Few-Shot versus Zero-Shot Detection of Novel Objects Using Pretrained Vision-Language Models: A Comparative Study on Kikoriki Characters

Nikita Andriyanov[1][0000-0003-0735-7697] and Vitaly Dementiev[2][0000-0002-4880-0432]

[1] Financial University under the Government of the Russian Federation, Leningradsky av. 49, 125167 Moscow, Russia
[2] Ulyanovsk State Technical University, Severny Venets str. 32, 432027 Ulyanovsk, Russia
nikita-and-nov@mail.ru

Abstract. Novel object detection remains challenging when target categories are absent from standard detection datasets and only limited annotations are available. This paper presents a controlled pilot comparison of zero-shot and few-shot detection using two pretrained vision–language models, BLIP and Grounding DINO. The study focuses on two visually distinctive cartoon characters from the Kikoriki/Smeshariki universe: Krash and Chiko. In the zero-shot setting, the models rely only on natural-language prompts, whereas in the few-shot setting they are adapted using 5 or 10 annotated images per category. The experiments are conducted on a small synthetic dataset with controlled backgrounds and manually verified bounding boxes. The results show that adding a small number of annotated examples improves detection performance in this controlled setting, particularly by reducing missed detections. However, because the evaluation subset contains only a limited number of test instances, the reported metrics should be interpreted as descriptive evidence rather than statistically conclusive estimates of general detection performance. The study therefore contributes a compact experimental case study of the annotation–performance trade-off for novel cartoon character detection and identifies the limitations that must be addressed in larger, more realistic benchmarks.

Keywords: Novel Object Detection, Few-Shot Learning, Zero-Shot Learning, Vision-Language Models, BLIP, Grounding DINO, Object Localization.

1 Introduction

Object detection is one of the fundamental tasks in computer vision and serves as a core component of numerous practical applications, including autonomous driving, video surveillance, robotics, medical image analysis, and human–computer interaction. The objective of object detection is not only to classify objects present in an image but also to accurately localize them using bounding boxes. During the last decade, remarkable progress has been achieved due to advances in deep learning, particularly through convolutional neural networks and transformer-based architectures ca-

pable of extracting highly discriminative visual representations from large-scale image datasets [1, 2].

Modern object detectors such as Faster R-CNN, YOLO, DETR, and DINO have demonstrated impressive performance on benchmark datasets including PASCAL VOC, MS COCO, and Open Images [3–6]. However, these approaches rely heavily on supervised learning and require thousands of manually annotated examples for each target category. Although such large-scale annotations are available for common object classes, collecting and labeling sufficient training data becomes prohibitively expensive for specialized domains or newly emerging categories.

To address this limitation, researchers have recently focused on reducing the dependence on annotated data. One promising direction is zero-shot object detection (ZSD), where models are expected to recognize and localize object categories that were not present during training [7]. Zero-shot methods typically leverage semantic representations derived from textual descriptions, attribute vectors, or language embeddings to establish relationships between seen and unseen categories. Despite substantial progress, zero-shot object detection remains challenging because the semantic gap between visual appearance and textual descriptions often leads to inaccurate localization and reduced detection performance [8].

The emergence of vision-language models (VLMs) has significantly expanded the capabilities of zero-shot recognition systems. Models such as CLIP and BLIP learn joint visual and textual representations from hundreds of millions of image–text pairs and exhibit strong generalization abilities across a wide range of downstream tasks [9, 10]. These models enable open-vocabulary recognition by grounding natural-language descriptions in visual content. Building upon these advances, Grounding DINO introduced a powerful framework that combines transformer-based object detection with language-guided grounding, allowing arbitrary object categories to be detected directly from textual prompts without task-specific retraining [11].

Although vision-language models have substantially improved zero-shot detection capabilities, their performance often remains insufficient when the target category differs significantly from objects encountered during pretraining. In many practical scenarios, a small number of annotated examples can be obtained with reasonable effort. This observation motivates the study of few-shot object detection (FSOD), where models are adapted to novel categories using only a limited number of labeled samples [12]. Recent surveys indicate that few-shot adaptation can significantly improve localization quality and classification accuracy while maintaining low annotation costs compared to fully supervised approaches [13].

