Garage: Generative Augmentation Framework for Transforming Object Representations in Images

Andrei Filatov[*1,3], **Daniil Dorin**[*2], **Nikita Barinov**[2,3], **Uliana Izmesteva**[2], **Igor Ignashin**[2], **Ilya Stepanov**[2], **Viacheslav Vasilev**[2,3], **Maxim Kurkin**[1,3], **Dmitry Yudin**[2,3], **Aibek Alanov**[3,5], **Sergey Zagoruyko**[1,4], **Denis Dimitrov**[3], **Andrey Kuznetsov**[3]

[1]Center for Artificial Intelligence Technology, [2]MIPT, [3]AIRI, [4]MTS AI, [5]HSE University

Correspondence: filatovandreiv@gmail.com

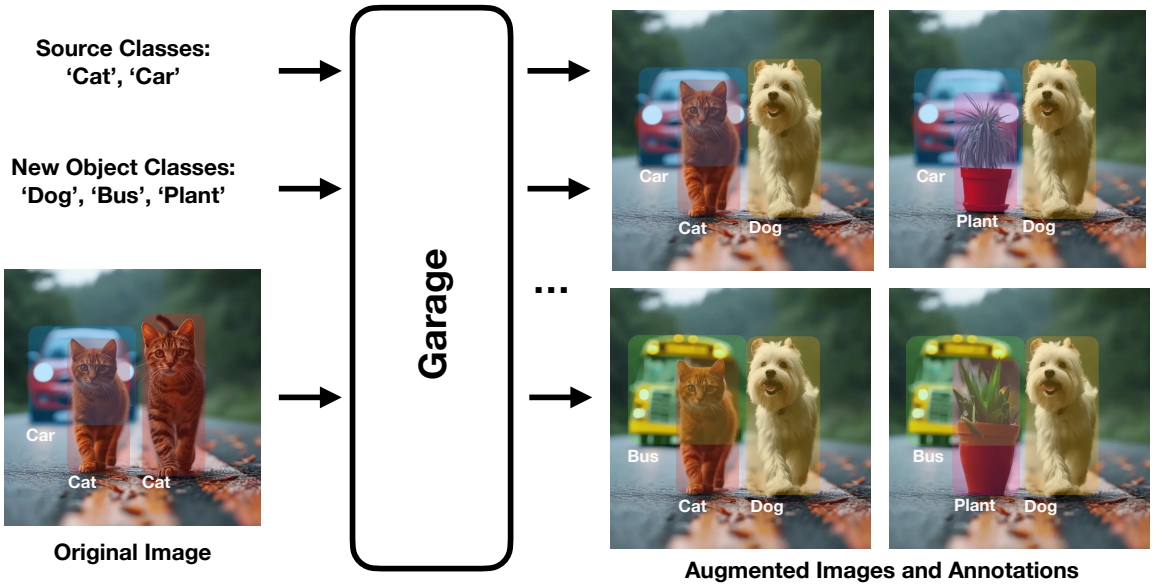

Figure 1: **Examples of generative data augmentation**. The original image (left) depicts a cat and a dog walking on a road with a car in the background. Variations include the substitution of the cat with a plant (top right), the dog with a plant (bottom right), and the background car with a bus (bottom left). These augmentations demonstrate how generative models can create diverse training data by modifying different elements of the original image, thereby enhancing the robustness of computer vision models.

Abstract

Data augmentation is essential for enhancing the performance of machine learning models, particularly in computer vision. Traditional methods such as rotations, shifts, and brightness adjustments are limited in their ability to provide significant semantic variations, often resulting in models that do not generalize well to new data. Language models, on the other hand, possess extensive knowledge about the world, its structure, and semantics. If this knowledge could be transferred to datasets, it would significantly enrich the visual diversity of the data and improve the quality of visual models. In this paper, we introduce **Garage**, a novel framework designed to overcome these limitations by allowing the replacement of object annotations in images. Using Vision-Language Models (VLMs), we obtain descriptions of images, and then, leveraging language models, we determine what can be replaced in the image and compose an extended prompt that generates data modifications. This way, we semantically augment the dataset in a meaningful way. **Garage** operates in two modes: interactive, where annotators can manually select and replace objects, and automatic, where objects are automatically replaced based on user-defined classes. We demonstrate the effectiveness of this framework by benchmarking models trained with and without the augmented data, showing improvements in performance of object detection.

