Recognize Any Regions

Haosen Yang[1]* **Chuofan Ma**[2] **Bin Wen**[3] **Yi Jiang**[3]† **Zehuan Yuan**[3] **Xiatian Zhu**[1]†

[1]University of Surrey [2]The University of Hong Kong [3]ByteDance

Abstract

Understanding the semantics of individual regions or patches of unconstrained images, such as open-world object detection, remains a critical yet challenging task in computer vision. Building on the success of powerful image-level vision-language (ViL) foundation models like CLIP, recent efforts have sought to harness their capabilities by either training a contrastive model from scratch with an extensive collection of region-label pairs or aligning the outputs of a detection model with image-level representations of region proposals. Despite notable progress, these approaches are plagued by computationally intensive training requirements, susceptibility to data noise, and deficiency in contextual information. To address these limitations, we explore the synergistic potential of off-the-shelf foundation models, leveraging their respective strengths in localization and semantics. We introduce a novel, generic, and efficient architecture, named *RegionSpot*, designed to integrate position-aware localization knowledge from a localization foundation model (e.g., SAM) with semantic information from a ViL model (e.g., CLIP). To fully exploit pretrained knowledge while minimizing training overhead, we keep both foundation models frozen, focusing optimization efforts solely on a lightweight attention-based knowledge integration module. Extensive experiments in open-world object recognition show that our *RegionSpot* achieves significant performance gain over prior alternatives, along with substantial computational savings (e.g., training our model with 3 million data in a single day using 8 V100 GPUs). RegionSpot outperforms GLIP-L by 2.9 in mAP on LVIS val set, with an even larger margin of 13.1 AP for more challenging and rare categories, and a 2.5 AP increase on ODinW. Furthermore, it exceeds GroundingDINO-L by 11.0 AP for rare categories on the LVIS minival set.

1 Introduction

Remarkable progress has been achieved in the realm of purpose-generic image-level Vision-Language (ViL) representation learning, as exemplified by foundation models like CLIP [24] and ALIGN [11]. These advancements have led to significant performance improvements across a diverse spectrum of vision and multi-modal downstream tasks [7, 44]. The efficacy of these approaches can be largely attributed to their utilization of extensive datasets, typically encompassing millions, if not billions, of training samples replete with rich information. In the pursuit of a more nuanced approach to visual analysis, researchers have also ventured into the realm of universal region-level (e.g., objects) comprehension. This is evident in recent research endeavors [7, 39, 23, 4, 43, 16, 19, 22]. A common approach to this involves learning the semantics of image regions by applying an image-level pre-trained model (e.g., CLIP) to cropped regions, followed by representational distillation using the output of a detection model [7, 39], as depicted in Figure 1(a). However, utilizing individual cropped

*This work was performed when Haosen Yang worked as an intern at ByteDance.

†Corresponding authors

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

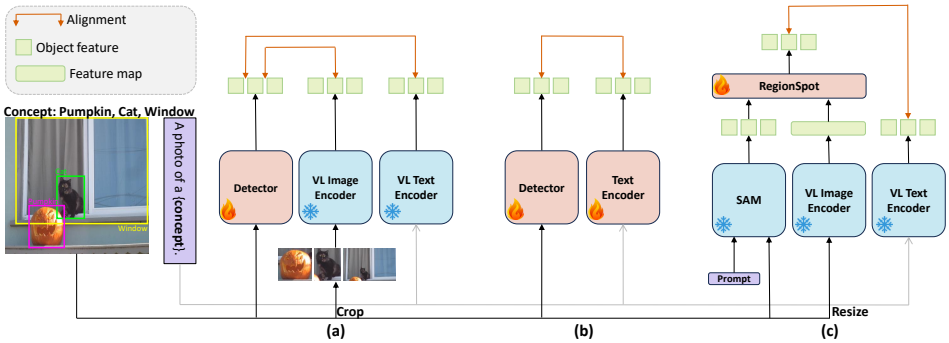

Figure 1: **Illustration of typical region-level visual understanding architecture**. (a) Learning the region recognition model by distilling image-level ViL representations from cropped regions and incorporating them into a detection model (*e.g.*, [7]). (b) Fully fine-tuning both vision and text models with a substantial dataset of region-label pairs. (c) Our proposed approach integrates pretrained (frozen) localization and ViL models, emphasizing the learning of their representational correlation.