The trade-off between zero-shot and few-shot learning is particularly important for categories that are absent from standard object detection datasets. Examples include fictional characters, proprietary industrial components, domain-specific medical findings, or newly designed products. In such cases, practitioners must decide whether textual descriptions alone are sufficient or whether collecting a small set of annotated examples can provide a meaningful improvement in detection performance.

Recent studies have also demonstrated the growing importance of foundation models and vision-language approaches across a variety of computer vision tasks. The integration of large language models with image understanding pipelines has shown

promising results for enhancing semantic image descriptions and visual content interpretation [14]. Likewise, foundation segmentation models such as Segment Anything have proven effective in industrial machine vision applications, enabling robust object localization with minimal task-specific adaptation [15]. Furthermore, recent research on satellite imagery analysis confirmed the practical potential of zero-shot detection for recognizing previously unseen objects in domains where annotated data are scarce or expensive to obtain [16]. These findings collectively suggest that pretrained vision-language models provide a strong basis for novel object detection, while also highlighting the need to systematically investigate the trade-off between zero-shot generalization and few-shot adaptation.

In this work, we investigate this question using two novel object categories represented by characters from the animated series Smeshariki: Krash, a blue rabbit with long ears, and Chiko, a purple hedgehog wearing glasses. These categories are intentionally selected because they are absent from conventional object detection benchmarks and therefore constitute a suitable testbed for evaluating the generalization capabilities of pretrained vision-language models. We compare the performance of BLIP and Grounding DINO under zero-shot conditions and after few-shot adaptation using 5 and 10 annotated images per category.

The proposed setting is intentionally controlled and should not be interpreted as a substitute for large-scale real-world novel object detection benchmarks. Instead, it is used as a compact pilot scenario in which the effect of a small number of annotated examples can be isolated under fixed object appearance, background complexity, and annotation quality. This design allows a focused comparison of zero-shot prompting and few-shot adaptation, while also requiring cautious interpretation of the resulting metrics.

The main contributions of this paper are as follows:

1) We formulate a controlled pilot scenario for evaluating novel cartoon character detection using pretrained vision–language models.

2) We compare zero-shot prompting and few-shot adaptation for BLIP and Grounding DINO under identical prompts and test conditions.

3) We report object-level detection metrics together with count-based interpretation to clarify the effect of the small test set.

4) We discuss the practical annotation–performance trade-off and the limitations of drawing broad conclusions from a small synthetic dataset.

The remainder of the paper is organized as follows. Section 2 reviews related work in zero-shot, few-shot, and vision-language-based object detection. Section 3 describes the dataset and experimental setup. Section 4 presents the evaluated models and adaptation procedure. Section 5 discusses experimental results, while Section 6 concludes the paper and outlines future research directions.

2 Related Works

Zero-shot object detection (ZSD) aims to recognize and localize object categories that are absent from the training set. Unlike conventional supervised detectors, zero-shot

approaches exploit semantic information derived from textual descriptions, attribute vectors, or language embeddings to establish relationships between seen and unseen categories. One of the first comprehensive formulations of this problem was introduced by Bansal et al. [7], who demonstrated that semantic representations can be used to transfer knowledge from known classes to novel ones. Subsequently, Rahman et al. [8] proposed a unified framework for simultaneously recognizing and localizing unseen concepts, further improving the generalization capabilities of zero-shot detectors.

Despite these advances, zero-shot object detection remains challenging because textual descriptions often fail to capture all visual characteristics of a target category. As a result, localization accuracy typically decreases when the semantic gap between training and target domains becomes large. This limitation is especially evident when dealing with highly specialized objects, fictional characters, or categories that are underrepresented in large-scale web datasets.

The emergence of vision-language models has significantly accelerated progress in open-vocabulary and zero-shot recognition tasks. CLIP introduced large-scale contrastive learning from image-text pairs and demonstrated remarkable transferability across a wide range of visual tasks [9]. Building upon similar principles, BLIP proposed a unified vision-language framework that combines image understanding and language generation through bootstrapped pretraining strategies [10].

More recently, Grounding DINO extended transformer-based object detection by incorporating language-guided grounding mechanisms directly into the detection pipeline [11]. Unlike traditional detectors that rely on a fixed set of predefined classes, Grounding DINO enables open-vocabulary object detection through natural-language prompts. Such capabilities make vision-language models particularly attractive for applications where annotated examples are unavailable or difficult to obtain.