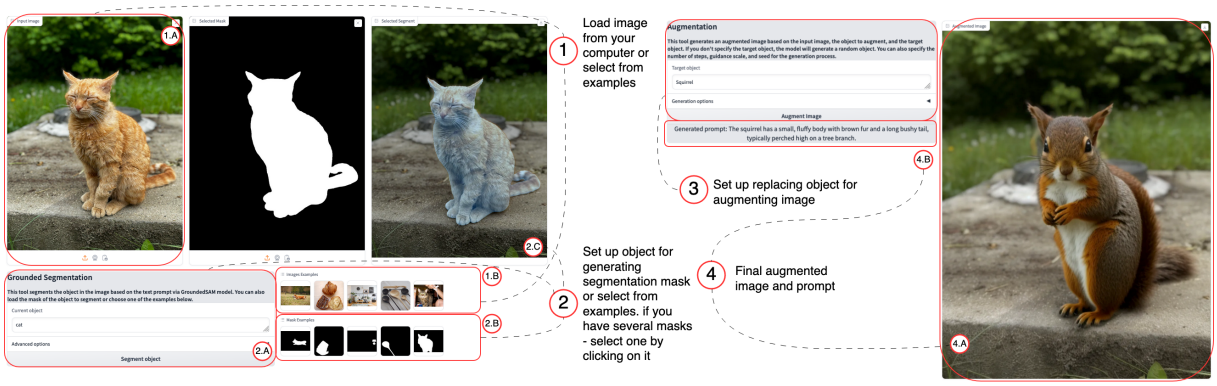

Figure 2: **Demo Visualization**.This demo is simple to use and provides users with flexible options for selecting objects and augmentation variants, allowing for easy generation of augmented images. The process begins with uploading an image from your computer or selecting from provided examples. Once the image is loaded, the user can set up the object for segmentation. This can be done either through a text prompt in the GroundedSAM model or by selecting a mask from given examples. Finally, the user configures the target object for augmentation. The tool will generate an augmented image based on the input image, the selected object, and the specified target object, ensuring flexibility and ease of use at each step.

1 Introduction

Data augmentation is a crucial tool in the arsenal of modern machine learning and computer vision researchers. It allows an increase in the volume of data, improving the overall performance of models by creating various variations of the original data. However, traditional augmentation methods, such as rotations, shifts, and brightness changes, are limited in their capabilities. They do not provide significant semantic extensions in the data, which could substantially enhance the training models. For example, when training a model for object recognition, standard augmentations such as rotation and scaling do not provide sufficient diversity in the objects within images. This can lead to models not generalizing well to new, unseen data.

In this paper, we propose a new framework **Garage** for data augmentation that allows to replace object annotations in images. This approach not only increases the amount of data but also enriches it semantically, which is important for improving the generalization capability of machine learning models.

We developed a framework that operates in two modes: interactive and automatic. In the interactive mode, the user can manually select the object to be augmented and specify the object to replace it with. This allows for more precise control over the augmentation process and adaptation to specific needs. In the automatic mode, the framework automatically selects objects in the image and replaces them with objects from user-specified classes. This enables the rapid generation of large volumes of data with diverse semantic combinations.

Our contributions are as follows:

- We present the first data augmentation framework **Garage** designed for replacing object annotations, providing significant semantic enrichment to datasets. **Garage** supports both interactive and automatic modes, catering to various user needs and datasets.
- We demonstrate experimental results that show improved generalization capabilities of models trained with our augmented data. Video demonstration is available[1]
- We offer an open-source implementation of our framework, allowing the research community to benefit from our advancements in data augmentation. We publish the code on GitHub [2] under the Apache 2 license and provide introduction video.