regions in this design leads to the loss of crucial contextual information, which can hinder recognition performance. [16] introduced an open-world detector with a fixed ViL model, bypassing knowledge distillation. However, the use of ROIAlign [9] for region feature extraction poses limitations. Furthermore, directly applying an image-level model to isolated local regions is less effective, as the model was pretrained on entire images encompassing both object regions and surrounding context. An alternative, albeit brute-force, approach revolves around constructing region-level representations from scratch, harnessing an extensive dataset that pairs regions with labels [18, 38, 41, 37] (Figure 1(b)). Nevertheless, this approach grapples with challenges such as the proliferation of noisy pseudo-labels and significant training costs. Furthermore, significant advancements have materialized in the realm of class-agnostic visual localization techniques, as illustrated by the notable work of SAM [14]. This approach is characterized by its distinctive feature—an integration of position-aware localization knowledge, which we consider a valuable complement to the inherent capabilities of ViL models. Expanding upon this conceptual framework, our research introduces an innovative architectural paradigm at the region level, herein referred to as `RegionSpot`. This framework seamlessly incorporates large pre-trained ViL and localization models within an efficient training regimen, obviating the necessity for an extensive repository of region-label pairings. Our methodology centers on the acquisition of the correlation between localization data extracted from 'local' regions by the localization model and the semantic representations encompassing the entirety of the image, derived from the ViL model. This strategic approach permits us to circumvent the conventional fine-tuning of both pre-trained models—wherein they remain 'frozen' during training—thereby safeguarding the integrity of their rich knowledge and ensuring its maximal utilization, all while mitigating the potential for performance degradation. To enact this cross-model correlation, we employ the cross-attention mechanism [32]. In this configuration, the localization feature assumes the role of the 'query', whereas the ViL feature assumes dual roles as both the 'key' and 'value'. This implementation effectively facilitates the fusion of semantic and localization information in a manner that is amenable to learning and yields substantive efficacy.

Our **contributions** are as follows: (1) We introduce the concept of integrating off-the-shelf foundation models to tackle region-level visual understanding. (2) To achieve this objective, we introduce a novel architectural paradigm called *`RegionSpot`*, which does not necessitate training from scratch. This approach excels in both optimization efficiency and data utilization. By circumventing the fine-tuning of both localization and Vision-Language (ViL) components, our architecture retains its openness and adaptability, welcoming the seamless integration of advancements in both domains. Extensive experimentation in the context of open-world object understanding confirms the superior performance of our method, even with a substantially smaller number of learnable parameters. Remarkably, *`RegionSpot`* surpasses the state-of-the-art GLIP-L by 2.9 in mAP, with an even more substantial advantage of 13.1 AP observed for the more intricate rare categories.

2 Related Work

Zero-shot in image recognition Zero-shot image recognition is the task of recognizing categories that have not been seen during training. In [5] and [10], the authors utilized visual attributes to facilitate knowledge transfer to unfamiliar categories. Researchers have also investigated the utilization of class hierarchies, similarities, and object parts to enhance knowledge transfer, as demonstrated in the works of [27, 1, 42, 35]. Recent research has focused on aligning latent image-text embeddings for classifying and describing visual content. [6] pioneered the establishment of a visual semantic space through deep learning. Subsequently, CLIP [24] and ALIGN [11] attained impressive results via contrastive learning with extensive collections of image-text pairs, showcasing exceptional performance across diverse benchmarks. In contrast to previous endeavors that primarily addressed image-level recognition, we focus on fine-grained recognition of visual elements at the regional level.

Zero-shot in region understanding In zero-shot object recognition, the aim is to enable object detectors to identify categories not encountered during training, such as [26, 31, 36]. Researchers have explored various methods to bridge the gap between known and unknown categories using pre-trained semantic or textual features [30, 25, 3], knowledge graphs [28, 34], and more. Inspired by the zero-shot capabilities of Vision-and-Language (ViL) like CLIP [24], several approaches have sought to integrate pretrained Vision-and-Language (ViL) models. For example, [39, 7, 4] proposed a method to distill learned image embeddings from CLIP for target detection by focusing on cropped proposal regions. Another approach, RegionCLIP [43] employs a multistage training strategy. It starts by generating pseudo-labels from captioning data and then proceeds with region-word contrastive pretraining before transferring the knowledge to the detection task. [18] took a novel approach by formulating object detection as a grounding problem and incorporating additional grounding data to enhance semantic alignment at both phrase and region levels. Their results demonstrated improved performance, even on fully-supervised detection benchmarks. [38] leveraged large-scale image captioning datasets and expanded their knowledge database using generated pseudo-labels, bolstering their detection capabilities. The use of generated pseudo-labels effectively extended the detectors' generalization ability.

However, these methods face computational challenges and are susceptible to training data inconsistencies and image-level distractions. Differing from these studies, we explore the synergistic benefits of foundation models SAM [14] and CLIP [24]. Leveraging their strengths in localization and semantics, we propose an innovative region recognition framework.

3 Method

Our objective is to employ efficiently a pretrained ViL model and a localization model, trained on extensive data, to achieve region-level representation and understanding. These representations facilitate robust object conceptualization, especially for open-world region recognition. To realize this, as shown in Figure 2(a) we formulate a new approach, named `RegionSpot`. In the following sections, we will begin with a brief introduction to the foundational models in Section 3.1, followed by a comprehensive explanation of our approach with focus on learning region-text alignment across two pretrained models in Section 3.2.

