only ONE motorcycle.
- Do NOT describe the environment.
- Do NOT specify lighting direction.
- Do NOT describe weather.
- Focus primarily on the motorcycle itself.
- Include only general integration cues such as:
  - realistic lighting
  - natural shadow
  - correct perspective
  - photorealistic integration

Each prompt must follow the prompt template.
Vary:

- Motorcycle type (sportbike, cruiser, cafe racer, dirt bike, touring motorcycle, electric motorcycle, naked bike, scrambler, etc.)
- Color (red, blue, green, solid colors, dual-tone, metallic shades, matte tones, etc.)
- Finish (matte, glossy, metallic paint, carbon fiber, chrome, brushed metal, etc.)
- Condition (brand new, lightly used, slightly worn, gently scuffed, moderately worn, subtle signs of use, etc.)
- Design details (fairings, exposed engine, saddlebags, knobby tires, aerodynamic bodywork, minimalist frame, custom exhaust, etc.)
- Viewing angle (front view, rear view, side profile, 3/4 front view, 3/4 rear view, slightly top-down view, low-angle view, eye-level view, etc.)

Generate 50 distinct prompts.
Output only the prompts as a numbered list. Do not include explanations. Do not include commentary.

**Prompt Template**

*"A realistic [motorcycle type/style], [color], [material/finish details], [condition], [design features], [viewing angle], highly detailed, accurate proportions, realistic lighting, natural shadow integration, correct perspective, photorealistic, seamless integration into the scene"*

Figure 8. System prompt used to generate motorcycle insertion prompts for the inpainting model.