2 Framework

2.1 Interactive augmentation

Creating annotations for object detection tasks is inherently labor-intensive, often requiring several minutes per image. Achieving a diverse dataset necessitates processing a substantial number of images, further compounding the effort. To streamline this process, our framework provides an interface

[1] Video Demonstration
[2] Github repository

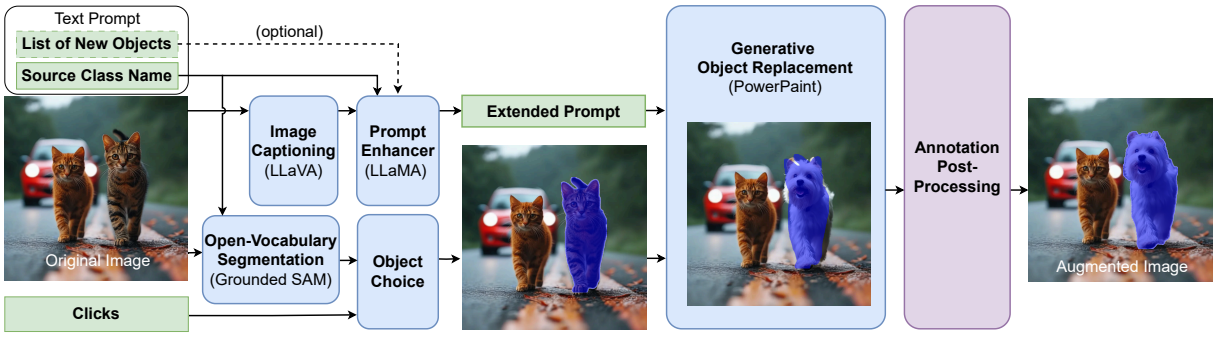

Figure 3: **Internal details of the framework**. The figure illustrates the sequential steps in the framework. It includes: **Object Choice:** Choosing the appropriate object based on user prompts or automatic choice. **Augmentation:** Generation of augmentation with additional user prompts or automatic choice. **Post-processing:** Removing generation artifacts and filtering incorrect examples. Each step is crucial for achieving high-quality results that meet user requirements.

designed to assist annotators in generating diverse annotations with ease. Interactive Augmentation

Our framework includes a Gradio demo (see Figure 2) that facilitates interactive image augmentation. The workflow for interactive augmentation is as follows:

- **Image Upload and Object Selection** The annotator uploads an image and selects the object to be replaced by either clicking on it or using a textual prompt to describe the object which passed to GroundedSAM (Ren et al., 2024) to extract object mask.
- **Class Specification** The annotator specifies the class with which the original object will be replaced by.
- **Prompt Extension and Image Generation** Our framework extends the provided prompt using LLaMA (Dubey et al., 2024) and generates the augmented image with the replaced object using PowerPaint (Zhuang et al., 2023), along with the new annotation.
- **Downloading Augmented Images:** Finally, the annotator can download the augmented images, thus significantly simplifying the annotation process and enhancing dataset diversity.

2.2 Automatic augmentation

In addition to manual augmentation, our framework offers a fully automatic augmentation mode. This mode is a powerful tool for creating a comprehensive augmented dataset, significantly enhancing the quality of image segmentation models.

Upon uploading the original dataset, users can specify which classes to add or augment. For instance, if the user wants to add a "capybara" class, they can specify this, and it will be incorporated into the dataset. The framework then automatically generates the necessary augmentations, resulting in an expanded dataset. This comprehensive system consists of three stages, detailed below.

2.3 Object choice

The first step involves selecting the object to augment. For automatic selection, we utilize LLaVA (Liu et al., 2024) to identify objects that can be replaced and provide image description. LLaVA extracts a list of objects from the image, from which a random object is selected for augmentation. Subsequently, we apply GroundedSAM (Kirillov et al., 2023) to generate the object's mask based on selected object. For LLaVA prompt details, please refer to Appendix A.