3.1 Foundation Models

Vision-language foundation models use contrastive learning to map visual and textual data into a shared embedding space through a contrastive loss. This technique, exemplified by CLIP with 400 million text-image pairs [24], and ALIGN with 1.8 billion pairs [11], aims to minimize the distances between paired images and texts while maximizing distances between unpaired ones.
Localization foundation models have been advanced significantly. A prominent example is the pioneering SAM model [14], which has been trained on the extensive SA-1B dataset, boasting more than 1 billion automatically generated masks—an unprecedented scale, surpassing existing segmentation datasets by a factor of 400. This dataset also comprises 11 million images.

SAM comprises three core modules: (a) Image encoder: Utilizing a ViT-based backbone, this module extracts image features, yielding image embeddings. (b) Prompt encoder: It encodes positional information from input points, boxes, or masks to facilitate the mask decoder. (c) Mask decoder:

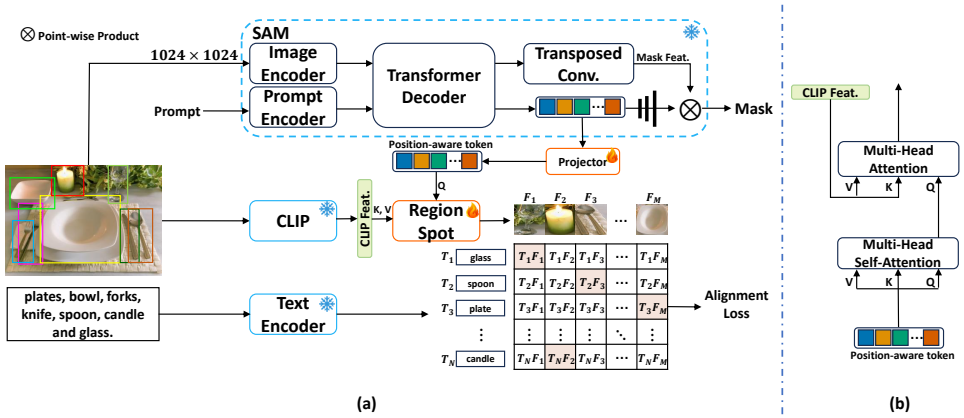

Figure 2: Overview of our proposed `RegionSpot`. (a) We integrate position-aware tokens from a localization model, such as SAM, with image-level feature maps extracted from a ViL model like CLIP. This integration yields region-level semantic tokens, which are then subjected to region text alignment. (b) Our cross-modal feature interaction design based on the attention mechanism.

This transformer-based decoder leverages both the extracted image embeddings and prompt tokens to make final mask predictions. One of SAM's remarkable features is its robust zero-shot generalization to novel data, obviating the need for domain-specific fine-tuning. Thanks to extensive training on a vast repository of prompt-text pairs, SAM demonstrates exceptional proficiency in object localization.

3.2 Region text alignment with frozen foundation models

In this section, we describe how we extract position-aware tokens from the localization foundation model and generate image-level semantic features using the ViL foundation model. We achieve inter-model association through a cross-attention mechanism that facilitates region text alignment.

Region-level position-aware tokens In our approach, we utilize manually-annotated object bounding boxes, denoted as $R = \{r_i\}, i = 1, .., N$, as regions of interest in the images. For each of these regions, represented as R, we extract position-aware tokens using the SAM model, denoted as $P = \{p_i\}, i = 1, .., N$.

As depicted in Figure 2, SAM employs a mask decoder to generate a mask based on a provided prompt. This process utilizes a transformer decoder, similar to the architecture in DETR [2], to generate an object token. This object token plays a crucial role in predicting the prompt mask, subsequently predicting dynamic MLP weights and performing a point-wise product with the mask features. We refer to this resulting token as "position-aware" because it encodes essential information about the object, including details about its texture and position. Following this, a projector is applied to map the output dimension of the position-aware token to the image-level feature space as discussed below.

Image-level semantic feature maps A single image can encompass multiple objects across numerous categories, capturing integrated context. We can conceptually view an image's feature map as a composition of region embeddings with varying structures. To fully capitalize the ViL model, we resize the input image to the required dimensions without cropping. Subsequently, we input this resized image into the ViL model, yielding the image-level semantic feature map denoted as V.

Relating position-aware tokens and semantic feature maps Our model, referred to as `RegionSpot`, efficiently establishes connections between region-level position-aware tokens and image-level semantic feature maps using the cross-attention mechanism [32]. In this mechanism, position-aware tokens P serve as queries, while semantic feature maps V take on the roles of both keys and values. This relationship is formulated as follows:

$$S = \text{Softmax}\left(\frac{F_p K_v^T}{\sqrt{C}}\right) V_v, \tag{1}$$

where F_p represents a transformation of P, K_v and V_v are derived from separate linear projections of V, and C is the projected feature dimension. This approach, well-established in the literature, has consistently demonstrated its effectiveness in information fusion. In our work, we extend its application to enhance region-level understanding in open-world scenarios. Specifically, we leverage this mechanism to integrate positional information with semantic content extracted from two distinct models at the regional level, while also considering the broader context from the entire image, as depicted in Figure 2(b).