Few-shot object detection seeks to reduce annotation requirements by adapting pretrained detectors using only a small number of labeled examples. Existing methods typically employ transfer learning, meta-learning, prototype learning, or feature reweighting strategies to generalize from a limited set of samples [12]. Compared to zero-shot approaches, few-shot methods benefit from direct visual examples of target categories, enabling more precise localization and classification.

Recent surveys indicate that few-shot object detection has become one of the most active research directions in visual recognition, particularly for industrial inspection, medical imaging, and remote sensing applications [13]. The main advantage of few-shot learning lies in its ability to achieve substantial performance gains while requiring only a fraction of the annotation effort associated with fully supervised training.

The practical applicability of foundation models has recently been demonstrated across several specialized domains. Andriyanov and Dementiev [14] investigated the integration of large language models with image caption generation and style transfer techniques, highlighting the potential of multimodal representations for enhanced visual understanding. In industrial computer vision, Andriyanov et al. [15] employed the Segment Anything Model (SAM) to develop a pallet loading control system, demonstrating that foundation models can be effectively adapted to real-world manufacturing environments with minimal task-specific modifications.

In the context of remote sensing, Andriyanov et al. [16] explored zero-shot detection for satellite imagery, showing that pretrained vision-language models can successfully identify previously unseen object categories even in highly specialized domains. These findings suggest that foundation models provide a promising basis for novel object detection tasks, while simultaneously motivating further investigation into the balance between zero-shot generalization and few-shot adaptation.

Although previous studies have separately investigated zero-shot detection, few-shot adaptation, and vision-language foundation models, a systematic comparison of zero-shot and few-shot object detection for completely novel categories remains limited. In particular, there is little evidence regarding how many annotated examples are required to significantly improve the performance of modern vision-language detectors when encountering unseen objects. This work addresses this gap by comparing BLIP and Grounding DINO under zero-shot, 5-shot, and 10-shot settings using previously unseen cartoon character categories.

3 Dataset and Controlled Experimental Setup

To evaluate the effectiveness of zero-shot and few-shot object detection, we constructed a dedicated benchmark containing two novel object categories that are absent from standard object detection datasets. The selected categories correspond to the animated characters Krash and Chiko from the Kikoriki universe. Krash is represented as a blue rabbit with long ears, while Chiko is represented as a purple hedgehog wearing glasses.

These categories were intentionally selected because they possess distinctive visual characteristics while remaining outside the distribution of common object detection benchmarks such as MS COCO or Open Images. Consequently, they provide a suitable testbed for evaluating the generalization capabilities of pretrained vision-language models.

Figure 1 illustrates representative examples of both categories together with an annotated test scene used during evaluation.

A synthetic dataset was created to ensure complete control over object appearance, scale, and annotation quality. For each category, multiple images with variations in position, orientation, and size were generated. Bounding-box annotations were manually verified to eliminate labeling inconsistencies.

The dataset was divided into three subsets:

- training subset for 5-shot adaptation;
- extended training subset for 10-shot adaptation;
- test subset used exclusively for evaluation.

No test images were included during adaptation, ensuring a fair comparison between zero-shot and few-shot settings.

The synthetic nature of the dataset provides full control over object placement, scale, and annotation quality, but it also limits the difficulty of the task. In particular, the images do not fully reflect real-world conditions such as cluttered backgrounds, partial occlusion, strong illumination changes, motion blur, or domain shift. There-

fore, the dataset is suitable for a controlled pilot analysis rather than for estimating real-world deployment performance.

Fig. 1. Examples of Krash and Chiko training images together with a labeled test scene.

The test subset contains four annotated object instances in total, two per category. All reported precision, recall, and F1-score values are therefore object-level descriptive metrics computed on this small test subset

Table 1 summarizes the main characteristics of the dataset.

Table 1. Dataset statistics.

Category	**Training images (5-shot)**	**Training images (10-shot)**	**Test objects**	**Average bounding-box size (px)**
Krash	5	10	2	116 × 116
Chiko	5	10	2	121 × 121
Total	10	20	4	118 × 118

Two pretrained vision-language models were evaluated: BLIP and Grounding DINO. Three experimental configurations were considered:

1) Zero-shot detection – only textual descriptions of target objects were provided.