2.4 Augmentation

When the object for augmentation is selected, LLaMA (Dubey et al., 2024) is employed to generate a suitable replacement. The user can provide a list of potential replacement objects, from which the model will make the final selection. If no list is provided, it will automatically select an object. Furthermore, the model receives a comprehensive image description from LLaVA to ensure that LLaMA fully understands the context.

To generate replacement we apply PowerPaint inpainting (Zhuang et al., 2023) to the object with the selected prompt. It is a common problem that generation of the object with short prompt usually is not great. So we apply the prompt extension,

using LLaMA, to the provided object prompt. In the end of this procedure we pass expanded prompt to PowerPaint to get the augmented image.

2.5 Post-processing

We apply two stage post-processing. On the first stage we use Alpha-CLIP (Sun et al., 2024) to filter out good generation. Alpha-CLIP acts as normal CLIP (Radford et al., 2021) but accepts mask which allows to calculate CLIP similarity with specific area on the image. If CLIP score is bigger than certain threshold we accept the synthesized augmentation. Otherwise, we generate image again. After the image passed CLIP scoring we apply SAM to get more precise object annotation for the image. It is necessary because inpainting often generate a new object not precisely by annotation, so we need to apply the correction to get the correct mask. One can see the mismatch between object and mask on the Figure 3

3 Experiments

To check efficiency of our augmentation framework we conducted experiments with augmented data on VOC (Everingham et al., 2010).

3.1 Addition of new class

One possible application of our framework is generating data for a class which is absent or has limited presence in an existing dataset. Augmenting with a specific object can help in cases where an object of a certain class is absent from the data, just by adding it in the training data. If the object is present, it can make the data sampling more diverse.

To verify this application, we conducted the following experiment. From VOC dataset, we removed data for certain classes. For VOC «*cat*» class. Then, we performed data augmentation on the remaining data to add the absent class to the sample. During the augmentation process, the proposed algorithm was applied to all images in the dataset that did not have an augmented class.

For images with multiple objects, a random object was selected from among the objects whose bounding boxes are within a relative area of no more than 0.5, if such objects are available. This was done to avoid overlapping of small objects during the generation process.

Next, we trained detection models (FasterRCNN (Ren et al., 2015), DETR (Carion et al., 2020), YOLOv10-N (Jocher et al., 2023)) on data with various augmentations of the absent class to examine the impact of our augmented data on the results. For training we used standard scripts from MMDetection (Chen et al., 2019) and Ultralytics (Jocher et al., 2023). The results are presented in Table 1. From our experiments, we observe that the use of our framework improves the quality of detection of absent class while maintaining the overall quality of the model at the same level. It is important to note that the detection quality for other classes did not change.

3.2 Knowledge transfer through expanding prompt

When using standard prompts, we observed a decline in the visual quality of generated images, particularly with shorter prompts, such as simple class labels. To address this issue, we employed extended prompts by leveraging the linguistic capabilities of the LLaMA model. For example, a simple instruction such as *cat* can be expanded into a more descriptive and contextually rich phrase, such as *The ginger tabby cat has a sleek body, pointy ears, and curious green eyes gazing around from its playful pose.* This extension introduces important contextual information, effectively bridging the gap between textual and visual representations.

To assess the importance of prompt extension in model training, we conducted an experiment by creating two datasets with identical images: one using only the basic prompt and the other using the extended prompt incorporating a textual description of the image. This approach allowed us to evaluate how knowledge transfer through extended prompts could enhance the quality of the generated images. The extended instructions provided the subsequent PowerPaint model with detailed information and nuances, enabling the creation of higher-quality, diverse, and visually consistent images according to the instruction.