Loss function In line with prior research, we generate text embeddings by processing category texts along with prompt templates, like *a photo of category in the scene*, using the text encoder. Then, we perform a dot product operation between each semantic token and its corresponding text features to calculate matching scores. These scores can be supervised using the focal loss [20].

Zero short inference Following [43], we focus on the more challenging region recognition task by utilizing human-annotated boxes or external region proposal generator. Inheriting the flexible prompting capability from SAM, our model allows for region recognition through prompting.

4 Experiments

Training data In pursuit of a robust training environment, we combined diverse datasets with varying label spaces. Our model's flexible architecture allowed us to seamlessly replace one-hot labels with class name strings. For training, we utilized publicly available detection datasets, comprising a total of approximately 3 million images. These datasets include Objects 365 (O365) [29], OpenImages (OI) [15], and V3Det (V3D) [33], each contributing uniquely to the diverse repository.

- Objects 365 (O365) is a large-scale object detection dataset featuring 365 distinct object categories across 0.66 million images. Our research employs an enriched version with over 10 million bounding boxes, averaging approximately 15.8 object annotations per image.
- OpenImages (OI) currently stands as the largest public object detection dataset, encompassing about 14.6 million bounding box annotations, equivalent to around 8 annotations per image.
- V3Det (V3D) distinguishes itself through a hierarchical organization, meticulously structuring up to 13,029 categories within a category tree.

Benchmark settings In our rigorous evaluation process, we utilized the extensive LVIS detection dataset [8], which encompasses 1203 categories and 19809 images reserved for validation. We do not prioritize the performance on COCO [21] which includes only 80 common categories covered by the Objects365 training dataset [29]. This limitation may not adequately assess a model's generalization in an open-world setting.

Since our current emphasis is not on object localization, we utilized ground-truth and class-agnostic bounding boxes from an existing detector to predict categories based on corresponding text descriptions, following the RegionCLIP approach [43]. Mean Average Precision (mAP) served as our evaluation metric.

Implementation details We train `RegionSpot` using AdamW [13] optimizer with the initial learning rate as 2.5×10^{-5}. All models are trained with a mini-batch size 16 on 8 GPUs. The default training schedule is 450K iterations, with the learning rate divided by 10 at 350K and 420K iterations. The training process unfolds in two sequential stages: (1) a warm-up phase leveraging the Objects365 to initiate the learning of region-word alignments, and (2) a phase of advanced learning for region-word alignments, utilizing a rich compilation from three diverse object detection datasets. The model is trained for 450K iterations at each stage. We implement several model variants: (1) `RegionSpot-Lite`: Integrating the base versions of both SAM and CLIP. (2) `RegionSpot-Pro`: Combining the SAM base with the more extensive CLIP large architecture. (3) `RegionSpot-Pro`$^{\uparrow 336}$: Further extending `RegionSpot-Pro` by using input image resolution of 336.

Table 1: Comparison of open-world zero-shot object recognition performance using ground-truth (GT) boxes, SAM proposals generate by automatic mask generator, and GLIP boxes on the LVIS dataset. * indicate finetune the CLIP with Adapter. The training time test on one V100 GPU

Method	Training Data	Proposals	Times	AP_r	AP_f	AP_{all}
CLIP-L w/ box	-	GT	-	40.6	59.2	48.7
CLIP-L w/ mask	-	GT	-	40.8	59.6	49.2
CLIP-L$_{\uparrow 336}$ w/ mask	-	GT	-	43.2	59.9	49.5
CLIP-L$_{\uparrow 336}$* w/ mask	O365, OI, V3D	GT	0.30k	46.8	63.2	53.1
`RegionSpot-Lite`	O365, OI, V3D	GT	0.18k	42.0	65.6	53.0
`RegionSpot-Pro`	O365, OI, V3D	GT	0.18k	50.6	68.8	56.6
`RegionSpot-Pro`$_{\uparrow 336}$	O365, OI, V3D	GT	0.20k	**55.4**	**68.6**	**59.9**
CLIP-L$_{\uparrow 336}$* w/ mask	O365, OI, V3D	SAM	0.30k	11.3	16.4	14.5
`RegionSpot-Pro`	O365, OI, V3D	SAM	0.18k	13.1	17.3	16.1
`RegionSpot-Pro`$_{\uparrow 336}$	O365, OI, V3D	SAM	0.20k	**14.3**	**19.2**	**18.2**
GLIP-T (B)	O365	GLIP-T(B)	57.5k	4.2	13.6	11.3
`RegionSpot-Lite`	O365	GLIP-T(B)	0.18k	12.7	15.7	14.1
GLIP-T	O365,GoldG,Cap4M	GLIP-T	92.1k	10.1	25.5	17.2
`RegionSpot-Lite`	O365, OI, V3D	GLIP-T	0.18k	20.0	24.2	21.1
GLIP-L	FourODs,GoldG,Cap24M	GLIP-L	120k	17.1	35.4	26.9
`RegionSpot-Pro`$_{\uparrow 336}$	O365, OI, V3D	GLIP-L	0.2k	**30.2**	**30.0**	**29.8**

Table 2: Evaluation of zero-shot object detection on the LVIS minival dataset.