2) 5-shot adaptation – five annotated examples per category were used for model adaptation.

3) 10-shot adaptation – ten annotated examples per category were used for model adaptation.

For zero-shot inference, the following textual prompts were employed:

- “blue rabbit with long ears” for Krash;

- “purple hedgehog with glasses” for Chiko.

The same semantic descriptions were retained during few-shot adaptation to preserve consistency across experimental settings.

Detection quality was assessed using standard object detection metrics. Precision and Recall were used to measure detection correctness and completeness, respectively. The harmonic mean of these metrics was reported as the F1-score. Localization quality was evaluated using the Intersection over Union (IoU) criterion. Because the test subset is very small, the reported metrics are not used to claim statistical significance. For count-based metrics, we additionally report true positives, false positives, and false negatives, and interpret precision and recall using Wilson 95% confidence intervals. The F1-score and mIoU are reported as descriptive summary values.

For a predicted bounding box (B_p) and a ground-truth bounding box (Bg), IoU is defined as

$$IoU = \frac{B_p \bigcap B_g}{B_p \bigcup B_g}. \quad (1)$$

A detection was considered correct when the IoU value exceeded 0.5. In addition to Precision, Recall, and F1-score, the mean IoU (mIoU) over all detected objects was computed to assess localization accuracy.

4 Models and Adaptation Strategy

BLIP (Bootstrapping Language-Image Pre-training) is a vision-language model designed to jointly learn visual and textual representations from large-scale image-text datasets [10]. The model combines an image encoder, a text encoder, and a multimodal fusion module, enabling a wide range of downstream tasks including image captioning, visual question answering, retrieval, and open-vocabulary recognition.

In the zero-shot setting, BLIP receives an input image together with a textual prompt describing the target object category. Detection is performed by identifying image regions that exhibit the highest semantic similarity to the textual description. For the purposes of this study, the following prompts were used:

- “blue rabbit with long ears” (Krash);
- “purple hedgehog with glasses” (Chiko).

Although BLIP was not originally designed as a dedicated object detector, its multimodal representations allow approximate object localization through language-guided attention mechanisms.

Since BLIP is not a dedicated object detector, its localization results should be interpreted as approximate language-guided localization rather than standard detector outputs. For this reason, the comparison with Grounding DINO focuses on the practical ability to recover target object regions under the same prompt and annotation conditions, rather than on architectural equivalence between the two models.

Grounding DINO extends transformer-based object detection by integrating language grounding directly into the detection architecture [11]. The model combines a visual backbone, a language encoder, and a transformer decoder capable of generating object proposals conditioned on textual prompts.

Unlike conventional detectors that operate on a predefined category set, Grounding DINO performs open-vocabulary detection and can localize arbitrary object categories specified by natural-language descriptions. This capability makes it particularly suitable for evaluating novel object categories that are absent from standard training datasets.

For both zero-shot and few-shot experiments, identical textual prompts were used to ensure a fair comparison with BLIP.

To investigate the impact of limited supervision, we adapted both models using two training configurations:

1) 5-shot adaptation (5 annotated images per category);

2) 10-shot adaptation (10 annotated images per category).

During adaptation, the majority of pretrained parameters remained frozen, while lightweight task-specific layers were optimized using the annotated examples. This strategy preserves the general visual knowledge acquired during pretraining while allowing the models to learn discriminative features specific to the novel categories.

The adaptation procedure was restricted to lightweight task-specific components, while the pretrained visual and language encoders were kept frozen. The following parameters were fixed across all few-shot experiments: confidence threshold = 0.25, IoU threshold = 0.5, optimizer = Adam, learning rate = 0.0001, number of epochs = 100, batch size = 4. These details are reported to improve reproducibility and to avoid interpreting the few-shot results as full model retraining.

The optimization objective combines localization and classification losses:

$$Loss = L_{cls} + \lambda L_{loc}\,. \tag{2}$$

where L_{cls} denotes the classification loss, L_{loc} denotes the localization loss, and λ controls the balance between the two terms.