As previously mentioned, generating an object with a short instruction usually turns out to be of poor quality. This effect is demonstrated on the Figure 4. Using the proposed model, images were augmented with a fixed seed. The resulting images show that the quality of generation decreases without additional expansion of the instruction. Also, we conducted additional experiment on how prompt expansion affects training object detection model. We train FasterRCNN model on dataset where was no cats with class prompt and extended prompt. The results are shown in Table. 4. From

Table 1: **Object detection results on Pascal VOC.**The results demonstrate that training on our data significantly improves model performance across various data percentages, as shown by the superior average precision (AP) values in the 'ours' rows compared to the 'original' ones. The percentages represent the ratio of cat pictures used from original dataset:

Dataset	Model	Data / Class Presence	0%	25%	50%	75%	100%
Pascal VOC	DETR	original	0.0	**57.5**	61.5	65.4	69.1
		ours	**5.3**	55.6	**62.2**	**66.3**	**70.3**
	YOLOv10-N	original	0.0	50.9	**54.8**	56.9	60.4
		ours	**31.5**	**51.3**	53.6	**57.9**	**61.3**
	Faster RCNN	original	0.0	74.5	77.4	75.5	76.6
		ours	**66.4**	**77.3**	**80.1**	**83.5**	**84.0**

the results we see that knowledge transfer using language model as a prompt expander improves the quality of augmented data and consequently detection model.

AP on *cat* category	
w/o expanded prompts	expanded prompts
64.6	**66.4**

Table 2: **Comparison of the quality of FasterRCNN training depending on instruction expanding.** Prompt extension improves the quality of the model. The results demonstrate that using expanded prompts significantly enhances the average precision (AP) for the cat category compared to using prompts without expansion.

4 Related Work

Data augmentation is a widely used technique in the field of machine learning and computer vision to enhance the diversity and volume of training datasets. Traditional data augmentation techniques, such as rotation, flipping, cropping, and color jittering, have been foundational in enhancing the robustness of models (Buslaev et al., 2020). For instance, (Krizhevsky et al., 2012) demonstrated the effectiveness of these techniques in their pioneering work on the ImageNet classification challenge. However, these methods often lack the sophistication to address more complex data distributions.

More recent approaches have leveraged generative models as a source for augmentation generation. (Alimisis et al., 2024) explores the use of advanced generative models to create diverse and realistic augmentations. (Yin et al., 2023) used text-to-text and text-to-image models to generate augmentations for image classification tasks, demonstrating significant improvements in model performance. (Fang et al., 2024) utilized a ControlNet adapter to generate augmentations, enhancing the dataset's variability and robustness. (Kupyn and Rupprecht, 2024) applied inpainting techniques to augment data while maintaining the same label annotations.

An alternative to augmentations that improve semantics is Open Vocabulary Detection (Wu et al., 2024), where a model learns to detect objects using a language model. While effective, this approach is limited by the vocabulary of objects it was trained on and does not allow for increasing the diversity of the dataset through refinements in the properties of detectable objects. Moreover, creating such architectures requires substantial resource investment, which can be a significant constraint when training an efficient model.

Our work differs from previous generative augmentation approaches by enabling the augmentation of images to generate new annotations for different classes, rather than merely enhancing the existing ones. This method enriches the dataset semantically and structurally, offering a more comprehensive augmentation strategy.

5 Limitations & Conclusion

In this paper, we presented **Garage**, an interactive and fully automated framework designed to address the challenges associated with data augmentation in machine learning and computer vision.