Method	Training Data	AP_r	AP
GLIP-L	FourODs,GoldG,Cap24M	28.2	**37.3**
GroundingDINO-L	O365,OI,GoldG,Cap4M,COCO,RefC	22.2	33.9
`RegionSpot-Pro`$_{\uparrow 336}$	O365,OI,V3DET	**33.2**	36.9

4.1 Zero-shot Inference for Region Recognition

Zero-shot object detection on LVIS Val v1.0 Results on the LVIS benchmark are presented in Table 1. With ground-truth bounding boxes as region proposals, our model substantially surpasses the CLIP baselines (which applies CLIP on image crops) by a large margin (*e.g.*, **48.7, 49.2** *vs*.**59.9**). For fair comparison, we finetune the CLIP withe adapter, our model substantially surpasses the CLIP by a large margin. Moreover, in simulation of real-world cases, we move forward to test our method with noisy region proposals generated from off-the-shelf proposal generator. We first employ SAM as a proposal generator, inputting dense grid points to automatically generate proposals. It can be seen that RegionSpot still consistently outperforms CLIP (*e.g.*, **14.5** *vs*.**18.2** on AP_{all}) in this case, demonstrating the robustness of our method.

To fully exploit the potential of our method and synergy with the advancements of open world object detection (OWD), we further utilize region proposals from state-of-the-art OWD models, *i.e.*, GLIP, as our region prompts. Comparing with GLIP-T trained solely on the objects365 dataset, we can observe a considerable performance gain achieved by RegionSpot (*e.g.*, **4.2** *vs*.**12.7** on AP_r and **11.3** AP *vs*.**14.1** on AP_{all}). After scaling up the training data and use 336 resolution image as input, our models maintains superior performances over their GLIP counterparts. For instance, `RegionSpot-Pro`$_{\uparrow 336}$surpasses GLIP-T by **17.6** AP_r with less training data, showing compelling scaling behavior with data scale and input resolution. For more extensive evaluation, we also utilize bounding boxes generated by GLIP-L as prompts. It is noteworthy that RegionSpot achieves an impressive **13.1** increase in $AP_\mathbf{r}$ compared to GLIP-L, even when trained on less data at higher efficiency. Despite using a noisy box, we were still able to achieve promising results, thanks to the robust localization ability of SAM. Additional experiments, including the ViLD protocol, can be found in the Appendix

Open vocabulary object detection under ViLD-protocal To thoroughly evaluate our method, we conducted experiments using the ViLD protocol [7], training on base categories and testing on novel ones with the LVIS AP metric. For fair comparsion, all the method training only use the LVIS-base dataset and use the RPN from RegionCLIP as proposal generator. We also adapted our training to the LVIS-base dataset. As shown in Table 3, RegionSpot demonstrates competitive performance.

It outperforms the similarly frozen-backbone F-VLM by 1.1 AP_r . When we compared to RegionCLIP, which benefits from additional caption pretraining, RegionSpot significantly outperforms the pretrained version of RegionCLIP by 2.6 when utilizing same RPN.

Table 3: Comparison under the ViLD protocol [7]. All methods use the ResNet50 backbone. * indicate pre-training with CC-3M

Method	Proposals	Trainable Backbone	AP_r	AP_{all}
ViLD	RPN	✓	16.1	*22.5*
RegionCLIP*	RPN	✓	17.1	*28.2*
Detic-ViLD	RPN	✓	17.8	*26.8*
F-VLM	RPN	✗	18.6	*24.2*
RegionSpot	RPN	✗	**19.7**	*25.0*

Zero-shot object detection on LVIS minival5k [12] To fully exploit the potential of our method, we report on MiniVal containing 5,000 images introduced in MDETR [12]. We use the output proposals from GLIP as the prompt. As shown in Table 2, although we use 9x less training data, our model maintains superior performances over GLIP-L by 5.0 on APr. Further, our method also surpasses Grounding DINO-L (which even uses a more advanced detector) by 11.0 in APr.

Zero-shot instance segmentation We evaluate the performance of instance segmentation in a zero-shot setting using the LVIS dataset [8]. By leveraging the output from GLIP as the prompt, we direct it to `RegionSpot` for mask and class prediction. The mask AP is evaluated using the released X-Decoder [45] and OpenSeeD [40], both of which are trained with mask-text pairs. Impressively, as indicated in Table 4, `RegionSpot` outstrips X-Decoder and OpenSeeD by margins of 14.1 and 3.9 in AP, respectively. These outcomes suggest that our proposed `RegionSpot` can effectively harness the foundational model's capabilities to achieve more accurate region understanding.

Zero-shot object detection on ODinW [17] This benchmark was designed to evaluate model performance in real-world scenarios. To accurately evaluate recognition capabilities, we filter the dataset to include only those with more than three categories. The AP for each dataset is reported in Table 5. Impressively, `RegionSpot-Pro`$^{\uparrow 336}$, utilizing GLIP-L proposals, surpasses GLIP-L by a margin of 2.5 AP, attributable to its precise region recognition. Furthermore, our method exceeds the performance of GroundingDINO, even though it employs a more advanced detector.