To improve robustness and reduce overfitting caused by the small training set, standard data augmentation techniques were applied, including horizontal flipping, random translation, and scale perturbation.

Figure 2 presents the overall evaluation pipeline. In the zero-shot setting, only a textual description of the target category is provided to the pretrained model. In the few-shot setting, a small set of annotated examples is additionally used to adapt the model before evaluation on the test set.

Figure 2 illustrates the overall evaluation pipeline adopted in this study. In the zero-shot setting (Fig. 2a), pretrained vision-language models, namely BLIP and Grounding DINO, receive only textual descriptions of the target categories (“blue rabbit with long ears” for Krash and “purple hedgehog with glasses” for Chiko). Object localization is then performed directly on the test images without any task-specific adaptation. In contrast, the few-shot setting (Fig. 2b) incorporates an addi-

tional adaptation stage. A small set of annotated training examples (5 or 10 images per category) is used to fine-tune lightweight model components while keeping most pretrained parameters frozen. The adapted models are subsequently evaluated on the same test images. Blue bounding boxes correspond to Krash, whereas red bounding boxes correspond to Chiko.

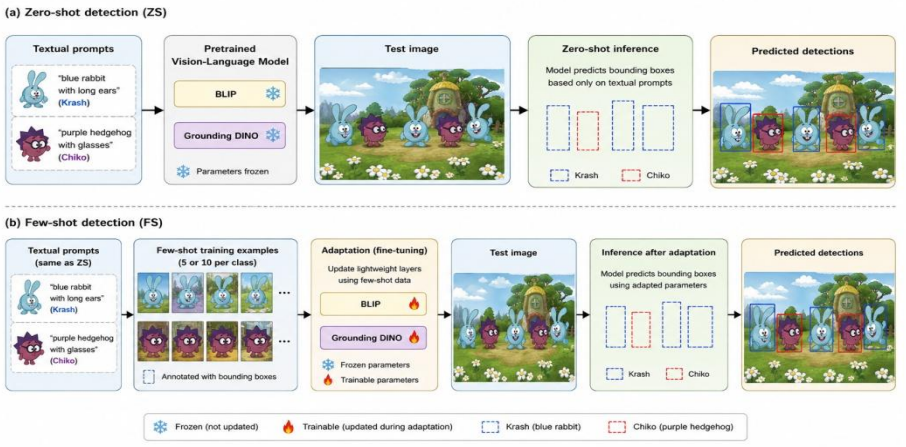

Fig. 2. Comparison of zero-shot and few-shot evaluation pipelines. The few-shot branch includes an additional adaptation stage using 5 or 10 annotated examples.

The same evaluation protocol, prompts, and test images were used across all experiments to ensure a consistent comparison between BLIP and Grounding DINO.

5 Results and Discussion

The performance of BLIP and Grounding DINO was evaluated under three experimental settings: zero-shot detection, 5-shot adaptation, and 10-shot adaptation. Detection quality was assessed using Precision, Recall, F1-score, and mean Intersection over Union (mIoU).

Given the small number of test instances, Table 2 should be read together with the underlying detection counts. For example, a recall of 0.50 corresponds to detecting two out of four ground-truth objects, while a recall of 1.00 corresponds to detecting all four objects in the controlled test subset.

Table 2 summarizes the obtained results.

The confidence intervals are necessarily wide because of the small test set. For example, detecting all four objects gives recall = 1.00, but the Wilson 95% confidence interval is approximately [0.51, 1.00]. Similarly, recall = 0.50 based on two detected objects out of four has an interval of approximately [0.15, 0.85]. Therefore, the perfect scores observed after few-shot adaptation should be interpreted as successful

detection of all objects in this particular controlled test subset, not as evidence of near-perfect general performance.

Table 2. Detection performance for different adaptation settings.

Method	Setting	Precision	Recall	F1-score	mIoU
BLIP	Zero-shot	0.67	0.50	0.57	0.69
BLIP	5-shot	1.00	1.00	1.00	0.90
BLIP	10-shot	1.00	1.00	1.00	1.00
Grounding DINO	Zero-shot	1.00	0.50	0.67	1.00
Grounding DINO	5-shot	1.00	1.00	1.00	0.93
Grounding DINO	10-shot	1.00	1.00	1.00	1.00

The results illustrate the limitations of pure zero-shot detection in this controlled pilot setting. Both models were able to use textual descriptions to identify some target instances, but the recall remained limited because only two of the four ground-truth objects were detected in the zero-shot setting.