We demonstrated the effectiveness of **Garage** through comprehensive experiments and benchmarking. The results show that models trained with data augmented by our framework exhibit improvements in performance compared. Also, by automating the augmentation process and providing an interactive user interface, **Garage** simplifies the task of data augmentation, making it accessible

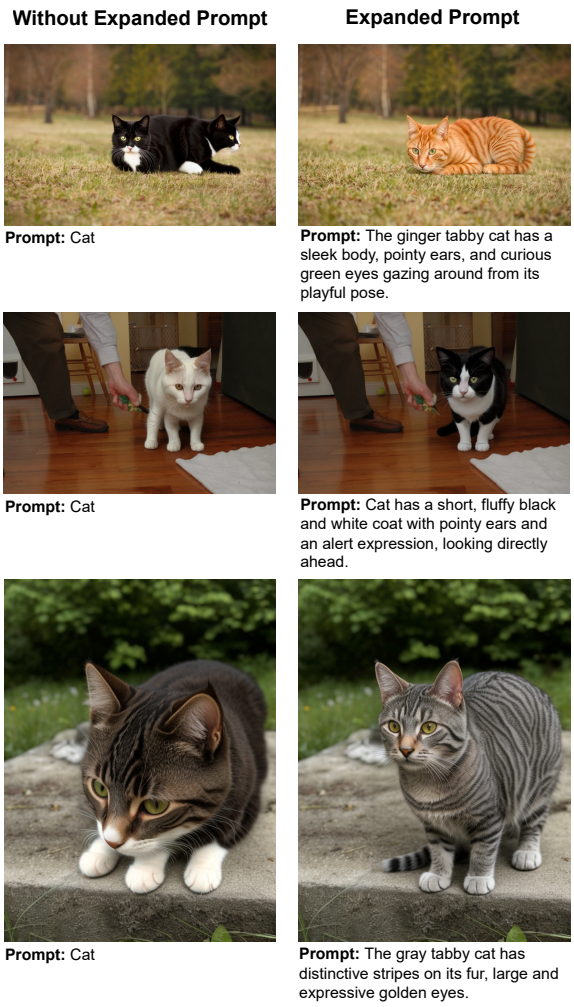

Figure 4: **Comparison of inpainting generation depending on prompt expansion.** The images on the left, generated using expanded prompts, exhibit significantly more detail and accuracy compared to those on the right, which were generated with minimal prompts. This showcases how our model performs better with detailed descriptions, capturing the nuances and specifics of the desired output more effectively.

and practical for researchers and practitioners.

Furthermore, **Garage**'s extensible architecture allows for easy integration of new augmentation techniques and customization according to specific needs. This flexibility ensures that our framework can adapt to various data types and domains, offering a robust solution for enhancing training datasets.

One limitation of our approach is the computational resources required, including the necessity of powerful GPUs and time to generate each augmented image. This can be a constraint in practical applications, especially when large datasets need to be processed. However, this limitation can be mitigated by using distilled models like SD3 Turbo (Sauer et al., 2024) in production, which are optimized versions of the original models. Distilled models can perform the same tasks more efficiently, reducing the computational load and time required for image generation, thereby making the Garage framework more practical and scalable for real-world use.

In conclusion, **Garage** represents a significant advancement in the realm of data augmentation, addressing key limitations of existing methods and providing a comprehensive, automated solution for creating high-quality, diverse, and balanced training datasets. We believe that our framework will contribute to the development of more robust and fair machine learning models, ultimately advancing the state of the art in the field.

6 Ethical considerations

This research introduces the **Garage** framework for data augmentation, raising several important ethical considerations. Ensuring data privacy is crucial, particularly when dealing with datasets containing personal information. Robust anonymization techniques and adherence to data protection regulations like GDPR are essential to protect individuals' identities. Additionally, data augmentation techniques must be monitored to prevent the introduction or amplification of biases related to race, gender, and socioeconomic status. Efforts should be made to identify and mitigate these biases, ensuring the development of fair and unbiased machine learning models.

The ability to replace object annotations and generate new images presents the risk of misuse, such as creating misleading or deceptive content. Safeguards and ethical guidelines are needed to prevent harmful applications of the **Garage** framework. Transparency is also key; researchers should provide clear documentation of methods, openly share code and datasets, and acknowledge limitations and potential ethical concerns. Addressing the environmental impact of computational resources used in training models is also important, with a focus on energy-efficient algorithms. Finally, despite advancements in automation, human oversight remains crucial to maintain ethical standards and quality control, ensuring responsible use and development of the **Garage** framework.