4.2 Ablation Study

We conducted an ablation study for `RegionSpot-BL` using the boxes generated by GLIP. Unless otherwise mentioned, training was performed on three different detection datasets.

Enhancement with CLIP vision embedding We conjugate that a key ingredient with `RegionSpot` is the use of semantic information from CLIP vision encoder. To validate this assertion, we began our evaluation without the CLIP feature and subsequently integrated the class token output from the CLIP vision encoder. Results in Table 6a demonstrate that: (1) The CLIP feature offers a significant boost, pushing the baseline without CLIP vision encoder to **22.1**, which suggests inherent semantic limitations in SAM. (2) More notably, the inclusion of CLIP enhances overall performance to **23.7**. This underscores the potential of the class token to encapsulate global information from the entire image.

Position-aware tokens selection in SAM The position-aware tokens are generated by intermediary module in the SAM. We examined various locations for this generation, specifically after the Prompt encoder, the Transformer decoder, and the MLP within the SAM. Results presented in Table 7b indicate that generating output tokens after the Transformer decoder yields the best performance.

Table 4: Evaluation of zero-shot instance segmentation on the LVIS minival dataset.

Method	AP_r	AP_c	AP_f	AP
X-Decoder	-	-	-	9.4
OpenSeed	-	-	-	19.6
`RegionSpot-Pro`$^{\uparrow 336}$	**21.5**	**25.0**	**23.2**	**23.5**

Table 5: Evaluation of zero-shot object detection on the ODinW dataset.

Method	Aerial. Drone.	Aquarium	PascalVOC	shellfish	vehicles	Avg.
GroundingDINO-T	10.3	17.5	55.7	29.5	58.5	34.3
GLIP-T	12.5	18.4	56.2	26.3	56.0	33.8
GLIP-L	7.1	26.9	61.7	68.9	57.3	44.4
`RegionSpot-Lite`	13.1	20.1	58.2	30.1	57.2	35.7
`RegionSpot-Pro`$^{\uparrow 336}$	**14.2**	**27.2**	**62.7**	**69.3**	**61.3**	**46.9**

Table 6: **Ablation experiments on LVIS. (a)** The effective of CLIP vision encoder; **(b)** Position-aware tokens selection ; **(c)** Depth of `RegionSpot`.

(a)

CLIP	AP
w/o CLIP vision.	8.0
+ CLIP feat. map	22.1
+ Class token	**23.7**

(b)

Position-aware tokens	AP
Prompt encoder	18.6
Transformer	**23.7**
MLP	20.4

(c)

Depth	AP
1	23.2
3	**23.7**
6	22.8

This observation is expected since tokens derived from the Prompt encoder are relatively undeveloped. Surprisingly, it can outperform GLIP (*i.e.*, **18.6** vs. **17.2**). Moreover, there is a performance decline after the MLP, which can be attributed to dimensional reduction.

Module architecture Another pivotal aspect of `RegionSpot` is the depth of model. To assess its impact, we experimented by varying the depth of our model. As indicated in Table 6c, it is imperative for the model to have a sufficiently large depth, such as 3 blocks, without being excessive.

Prompt engineering Finally, we carried out an ablation study focusing on prompt engineering, incorporating both box prompts in SAM and text prompts in the text encoder. As evidenced by the results in Table 7a: (1) Leveraging multiple boxes as prompts in SAM boosts performance, achieving an AP of 22.1. This enhancement is credited to the self-attention mechanism of `RegionSpot`, which adeptly fuses information from varied regions. (2) Further utilizing text prompts results in a modest performance boost, specifically an increase of 1.6 AP.

Ablation study of SAM model We conjugate that a key ingredient is the position-aware information from SAM. We evaluating the impact of different SAM model sizes, such as ViT-L, is essential. We conducted experiments with varying SAM model sizes. As shown in the Table 7b, our findings are summarized as follows: (1) Impact of SAM Model Size: Our results indicate that the use of larger SAM models (e.g., SAM-L) improves mask AP due to the higher quality of mask generation. However, for box AP, the improvement is not significant. This is because the SAM mask token primarily contributes position-aware knowledge, which is already sufficiently captured by ViT-B and ViT-L. (2) Choice of SAM Model: Given our focus on region recognition, we opted for SAM-B, balancing performance and computational efficiency.

4.3 Visualization

Result visualization In Figure 3, we present the results of bounding region recognition on the LVIS [8] dataset, comparing between GLIP and `RegionSpot`. To assess the zero-shot recognition capability, we employ the same bounding boxes for both models. As observed, `RegionSpot` can distinguish even subtle differences, recognizing smaller objects like "lemon" and "tennis ball" and similar objects like "lantern" and "fireplug". Notably, `RegionSpot` stands out in terms of the accuracy of its label predictions, especially within the category of rare classes.