The introduction of five annotated examples per category improved the observed detection results for both models. In the evaluated test subset, all four annotated objects were detected after few-shot adaptation. However, because this result is based on a very small number of test instances, it should be considered an encouraging pilot observation rather than a statistically robust performance estimate.

Further increasing the number of training examples from five to ten resulted in only marginal improvements. Increasing the number of annotated examples from five to ten did not change the object-level F1-score on this test subset and improved or stabilized mIoU. This suggests that, under the controlled conditions considered here, most of the observed gain was obtained after the first few annotated examples. Larger test sets are required to determine whether this pattern holds more generally.

Figure 3 presents representative detection examples obtained in the zero-shot and few-shot settings. In the zero-shot configuration, the models occasionally produced shifted bounding boxes or failed to detect one of the target categories. Such behavior can be attributed to the ambiguity of natural-language descriptions and the absence of category-specific visual examples during training.

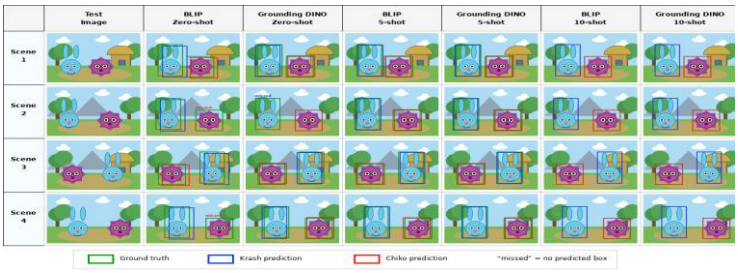

Fig. 3. Qualitative comparison between zero-shot and few-shot detection results. Green boxes indicate ground-truth annotations, while predicted detections are shown using category-specific colors.

After few-shot adaptation, the localization quality improved considerably. The predicted bounding boxes became more tightly aligned with object boundaries, and previously missed detections were successfully recovered. This observation confirms that the adaptation process enables the models to learn visual characteristics that cannot be fully expressed through textual prompts alone.

To investigate the relationship between annotation effort and detection quality, Figure 4 illustrates the variation of F1-score as the number of annotated examples increases.

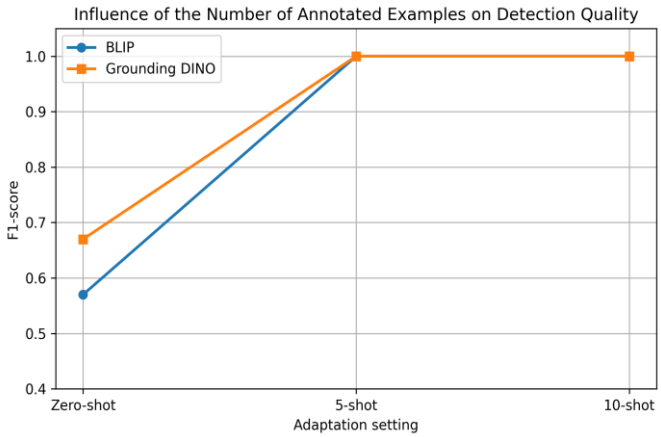

Fig. 4. Influence of the number of annotated examples on performance.

The most significant performance improvement occurs when moving from the zero-shot setting to the 5-shot configuration. This behavior suggests that a small number of representative examples provides sufficient information for the adaptation layers to establish robust category-specific visual prototypes. Increasing the number of examples to ten further stabilizes localization accuracy but yields comparatively smaller gains.

From a practical perspective, these results are important because they indicate that collecting a limited set of labeled samples may be considerably more cost-effective than relying solely on zero-shot detection or performing full-scale supervised training.

The experimental results support the hypothesis that few-shot adaptation significantly improves novel object detection performance compared with purely zero-shot inference. While modern vision-language models possess strong semantic understanding capabilities, textual descriptions alone are often insufficient for accurate localization of categories that are absent from the pretraining distribution.