Acknowledgments

References

Panagiotis Alimisis, Ioannis Mademlis, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, and Georgios Th Papadopoulos. 2024. Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions. *arXiv preprint arXiv:2407.04103*.

Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. 2020. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125.

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

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. 2019. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338.

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

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. 2023. Yolo by ultralytics.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Orest Kupyn and Christian Rupprecht. 2024. Dataset enhancement with instance-level augmentations. *arXiv preprint arXiv:2406.08249*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

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

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. 2024. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015*.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029.

Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. 2024. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuwei Yin, Jean Kaddour, Xiang Zhang, Yixin Nie, Zhenguang Liu, Lingpeng Kong, and Qi Liu. 2023. Ttida: Controllable generative data augmentation via text-to-text and text-to-image models. *arXiv preprint arXiv:2304.08821*.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. 2023. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint arXiv:2312.03594*.

A Example Appendix

A.1 Prompts

A.1.1 LLaVA Prompt image caption.

The prompt is designed to guide the assistant in generating a comprehensive caption for an image. The key elements to be included in the caption are the details of objects within the image, their relative positions, and their quantities. This structured approach ensures that the generated caption is thorough and informative, providing a detailed understanding of the visual content.

USER: <image> Provide a detailed caption for this image. Include details about the relative position of objects in the picture and their number.

A.1.2 LLaMA Prompts for selecting a new object

This prompt is tailored to instruct the assistant in identifying a suitable replacement object within a given scene. The assistant is prompted to suggest a new object that is distinctly different from the existing one, based solely on the provided scene description and current object. This ensures the replacement maintains the context and coherence of the scene while introducing variety.

USER: Imagine you are an object replacer. Your task is generating a replacement object instead of the existing object on the scene. It's important that the new object is not the same as the existing one. I will give you a description of the scene and the existing object. You must give me an object which could be depicted instead of the existing object. So, image description: {ImageDescription}, existing object: {CurrentObject}. You should return only a name of the new object and nothing else.

If a list of potential objects is submitted:

USER: Imagine you are an object replacer. Your task is generating a replacement object instead of the existing object on the scene. It's important that the new object is not the same as the existing one. I will give you a description of the scene, existing object, and a list of potential new objects. You must give me an object from the list of potential new objects which could be depicted instead of the existing object. The new object should fit well into the picture in place of the existing object. The new object should be approximately the same size as the existing object. If no object from the list fits into the picture, return the existing object. The image should remain believable after replacement. So, image description: {ImageDescription}, existing object: {CurrentObject}, a list of potential new objects: {NewObjectsList}. You should select and return only the name of the new object from the provided list, which fits into the picture to replace the existing one.

A.1.3 Prompt for expansion using LLaMA

USER: Imagine that you want to describe the {NewObject}'s appearance to an artist in one sentence, under 15 words. Mention {NewObject} in the description for clarity. Focus solely on the realistic description of the {NewObject}, ignoring any external elements or surroundings. For example, if the object is an animal, the description should include the animal's color, size, breed, pose, view direction, etc. If the object is a vehicle, the description should include the vehicle's brand or model, color, size, type, etc. If the object is a person, the description should include the person's age, gender, height, weight, hair color, eye color, clothing, pose, etc. Do not add anything extra to the visual description that is not directly related to {NewObject}.

A.2 Generation examples

All 20 classes from the PascalVOC dataset have been augmented in Figure 5.