Table 7: **Ablation experiments on LVIS. (a)** The effective of propmpt engineering; **(b)** The effective of SAM

(a)

Prompt	AP_r	AP_f	AP
baseline	19.6	21.2	18.5
w/ mutiple boxes prompt	23.2	25.0	22.1
w/ text prompt	**24.9**	**25.5**	**23.7**

(b)

SAM	box AP_r	mask AP_r
ViT-B	24.9	22.8
ViT-L	24.7	23.6

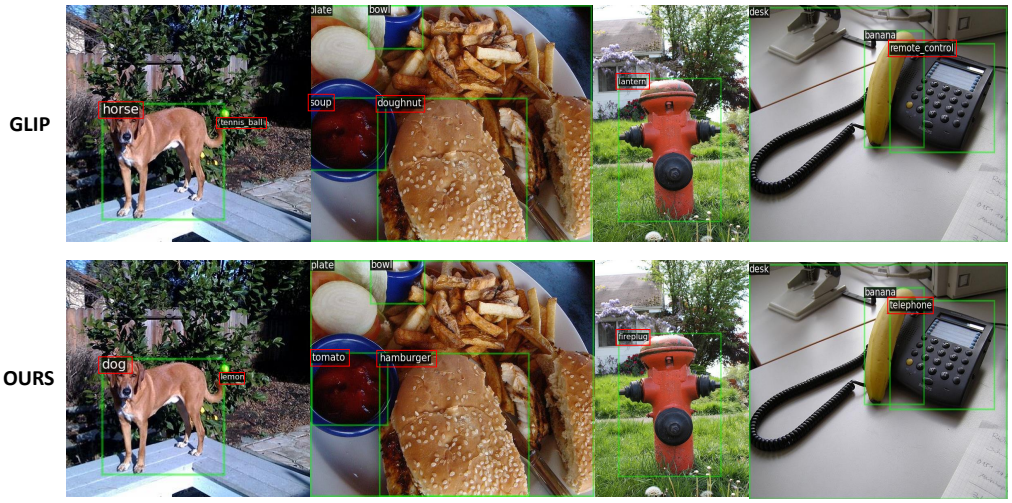

Figure 3: Qualitative prediction results of GLIP-T [18] (first row) and `RegionSpot` (second row) on the LVIS dataset [8]. Our model recognizes the objects more accurately. Best viewed when zooming-in.

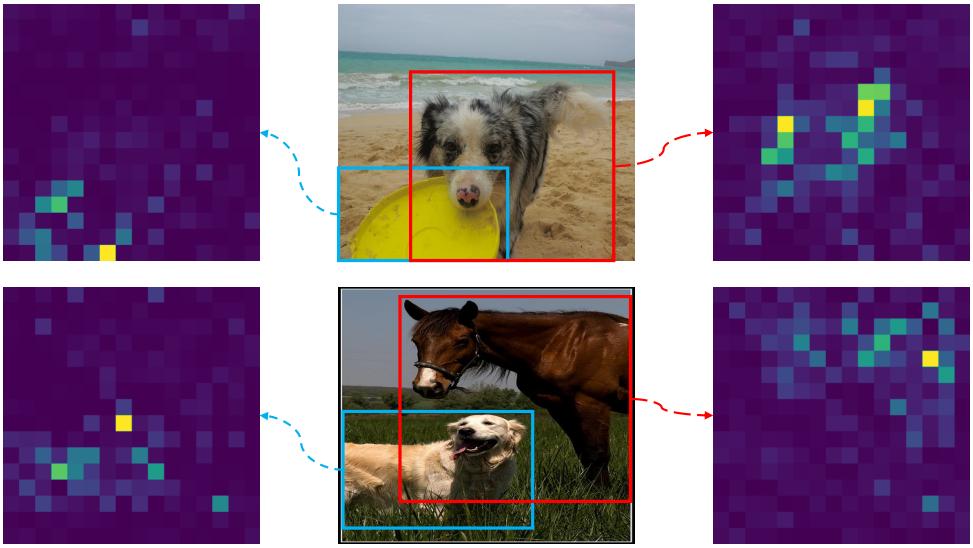

Figure 4: Cross-attention maps in `RegionSpot`. These maps show that the position-aware token aligns effectively with the semantic feature map of the entire image. In each row, the blue and red boxes are corresponding to the left and right maps respectively.

Model behavior visualization To gain more intuitive understanding on the effect brought by our `RegionSpot`, we examine the cross attention map on LVIS [8]. We take the output tokens as 'query' and CLIP feature map as 'key' and 'value'. For clearer visualization, we omit the class token from the CLIP semantic feature. The resulting attention map clearly depicts the correspondence between the position-aware tokens generated by SAM and the feature map produced by CLIP. Using this arrangement, we gain a visual insight into how `RegionSpot` establishes connections between distinct features. As depicted in Figure 4, the attention maps vividly showcase `RegionSpot` capability to seamlessly incorporate both SAM and CLIP. Such visualizations serve a dual purpose: they highlight the efficacy of our method, and simultaneously, shed light on the intricate mechanisms underpinning the `RegionSpot`.