Grounding DINO demonstrated superior zero-shot localization performance, which is consistent with its architecture specifically designed for language-guided object grounding [11]. BLIP, on the other hand, was originally developed as a general-purpose vision-language model and therefore exhibited weaker localization capabilities in the absence of adaptation.

Nevertheless, after introducing only a small number of annotated examples, the performance gap between the two models nearly disappeared. This finding suggests that the quality of pretrained visual representations is sufficiently strong in both architectures and that the primary challenge lies in adapting these representations to category-specific visual features.

Overall, the obtained results indicate that few-shot adaptation offers an effective compromise between annotation cost and detection accuracy, making it a promising strategy for practical deployment in domains where large annotated datasets are unavailable.

The main limitation of this study is the very small evaluation subset, which contains only four annotated test objects. As a result, the numerical metrics are highly sensitive to a single missed or false detection. The second limitation is the synthetic and visually controlled nature of the images, which reduces background clutter, occlusion, illumination variability, and domain shift. Third, only two visually distinctive cartoon categories were evaluated. Therefore, the results should not be generalized to arbitrary novel object detection tasks without further experiments on larger and more diverse datasets. Future work should include additional categories, more test scenes, real or semi-real backgrounds, prompt variation analysis, and repeated evaluation across multiple random few-shot splits.

6 Conclusions

This paper presented a controlled pilot comparison of zero-shot and few-shot detection for two novel cartoon character categories using BLIP and Grounding DINO. In the zero-shot setting, both models detected only part of the target objects, indicating that textual prompts alone were insufficient for complete localization in the evaluated scenes. Adding 5 or 10 annotated examples per category improved the observed results, and all four test objects were detected after few-shot adaptation. At the same time, the small size of the test subset and the synthetic nature of the images prevent strong statistical or real-world generalization claims. The results should therefore be interpreted as preliminary evidence that few-shot adaptation can be useful for controlled novel character detection, rather than as proof of near-perfect performance. Future work will extend the evaluation to larger datasets, more categories, multiple random few-shot splits, and more challenging scenes with clutter, occlusion, and domain shift.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25 (NeurIPS 2012), Lake Tahoe, NV, USA, pp. 1097–1105 (2012)
2. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, pp. 6000–6010 (2017)

3. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6), 1137–1149 (2017)
4. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of CVPR 2016, Las Vegas, NV, USA, pp. 779–788 (2016)
5. Carion, N., Massa, F., Synnaeve, G., et al.: End-to-end object detection with transformers. In: Proceedings of ECCV 2020, Glasgow, UK, pp. 213–229 (2020)
6. Zhang, H., Wang, Y., Dayoub, F., Sünderhauf, N.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: Proceedings of ICLR 2023, Kigali, Rwanda (2023)
7. Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A.: Zero-shot object detection. In: Proceedings of ECCV 2018, Munich, Germany, pp. 384–400 (2018)
8. Rahman, S., Khan, S.H., Porikli, F.: Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In: Proceedings of ACCV 2018, Perth, Australia, pp. 547–563 (2018)
9. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of ICML 2021, Virtual Event, pp. 8748–8763 (2021)
10. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of ICML 2022, Baltimore, MD, USA, pp. 12888–12900 (2022)
11. Liu, S., Zeng, Z., Ren, T., et al.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv:2303.05499 (2023)
12. Wang, X., Huang, T.E., Gonzalez, J., et al.: Frustratingly simple few-shot object detection. In: Proceedings of ICML 2020, Virtual Event, pp. 9919–9928 (2020)
13. Yan, X., Han, M., Li, Y., et al.: Few-shot object detection: Research advances and challenges. Artificial Intelligence Review 57, 112 (2024)
14. Andriyanov, N., Dementiev, V.: Image caption extending using LLM and style transfer. In: Communications in Computer and Information Science, pp. 113–126. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-87663-9_9
15. Andriyanov, N., Mikhailova, S., Fao, X.: Using the Segment Anything Model to develop control pallet loading system. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-2/W9, 19–25 (2025). https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-19-2025
16. Andriyanov, N., Tashlinsky, A., Dementiev, V.: Zero-shot detection in satellite images. In: 2025 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), Moscow, Russia, pp. 1–5 (2025). https://doi.org/10.1109/SYNCHROINFO65403.2025.11079342