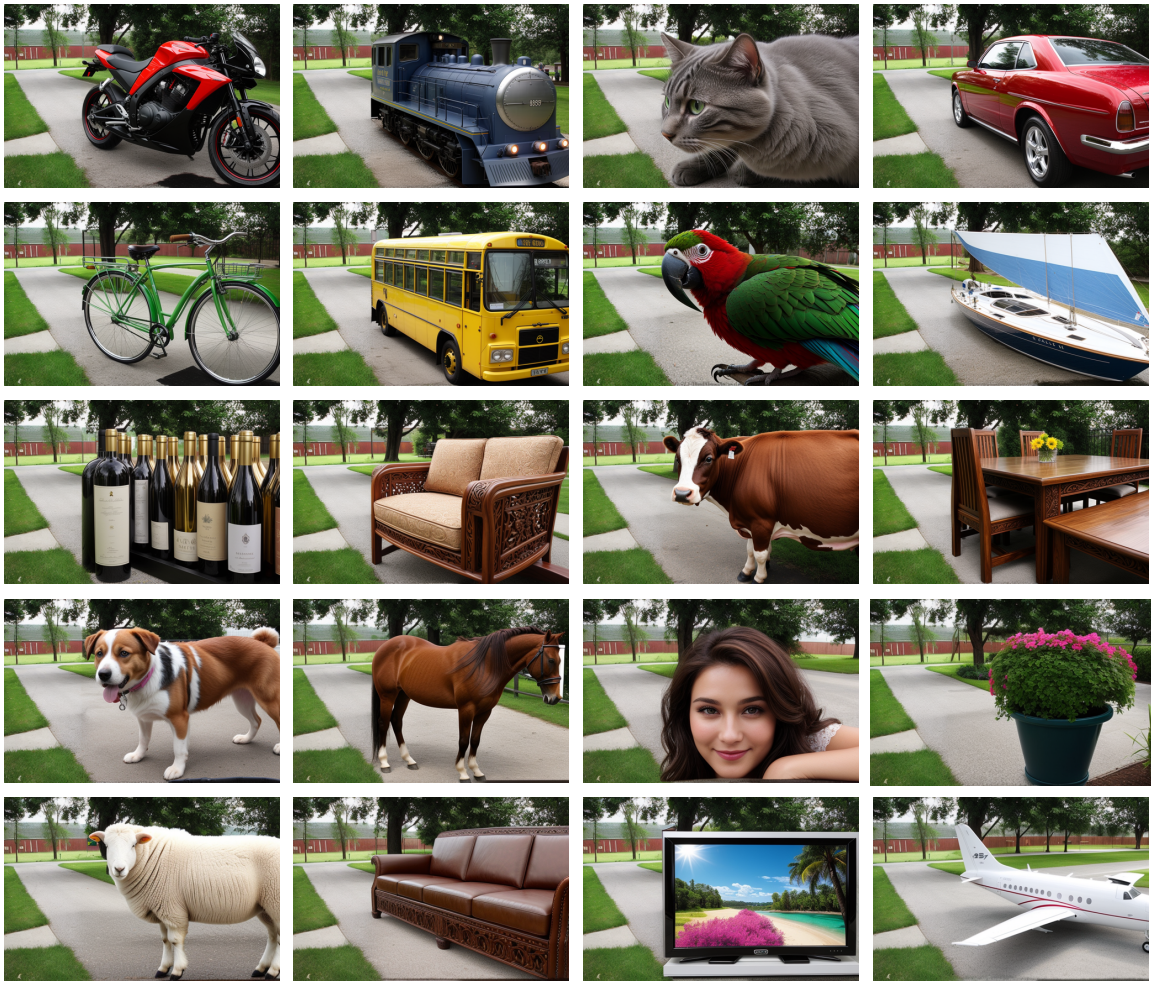

Figure 5: **Augmentation PascalVOC categories.**This figure illustrates various object classes from the PascalVOC dataset, each augmented using our Framework. The categories include: (top row) motorcycle, train, cat, car; (second row) bicycle, bus, bird, boat; (third row) bottle, chair, cow, dining table; (fourth row) dog, horse, person, potted plant; (bottom row) sheep, sofa, television, and airplane. Each object class is depicted in a lifelike manner, showcasing the potential of generative augmentations to produce realistic and diverse instances for training machine learning models.