5 Conclusions and limitations

In this study, we introduce `RegionSpot`, a novel and efficient framework leveraging frozen vision and vision-language foundation models for region recognition, eliminating the need for training from scratch. To fully exploit knowledge in pretrained models and minimize the training overhead, we keep both foundation models frozen and focus optimization efforts solely on a lightweight attention-based knowledge integration module. Extensive experiments in the context of open-world object understanding confirms the superior performance of our method, even with a substantially smaller number of learnable parameters, which distinguishes our method and enables efficient training. Impressively, *RegionSpot* outperforms the leading GLIP-L by 2.9 in mAP, and this lead grows to 13.1 when considering complex rare categories. While our method advances open world region understanding, it still not unleash potential capabilities from the foundmental models, such as the automatic localization ability from SAM, which could reduce reliance on external region proposal mechanisms for object detection and enhance versatility. We leave this for further investigation.

References

[1] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 59–68, 2016.

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

[3] Soravit Changpinyo, Wei-Lun Chao, and Fei Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *Proceedings of the IEEE international conference on computer vision*, pages 3476–3485, 2017.

[4] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022.

[5] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1778–1785. IEEE, 2009.

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[10] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. *Advances in neural information processing systems*, 27, 2014.

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[12] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[15] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18, 2017.

[16] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.

[17] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35:9287–9301, 2022.

[18] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[19] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[22] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *arXiv preprint arXiv:2310.16667*, 2023.

[23] Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14074–14083, 2022.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[25] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58, 2016.

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

[27] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR 2011*, pages 1641–1648. IEEE, 2011.

[28] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. In *CVPR 2011*, pages 1481–1488. IEEE, 2011.

[29] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

[30] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.

[31] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[33] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. *arXiv preprint arXiv:2304.03752*, 2023.

[34] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018.

[35] Guo-Sen Xie, Li Liu, Fan Zhu, Fang Zhao, Zheng Zhang, Yazhou Yao, Jie Qin, and Ling Shao. Region graph embedding network for zero-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 562–580. Springer, 2020.

[36] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023.

[37] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23497–23506, 2023.

[38] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *Advances in Neural Information Processing Systems*, 35:9125–9138, 2022.

[39] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022.

[40] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023.

[41] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022.

[42] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2002–2010, 2017.

[43] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.

[45] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023.

A Additional Experimental Result

Training efficiency. To illustrate the training efficiency of `RegionSpot`, we benchmark its GPU training hours against RegionCLIP and GLIP, as showcased in Table 8. Even though we utilize the ViT-Large as our backbone, our model achieves faster training. This efficiency can be attributed to our frozen approach and processing of images at a reduced resolution of 224x224 for the CLIP Large. All benchmarks were executed in a consistent hardware environment, leveraging eight NVIDIA V100 GPUs. In stark contrast, GLIP necessitates an extensive 92K GPU hours, a whopping 436 times more than our approach, mainly due to its exhaustive fine-tuning of both Vision and Text models. Interestingly, even when RegionCLIP adopts a smaller backbone akin to ours, it still requires 4.6K GPU hours.

Table 8: Comparisons with the training efficiency.

Method	Training data	Train time (GPU hours)	Learnable Param (M)
RegionCLIP	CC3M	4.6K	-
GLIP-T	O365, GoldG, Cap4M	92.1K	-
GLIP-L	FourODs,GoldG,Cap24M	120K	289
GDINO-L	O365,OI,GoldG,Cap4M,COCO,RefC	no released	341
`RegionSpot-Pro`	O365, OI, V3D	**0.2K**	35

Table 9: Effect of increasing the detection training data.

Data	AP_r	AP_c	AP_f	AP
Objects365	11.9	13.6	20.2	15.9
Objects365 + OpenImages	16.4	18.2	23.7	20.1
Objects365 + OpenImages + V3DET	**24.9**	**21.6**	**25.5**	**23.7**

Benefit of increasing the detection training data Table 9 showcases the performance improvements observed when augmenting the training data size. Through our proposed framework, integrating additional detection data from diverse sources consistently enhances the capabilities rooted in pre-trained knowledge. Compared to training with only Objects365, including OpenImages effectively improves the overall AP from 15.9 to 20.1. The inclusion of V3Det further propels the performance, achieving an impressive overall AP of 23.7. This improvement is particularly significant for rare categories, with an increase from 16.4 to 24.9 (a gain of +8.5 AP), attributable to its extensive vocabulary.

B More visualizations on LVIS

Figure 5 provides more examples on LVIS [8].

Figure 5: More visualizations in comparison with GLIP. Best viewed when zoomed-in.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Refer to the **Abstract** and **Introduction** in Sec. and Sec. 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Refer to Sec. 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical results in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The result is reproducible by training the framework under our instructions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be available after being accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Refer to the **Implement Details** in Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Refer to the **Implement Details** and **Inference Strategy** in Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Refer to the **Broader Impacts** in Sec. 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The code and models we use and will be released do not have high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The assets are properly credited and the license and terms of use explicitly mentioned and properly respected in this paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.